KARIMI, FIROOZEH, Ph.D. Urban Expansion Modeling Using Machine Learning Algorithms. (2021)
Directed by Dr. Selima Sultana. 129 pp.

Modeling and simulating urban expansion is required for assessing and predicting the consequences of the current urban growth patterns. Given the dynamic and convoluted nature of the urban expansion process and the necessity of handling continuous and categorical variables, non-normal distributed data, and non-linear relationships, urban expansion modeling is challenging. It is also critically important to find an appropriate method for modeling and simulating urban expansion in order to meticulously identify spatiotemporal variables and predicting the direction of land use/land cover (LULC) changes. To handle these issues effectively and enhance the quality of urban expansion prediction, the capabilities of machine learning methods are explored in this dissertation. Machine learning methods are relatively unknown in urban expansion modeling and have not been evaluated thoroughly in the current literature. The machine learning methods allow the exploration of a variety of data sampling strategies, predictor variables, and model configurations to enhance the accuracy and predictability of urban expansion modeling. The models are calibrated using spatiotemporal data of 2001-2016 and are applied to simulate future urban developments for two urbanized counties—Guilford and Mecklenburg in NC, USA. The accuracy and reliability of the models are evaluated by apposite evaluation metrics. Distance to highways is recognized as the most important predictor variable in both study areas, however, the importance of the predictor variables varies in different geographic contexts and with different methods. A comparative study on machine learning methods demonstrated that the random forest (RF) model is a fast, high-performance, and accurate model with low uncertainty; therefore, it can be effectively utilized to evaluate a wide range of urban development scenarios and support decision-making to accomplish the goal of implementing environmentally sustainable development. Sustainable urban growth management in addition to sophisticated and elaborative

models requires different urban growth scenarios. An integration of random forest and cellular automata (RF-CA) is proposed to simulate urban development under three urban growth scenarios, including current trends, controlled urban development, and environmentally sustainable urban development. While current trends allow the urban fringe to be uncontrollably developed, the controlled and environmentally sustainable urban development scenarios constrain future developments and reduce the environmental implications. The results show that the current urban development in the study area for 2021 and 2026 will appear near current or newly built urban clusters or adjacent to the major roads, however, the controlled and environmentally sustainable urban development scenarios are much higher compact and minimize environmental costs.

Keywords: Urban expansion, environmentally sustainable development, machine learning, decision tree, random forest, support vector machine, North Carolina

URBAN EXPANSION MODELING USING MACHINE LEARNING ALGORITHMS

by

Firoozeh Karimi

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2021

Approved by

Selima Sultana Selima Sultana
Committee Chair

To my beloved spouse, Ali Shirzadibabakan,

without whom I could not have come this far.

and

my daughter Nila,

without whom I cannot imagine my life.

APPROVAL PAGE

This dissertation, written by FIROOZEH KARIMI, has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair	Selima Sultana
Committee Members	Paul Knapp
	Shan Suthahran
	Wenliang Li

03/22/2021

Date of Acceptance by Committee

03/12/2021

Date of Final Oral Examination

ACKNOWLEDGMENTS

I would like to express my gratitude and appreciation to my advisor, Dr. Selima Sultana, for all the support and encouragement she gave me throughout my Ph.D. work. Without her guidance and constant feedback, this project would not have been successful. I would also like to thank my committee members, Dr. Paul Knapp, Dr. Shan Suthaharan, and Dr. Wenliang Li for their continuous guidance and support throughout this process. This dissertation would not have been possible without their support and feedback. I would also like to thank all professors at the Department of Geography, Environment, and Sustainability (GES) whose classes gave me lots of benefits and enjoyment. I would also like to thank all my friends at GES whose support made this journey easier. Thanks Joyce, Michele, Nichole, Jennifer, Rojer, Pankaj, Shane, Jesse, and Mehrnaz, for their thoughts, well-wishes, and advice.

I must also thank all my friends in the U.S. who supported me along the way. Many thanks to Farhad and Majedeh and their lovely kids Tara and Rayan who were always helpful in numerous ways during the past four years. I would like to express my appreciation to the Rupe family, Jamie, Rachel and Sidney who supported me and my family. Most importantly, special thanks to Roshanak and Kimia who have done a lot to keep me going.

I would like to express my deepest gratitude to my family, my mother, my father, and my sisters, for loving and supporting me to no end and showing me the true meaning of the “death of distance”. Thanks should also go to my in-laws, especially my mother-in-law and my father-in-law for their support. I would like to thank my husband, Ali, for

taking this long journey with me, encouraging me to keep dreaming, and being a true friend when I struggled. No words can describe my love and thanks to him. Lastly but most importantly, I would like to thank my baby Nila for being patient with me and giving me love.

TABLE OF CONTENTS

Page

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER

I. INTRODUCTION 1

1.1 Research Background and Motivations 1
1.2 Research Objectives 6
1.3 Synopsis of Dissertation 6

II. AN ENHANCED SUPPORT VECTOR MACHINE MODEL FOR URBAN EXPANSION PREDICTION 8

2.1 Introduction 8
2.2 A Background on SVM-based Urban Expansion Modeling 14
2.3 Data and Methodology 19
2.3.1 Study Area 19
2.3.2 Predictor Variables 20
2.3.3 Data Collection and Preparation 23
2.3.4 Data Sampling 25
2.3.5. Model Implementation 30
2.3.6. Model Evaluation 32
2.4 Results and Discussion 37
2.5 Conclusion 46

III. A COMPARATIVE STUDY ON MACHINE LEARNING ALGORITHMS FOR URBAN GROWTH PREDICTION 48

3.1 Introduction 48
3.2 Machine-Learning-Based Urban Expansion Modeling 53
3.2.1 Classification And Regression Trees (CART) 54
3.2.2 Random Forest (RF) 55
3.2.3 Support Vector Machine (SVM) 56
3.2.4 Logistic Regression (LR) 57
3.2.5 Artificial Neural Network (ANN) 58
3.3 Data and Methodology 61
3.3.1 Study Area 61
3.3.2 Data 62
3.3.3 Methodology 64
3.4 Results and Discussion 71
3.5 Conclusion 78

IV. A SCENARIO-BASED SIMULATION OF URBAN GROWTH BY COUPLING RANDOM FOREST AND CELLULAR AUTOMATA 80

4.1 Introduction 80
4.2 Random Forest-Cellular Automata (RF-CA) Model 86
4.3 Case Study 89
4.4 Implementation 95
4.5 Results 102
4.6 Conclusion 112

V. CONCLUSION 114

REFERENCES 118

LIST OF TABLES

Page

Table 2.1: A Summary of Considered Predictor Variables for Urban Expansion Modeling 23

Table 2.2: The Training and Testing Datasets 25

Table 2.3: The Number of Built and Unbuilt Land Cells in Guilford County for the Years 2001, 2006 and 2011 30

Table 2.4: A Summary of LULC Change in Guilford County Over 2001-2011 30

Table 2.5: The Built-unbuilt Confusion Matrix 33

Table 2.6: The Changed-unchanged Confusion Matrix 34

Table 2.7: PCPC and PCPU Values for Three Sampling Methods 38

Table 2.8: The Significance of Predictor Variables 39

Table 2.9: The Performance Evaluation Results of the Final Model 43

Table 2.10: The Specifications of the Final Model 43

Table 3.1: Summary of the Strengths and the Limitations of the Models 61

Table 3.2: Predictor and Target Variables Utilized for Urban Expansion Modeling 65

Table 3.3: The Performance Obtained for the Models From the Cross-tabulation of the Overlay Analysis of Actual and Predicted Urban Development Map of Mecklenburg County, NC 76

Table 3.4: Comparison of the Methods Concerning the Average Accuracy, the Number of Hyperparameters, Run Time, and the Need for Data Preparation 76

Table 3.5: Comparison of the Methods Concerning the Number of FN and the Number of FP Predicted Land Cells for 2011 and 2016 77

Table 4.1: The Classification of the Developed Urban Area and the Natural Environment 93

Table 4.2: Target and Predictor Variables for Urban Expansion Modeling 96

Table 4.3: The Performance Achieved from the Cross-tabulation of the Overlay Analysis of Actual and Predicted Urban Development Map of Mecklenburg County, NC for 2011 and 2016 107

Table 4.4: The Number of TN, TP, FN, and FP Land Cells for 2011 and 2016 for the RF and RF-CA Model 107

Table 4.5: The Environmental Cost Index for Simulated Urban Developments Under Current Trends, Controlled Urban Development, and Environmentally Sustainable Urban Development Scenarios for 2021 and 2026 111

LIST OF FIGURES

Page

Figure 2.1: A Linear, Binary SVM Classifier and the Optimal Separating Hyperplane H, Lying Between and Parallel with H1 and H2 18

Figure 2.2: The Location of Guilford County in North Carolina 20

Figure 2.3: The LULC Map of Guilford County in (a) 2001, (b) 2006, and (c) 2011 27

Figure 2.4: The Distance Map to a) City Centers, b) Built Areas, c) Highways, d) Main Roads, e) Streets, f) Railroads, g) Green Spaces, and h) Water Bodies in Guilford County in 2001. 28

Figure 2.5: The Number of Neighboring Cells with the LULC Type of a) Forest, b) Water body, c) Wetland, d) Low-intensity Developed Area, e) Medium-intensity Developed Area, f) High-intensity Developed Area, g) Developed Open Space, h) Potential LULC for Urban Expansion Including Agricultural Land, Shrub, Herbaceous and Pasture, and i) Barren Land in Guilford County in 2001. 29

Figure 2.6: The Population Density Map in Guilford County in 2001 30

Figure 2.7: The Diagram of SVM-based Urban Expansion Modeling 36

Figure 2.8: The Training Accuracy for Different Configurations of SVM Models (2001-2006) 40

Figure 2.9: The Testing Accuracy, Precision, Sensitivity, and Specificity Using the Built-unbuilt Confusion Matrix 41

Figure 2.10: The Testing Accuracy, Precision, Sensitivity, and Specificity Using the Changed-unchanged Confusion Matrix 41

Figure 2.11: The Results of PCP, PCPB, PCPUB, PCPUC, and PCPC Evaluations of the Developed SVM models (2006, 2011) 43

Figure 2.12: The Binary Classification of Real and Predicted LULC Maps in 2011 45

Figure 2.13: The Binary Classification of Simulated LULC Map in 2016 45

Figure 3.1: The Process of Machine-learning-based Urban Expansion Modeling 60

Figure 3.2: The Map of the Study Area (Mecklenburg County, NC, USA) 63

Figure 3.3: The Diagram of Machine Learning-based Urban Expansion Modeling 70

Figure 3.4: Predictor Variables Importance Based on CART Model 77

Figure 3.5: Predictor Variables Importance Based on RF Model 77

Figure 4.1: The Location Map of the Study Area, Mecklenburg County 91

Figure 4.2: The Urban Development in Mecklenburg County in the Study Period (2001-2016) 92

Figure 4.3: Rate of Development in the Urban Area in Mecklenburg County for 2001-2006, 2006-2011, and 2011-2016 93

Figure 4.4: The Flowchart of the RF-CA Model for Simulating Urban Development 98

Figure 4.5: Constraint Layer for Exclusions of (a) Current Trends, (b) Controlled Urban Development, and (c) Environmentally Sustainable Urban Development Scenarios for Mecklenburg County, NC 101

Figure 4.6: Random Sampling Rate and Prediction Error 103

Figure 4.7: The Value of Kappa Statistics with Respect to the Number of Trees (k) and the Number of Random Split Variables (m) 104

Figure 4.8: Transition Potential Map of Mecklenburg County, NC for 2011 106

Figure 4.9: The Predicted Map of Mecklenburg County, NC for 2011 Consideing True Positive, True Negative, False Positive and False Negative Land Cells 107

Figure 4.10: The Importance of Predictor Variables 108

Figure 4.11: The Predicted Urban Development Map of Mecklenburg County, NC for 2021 and 2026 109

Figure 4.12: Simulated Urban Development Map for (a) 2021 and 2026 Under Controlled Development Scenario, (b) 2021 and 2026 Under Environmentally Sustainable Urban Development Scenario 111

CHAPTER I

INTRODUCTION

1.1 Research Background and Motivations

Urban growth, the most conspicuous indication of land-use/land-cover (LULC) change caused by human activities (B. Huang et al. "Support Vector Machines for Urban Growth Modeling"), has been identified as an important element of environmental risk (Martellozzo et al.). In other words, urban growth causes the conversion of the natural environment to agricultural lands and finally to urban land uses such as residential, recreational, transport, commercial, and industrial land uses (Clarke et al.; del Mar López et al.; Meyer and Turner). It is undeniable that urban growth is a complex and dynamic spatial-temporal process (Sultana and Weber "The Nature of Urban Growth and the Commuting Transition: Endless Sprawl or a Growth Wave?") which is the result of several factors such as continuous urbanization (Weber and Puissant), population growth (Meyer and Turner), economic growth (Black and Henderson), industrialization (Kelley and Williamson), transportation development (Duranton and Turner), government developmental policies (Darin-Drabkin), and development and property tax (Bengston et al.). Currently, almost half of the world's population lives in urban areas and United Nations has forecasted that 67.2% of the world population will live in urban areas in 2050. The global urban footprint will increase approximately 40–67% until 2050 relative to 2013, and this trend will continue to a growth ratio of more than 200% by 2100

(X. Li et al.). The rapid and low-density urban growth has broad impacts on the environment such as diminishing them in size, resulting in habitat fragmentation (Nagendra et al.), and disturbing wildlife (Marzluff; McKinney), as well as generating damaging effects through such sources as pollutions (Shukla and Parikh), hydrological complications (Lindh; Williams), deforestation and agricultural fields loss (Masri; Yankson and Gough), and regional and global warming (Alcoforado and Andrade; Stone Jr The City and the Coming Climate: Climate Change in the Places We Live). Broad conversion of native vegetation to agricultural lands to provide food for the growing population has occurred at an unprecedented pace in the last century, which has led to diminishing the natural environment in size (Armesto et al.). Expanding the urban area into the natural environment, decreasing domestic habitat area, and increasing the extent of forest-opening boundaries (H. Li et al.) has led to habitat fragmentation. These circumstances result in species loss, on both a local and global level (McCauley et al.). Urban growth and transportation development are interdependent and by expanding urban area construction of new transport networks is inevitable (Sultana "Land Use and Transportation"). As a result, vehicle traffic introduces toxic metals into the urban soils, water, and air (Mireles et al.). Also, population concentration and industries, as the main sources of smog, causes pollution (Romero et al.). Extensive increases of impervious surfaces which are mainly artificial structures such as roads, sidewalks, roofs, parking lots, airports, and turfgrass can dramatically increase the speed and amount of runoff and other aspects of the water cycle, therefore have tremendous impacts on basin hydrology and water quality (McDonald et al.). Urban growth converts the natural land cover to

agricultural lands and then to urban land uses, the whole process leads to forest and agricultural field loss (Richards and VanWey). Urban heat island (UHI) and climate change are two significant outcomes of uncontrolled growth on a regional and global scale, respectively. UHI, a higher temperature in an urban area compared to its surrounding rural or less developed areas (Coseo and Larsen; Memon et al.; Nuruzzaman). UHI causes various adverse effects on the urban environment, society, and economics such as an increase in energy consumption, deterioration of livability and safety of the urban environment, elevation in ground-level ozone, and even an increase in heat-related mortality rate (Coseo and Larsen; Kim and Guldmann; Memon et al.; Mirzaei and Haghighat; Nuruzzaman; Onishia et al.; Stone Jr "Urban Heat and Air Pollution: An Emerging Role for Planners in the Climate Change Debate"; Tomlinson et al.). Urban growth affects climate change patterns by emitting greenhouse gases through deforestation and plant clearance (Stone Jr The City and the Coming Climate: Climate Change in the Places We Live). But the impacts of the expansion do not end to the aforementioned problems. In addition to environmental impacts, urban growth also leads to unbalanced economic growth which jeopardizes productivity, and personal and public finances (Ekins). Urban growth endangers the security, livability, and social equity of city dwellers by changing the size and form of urban areas (Bramley and Power; Lin and Yang).

Thus, a precise perception of the location, direction, scale, type, causes, and consequences of urban growth is required for most urban purpose projects, future planning, and sustainable development (Pradhan; Yao et al.). Urban growth models act as

essential tools for understanding the dynamic process of urban expansion, assessing causal factors, examining the consequences of planning policies, supporting the planning and decision-making, and determining the efficiency of the plans before implementation (Hosseinali et al.). But, urban growth modeling is not sufficient for maintaining sustainable development. Sustainable development is the cities capability to decrease the environmental impacts of urban activities while protecting a stable economy and enhancing social equity and quality of life in urban areas (Haughton and Hunter; Newman and Kenworthy). While rapid and horizontally urban growth in an uncontrolled and unorganized manner exacerbates the mentioned problems (Jiang et al.; Sisodia et al.), managed and compact growth is a sustainable form of development advocated by many researchers (Burton et al. The Compact City: A Sustainable Urban Form? ; Burton et al. "The Compact City and Urban Sustainability: Conflicts and Complexities"; Nurul). To attain environmental urban sustainability and to solve the problems caused by urban expansion, sustainable urban development strategies should be embedded with urban growth and LULC change modeling approaches. Some operational strategies for environmentally sustainable urban development are: (1) not to convert good-quality agricultural lands to urban lands excessively; (2) to control the amount of land conversion based on available lands and population growth; (3) to guide land conversion to less important sites from sustainability point of view; and (4) to retain compact development patterns.

Over the last three decades, various types of models and methods have been developed based on remote sensing (RS) and geographic information system (GIS)

techniques with the general purpose of understanding the complexity of urban growth (Aburasa et al.; Musa et al.). GIS plays an important role in storing, managing, and preparing the data layers. Most of the GIS-based models are cell-based and the data layers are tessellated to form a grid of cells, where the values of each cell exhibit its spatial attributes (Batty et al.; Michalak). RS provides the fundamental required data for urban growth modeling through satellite images (Chen et al.). The conventional models such as cellular automata (CA) have demonstrated different levels of success in various case studies; but, their shortcomings limit their effectiveness in urban expansion and LULC change modeling (Batty et al.; Clarke et al.; de Noronha Vaz et al.; Feng et al.; Wu and Martin). In recent years, machine learning methods, as an important part of artificial intelligence, have been taken researchers' attention in related studies because they are learning algorithms that implement the modeling process automatically without human assistance or expertise (Suthaharan). The increased use of non-parametric and supervised machine learning models in urban growth and LULC change studies is because of their effectiveness and reliability, which have been proven by most previous studies (Aburas et al.). These machine learning methods address both continuous and categorical variables, non-linear relationships, noisy and complicated data, and the existence of outliers in the training dataset, also avoid overfitting and ensure good generalization performance (B. Huang et al. "Support Vector Machines for Urban Growth Modeling"; B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines"). In urban growth studies, these models are assessed for their effectiveness and, there is still the debate on the effectiveness and reliability of the

machine learning methods for modeling, simulating, and predicting future urban growth patterns. In addition, these machine learning methods can be coupled with dynamic models such as CA to predict alternative future patterns of urban growth (Shafizadeh-Moghadam et al.).

1.2 Research Objectives

This doctoral dissertation research aims to model urban expansion patterns more effectively and efficiently by using machine learning methods. Very specifically this dissertation intends to address the following three research questions using a mixture of historical LULC maps, social, and physical data:

1. How can machine learning improve urban expansion modeling?
2. What is the best machine learning method to model urban expansion patterns?
3. How can urban expansion modeling support environmentally sustainable urban development?

1.3 Synopsis of Dissertation

The dissertation is organized as follows: Chapter I is an introduction to the study and outlines the research problem and objectives. Chapter II explores the capabilities of the support vector machine (SVM) method with the emphasis on the process of machine learning-based urban expansion prediction and evaluating the importance of causal factors. Chapter III examines five machine learning methods to thoroughly understand the performance of the models and their strengths and limitations. Chapter IV examines urban growth models to simulate urban development under three urban growth scenarios,

including current trends, controlled urban development, and environmentally sustainable urban development. Chapter V draws the conclusions, indicates the limitations, and provides potential future research.

CHAPTER II

AN ENHANCED SUPPORT VECTOR MACHINE MODEL FOR URBAN EXPANSION PREDICTION[1]

2.1 Introduction

Urban growth is a dynamic process (Bruegmann; Sultana and Weber "The Nature of Urban Growth and the Commuting Transition: Endless Sprawl or a Growth Wave?") resulting in the expansion of urban areas into surrounding natural areas (Adams; Blumenfeld; Puertas et al.; Wehrwein). The adverse effects of urban expansion such as reducing natural areas and habitat fragmentation (Nagendra et al.), increasing air, water, and soil pollutions (Shukla and Parikh), exacerbating hydrological problems (Lindh; Williams), destroying forests and agricultural fields (Masri; Yankson and Gough), disturbing natural and wildlife (Marzluff; McKinney), and intensifying regional and global warming (Alcoforado and Andrade; Stone Jr The City and the Coming Climate: Climate Change in the Places We Live) are great concerns among researchers, practitioners, and decision-makers (Agarwal et al.; P.H. Verburg et al.). That is why there is always a major interest to study and understand the processes of urban expansion from

[1] Karimi, F., Sultana, S., Babakan, A. S., & Suthaharan, S. (2019). An enhanced support vector machine model for urban expansion prediction. *Computers, Environment and Urban Systems*, *75*, 61-75. https://doi.org/10.1016/j.compenvurbsys.2019.01.001

many perspectives (Gober and Burns; Greene; Hart; Theobald) and how it influences the physical environment (Swenson and Franklin). Likewise, modeling and simulating urban expansion patterns have been a long tradition in geography and planning fields, which not only allow to assess the efficiency of a plan before implementing it but also help to forecast its consequences after implementation (Batty et al.; Clarke et al.; de Noronha Vaz et al.; Feng et al.; Wu and Martin). Hence, finding an appropriate method for modeling and simulating urban expansion becomes critically important for managing sustainable urban development (Hersperger et al.; Turner et al.).

It is undeniable that urban expansion modeling is a convoluted process requiring a deep historical understanding of urban growth and policies in the particular geographic context in order to meticulously perceive spatiotemporal relationships between predictor variables and land use/land cover (LULC) change (Clarke et al.; Pijanowski et al. "Using Neural Networks and Gis to Forecast Land Use Changes: A Land Transformation Model"). Over the last three decades, a variety of models and methods have been developed to understand the complexity of urban growth processes (Aburasa et al.; Musa et al.). The integration of geographic information system (GIS) in modeling urban expansion became essential in the late 20th century (Batty et al.; Michalak) for capturing the spatiotemporal changes of predictor variables. Various urban growth and LULC change models including cellular automata (CA) (Batty et al.; Clarke et al.; de Noronha Vaz et al.; Feng et al.; Wu and Martin), regression models (Z. Hu and C.P. Lo; Liao and Wei; Mom and Ongsomwang; Tahami et al. "Virtual Spatial Diversity Antenna for Gnss Based Mobile Positioning in the Harsh Environments"; Tahami et al. "The Preliminary

Study on the Prediction of a Hurricane Path by Gnss Derived Pwv Analysis"; Tayyebi et al. "Predicting the Expansion of an Urban Boundary Using Spatial Logistic Regression and Hybrid Raster-Vector Routines with Remote Sensing and Gis"), artificial neural networks (ANNs) (Mohammady and Delavar; Pijanowski et al. "Using Neural Networks and Gis to Forecast Land Use Changes: A Land Transformation Model"; Pijanowski et al. "A Land Transformation Model: Integrating Policy, Socioeconomics and Environmental Drivers Using a Geographic Information System"; Pourebrahim et al.; Tayyebi et al. "An Urban Growth Boundary Model Using Neural Networks, Gis and Radial Parameterization: An Application to Tehran, Iran"; Tian et al.), agent-based models (ABMs) (Babakan and Taleai; Hosseinali et al.; J. Li et al.; Murray-Rust et al.; Shirzadi Babakan and Alimohammadi; Shirzadi Babakan et al.), and tree-based models (Shafizadeh-Moghadam et al.; Tayyebi and Pijanowski) have demonstrated significant accomplishments in various case studies, yet their drawbacks limit their efficiency in urban expansion modeling (Musa et al.).

In recent years, support vector machine (SVM), one of the most effective Machine Learning (ML) techniques, has been attracted the attention of many researchers in geospatial analysis. While in a majority of geospatial studies, SVM has been applied to the classification of remotely sensed data (C. Huang et al.; Huang and Zhang; Muñoz-Marí et al.), it has recently been used for modeling LULC changes (B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines"; Samard zic´ -Petrovic´ et al.; Samardži´c-Petrovi´c et al.). SVM is not only able to effectively consider both continuous and categorical variables, non-normal distributed data, non-

linear relationships, noisy and complex data, and training datasets with outliers, it can also avoid overfitting and ensures good generalization performance (B. Huang et al. "Support Vector Machines for Urban Growth Modeling"; B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines"). These characteristics of SVM can be very useful for modeling urban expansion patterns. B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines" is the first scholar proposed the application of SVM method for LULC modeling and developed an SVM model to address the issue of dealing with unbalanced LULC data. They considered 9 causal factors including population, distance to roads and facilities, and surrounding land uses to model LULC changes in Calgary, Canada over the periods of 1985-90, 1990-92, 1992-99, 1999-2000, and 2000-01 and found that the unbalanced SVM can achieve high and reliable results for LULC change modeling. However, they trained and evaluated their model in the same period which considerably leads to obtaining overestimated and biased results. In other words, there is no assurance to achieve such high-accurate results if a model is developed in one period and applied to predict LULC changes in the next period. To realistically assess the predictive power of a model, it is essential to train and evaluate the model in consecutive periods. Moreover, in their evaluations, they did not consider the capability of the model for predicting unchanged land cells.

In another similar study, B. Huang et al. "Support Vector Machines for Urban Growth Modeling" used SVM to model urban expansion of New Castle County, Delaware, during 1984-1992, 1992-1997, and 1997-2002. They compared the results of

SVM with the results of a binomial logistic regression (BLR) model and demonstrated the better performance of SVM. Rienow and Goetzke integrated the SLEUTH model and SVM to compare the results with a combination of SLEUTH and BLR models. Their results showed that the performance of SLEUTH increases by coupling with BLR-based or SVM-based probability maps. They also demonstrated that not only the SVM approach requires fewer variables than the BLR model but also exhibits lower uncertainty. Samard zic´ -Petrovic´ et al. used a balanced data sampling method as a solution for addressing the issue of handling unbalanced datasets in LULC change modeling. They developed a SVM model for LULC change modeling in the Municipality of Zemun, Republic of Serbia, using LULC data in the years 2001, 2003, 2007, and 2011. Although unlike B. Huang et al. (2009, 2010), they trained and evaluated their model in two separate time intervals, they just used balanced sampled datasets for this purpose. Additionally, their approach did not examine various kernel functions; instead, the RBF function was used as the standard kernel function. In a recent study, Samardži´c-Petrovi´c et al. examined the effectiveness of three ML techniques including Decision Trees (DT), Neural Networks (NN), and SVM for land-use change modeling. They applied these techniques to the same case study in three urban districts in Belgrade, the capital of Republic of Serbia, using historical LULC data sets comprised of nine land-use classes. Their results indicated that all three ML techniques can be effectively used for short-term land-use change forecasting, but the SVM model showed the highest prediction accuracy. Yet, utilization of SVM method in urban modeling is still at infancy.

The major question of this study is whether utilizing a proper sampling strategy and regulating SVM can improve the accuracy and predictability of urban expansion modeling. To answer this question, an appropriate SVM-based urban expansion model is developed by investigating a variety of sampling approaches, predictor variables, kernel functions, and SVM parameters in Guilford County, NC, over the period of 2001-2011. The performance and prediction accuracy of the model is also evaluated by apposite evaluation metrics particularly developed for LULC change case studies. This study contributes to the literature in several ways. First, the effectiveness of three sampling methods including random sampling (Cheng and Masser), balanced sampling (Marjanović et al.), and sampling all the changed cells, developed for the first time in this study, are employed to create an appropriate sample training dataset. The effects of these sampling methods on the performance of the SVM-based urban expansion model are examined. Second, nineteen predictor variables, four of which are first introduced in this study, are classified into three main categories of proximity, neighborhood, and site-specific characteristics. Then, a comprehensive combination of the most significant predictor variables is determined using an information gain metric defined based on the entropy concept (Shannon). Third, by simultaneously regulating the SVM's penalty parameter, kernel function, and kernel's parameter, various configurations of the model are evaluated to find the most efficient configuration of SVM-based urban expansion model. Fourth, novel goodness-of-fit metrics are proposed to specifically evaluate the performance of SVM model for LULC change modeling. Finally, to achieve more realistic and reliable results, contrary to most previous studies, the predictability and

performance accuracy of the model is evaluated in the entire study area over a separate period.

The rest of this paper is organized as follows: first, SVM-based urban expansion modeling is elucidated. Then the data and methodology including the study area, predictor variables, data collection and preparation, data sampling, and model development are elaborated. Finally, experimental results together with a number of implications for future studies are presented.

2.2 A Background on SVM-Based Urban Expansion Modeling

Urban expansion is probably the most conspicuous indication of LULC change induced by human (B. Huang et al. "Support Vector Machines for Urban Growth Modeling") and mostly occurs at the fringe of an urban area (Sultana and Weber "The Nature of Urban Growth and the Commuting Transition: Endless Sprawl or a Growth Wave?") where lands are converted from their previous LULC to urban land use. There are various geospatial factors affecting this complex process in a non-linear way that is regarded as data layers in a GIS-based urban expansion modeling. GIS plays a prominent role in preparing, managing, analyzing, and presenting geospatial data layers. Most of the GIS-based urban expansion models are cell-based; namely, data layers are presented as grids of cells in which each cell representing an area with specified attributes. The goal of urban expansion modeling is to model LULC type at time *t+1* according to LULC and other characteristics at time t. That is, an appropriate function $f(x^t,y^{t+1})$ should be found to model the most probable LULC class at the next time (y_i^{t+1}) for a cell at the previous time (x_i^t). Afterward, the effectiveness and prediction accuracy of the model should be

evaluated for the next time interval (x^{t+1},y^{t+2})(Samard zic´ -Petrovic´ et al.). If the model reflects LULC changes correctly and the past LULC change patterns persist in the study area, it can be applied to simulate LULC in the future (x^{t+2},y^{t+3}). In other words, although the past LULC change patterns are informative for simulation of LULC changes in future, there is no guarantee that patterns occurred in the past will be replicated in future. The aforementioned function can be SVM, a supervised non-parametric ML technique, which uses binary LULC classification and past LULC change patterns to train the model (Suthaharan).

SVM was first proposed by Vapnik and Lerner, and then, Boser et al. enhanced it by inspiration from statistical learning theory. The standard SVM technique was introduced as a binary classification tool, but it can be upgraded to an n-class classification method by regarding a sequence of n or n(n-1)/2 binary classifications (Belousov et al.). SVM projects input data into the Hilbert space where an optimal separating hyperplane is used for classification (Yang et al.). By maximizing the hyperplane separating the two classes, binary SVM minimizes the upper bound of generalization error (B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines"; C. Huang et al.; Samard zic´ -Petrovic´ et al.). This capability can be regarded as the approximate implementation of Structural Risk Minimization (SRM), which grants SVM a good generalization performance, independent of the distribution of data.

A training dataset consisting of n data points that are separable into two classes can be represented by ˝Yسي ὼ ȟώ ȟὼ. ȟώ. ȟȣ ȟὼ ȟώ ᴺ ὢعي ὢȟὢ ˙Yᶜȟὢ ρȟطρ.

The goal of the classification method is to find a classifier y = f(x), which is a projection from X to Y based on data in T. Any points outside the training set T but inside ὣعيὣ will be classified correctly by the defined projection (Vapnik). Suppose the problem of separating the training set into two classes of +1 and -1, ῎Ọ ɖʯ ȣὼطى ὣسيطى ρand ῎Ọ. ɖʯ ȣὼطى ὣسيطىɋ Ø ᴺ ˙Ƴᶜ ȟὣ ᴺ ˙Ƴ would be possible hyperplanes such that the majority of class 1 instances lie above H1 (ύx+b> 1) and the majority of class -1 fall below H2 (ύx+b <-1), where the points located on H1 and H2 are defined as support vectors and are responsible for determining the optimal separating hyperplane H:w.x+b=0 (Statnikov) (Figure 2.1). The distance between H1 and H2 can be denoted by $\frac{\cdot}{ᶊ ᶋ}$; therefore, the maximization of the distance between H1 and H2 can be obtained by minimizing the norm of w, leading to a constrained optimization problem. As Figure 2.1 shows, all training sample data may not be linearly separated by a hyperplane perfectly. To consider misclassification errors, a penalty parameter c for the instances falling off the margin, and also nonnegative slack variables ξi are incorporated into the problem. Slack variables represent the distances between the misclassified points and the initial hyperplane. The penalty parameter c makes a trade-off between the margin size and the number of misclassified training points; whereas larger c provides smaller misclassifications, which also leads to smaller margin size. As a result, it is a constrained optimization problem, a quadratic programming problem with inequality constraints, which is presented by Eq. 1 (Vapnik):

Minimize ῝Ω ύ ȟ, سي-̣̊ ǼύǼ طى ὣΒϛ ,

Subject to ώ ύ ȣὺ طى ώ طى , طي ρ جي π ȟ , جي ρȟΏ سي ρȟȣ ȟὲ (1)

Where ụ are the positive slack variables and c is the penalty parameter. However, the goal is to find an optimal hyperplane to minimize the misclassification errors and maximize the margin size simultaneously. The most common way to deal with such problems, which is hard to solve directly, is the use of Lagrange multipliers to transfer the problem from the primal space to a dual space. Introducing n nonnegative Lagrange multipliers a1,a2,…, an ≥ 0 associated with the inequality constraints defined in Eq. 1 results in Eq. 2 (Vapnik):

Maximize ΰ ώ سي Bζ طي– Bζ Bζ ώώὼȣὺ

Subject to: Bζ ώ سي πȟπ جى ώȟΏ سي ρȟȣ ȟὲ (2)

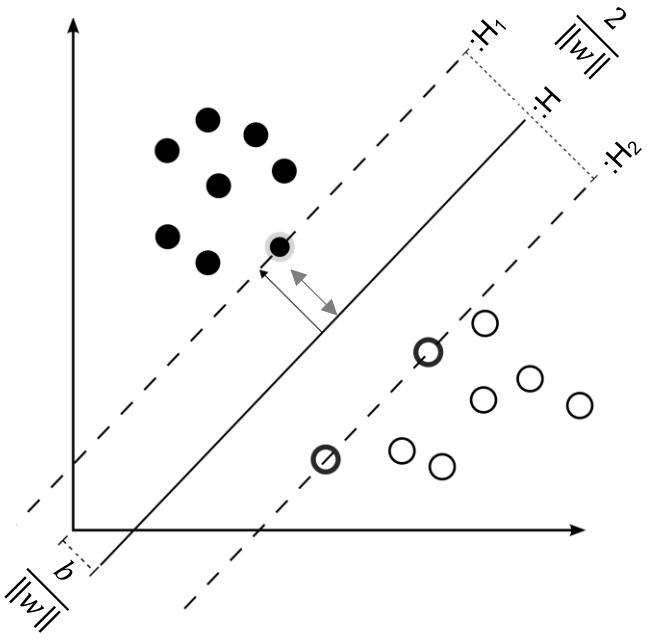


Figure 2.1: A Linear, Binary SVM Classifier and the Optimal Separating Hyperplane H, Lying Between and Parallel with H1 and H2

To address non-linearity, data can be mapped to a higher dimensional space created using a mathematical projection and known as the kernel trick (Statnikov). Because in this optimization problem, only the dot product of two vectors appears in the feature space, by replacing x with its mapping in the feature space, the kernel function k can be defined as $k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$. Using a kernel function, the optimization function accounts to maximizing Eq. 3 (Chapelle et al.):

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \quad (3)$$

Where common kernel functions are the linear function $k(x_i, x_j) = x_i \cdot x_j$ radial basis function (RBF) $k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, and polynomial function $k(x_i, x_j) = (x_i \cdot x_j + 1)^q$, where γ and q are kernel parameters (Chapelle et al.).

2.3 Data and Methodology

2.3.1 Study Area

In this study, the developed SVM model is applied to model urban expansion in Guilford County, North Carolina (NC), USA, over the period of 2001-2011. Guilford County, located in the piedmont area and a part of the Metropolitan Statistical Area of Greensboro- High Point, is one of the ten most urban and third most populous County in NC (Carolina Population Center) (Figure 2.2). The estimated population of Guilford County was 421,048, 489,557, and 521,330 respectively in 2000, 2010, and 2016, making it one of the fastest-growing counties in NC (U.S. Census Bureau). The area of this county is about 170324 ha (U.S. Census Bureau), of which 31% was built lands and 69% was natural lands in 2001. The percentage of built lands in this county was grown 6% from 2001 to 2006, and 3% from 2006 to 2011 (USGS "The National Map"). Given NC is one of the fastest-growing states in the United States (U.S. Census Bureau), no doubt Guilford County will continue to grow (News and Records) and built areas will imperil the natural environment and underscore the necessity of studying urban expansion patterns in this county. Developing an efficacious urban expansion model enables urban planners and decision-makers in Guilford County to scrutinize LULC change patterns and make proper plans regarding the preservation of crucial natural lands and the management of urban expansion patterns.

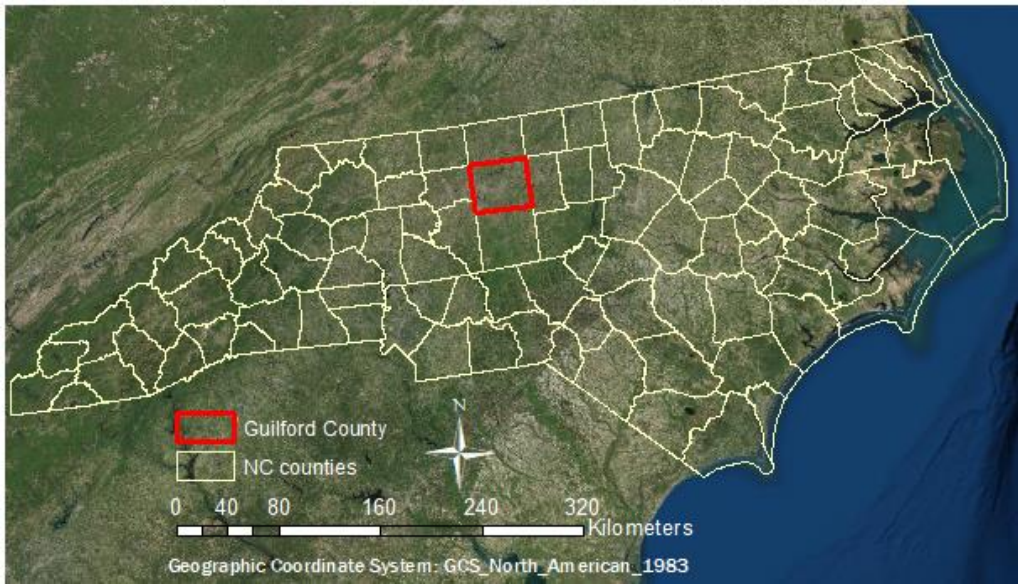


Figure 2.2: The Location of Guilford County in North Carolina

2.3.2 Predictor Variables

Urban expansion converts vacant or agricultural lands into urban built lands (del Mar López et al.). It is evident that such LULC changes are influenced by an intricate combination of various social, economic, and environmental factors and cannot be simplistically attributed to any single group of factors (Bhatta). The relationship between these factors and LULC change patterns can be illustrated by urban expansion models. According to the summary of the most widely used factors in previous studies conducted by Musa et al., the historical spatiotemporal LULC change patterns in the study area, and data availability, 19 predictor variables are considered under three main categories to model urban expansion in this study (Table 2.1):

1) Site-specific characteristics: One of the significant factors considered in this category is the population density of a cell which leads to striking LULC changes (Meyer

and Turner). Another factor in this category is the current LULC type of a cell which considerably affects its future LULC type (Samard zic´ -Petrovic´ et al.).

2) Proximity characteristics: proximity measures the minimum Euclidean distances to various influential factors (Rienow and Goetzke) on LULC change such as the nearest city center or downtown area, urban facilities, transportation network infrastructures (road/railway), green spaces, and water bodies (Musa et al.). Because of more employment opportunities at downtowns and city centers, proximity to city centers is a consequential factor in LULC change modeling (Deng and Srinivasan). Proximity to urban areas is also reported as an important factor in urban expansion; closest lands to urban areas have more potential for urbanization due to less monetary costs required for connecting to urban utilities such as water and sewer (Pijanowski et al. "Calibrating a Neural Network-Based Urban Change Model for Two Metropolitan Areas of the Upper Midwest of the United States"). Because of less commuting costs and more mobility, new residential areas predominantly emerge in proximity to transportation networks (Cervero and Landis; Horner and Schleith; H. Kim et al.); therefore, proximity to transportation facilities such as roads and railways is a driving factor for LULC change (Babakan and Taleai; Shirzadi et al.; Shirzadi Babakan and Alimohammadi; Shirzadi Babakan et al.). Proximity to green spaces and water bodies are other momentous variables in LULC change studies; while they play a restricting role in urban expansion due to imposing some limitations for building new residential properties, they are considered a boon by inhabitants due to their benefits on mental and physical health and providing a suitable place to support various social and recreational activities (Dadvand et al.; Pijanowski et

al. "Calibrating a Neural Network-Based Urban Change Model for Two Metropolitan Areas of the Upper Midwest of the United States").

3) Neighborhood characteristics: In addition to its own LULC type, the neighboring lands' LULC types of a cell have substantial effects on its future LULC type. In general, the potential of changing a cell's LULC to the built type increases when its neighborhood is mostly developed. On the other hand, if a majority of neighboring lands of a cell are undeveloped, the probability of LULC change in that cell will decrease (P. H. Verburg et al.; White and Engelen). In this study, the numbers of a variety of developed and undeveloped LULC types such as wetland, forest, water, barren, built lands with different densities, and potential lands for urban expansion including agricultural lands, shrubs, herbaceous and pastures are calculated in a 3*3 Moore's neighborhood of each cell and used to enhance the efficiency of LULC change modeling.

Irrelevant and redundant predictor variables affect the modeling results through overfitting and poor generalization (Y. Kim et al.). Feature selection is a pre-processing stage to determine the most significant predictor variables (Chandrashekar and Sahin; H. Peng et al.) and information gain is a frequently used supervised feature selection algorithm in ML classification studies (Azhagusundari and Thanamani; Lee and Lee; Yang and Pedersen). Information gain measures how much a predictor variable is important and relevant to the target variable (Yang and Pedersen). The information gain function originated from information theory (Shannon) and it is based on the notion of entropy. Entropy indicates the uniformity of the system, the more chaotic, the higher the value of entropy (Yang and Pedersen). Entropy is calculated using Eq. 4 (Shannon):

$$Entropy(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \qquad (4)$$

Where pi is the proportion of instances belonging to the target variable m in the dataset D. Information gain of a predictor variable is the reduction in the entropy that is archived by that variable.

Table 2.1: A Summary of Considered Predictor Variables for Urban Expansion Modeling

Predictor Category	Predictor Variable
Site-specific characteristics	The current LULC type of a cell
	The population density of a cell
Proximity characteristics	The distance to city centers
	The distance to urban built areas
	The distance to highways
	The distance to major roads
	The distance to streets
	The distance to railroads
	The distance to water bodies
	The distance to greens spaces
Neighborhood characteristics	The number of potential lands for urban expansion including agricultural lands, shrubs, herbaceous and pastures
	The number of water body cells in the neighborhood of a cell
	The number of forest cells in the neighborhood of a cell
	The number of wetlands cells in the neighborhood of a cell
	The number of barren land cells in the neighborhood of a cell
	The number of developed, open-space cells in the neighborhood of a cell
	The number of low-intensity developed cells in the neighborhood of a cell
	The number of medium-intensity developed cells in the neighborhood of a cell
	The number of high-intensity developed cells in the neighborhood of a cell

2.3.3 Data Collection and Preparation

As the necessity of using historical LULC data in urban expansion modeling, Guilford County's LULC data were collected from the National Land Cover Database (USGS "The National Map") at the spatial resolution of 30 meters for the years of 2001 (C. Homer et al.), 2006 (Fry et al.), and 2011 (C.G. Homer et al.) (Figure 2.3) with the overall accuracy of 79% (J. D. Wickham et al. "Thematic Accuracy of the Nlcd 2001

Land Cover for the Conterminous United States"), 78% (J. D. Wickham et al. "Accuracy Assessment of Nlcd 2006 Land Cover and Impervious Surface") and 83% (J. Wickham et al.), respectively. In these LULC maps, the built environment is classified into developed open spaces, low-intensity, medium-intensity, and high-intensity developed areas and the natural or unbuilt environment includes the classes of open water, barren land, deciduous forest, evergreen forest, mixed forest, shrub and scrub, herbaceous, and hay and pasture. As shown in Figure 2.3, the urban expansion mostly happened along the boundary of existing urban areas; therefore, the proximity to urban facilities seems to play a major role in urban expansion.

In addition to LULC data, vector data of transportation networks were collected from TIGER files (U.S. Census Bureau "Tiger/Line Shapefiles and Tiger/Line Files") and prepared for Guilford County over the years of 2001, 2006, and 2011. The vector data of built areas, city centers, green spaces, and water bodies are extracted from LULC maps and used to produce the required proximity raster maps. As shown in Figure 2.4, the proximity maps display the Euclidian distance of each cell to the closest facility of interest as grayscale images where darker gray values represent shorter distances from the facility. Furthermore, the neighborhood raster maps displaying the number of cells with the LULC type of interest within the 3*3 Moore's neighborhood of each cell are produced using ESRI ArcGIS 10.3 software as grayscale images where lighter gray values represent the higher number of neighboring cells (Figure 2.5). Population data were gathered from (U.S. Census Bureau) at the scale of census tracts for the decennial census years of 2000 and 2010. Because of the unavailability of census population data

for 2001, 2006, and 2011, a simple population estimation model is used to approximate population data in these years from the available decennial census data in 2000 and 2010. Finally, a population density map representing the number of inhabitants per cell is produced as a grayscale image where lighter gray values show higher population density (Figure 2.6).

To train and evaluate a binary SVM-based LULC change model, the prepared data layers are used as predictor variables at time t to model the binary classification map of LULC including the two classes of built and unbuilt at time t+1. In other words, at first, the 2001 data layers are considered as predictor variables, and a binary LULC classification map of the two classes of built and unbuilt in 2006 is used to train the model; then the 2006 data layers and the binary classification map of LULC in 2011 are utilized as a testing dataset to evaluate the effectiveness and prediction accuracy of the model (Table 2.2).

Table 2.2: The Training and Testing Datasets

	Predictor variables	Label
Training Dataset	2001 data layers	binary classification of LULC in 2006
Testing Dataset	2006 data layers	binary classification of LULC in 2011

2.3.4 Data Sampling

The LULC change process neither occurs randomly over the whole study area nor uniformly among all types of LULC. While some LULC types such as wetlands and water bodies remain immutable for a very long time, a number of LULC types such as agricultural lands and pastures indicate more potential to change to an urban area in a

short time. Table 2.3 shows the number of urban built and unbuilt land cells over the total number of 1,892,491 land cells in the study area for the years 2001, 2006, and 2011. Table 2.4 provides the number of changed and unchanged land cells together with the rate of LULC changes and the rate of increase in urban built areas that have been occurred during the study period. Clearly, only LULC of a small portion of the study area has been changed over the period of 2001-2011 (Table 2.4), which may make the SVM model biased considering the fact that a majority of lands do not contribute to the LULC change process. This critical issue is punctiliously resolved by applying three sampling strategies to create a training dataset. First, the training dataset is created by a random selection of 5%, 10%, and 20% of the whole cells. Second, the training dataset is created by selecting all the changed cells. Finally, a balanced sampling strategy is applied to create the training dataset from all the changed cells and an equal number of unchanged cells randomly selected over the whole study area. The results of applying these sampling strategies and their efficiency for urban expansion modeling in Guilford County are discussed in detail in the results and discussion section.

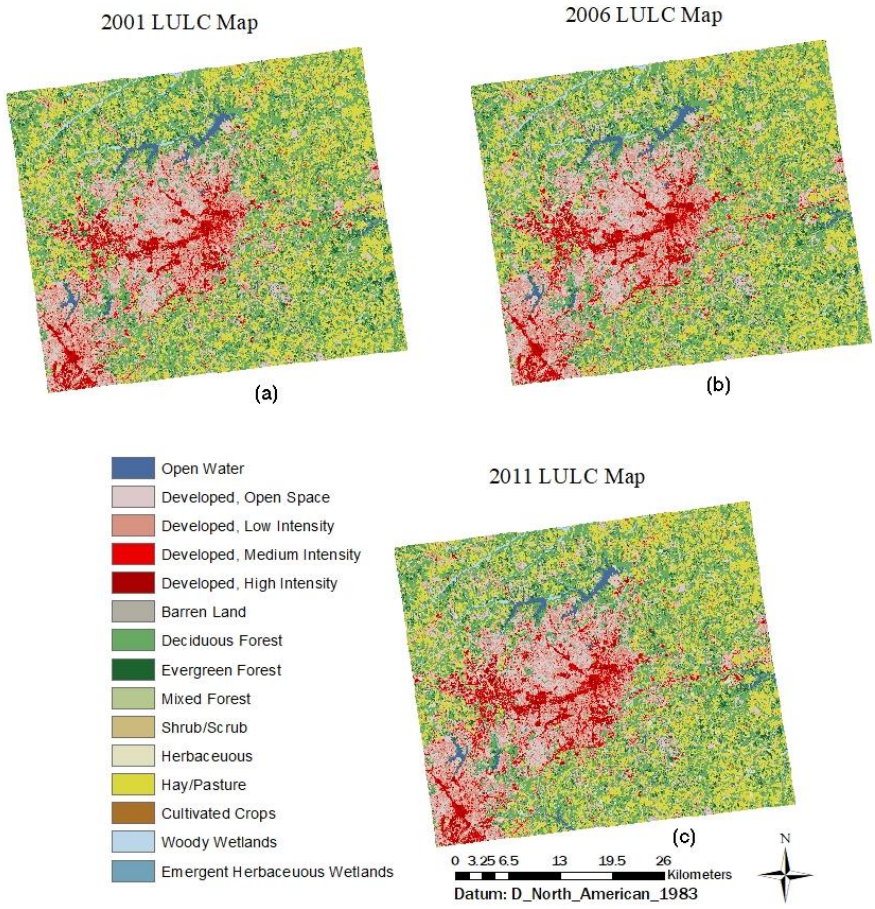


Figure 2.3: The LULC Map of Guilford County in (a) 2001, (b) 2006, and (c) 2011

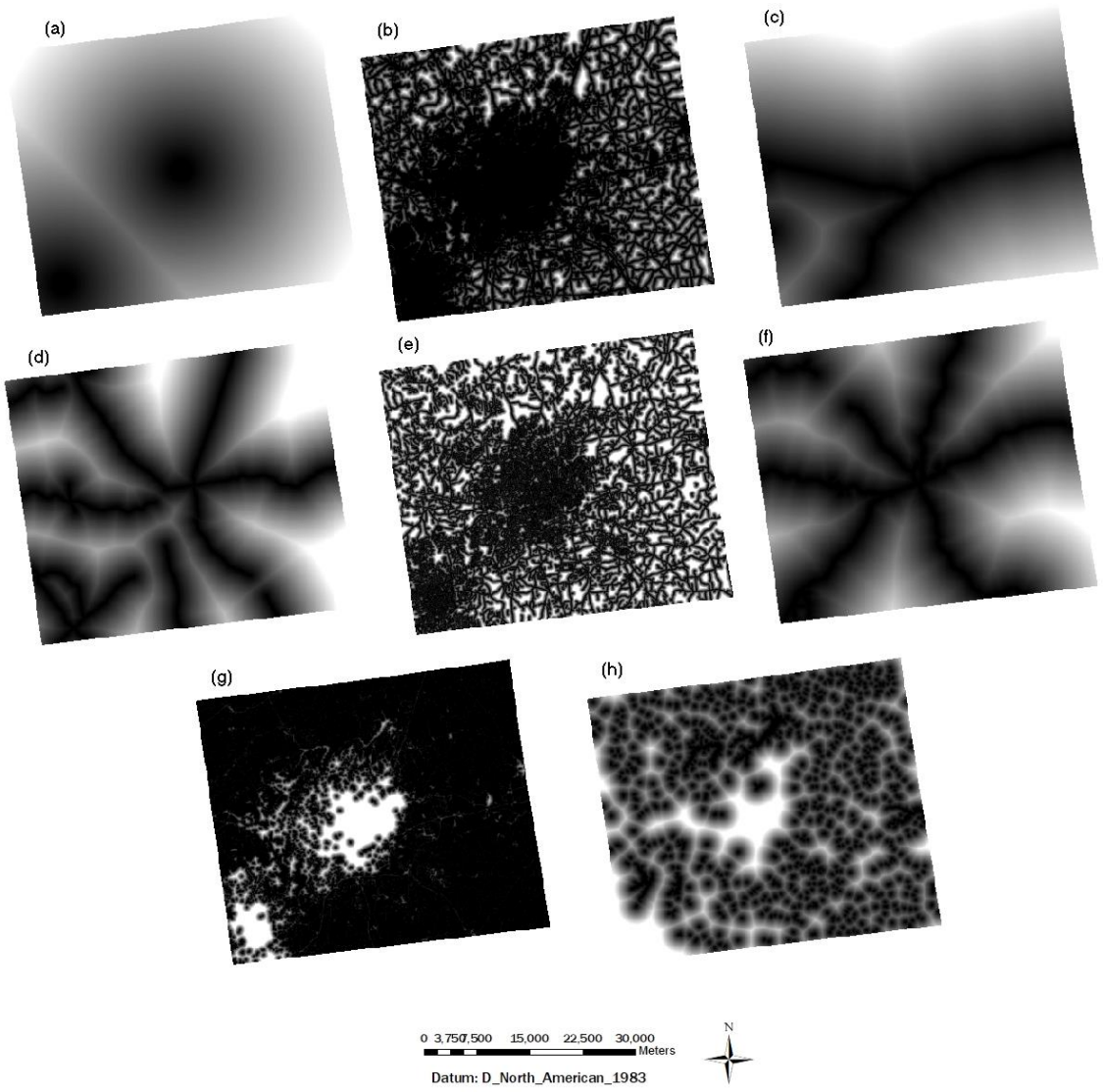


Figure 2.4: The Distance Map to a) City Centers, b) Built Areas, c) Highways, d) Main Roads, e) Streets, f) Railroads, g) Green Spaces, and h) Water Bodies in Guilford County in 2001.

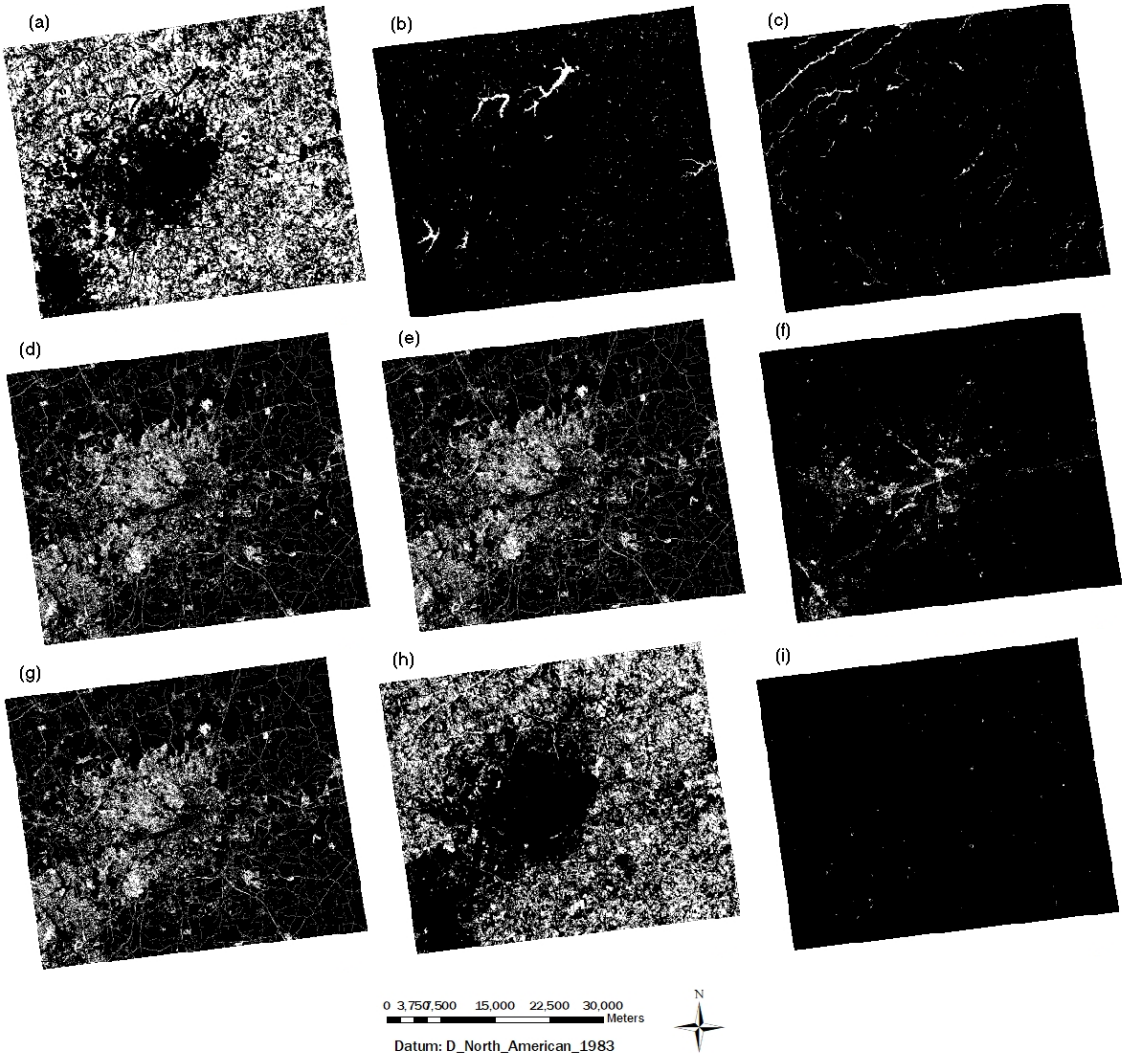


Figure 2.5: The Number of Neighboring Cells with the LULC Type of a) Forest, b) Water body, c) Wetland, d) Low-intensity Developed Area, e) Medium-intensity Developed Area, f) High-intensity Developed Area, g) Developed Open Space, h) Potential LULC for Urban Expansion Including Agricultural Land, Shrub, Herbaceous and Pasture, and i) Barren Land in Guilford County in 2001.

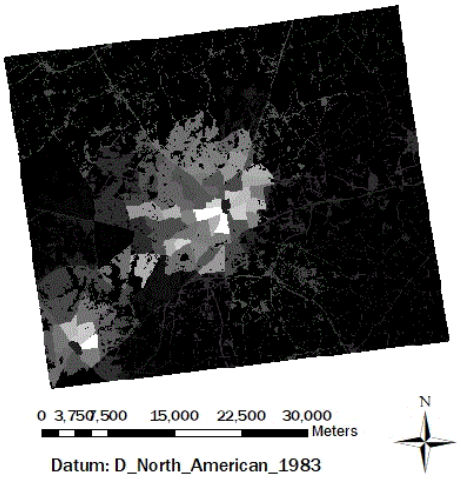


Figure 2.6: The Population Density Map in Guilford County in 2001

Table 2.3: The Number of Built and Unbuilt Land Cells in Guilford County for the Years 2001, 2006 and 2011

	2001	2006	2011
The number of built land cells	589,929	624,651	644,610
The number of unbuilt land cells	1,302,562	1,267,840	1,247,881

Table 2.4: A Summary of LULC Change in Guilford County Over 2001-2011

	2001-2006	2006-2011
The number of changed cells	34,722	19,959
The number of unchanged cells	1857769	1872532
The rate of change	2%	1%
The rate of increase in urban built areas	6%	3%

2.3.5. Model Implementation

The SVM-based urban expansion model is developed using MATLAB 2017. All the prepared raster data layers for Guilford County are converted to Ascii files to make them readable by MATLAB Software. To train the model, nineteen data layers are prepared for 2001 and the binary classification of the LULC map in 2006 are considered. As the study area is composed of 1,892,491 30 by 30-meter cells, the training dataset

includes 1,892,491 observations and 19 features in 2001, and 1,892,491 binary LULC labels (-1,1), 1 for built and -1 for unbuilt cells, in 2006. However, by applying a sampling method, not only the number of observations in the training dataset decreases, and subsequently the computational performance increases considerably, but also the prediction accuracy of the model is improved. For this purpose, different sampling strategies are implemented to recognize the best sampling strategy to create the training dataset. Finally, the cell values of features including proximity, neighborhood, and population density data layers are standardized (mean 0, variance 1), because the determination of optimal hyperplane in SVM is significantly influenced by the scale value of input features. Therefore, the standardization of feature values makes the significance of all the features equal in the SVM model. Similarly, the 2001 LULC type, which is used as a feature data layer in the training dataset, is converted to a dummy variable due to its categorical scale value.

The configuration of a SVM model including regularizing the parameter c, selecting a kernel function type, and regulating the kernel function's parameter has substantial effects on the model's performance and should be specifically considered for each case study. The regularization of parameter c is used to control the trade-off between the empirical risk and the model complexity. A larger c value leads to a more complex model that decreases the empirical risk and thus tends to overfit the training dataset by a decision surface which is more influenced by local support vectors. On the other hand, smaller c values produce smoother surfaces and result in simpler models that reduce the model complexity but may not effectively consider the underlying LULC

change patterns. As a consequence, an optimal c value should be identified to trade off the model complexity and the empirical risk in order to attain the best generalization performance. In this study, the c values of 0.1, 1, 10, and 100 are tested to achieve the best performance of the SVM model.

In addition to c value, the efficiency of a kernel function in converting the nonlinear boundary to a linear one profoundly influences the performance of a SVM model. In this study, the linear function, radial basis function (RBF), and polynomial function are tested to select the most suitable kernel function for addressing the nonlinearity issue in the SVM-based urban expansion modeling. While in the linear kernel, there is no parameter to be set, in the RBF kernel, the number of support vectors decreases by increasing the kernel's parameter γ. Therefore, when parameter c is kept constant, raising the value of parameter γ up to a certain threshold leads to a more complex model because the decision surface's shape is more influenced by local support vectors. In the polynomial kernel, increasing the kernel parameter q brings about a better generalization, but when it excessively increases, the performance of the model decreases due to overfitting the SVM model. In this study, the γ values of 1, 2, and 3 and the q values of 1, 2, and 3 are explored to determine the best values for these parameters. As a result, an efficacious SVM model is secured by regulating the configuration of SVM model. Figure 2.7 shows the entire process of SVM-based urban expansion modeling.

2.3.6. Model Evaluation

The developed SVM model is evaluated using various accuracy metrics. First, a training accuracy, the classification accuracy based on the training dataset (2001-2006

dataset), is evaluated to investigate the model's ability to demonstrate the existing LULC change patterns in the training dataset. The training accuracy evaluates the stability and generalization performance of the model (Suthaharan). In the next step, to validate the applicability of the model, its performance is tested using a dataset related to another period (2006-2011 dataset). The testing performance including the computation of testing accuracy, precision, sensitivity, and specificity is performed using a confusion matrix (Suthaharan). The definition of the confusion matrix is presented in Tables 2.5 and 2.6. In this study, two confusion matrices are produced using built and unbuilt, and also, changed and unchanged cells. In the confusion matrix based on built and unbuilt cells, if a cell with the real-world label of built is correctly classified as built by the model, then it calls a True Positive (TP) and if it is incorrectly classified as unbuilt, then it calls a False Negative (FN). Similarly, if a cell whose real label is unbuilt is incorrectly classified as built by the model, it calls False Positive (FP) and if it is correctly classified as unbuilt, then it calls True Negative (TN) (Table 2.5). In a similar manner, a confusion matrix based on changed and unchanged cells is specifically designed to investigate the testing performance of the LULC change model for the first time in this study (Table 2.6).

Table 2.5: The Built-unbuilt Confusion Matrix

predicted / observed	built	unbuilt
built	True positive	False negative
unbuilt	False positive	True negative

Table 2.6: The Changed-unchanged Confusion Matrix

observed \ predicted	changed	unchanged
changed	True positive	False negative
unchanged	False positive	True negative

After the creation of confusion matrices, the testing accuracy, precision, sensitivity, and specificity are calculated using the following equations (Suthaharan):

$$\text{Testing accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

$$\text{Testing precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{Testing sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

$$\text{Testing specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (8)$$

Testing accuracy shows the performance of the model based on the proportionality between the FP and the TP. A high testing accuracy indicates that the classification is highly accurate, and therefore, the classification errors of FN and FP are negligible. Testing precision exhibits the performance of the model based on the proportionality between FP and TP. A high precision indicates that TP is high together with low values of FN. Testing sensitivity represents the performance of the model regarding the proportionality between FN and TP. A high sensitivity demonstrates that the classification of TP is highly sensitive to FP. Testing specificity presents the

performance of the model according to the proportionality between TN and FP. A high testing specificity means that while TN is high, FN is also significant (Suthaharan).

Because of the small number of changed cells in the study area, the aforementioned evaluations are not able to reveal the realistic performance of the model (Jantz et al.). In other words, the main objective of a SVM-based urban expansion model is not the classification of built and unbuilt cells but is the prediction of the LULC change process, change from unbuilt cells to built cells. Therefore, the model can be claimed to show high performance when it is able to accurately predict the LULC change process over time. To evaluate the realistic performance of the SVM-based urban expansion model, a novel combination of five evaluation metrics including the percentage of correctly predicted cells (PCP), the percentage of correctly predicted cells as built (PCPB), the percentage of correctly predicted cells as unbuilt (PCPUB), the percentage of correctly predicted cells as unchanged (PCPUC), and the percentage of correctly predicted cells as changed (PCPC) are defined in this study (Eq.8-12). PCPC and PCPUC are the most important metrics for evaluating the performance of an urban expansion model because they consider the LULC change process.

$$PCP = \frac{\text{correctly predicted cells}}{\text{total cells}} \times 100 \quad (9)$$

$$PCPB = \frac{\text{correctly predicted built cells}}{\text{total built cells}} \times 100 \quad (10)$$

$$PCPUB = \frac{\text{correctly predicted unbuilt cells}}{\text{total unbuilt cells}} \times 100 \quad (11)$$

$$PCPUC = \frac{\text{correctly predicted unchanged cells}}{\text{total unchanged cells}} \times 100 \quad (12)$$

PCPC= [illegible] (13)

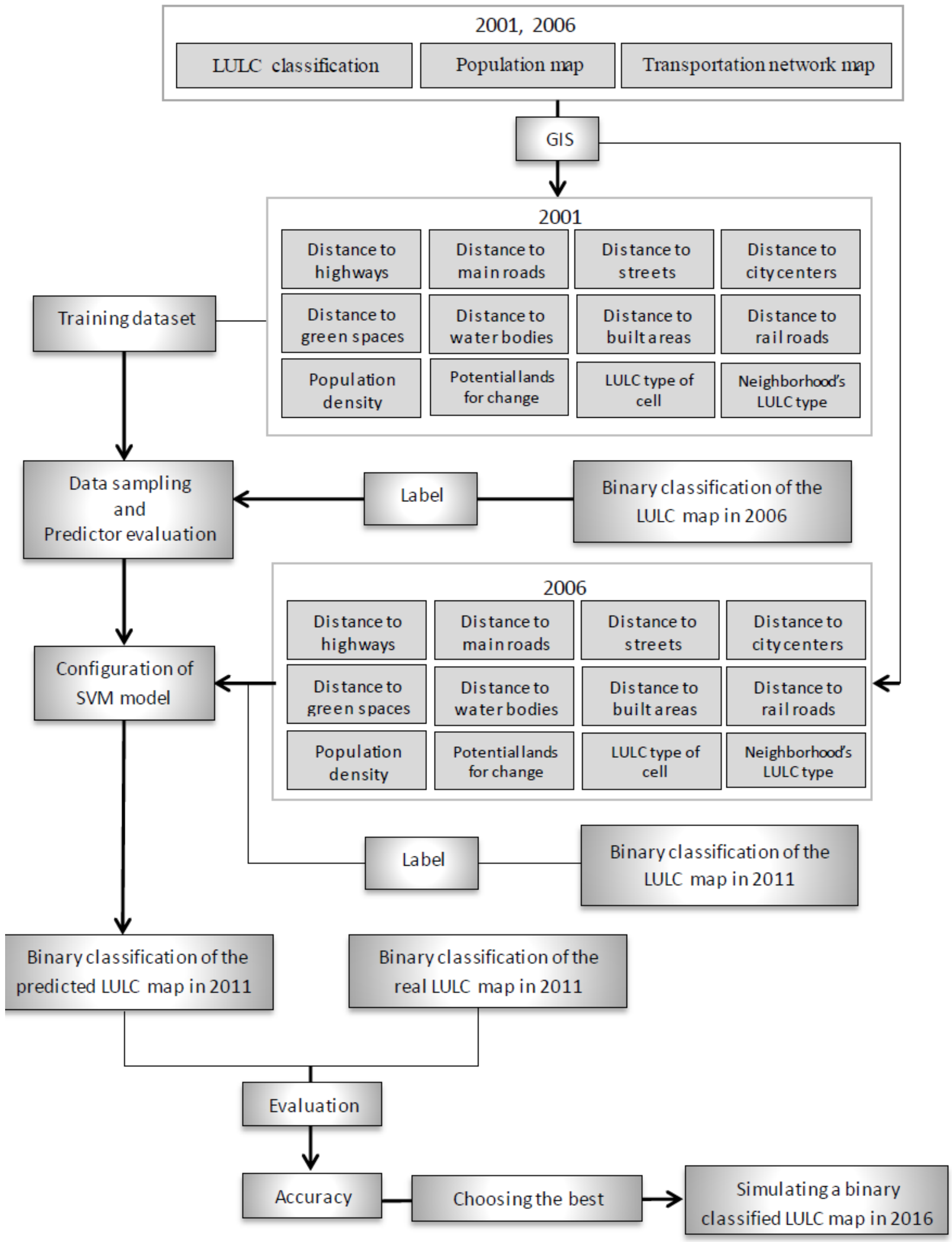


Figure 2.7: The Diagram of SVM-based Urban Expansion Modeling

2.4 Results and Discussion

The SVM model is developed using the value of parameter c equal to 1 and linear kernel. As mentioned before, the major evaluation metrics for choosing the best sampling method are PCPUC and PCPC; therefore, the values of these metrics obtained by applying three sampling methods are presented in Table 2.7. As illustrated in Table 2.7, the SVM model based on the sampling of all changed cells and random sampling does not predict changed and unchanged cells properly. In the sampling of all changed cells, because only the changed cells are considered for training the model, the ability of the model to predict unchanged cells is very low. On the other hand, in random sampling, the probability of taking into account changed cells is significantly low due to the small number of changed cells compared to the number of unchanged cells; hence, the ability of the model to predict changed cells considerably decreases. As a result, the balanced sampling method, selecting all the changed cells and the same number of unchanged cells randomly is utilized to train the SVM model in all the remaining experiments. The balanced sampling just contains about 4% of the training dataset that means with a few numbers of cells, the SVM model can be trained efficiently. To evaluate the effects of stochasticity on the results, the model is run with all three sampling methods several times. The results show that the outcomes of LULC change prediction do not change significantly in different runs of the model.

Table 2.7: PCPC and PCPU Values for Three Sampling Methods

	Random sampling			Changed cells sampling	Balanced sampling
	5% of cells	10% of cells	20% of cells		
PCPC (%)	1	3	8	98	67
PCPUC (%)	100	100	100	1	38

After choosing the best sampling method, the next step is the evaluation of predictor variables. Selecting the most informative features by eliminating those with little significance from the model not only enhances the accuracy of the model but also reduces the model's complexity and the required time for training and testing the model (Y. Kim et al.). For this purpose, the predictor variables are ranked using the Information Gain metric (Shannon). As presented in Table 2.8, the current LULC type is the most significant predictor, following with distance to highways, neighboring with medium-intensity developed areas, neighboring with potential lands for urban expansion, and distance to water bodies as the next high-ranked predictors. Because of the low significance of neighboring with forests, neighboring with developed open spaces, neighboring with water bodies, neighboring with wetlands, and neighboring with barren lands, these predictor variables are not considered in the modeling process.

To improve the prediction accuracy, the model is configured by adjusting the parameter c, applying different kernel functions, and adjusting the kernel's parameter. For this purpose, a combination of the c values of 0.1, 1, 10, and 100 with various configurations of kernel functions including the linear kernel with no kernel parameter, the RBF kernel with the γ values of 1, 2, and 3, and the polynomial kernel with the q values of 1, 2, and 3 are tested to select the most efficient SVM model. The best model

should demonstrate a good balance between PCPC and PCPUC and acceptable values for other evaluation metrics. Figure 2.8 shows the training accuracy of all the configured SVM models for the period of 2001-2006. As illustrated in this diagram, the training accuracy resulted from the RBF kernel is conspicuously higher than other kernel functions. In general, excluding c = 0.1, by raising the values of c and the RBF kernel's parameter γ, the training accuracy increases. As a result, the training accuracy is approximately 100% in all the configurations of the model using the RBF kernel and the c values of 10 and 100. On the other hand, all the configurations of the model by polynomial kernel show low training accuracies. Finally, the linear kernel function presents a moderate training accuracy of about 75% for all the c values, except for c = 100 that the accuracy decreases to 60%.

Table 2.8: The Significance of Predictor Variables

Predictor variable	Rank	Weight
Current LULC type	1	0.21
Distance to highways	2	0.18
Neighboring with medium-intensity developed areas	3	0.17
Neighboring with potential lands for urban expansion	4	0.17
Distance to water bodies	5	0.16
Distance to streets	6	0.15
Distance to major roads	7	0.15
Distance to urban built areas	8	0.14
Distance to green spaces	9	0.11
Neighboring with low-intensity developed areas	10	0.10
Population density	11	0.08
Distance to city centers	12	0.08
Neighboring with high-intensity developed areas	13	0.05
Distance to railroads	14	0.02
Neighboring with forests	15	0.00
Neighboring with developed open spaces	16	0.00
Neighboring with water bodies	17	0.00
Neighboring with wetlands	18	0.00
Neighboring with barren lands	19	0.00

The testing accuracy, precision, sensitivity, and specificity of the developed SVM models are examined using both built-unbuilt and changed-unchanged confusion matrices (Figures 2.9 and 2.10). Similar to the training accuracy (Figure 2.8), the RBF kernel generally results in better testing performances than other kernel functions to predict both built-unbuilt and changed-unchanged cells. As shown in Figures 2.8 and 2.9, the RBF-based configurations present noticeably higher testing accuracy, precision, and specificity compared to other configurations of the model. In addition, while the testing sensitivity of predicting built-unbuilt cells is very high for all the configurations of the model, the RBF-based configurations show significantly better testing sensitivities for predicting changed-unchanged cells. Particularly, for all the c values, the testing sensitivity increases by raising γ values.

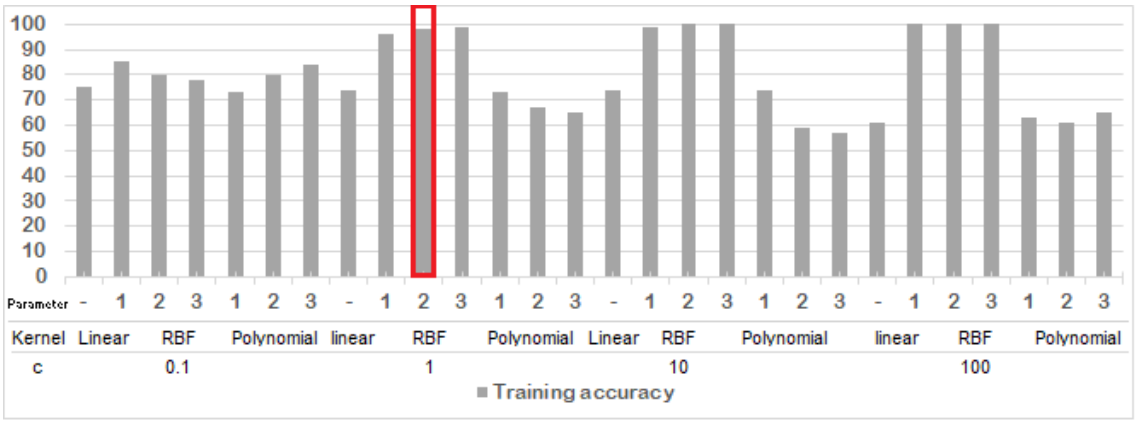


Figure 2.8: The Training Accuracy for Different Configurations of SVM Models (2001-2006)

In the prediction of built-unbuilt cells, besides the testing sensitivity which is remarkably higher than other metrics and changes a little around 100% in all the configurations of the model, the testing accuracy, precision, and specificity follow nearly

the same pattern. However, in all the configurations, the testing accuracy is higher than the testing precision and specificity (Figure 2.8). On the other hand, in the prediction of changed-unchanged cells, the testing accuracy and specificity are approximately equal and the testing precision slightly changes in all the configurations of the model (Figure 2.9).

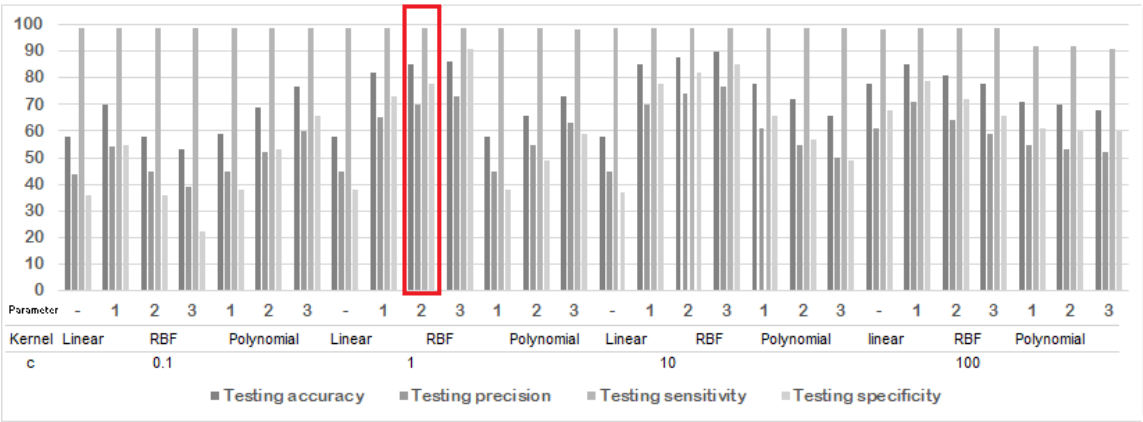


Figure 2.9: The Testing Accuracy, Precision, Sensitivity, and Specificity Using the Built-unbuilt Confusion Matrix

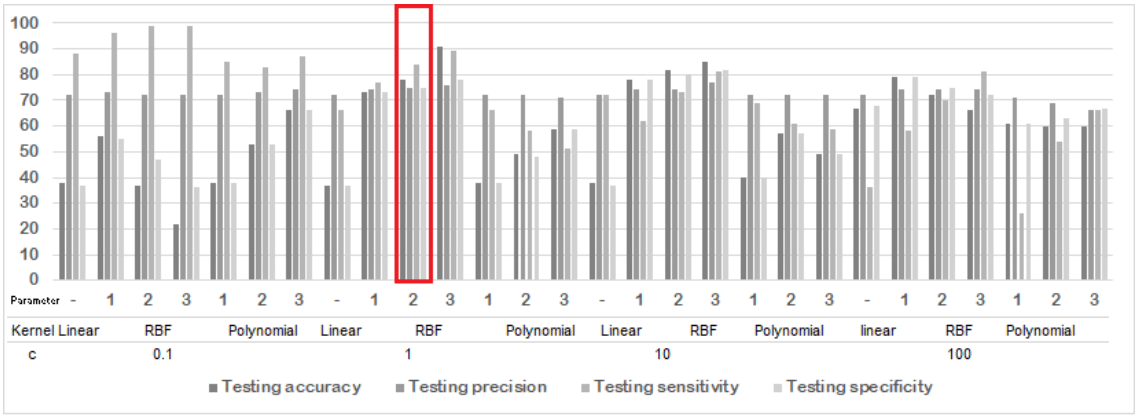


Figure 2.10: The Testing Accuracy, Precision, Sensitivity, and Specificity Using the Changed-unchanged Confusion Matrix

The PCP, PCPB, PCPUC, PCPUC, and PCPC evaluations of the developed SVM models show that similar to previous evaluations, the RBF kernel generally leads to more reliable results in comparison to other kernel functions (Figure 2.11). As indicated in Figure 2.10, excluding PCPB which is notably higher than other metrics and slightly changes around 100% in all the configurations of the model, PCP, PCPUB, PCPUC, and PCPC approximately pursue the same pattern. In addition, PCPUB and PCPUC are almost equal in all the configurations of the model. Also, by raising γ, while PCPC increases in all the RBF-based models, PCPUB and PCPUC decrease in all the RBF-based models except in the case of c=1.

Because of changing a small number of cells in the study area, PCPC is the most effective metric for the evaluation of change predictability of a SVM-based urban expansion model. However, other evaluation metrics are also consequential and should be taken into account to select the best model. Considering all the evaluation metrics, especially PCPC and the computational complexity of the models, the SVM model regularized by the c value of 1 and the RBF kernel function with the γ value of 2 is selected as the best SVM-based urban expansion model in Guilford County. The performance evaluations of the final SVM-based model are highlighted by red boxes in Figures 2.8, 2.9, 2.10, and 2.11. Also, Table 2.9 and 2.10 present the specifications and performance evaluation results of the final model.

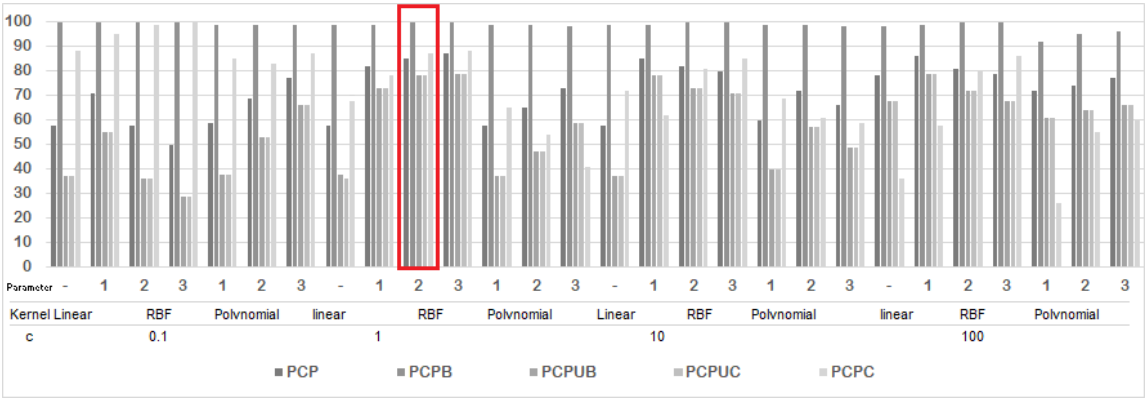


Figure 2.11: The Results of PCP, PCPB, PCPUB, PCPUC, and PCPC Evaluations of the Developed SVM models (2006, 2011)

Table 2.9: The Performance Evaluation Results of the Final Model

Performance evaluation		Value (%)
Training accuracy		98
built-unbuilt	Testing accuracy	85
	Testing precision	70
	Testing sensitivity	99
	Testing specificity	78
changed-unchanged	Testing accuracy	78
	Testing precision	76
	Testing sensitivity	89
	Testing specificity	78
PCP		85
PCPB		100
PCPUB		78
PCPUC		78
PCPC		87

Table 2.10: The Specifications of the Final Model

Sampling method	Balanced sampling
Selected predictor variables	Current LULC type, Distance to highways, Neighboring with medium-intensity developed areas, Neighboring with potential lands for urban expansion, Distance to water bodies, Distance to streets, Distance to major roads, Distance to urban built areas, Distance to green spaces, Neighboring with low-intensity developed areas, Population density, Distance to city centers, Neighboring with high-intensity developed areas, Distance to railroads
Penalty parameter c	1
Kernel function	RBF
Kernel parameter	2

To compare to the real binary classification map of LULC, the final SVM-based urban expansion model is applied to produce a binary classification map of LULC in 2011 (Figure 2.12). As illustrated in this Figure, the predicted LULC map is conspicuously compatible with the real LULC map that demonstrates the high efficiency of the model to predict urban expansion. After substantiating the model's predictability performance, the model can be applied to simulate urban expansion in the future assuming that the past urban expansion patterns will similarly continue in the future. For instance, in this study, the model is used to simulate urban expansion in 2016 (Figure 2.13) in the case that this data is not available. Since the NLCD LULC map of the study area in 2016 is not available, the binary LULC map can be useful in planning and development studies. By using predictor variables of 2016, the SVM-based urban expansion model can predict the binary LULC map of 2021 and any further binary maps by a 5-year interval.

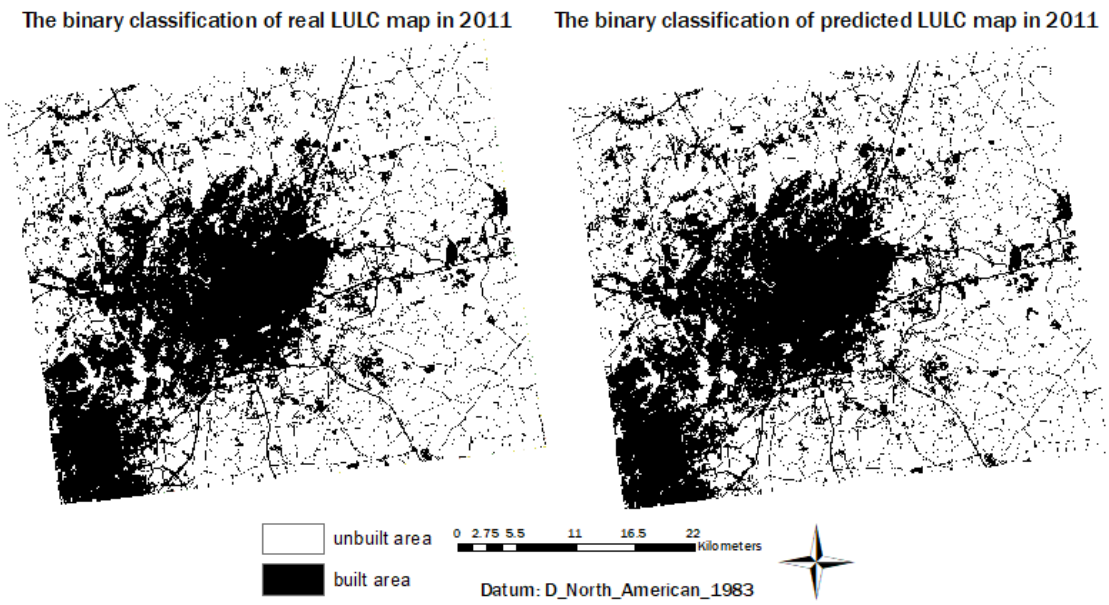


Figure 2.12: The Binary Classification of Real and Predicted LULC Maps in 2011

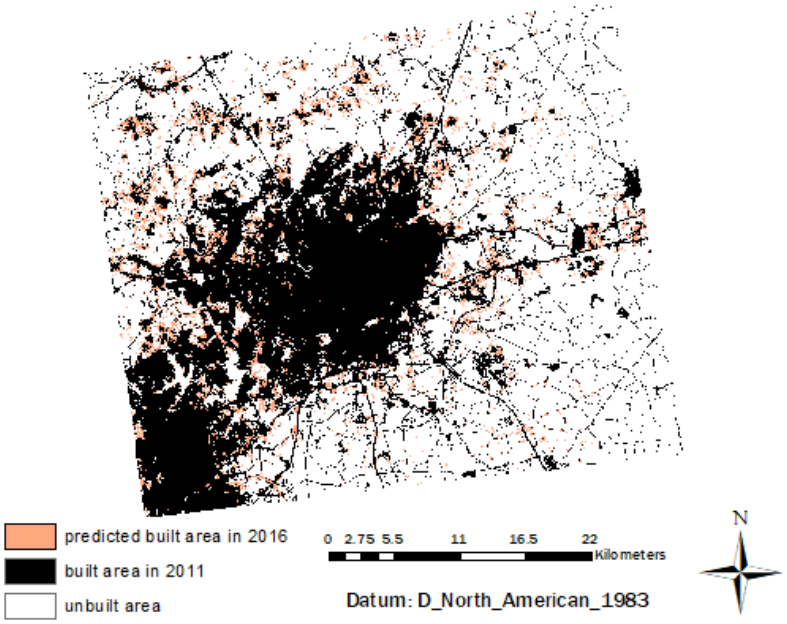


Figure 2.13: The Binary Classification of Simulated LULC Map in 2016

2.5 Conclusion

The capability of the SVM method for modeling urban expansion is explored in this study. Various SVM models are evaluated to select the most efficient urban expansion model in Guilford County, NC, over the period of 2001-2011. The modeling process includes the exploration of different sampling methods, the examination of a variety of predictor variables, the investigation of SVM parameterization and kernel regulation, and the development of various evaluation metrics. The application of three sampling methods including random sampling, sampling of all changed cells, and balanced sampling reveals the significant effects of these methods on the performance of the model. The findings demonstrate that the balanced sampling method produces more reliable results due to considering all changed cells and making a balance between the changed and unchanged cells in the sampling dataset. In addition, a variety of predictor variables are considered in three main categories of proximity, neighborhood, and site-specific characteristics, and their significance is examined using the information gain metric. The results show that the elimination of five insignificant variables including neighboring with forests, neighboring with developed open spaces, neighboring with water bodies, neighboring with wetlands, and neighboring with barren lands from the model leads to less complexity and more accuracy of the model.

Afterward, several configurations of the SVM model are developed using different values of the penalty parameter c, kernel functions, and kernel's parameters. The evaluation of the model configurations elucidates that the kernel function and other model parameters substantially affect the performance of the model; hence, they must be

specifically regularized for each case study to maximize the efficiency of the SVM-based urban expansion model. To conduct a more realistic and reliable evaluation, novel performance evaluation metrics are meticulously defined for the application of the SVM model to urban expansion modeling. The evaluation results show that the SVM model developed using the c value equal to 1 and the RBF kernel with the γ value equal to 2 is the most efficient model to predict urban expansion in this study. The comparison of predicted and real LULC maps demonstrates the high predictability and accuracy of the SVM technique to model urban expansion. The results show the overall training accuracy of 98%, the testing accuracy of 85% for built-unbuilt land cells, and the testing accuracy of 78% for changed-unchanged land cells. Moreover, the percentage of correctly predicted cells as unchanged is 78%, and the percentage of correctly predicted cells as changed is 87%. The results substantiate the striking efficacy, reliability, and predictability of the developed SVM-based urban expansion model.

The SVM-based urban expansion model can be utilized to evaluate the impacts of urban expansion on habitat fragmentation, environmental pollutions, hydrological issues, wildlife disturbance, deforestation, destruction of agricultural fields, and regional and global warming. Therefore, the model would remarkably help urban planners, environmental policymakers, and geographers to ameliorate activities regarding the interaction between the natural and built environments. Investigating more predictor variables, developing a multi-class SVM model, and comparing the performance of SVM with other ML methods such as decision tree, random forest, and deep learning are suggested for future studies to enhance the efficacy of urban expansion modeling.

CHAPTER III

A COMPARATIVE STUDY ON MACHINE LEARNING ALGORITHMS FOR URBAN GROWTH PREDICTION[2]

3.1 Introduction

In the last century, urbanization and population growth have made considerable changes to the natural environment (Mohammady and Delavar; Taravat et al.). Currently, almost half of the world's inhabitants live in urban areas resulting from continued urbanization and population growth (B. Cohen). United Nations has forecasted that 67.2% of the world's inhabitants will live in urban areas in 2050. Urban growth is faster than urban population growth, which means expansion of cities as low-density areas (Mohammady and Delavar; Taravat et al.). The rapid and low-density urban growth converts the natural environment and open spaces to urban land-uses, which makes concerns among urban planners and geographers to understand when and where urban growth occurs and how it influences the natural environment (Karimi et al. "An Enhanced Support Vector Machine Model for Urban Expansion Prediction"). Predicting the complex process of urban expansion is essential for them to make proper decisions for future urban development (Yao et al.). Urban growth prediction needs to consider comprehensive historical information of land conversion to precisely understand this

[2] Karimi, F., Sultana, S. (2021). A comparative study on machine learning algorithms for urban expansion prediction. *Landscape and urban planning.* In review.

complex process and spatiotemporal relationships (Clarke et al.; Pijanowski et al. "Using Neural Networks and Gis to Forecast Land Use Changes: A Land Transformation Model"). Over the last three decades, various methods such as cellular automata (Batty et al.; Liu and Phinn) and Markov chain (Baja and Arif; Myint and Wang) have been utilized to model and predict urban growth based on remote sensing and geographic information system (GIS) techniques. The models aim to learn the spatial process of urban expansion within a specific time toward the determination of future policies of urban development (Amato et al.). Hence, researchers are constantly looking for novel approaches to promote urban expansion predictions for effective planning.

Recently, the use of machine learning (ML) models, including decision tree (DT) (Karimi et al. "Urban Expansion Modeling Using an Enhanced Decision Tree Algorithm"; Samardžić-Petrović et al. "Exploring the Decision Tree Method for Modelling Urban Land Use Change"), random forest (RF) (Kamusoko and Gamba; Shafizadeh-Moghadam et al.), support vector machine (SVM) (B. Huang et al. "Support Vector Machines for Urban Growth Modeling"; B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines"; Karimi et al. "An Enhanced Support Vector Machine Model for Urban Expansion Prediction"; Samard zic´ -Petrovic´ et al.), logistic regression (LR) (Z. Hu and C.P. Lo; Mom and Ongsomwang; Tayyebi et al. "A Spatial Logistic Regression Model for Simulating Land Use Patterns: A Case Study of the Shiraz Metropolitan Area of Iran"), and artificial neural network (ANN) (Mohammady and Delavar; Pourebrahim et al.; Tayyebi et al. "An Urban Growth Boundary Model Using Neural Networks, Gis and Radial Parameterization: An

Application to Tehran, Iran"; Tian et al.) has increased in due to their effectiveness and accuracy. ML methods are learning algorithms that implement the modeling process automatically without the need for human assistance or expertise for modeling and prediction using these methods (Suthaharan). They can handle categorical and continuous variables, identify non-linear relationships, and cope with noisy and complex data. In urban growth studies, these models are assessed for their effectiveness and, there is still debate about which ML model is more reliable, accurate, and proper for modeling and predicting future urban growth patterns and land-use\land-cover (LULC) changes.

Based on the literature, the DT algorithms have significant potential for urban growth modeling (Samardžić-Petrović et al. "Exploring the Decision Tree Method for Modelling Urban Land Use Change"; Shafizadeh-Moghadam et al.). Samardžić-Petrović et al. "Exploring the Decision Tree Method for Modelling Urban Land Use Change" demonstrated that DT is an efficient method for LULC change modeling. Later, Shafizadeh-Moghadam et al. confirmed that the DT is a reliable method for LULC change modeling. Tayyebi and Pijanowski found the tree model's ability to simulate multiple land use classes. Karimi et al. "Urban Expansion Modeling Using an Enhanced Decision Tree Algorithm" focused on exploring different configurations of stopping rules for the tree model for urban expansion modeling and demonstrated a remarkable performance.

The RF method mostly has been used for land-cover classification in geospatial studies (Belgiu and Drăguţ), however, some studies have investigated RF for urban growth studies. Kamusoko and Gamba combined RF with cellular automata (CA) to

model urban growth. They demonstrated that the RF-CA model is approximately reliable at allocating land conversion. Shafizadeh-Moghadam et al. also combined RF with CA and highlighted the RF ability to identify the importance of variables during the model building process in addition to its high accuracy.

SVM has recently been applied for modeling land-use changes. B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines" for the first time proposed the SVM method to analyze LULC change and found that the unbalanced SVM is accurate and reliable for LULC change modeling. Samard zic´ -Petrovic´ et al. used the SVM method for LULC change modeling, and their results showed that the SVM-based models perform precisely by balanced data sampling, reducing datasets to informative variables, and adequately recognizing the optimal learning parameters. Karimi et al. "An Enhanced Support Vector Machine Model for Urban Expansion Prediction" found the SVM method reliable for modeling and predicting urban expansion.

Several studies have been conducted using the ANN method. Pijanowski et al. "A Land Transformation Model: Integrating Policy, Socioeconomics and Environmental Drivers Using a Geographic Information System" developed a packaged urban growth model by coupling ANN and GIS to model LULC changes. Later, Pijanowski et al. "Calibrating a Neural Network-Based Urban Change Model for Two Metropolitan Areas of the Upper Midwest of the United States" parameterized ANN-based models for two different case studies and developed different types of models to evaluate ANN. Tayyebi et al. "An Urban Growth Boundary Model Using Neural Networks, Gis and Radial

Parameterization: An Application to Tehran, Iran" developed an urban growth boundary model (UGBM) using ANN, GIS, and remote sensing to model and simulate the complicated geometry of the urban boundary. Mohammady and Delavar studied urban sprawl by proposing an urban sprawl model utilizing ANN and adaptive neuro-based fuzzy inference system (ANFIS) methods with remote sensing data and GIS spatial analyses.

Regression models have been used by several studies. Tayyebi et al. "A Spatial Logistic Regression Model for Simulating Land Use Patterns: A Case Study of the Shiraz Metropolitan Area of Iran" presented an urban expansion model which uses LR as a mean to model and predict urban growth pattern. Later, Tayyebi et al. "Predicting the Expansion of an Urban Boundary Using Spatial Logistic Regression and Hybrid Raster-Vector Routines with Remote Sensing and Gis" developed an urban growth boundary model by using spatial logistic regression, remote sensing, and GIS to simulate the geometry of a dynamic urban boundary that expands in different directions over decadal periods. In another research, LR was applied to model urban growth in GIS software, and discovered that the relationship between urban growth and the predictor variables is challenging and problematic (Z. Hu and C.P. Lo). Mom and Ongsomwang used satellite images and employed a LR model to discover the predictor variables of urban growth and predict urban growth trends.

Among the studies using ML methods for urban growth modeling and prediction, just a few of them compared different ML models (Samardžić-Petrović et al. "Machine Learning Techniques for Modelling Short Term Land-Use Change"; Shafizadeh-

Moghadam et al.); however, these studies compared the models only from the aspect of accuracy by different goodness-of-fit metrices. In this study, all the models are evaluated on the same case study and compared from the aspects of accuracy, the number of required hyperparameters to be adjusted, run time, the need for data preparation, and the number of false-positive and false-negative land cells in the prediction. In the following, first, ML-based Urban expansion modeling, ML methods are explained, and the advantages and disadvantages of each model are summarized. Next, the case study, data, and methodology are described. Then, the experimental results are illustrated and discussed. Finally, a summary and the conclusion are presented.

3.2 Machine-Learning-Based Urban Expansion Modeling

Urban expansion modeling aims to model LULC map at time t according to LULC map and some predictor variable layers at time t-1. Therefore, a suitable function should be determined to model the most probable LULC map at time t for a cell at time t-1. Afterward, the effectiveness, reliability, and prediction accuracy of the model should be assessed for the next time intervals (Karimi et al. "An Enhanced Support Vector Machine Model for Urban Expansion Prediction"; Samard zic´ -Petrovic´ et al.). If the model indicates LULC changes precisely for times t+1, and the past relationships insist in the study area, it can be applied to predict LULC change in the future at the same time intervals (Figure 3.1). This function can be RF, SVM, DT, ANN, and LR, which can use past LULC maps and spatial variables to train the model and predict binary LULC maps for the future (Karimi et al. "An Enhanced Support Vector Machine Model for Urban Expansion Prediction"). These methods are supervised ML methods, but the first three

are non-parametric and the two latter are parametric methods. The following sections briefly introduce these methods, and Table 3.1 summarizes their strengths and limitations.

3.2.1 Classification And Regression Trees (CART)

CART (Breiman et al.) is a binary DT method that uses several if-then rules (Debeljak and Dzeroski) to divide a complex dataset into one of the target variable categories based on the value of the predictor variables (Suthaharan). To construct a tree, the DT classifies the training dataset by sorting it from the root to some leaf nodes descendingly (Tan et al.). The DT algorithm is implemented in the whole training dataset repetitively and divides it into subsets based on the splitting rules. The tree is built when stopping rules are met (Quinlan). The minimum number of records in a leaf node, the minimum number of records in a parent node (the node before splitting), and the maximum number of splits are the most common stopping rules (Singh and Gupta; Song and Ying). The final purpose of splitting is to determine proper variables and their corresponding thresholds to maximize the homogeneity of subsets. CART uses Gini Index and towing criteria as splitting rules (Delen et al.; Singh and Gupta).

CART is easy to understand and interpret, requires little data preparation, makes no statistical assumptions as it is strictly non-parametric, and analyzes data with different measurement scales (Qin et al.). The performance of CART in dealing with large datasets is excellent (Friedl and Brodley; Pal and Mather). It is able not only to recognize the most important predictor variables and their relative weights (Debeljak and Dzeroski) but also produces visualizations of the relationships between the variables (Delen et al.). One of

the most important disadvantages of CART is that a small change in the dataset can cause a large change in the structure of the DT and lead to instability. Although CART easily handles the splits, CART may not catch the correct structure of the dataset if the structure is complex (Timofeev).

3.2.2 Random Forest (RF)

RF (Breiman "Random Forests") consists of a combination of DT models, in which each model assigns the most prevalent class to the input with a single vote (Breiman "Bagging Predictors"; Dietterich). The basic premise of a combination of models is that a set of models perform better than a single model (Dietterich) and much information can be garnered by selecting random samples from the training dataset (Breiman "Random Forests"). A RF grows trees from various training subsets to enhance diversity (Breiman "Bagging Predictors"; Gislason et al.; Pal; Suthaharan). Thus, greater classification stability and more classification accuracy are achieved (Breiman "Random Forests"). RF method requires the regulation of two hyperparameters for building a model. A constant number of, m predictor variables selected randomly at each node, and each subset is classified by a k number of trees. In RF, the generalization error converges as the number of trees enhances, thus, the model does not overfit the data (Breiman "Bagging Predictors"). Decreasing the number of predictor variables (m) causes each tree of the model to be less strong, and reduces the computational complexity of the algorithm and the correlation between trees, which increases the accuracy of the model and leads to the reduction of the generalization error (Breiman "Random Forests").

RF is a fast algorithm and can handle multicollinearity and high-dimensional data (Belgiu and Drăguţ). It performs efficiently on huge datasets, can handle a large number of variables without variable deletion, estimates predictor variables importance, produces an internal unbiased estimate of the generalization error, and computes proximities between pairs of cases that can be used in locating outliers which makes it relatively robust to outliers and noise (Breiman "Random Forests"; Rodriguez-Galiano et al.). A disadvantage of RF compared to a simple tree is that individual trees cannot be examined separately, thus becoming a black-box approach (Wiesmeier et al.). Also, RF strongly depends on the input dataset, especially the quality of spatial sampling; high-quality data leads to minimizing extrapolation problems and any type of bias in data (Ließ et al.).

3.2.3 Support Vector Machine (SVM)

SVM (Boser et al.; Vapnik and Lerner) was initially presented as a binary classification method, but it can be promoted to an n-class method (Belousov et al.). SVM projects input data into the Hilbert space where an optimal separating hyperplane is utilized for classification (Yang et al.). A binary SVM minimizes the upper bound of generalization error by maximizing the hyperplane separating the two classes, (B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines"; C. Huang et al.; Samard zic´ -Petrovic´ et al.). This reduces generalization error, independent of the data distribution (B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines"). In order to address non-linear datasets, the instances are mapped to a feature space of very high dimension with a particular class of functions called kernels (Cristianini and Shawe-Taylor; Samard zic´ -Petrovic´ et al.;

Statnikov; Suthaharan). Common kernel functions for a SVM model are the linear function, radial basis function (RBF), and polynomial function with their kernel parameters (Chapelle et al.). To consider the misclassification error of the data falling off the margin, a penalty parameter c is considered, which makes a trade-off between the margin size and the number of misclassified training data; whereas larger c gives smaller misclassifications and reduces the margin size. The goal is to find an optimal hyperplane to minimize the misclassification errors and simultaneously maximize the margin size.

SVM can consider non-normal distributed data, and training datasets with outliers, it can also avoid overfitting and guarantees good generalization performance (B. Huang et al. "Support Vector Machines for Urban Growth Modeling"; B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines"). On the other hand, SVM cannot evaluate the relative importance of predictor variables. It needs standardization of data with different scales and hot encoding of the categorical data. It requires large memory and the training process is time-consuming dealing with huge datasets (Zhang et al.).

3.2.4 Logistic Regression (LR)

LR is a statistical model for solving binary problems that was developed by Cox, which tessellates the data layers to form a grid of cells. In LR, the target variable is a categorical binary value of 1 or 0 and the output of LR is likelihood values, which specify the probability of the occurrence of a specific class based on the predictor variables (Tayyebi et al. "A Spatial Logistic Regression Model for Simulating Land Use Patterns: A Case Study of the Shiraz Metropolitan Area of Iran"). Mathematically, a LR estimates

a multiple linear regression and is based on the central mathematical concept of logit, the natural logarithm of an odds ratio (Hosmer Jr and Lemeshow, 2000). The plot of linear regression for one feature and one dichotomous outcome is two parallel lines which is a linear plot in the middle and curved at the ends, by computing the mean of the two outcomes. It is not easy to describe this S-shaped plot with a linear equation as the ends are not linear and the errors are not normally distributed or constant across the entire range of data (C. Y. J. Peng et al.). The key to solving this problem is LR as it applies the logit transformation to the target variable and predicts the logit of outcomes from features (Peng et al., 2002). The conditional mean of the dichotomous outcome in LR is based on the binomial distribution which is the only assumption of LR and denotes that there is the same probability across the range of feature values.

LR describes the effect of predictor variables by determining the coefficients of them (Menard). Interpretability of this model for gaining knowledge of the processes and driving the change of spatial patterns is desirable (Z. Hu and C.P. Lo; Triantakonstantis and Mountrakis). This model considers factors like spatial effects, autocorrelation, and heterogeneity (J.J. Arsanjani et al.). The disadvantage of LR is that data cannot deviate from the normal distribution and it is less effective in modeling spatial-temporal data (Westreich et al.).

3.2.5 Artificial Neural Network (ANN)

ANN is one of the most popular artificial intelligence methods that identify intricate patterns in data (Skapura). This method comprises many non-linear computational components working in parallel and arranged in patterns inspired

by biological neural nets (Lippmann). ANN has traditionally been composed of an input layer, one or more hidden layers, and an output layer, designing a multilayer perceptron (Rosenblatt). This network is trained in three phases: the feed-forward, the backpropagation, and weights adjustment (Basheera and Hajmeer). In the feed-forward phase, the data broadcasts to each of the hidden layers with multiple weighted summations occurring before reaching the output layer then an activation function computes the output value. The back-propagation phase randomly chooses the primary weights and then compares the estimated output with the actual output. After giving all the observations to the network, the weights are adjusted according to a generalized delta rule, and the total error is distributed among the various nodes in the network. The process of feeding-forward the signals and back-propagating the errors is repeated iteratively until a high performance is achieved.

ANN has the ability to deal with a large number of data, is a fast processing approach, can conduct pattern cognition, parallel processing, and data fusion (Basheera and Hajmeer; Mohammady and Delavar). One of the most important disadvantages of ANN is the disability of dealing with uncertainty, which is an inescapable part of spatial phenomena. A combination of ANN with fuzzy logic can be one of the best solutions to overcome this shortage (Mohammady and Delavar). ANN needs data preparation of the input data through scaling and hot encoding. ANN might suffer from multiple local minima, therefore, has challenges with generalization and may construct models that overfit the data (B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines").

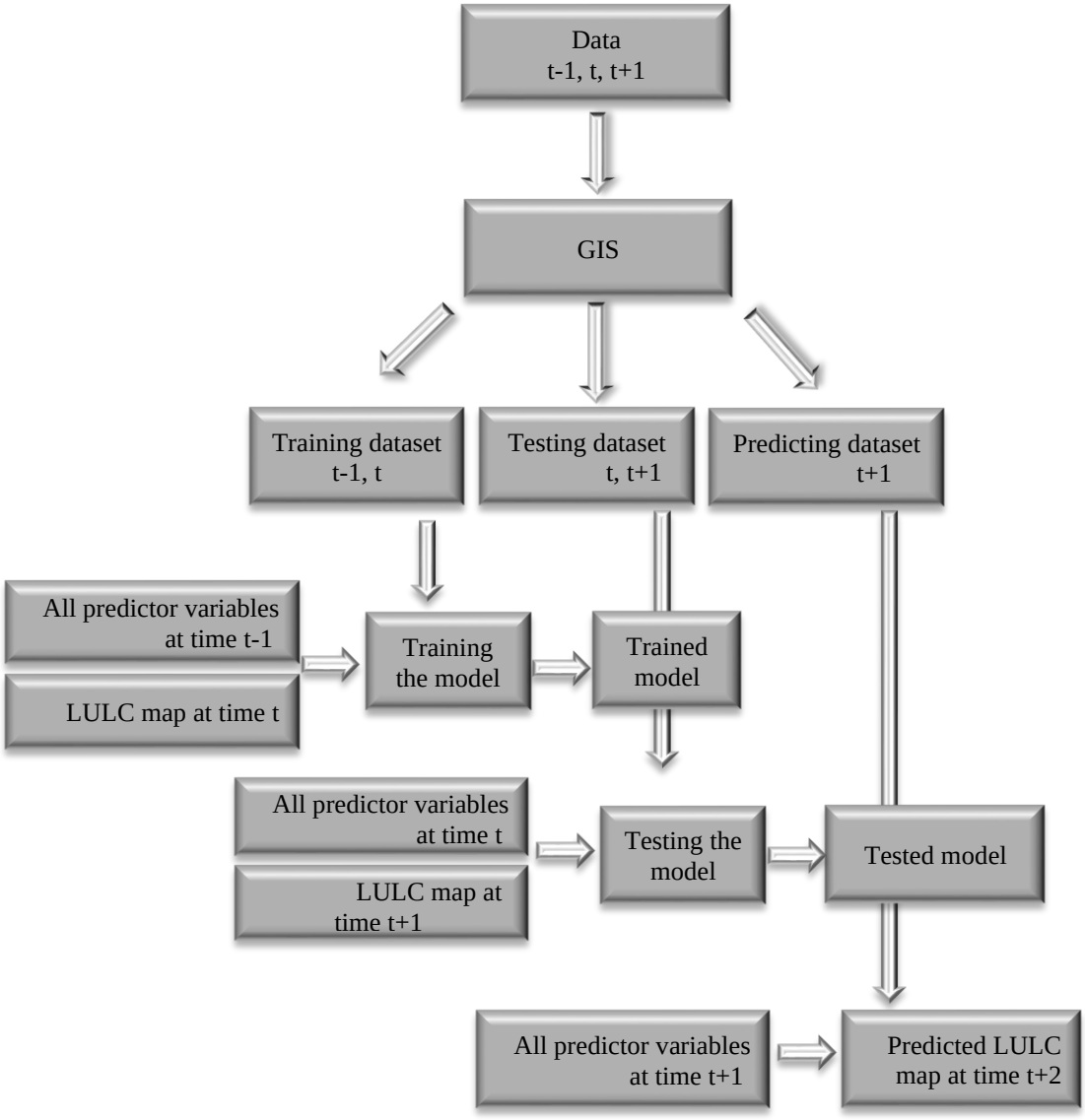


Figure 3.1: The Process of Machine-learning-based Urban Expansion Modeling

Table 3.1: Summary of the Strengths and the Limitations of the Models (Breiman "Random Forests"; Cheng et al.; B. Huang et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines"; Musa et al.; Qin et al.; Rodriguez-Galiano et al.; Timofeev; Torrens and O'Sullivan; Wiesmeier et al.; Zhang et al.)

Model	Strength	Limitation
CART	¶ easy to understand and interpret ¶ requires little data preparation ¶ makes no statistical assumptions ¶ handles large datasets ¶ estimates predictor variables importance	¶ may have unstable trees ¶ splits only by one variable ¶ a binary classification ¶ difficulty with complex structures
RF	¶ fast algorithm ¶ handles multicolinearity and high dimensional data ¶ runs efficiently on large datasets ¶ handles a large number of variables ¶ estimates predictor variables importance ¶ estimates the generalization error ¶ requires little data preparation	¶ is a black-box approach ¶ high sensitivity to input data quality
SVM	¶ handles non-normal distributed data ¶ deals with training datasets with outliers ¶ avoids overfitting ¶ establish good generalization performance	¶ cannot evaluate the relative importance of predictor variables ¶ sensitive to the scale of input data ¶ requires large memory ¶ the training process is long dealing with large datasets
LR	¶ examine the relationships of predictor variables ¶ examine factors such as spatial effects, autocorrelation, and heterogeneity	¶ data cannot deviate from the normal distribution ¶ less effective in modeling spatial-temporal data ¶ a binary classification ¶ difficulty with non-linear datasets
ANN	¶ deal with a large volume of data ¶ fast processing ¶ pattern cognition ¶ parallel processing ¶ data fusion ¶ handling noisy data	¶ disability of dealing with uncertainty ¶ difficulties with generalization and overfitting the data ¶ is a black-box approach ¶ might suffer from multiple local minima

3.3 Data and Methodology

3.3.1 Study Area

The five ML models were applied and assessed in Mecklenburg County, NC, USA (Figure 3.2). Some characteristics of this study area including rapid population growth, developing transportation network and economic growth have made it to be

expanding over time (Swain). Due to the rapid population growth, Mecklenburg County is the most populous county and the first county exceeding 1 million population in NC (U.S. Census Bureau). On the other hand, Charlotte, the major city and commercial hub of the state seat in Mecklenburg County. Charlotte is the third-fastest-growing major city in the United States (Balk), Uptown Charlotte is the central business district (CBD) of the county and the transportation network has been developing in this city such as the interstate highways (I-485) around the city of Charlotte (U.S. Census Bureau "Tiger/Line Shapefiles and Tiger/Line Files"). Therefore in this county, the area of the natural environment has been decreasing by the expansion of the urban area in a low density and dispersed pattern (MRLC) which highlights the importance of urban growth modeling.

3.3.2 Data

The required data was collected based on the data availability and data used in previous studies (Bhatta; Li and Yeh "Calibration of Cellular Automata by Using Neural Networks for the Simulation of Complex Urban Systems"; White and Engelen) and then prepared using ESRI ArcGIS 10.3 software to create predictor variable layers fixed at a ground resolution of 30m2. The LULC maps were collected from the national land cover database (NLCD) (USGS "The National Map") at the spatial resolution of 30 meters for the years 2001, 2006, 2011, and 2016 to provide historical information on LULC changes in the study area. In these LULC maps, the urban area includes developed open spaces, low-intensity, medium-intensity, and high-intensity developed areas and the natural environment involves open water, barren land, deciduous forest, evergreen forest, mixed forest, shrub, and scrub, herbaceous, and hay and pasture. Additionally, vector data of

transportation networks were collected from (U.S. Census Bureau "Tiger/Line Shapefiles and Tiger/Line Files"). The vector data of built areas, city centers, green spaces, and water bodies were extracted from LULC maps and used to produce the proximity raster maps. The digital elevation model (DEM) of the study area was acquired from “the national map data download and visualization services” (USGS "Elevation Products (3dep)"). Population data were gathered from (IPUMS-USA) at the scale of census tracts for the years 2000 and 2010.

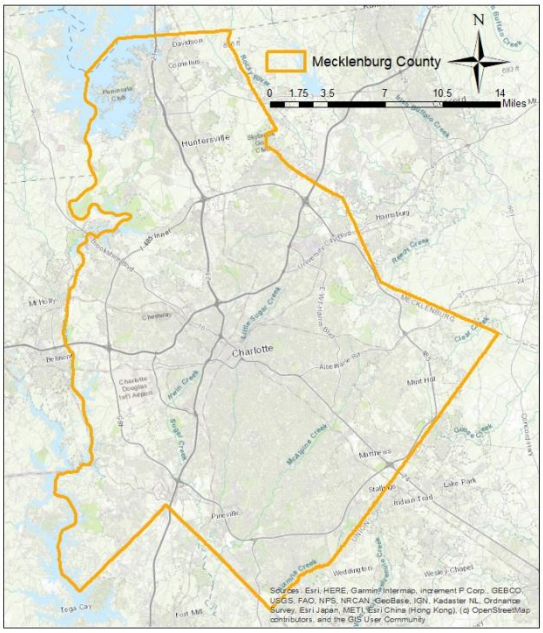


Figure 3.2: The Map of the Study Area (Mecklenburg County, NC, USA)

All the data were prepared for the study area over the years 2001, 2006, 2011, and 2016. The target variable is the urban development in the study period. The predictor variables are categorized into proximity, neighboring, physical and social variables (Table 3.2). Proximity variables were acquired by calculating the distances to CBD,

suburb towns, highways, major roads, railways, and urban areas. These variables play an essential role in urban development, as a higher development probability occurs near transportation networks and urban areas (Kucsicsa and Grigorescu). Neighborhood variables are important spatial components that highlight the dynamics of different change circumstances. The number of developed land cells about a cell is important in determining the conversion of a cell, as there is a higher development probability if a cell has surrounded by a larger number of developed land cells. Furthermore, land change mostly occurs by converting agricultural lands and open spaces to urban areas which are potential lands for development (Karimi et al. "Land Suitability Evaluation for Organic Agriculture of Wheat Using Gis and Multicriteria Analysis"). The neighboring variables were calculated in the Moore neighborhood of 5×5. The physical variables also affect the development probability in an urban growth model. The development probability of a cell is affected by the type of LULC of it. Terrain features such as slope and elevation exert constraints to urban development. The layer of the slope was generated from DEM. Finally, a population density map denoting the number of residents per cell is produced as a social variable. Because of the unavailability of census data in 2001, 2006, 2011, and 2016, a population estimation model is used to approximate population data in these years from the available census data in 2000 and 2010.

3.3.3 Methodology

The process of implementing the models in the study area is presented in Figure 3.3. After collecting the data, the raster layer of predictor and target variables, composed of 1,569,230 30-by-30-meter land cells, are prepared for 2001, 2006, 2011, and 2016.

The layers are converted to Ascii files for the next steps. Then, a sampling strategy based on the number of developed and undeveloped land cells during the study period is considered to create the training dataset over the 2001-2006 period. Thus, all the developed land cells and a different number of undeveloped land cells selected randomly are combined. Next, the models are developed using Python 3.7 and configured by different values of their hyperparameters, based on the influences on the performance and complexity of the algorithms.

Table 3.2: Predictor and Target Variables Utilized for Urban Expansion Modeling

Variables	Acquisition Method	Value ranges
Target variable: Urban development	Binary LULC map	1: Developed urban area 0: Natural environment
Proximity Predictor variables Distance to CBD (km) Distance to suburb towns (km) Distance to the nearest highway (km) Distance to the nearest major road (km) Distance to the nearest streets (km) Distance to the nearest railway (km) Distance to the nearest urban area (km) Distance to the nearest greenspaces (km) Distance to the nearest water bodies (km)	Euclidean distance of ArcGIS	
Neighboring Predictor variables Number of developed cells Number of potential lands for development The most frequent LULC type	The focal tool of ArcGIS	
Physical Predictor Variables LULC type Slope (%) Elevation (m)	 NLCD maps Slope tool of ArcGIS from DEM DEM	
Social variable Population density	 Census bureau	

The CART model is adjusted by three stopping rules including the minimum number of records in a leaf node, the minimum number of records in a parent node, and the maximum number of splits. Lower values for CART's hyperparameters lead to a less complex model, and the higher values may build a model with higher performance. To achieve the best performance of the model, different combinations of the minimum number of records in a leaf node with the 1 to 9 values, the minimum number of records in a parent node parameter with 2 to 10 values, and the maximum number of splits parameter with values 1 to 10000 are evaluated. The minimum number of records in a leaf node value should be always less than the value of the minimum number of records in a parent node. One of the most conspicuous strengths of the Due to the ability of the DT algorithm to identify the predictor variables importance (Karimi et al. "Urban Expansion Modeling Using an Enhanced Decision Tree Algorithm"), the CART model selects the most significant predictor variables based on the model configuration to generate the tree.

For configuration of the RF model, the RF model hyperparameters including the number of input variables selected at each node split, m, by the values of 1 to 16 and the total number of trees included in the model, k, by the values of 1 to 1000 are adjusted. By enhancing the number of trees, the generalization error converges, and overfitting does not occur. Reducing the number of predictor variables causes each tree of the model to be less intense, but, degrades the computational complexity of the algorithm and the correlation between trees, and increases the model accuracy. Optimization of m by keeping a large and constant k may minimize the generalization error and lead to a robust

RF model (Breiman "Random Forests"). Thus, hyperparameters optimization is essential for generalization error minimization. Like DT, the RF model can determine the most significant predictor variables in the process of building the model (Suthaharan).

For the SVM model configuration, the hyperparameters including the penalty parameter c, the kernel function, and the kernel function's are regularized. A larger c value leads to a more complex model that leads to overfitting the training dataset. Hence, smaller c values produce simpler models but may reduce the accuracy. The 0.1, 1, 10, and 100 values for c are tested to obtain the best model performance. In this study, the radial basis function (RBF) (Karimi et al. "An Enhanced Support Vector Machine Model for Urban Expansion Prediction") with the γ values of 1, 2, and 3 are tested to solve the non-linearity concern in the modeling. By keeping the hyperparameter c constant, and raising the value of parameter γ a more complex model is achieved.

The configuration of the LR model hyperparameters including the maximum iteration, a solver, the inverse of regularization strength, and the tolerance for stopping criteria highly affects the model performance. The maximum iteration is the maximum number of iterations taken for the solvers to converge. The solver is an algorithm to use in the optimization problem. For inverse of regularization strength, values close to 1.0 shows very little penalty, and values close to zero indicate a strong penalty. For model configuration, the maximum iteration values of 1 to 2000, the solver equal to limited-memory broyden-fletcher-goldfarb-shanno (lbfgs), stochastic average gradient (sag), and a variant of sag (saga), the inverse of regularization strength float values of 0 to 1 and the tolerance for stopping criteria values of 0.0001 to 0.1 are set.

For configuring the ANN model, the hyperparameters including the number of hidden layers and neurons in each layer, an activation function for the hidden layer, and a learning rate for weight updates are set, a solver for weight optimization, and the maximum number of iterations. The number of hidden layers is set from 2 to 200 and the number of neurons in each layer is set to 2 to reach the highest accuracy, the activation function is set to identity, logistic, tanh, and Relu, the learning rate for each network weight is set from 0.0001 to 0.1, the solver is set to broyden-fletcher-goldfarb-shanno (bfgs), stochastic gradient descent (sgd), and Adam, and the maximum iteration is set from 1 to 2000. The solver algorithm updates network weights iteratively based on training data; it iterates until convergence. Adam is a method for stochastic optimization.

For configuring the models, the trained models are implemented on the testing datasets of 2006-2011 and 2011-2016 periods over the whole study area. The performance of the models to predict future patterns of urban expansion is tested by comparing the actual and predicted binary LULC maps of 2011 and 2016 and conducting the evaluation strategies. The process of training and testing is conducted repetitively to find the best hyperparameters for each model. Finally, the best model is chosen, and it can be used to predict future urban expansion patterns with an interval of 5 years (e.g., 2021 and 2026).

For evaluating the models, after the creation of the confusion matrix (Suthaharan) and determining the number of true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) values, the training accuracy, testing accuracy, precision, negative predictive value (NPV), sensitivity, specificity, F-score, MCC, Kappa statistics

and AUC are calculated (Karimi et al. "An Enhanced Support Vector Machine Model for Urban Expansion Prediction"; Suthaharan). Training and testing accuracy show the performance of the model on the training dataset (2001-2006) and testing datasets (2006-2011 and 2011-2016), respectively, based on the number of the land cells to be predicted correctly. A high training and testing accuracy indicate that the modeling is highly accurate. Precision is the percentage of correctly predicted land cells as developed. A high value of precision shows that TP is high together with low values of FP. NPV is the percentage of correctly predicted cells as undeveloped. A high value of NPV shows that TN is high together with low values of FN. Sensitivity depicts the performance of the model concerning the proportionality between TP and FN. A high sensitivity shows a higher value of TP while FN is negligible. Specificity demonstrates the performance of the model regarding the proportionality between TN and FP. A high specificity value shows that TN is high, FN is low. F-score is a single metric that combines testing sensitivity and precision using the harmonic mean (Sokolova et al.). Matthews correlation coefficient (MCC) (Matthews) shows the correlation between the reference and predicted cells. A MCC equal to +1 indicates a perfect prediction, when it is equal to -1 demonstrates absolute dissimilarity between prediction and reference land cells, and zero means that no better than random prediction. Cohen's kappa coefficient (J. Cohen) is a measure of how well the model performed as compared to how well it would have performed simply by chance. Equally arbitrary guidelines characterize Kappa over 0.75 as excellent, 0.40 to 0.75 as fair to good, and below 0.40 as poor (Landis and Koch). Receiver Operator Characteristic (ROC) is defined as the sensitivity plotted against [1 –

specificity]. The balance between sensitivity and specificity is a demonstration of the model performance (Evans et al.). The Area Under the Curve (AUC) indicates the area under a ROC curve, ranging from 0 to 1, and 0.5 indicates no discrimination and 1.0 perfect classification (Fawcett 2006).

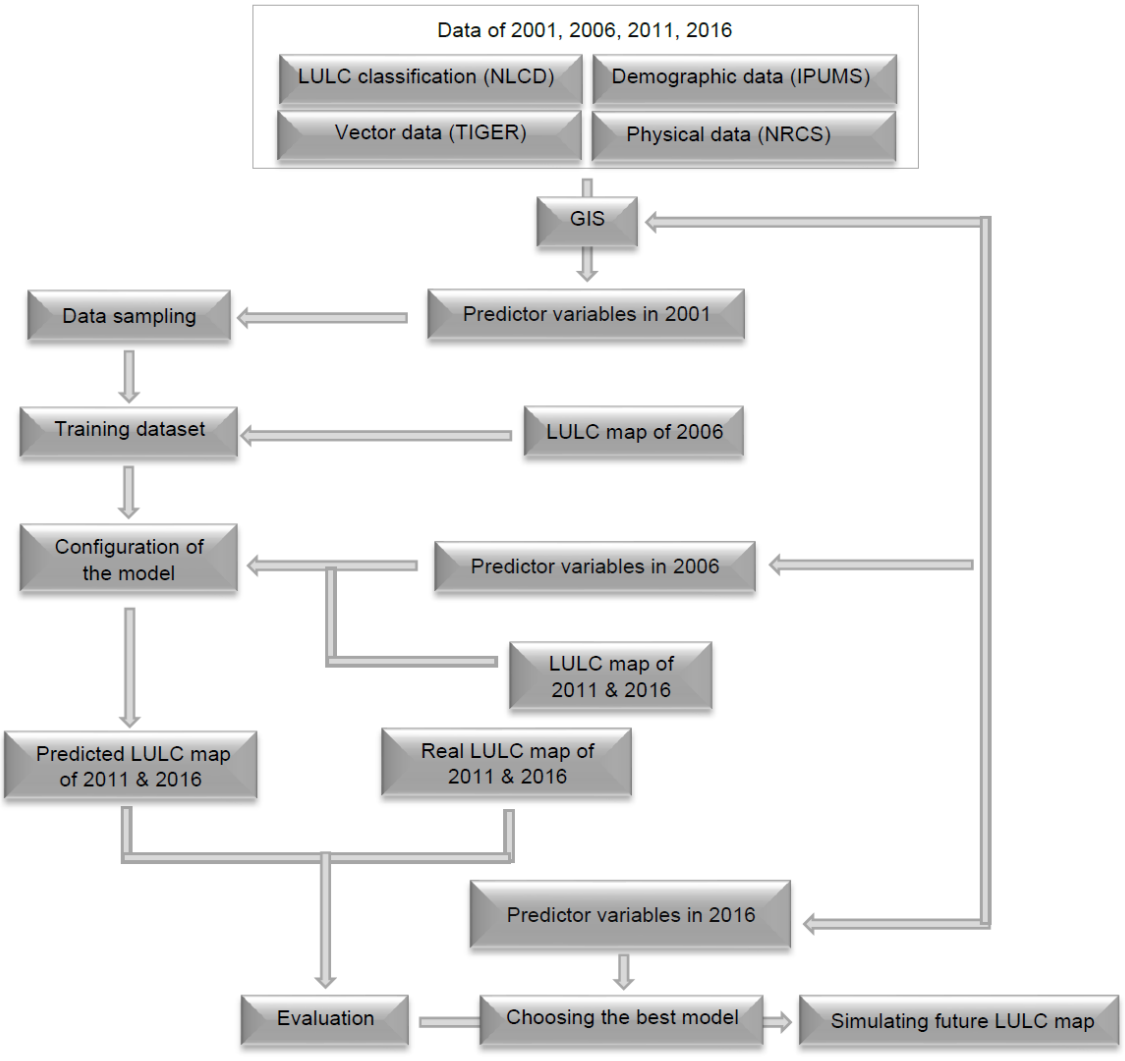


Figure 3.3: The Diagram of Machine Learning-based Urban Expansion Modeling

3.4 Results and Discussion

In this study, five machine learning models are applied to the same case study to assess their capability for modeling and predicting urban growth. For each model, first, the sampling strategy based on all the developed land cells and a different number of undeveloped land cells randomly selected over the whole study area is selected. The sample dataset is selected from the training dataset, 2001-2006 period, and the performance of the models based on the sampling is evaluated over a testing dataset, 2006-2011 period. For all the models, considering all the developed and different amounts of undeveloped land cells in the 2001-2006 period for the training dataset, increasing the undeveloped land cells in the sample increases the overall accuracy. However, this means better model accuracy for predicting undeveloped land cells and worsen the accuracy for predicting developed land cells. Then, the models are configured considering their effective hyperparameters for the best performance, lowest complexity, and least run time of the algorithms. At the same time, the number of needed hyperparameters and the need for data preparation are important. A variety of goodness-of-fit metrices are applied to evaluate the models over the 2006-2011 and 2011-2016 periods. The number of FN and FP land cells in prediction is important as well, as a high accuracy may not show the performance of the model due to a large number of land cells. Table 3.3 presents the performance obtained for the models from the cross-tabulation of the overlay analysis of reference and predicted urban development map of Mecklenburg County, NC for 2011 and 2016. As presented all the goodness-of-fit metrices show acceptable values for all the models, due to the lower number of developed cells in the

study period than the whole land cells in the study area. Therefore, there should be other criteria to compare the models. Also, the performance of all the models decreased for 2011-2016 compared to the 2006-2011 period. Table 3.4 summarizes the comparison of the methods from the perspective of accuracy, the number of hyperparameters to be configured, run time, and the need for data preparation. Also, Table 3.5 compares the models by the number of TN, TP, FN, and FP land cells for the prediction of 2011 and 2016.

For the CART model, as the testing accuracy and precision are high when the minimum number of records in a leaf node equals 1, the minimum number of records in a parent node equals 2, the maximum number of splits equals 20000, and the number of undeveloped cells is twice of the developed cells in the training dataset, and both values are highly acceptable, this sampling strategy is selected for further analysis of CART-based urban expansion modeling in the study area. By changing the hyperparameters of CART simultaneously, it is realized that a higher amount of the maximum number of splits leads to better results, however changing the minimum number of records in a leaf node and the minimum number of records in a parent node does not change the performance dramatically. The highest performance of the model is achieved by the maximum number of splits equal to 10000, where a higher number does not change the results. To prevent the model to be complex the minimum number of records in a leaf node and the minimum number of records in a parent node is set to 1 and 2, respectively. With this regulation, the training accuracy equals 1.00 and the testing accuracy equals

0.96. The CART algorithm is fast, and it needs no data preparation. The number of FN and FP is relatively low.

For the RF model, as the testing accuracy and precision are high when the number of trees, k, equals 1000, the number of input variables selected at each node split, m, equals 1, and the number of undeveloped cells is three of the developed cells in the training dataset, and both values are highly acceptable, this sampling strategy is selected for further analysis of RF-based urban expansion modeling in the study area. It is realized that a higher amount of k and a lower amount of m leads to better results, and the best-tuned hyperparameters for the RF model obtained 1 and 206 for m and k, respectively. The training accuracy equals 1.00 and the testing accuracy equals 0.99. The algorithm is fast, and it needs no data preparation. The number of FN and FP for both validation periods is lower than the other methods.

For the SVM model, as the testing accuracy and precision are high when the balanced sampling method is selected as the sampling strategy, in which all the developed land cells and the same number of undeveloped land cells selected randomly, this sampling method is utilized to train the SVM-based urban expansion model in all the remaining experiments. To improve the prediction accuracy, the model is configured by regulating parameter c and applying different RBF kernel's parameter. The penalty parameter equal to 1, and the value of 2 for the kernel's parameter leads to the highest accuracy and precision. The training accuracy equals 0.98 and the testing accuracy equals 0.96. The SVM algorithm is slow and takes a long time to deliver the results. It needs hot

encoding of the categorical variables and scaling the continuous variables of input data. The number of FN and FP is relatively low for this method.

For the LR model, as the testing accuracy and precision are high when iteration equals 1000, the inverse of regularization strength equal values of 1, the solver equal to lbfgs, and the tolerance for stopping criteria values of 0.1 and the number of undeveloped cells is twice of the developed cells in the training dataset, and both values are highly acceptable, this sampling strategy is selected for further analysis of LR-based urban expansion modeling in the study area. The best-tuned hyperparameters for the LR model are the maximum iteration equal values of 500, the inverse of regularization strength equals 2, the solver equal to lbfgs, and the tolerance for stopping criteria values of 0.001, however, in this model changing the hyperparameters did not change the accuracy meaningfully. The training accuracy equals 0.92 and the testing accuracy equals 0.95 which shows the underfitting problem in this model. The algorithm is fast, and it needs hot encoding and scaling of data. The number of FN and FP is lower than the other methods, which demonstrates that LR is not as good as other methods for the aim of urban expansion modeling.

For the ANN model, as the testing accuracy and precision are high when the optimizer is set to bfgs, the activation function is set to Relu, the number of hidden layers is set to 100, the number of neurons in each layer is set to 2 to reach the highest accuracy, and a learning rate for each network weight is set to 0.1 and the maximum iteration is set to 1000 and the number of undeveloped cells is four times of the developed cells in the training dataset, and both values are highly acceptable, this sampling strategy is selected

for further analysis of ANN-based urban expansion modeling in the study area. The best-tuned hyperparameters for the ANN model are the optimizer is set to Adam, the activation function is set to Relu, the number of hidden layers is set to 60, the number of neurons in each layer is set to 2, a learning rate for each network weight is set to 0.01, and the maximum iteration is set to 1000 to reach the highest accuracy. Training accuracy equals 1.00 and the testing accuracy equals 0.98. The algorithm is fast, and it needs hot encoding and scaling of data. The number of FN and FP is lower than the LR model but higher than the other three models.

Among the models, RF showed the highest performance, the lowest number of hyperparameters to be set, a low run time, and no need for data preparation. For RF, the number of FN and FP is low relative to other models. Obviously, the RF model is superior to DT, SVM, LR, and ANN. However, DT model has remarkable characteristics, which make it considerable for urban growth studies. DT has the lowest run time and needs no data preparation, but the number of FP and FN is high compare to the RF method. Among the models, SVM has the highest run time, and it needs a long time to show the results. For the SVM model, although the accuracy is high, the number of FN and the number of FP are higher than the RF and DT models'. LR shows the lowest performance and the number of hyperparameters to be set is relatively high. The performance of the ANN model is high, though it needs the highest number of hyperparameters to be set and the number of FN and the number of FP are higher than the RF and DT models'.

Table 3.3: The Performance Obtained for the Models From the Cross-tabulation of the Overlay Analysis of Actual and Predicted Urban Development Map of Mecklenburg County, NC

Models	Period	Testing accuracy	Precision	NPR	Sensitivity	Specificity	F-score	MCC	Kappa	AUC
CART	2006-2011	0.98	0.99	0.96	0.97	0.99	0.98	0.96	0.96	0.98
	2011-2016	0.95	1.00	0.87	0.93	1.00	0.97	0.90	0.90	0.97
RF	2006-2011	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99
	2011-2016	0.99	0.99	0.99	0.98	0.98	0.99	0.98	0.98	0.99
SVM	2006-2011	0.97	0.98	0.95	0.97	0.97	0.98	0.94	0.94	0.97
	2011-2016	0.95	0.97	0.92	0.96	0.94	0.96	0.89	0.89	0.95
LR	2006-2011	0.96	0.98	0.92	0.95	0.97	0.97	0.91	0.91	0.96
	2011-2016	0.96	1.00	0.89	0.94	1.00	0.97	0.92	0.91	0.97
ANN	2006-2011	0.98	0.98	0.98	0.99	0.97	0.98	0.96	0.96	0.98
	2011-2016	0.98	0.98	0.98	0.99	0.96	0.98	0.95	0.95	0.97

Table 3.4: Comparison of the Methods Concerning the Average Accuracy, the Number of Hyperparameters, Run Time, and the Need for Data Preparation

Models	Average accuracy	The number of hyperparameters	Run time (Second)	Data preparation
CART	0.96	3	43	No
RF	0.98	2	1249	No
SVM	0.96	3	21,485	Hot encoding and scaling
LR	0.96	4	209	Hot encoding and scaling
ANN	0.98	5	298	Hot encoding and scaling

As mentioned before, CART and RF are able to estimate predictor variables' importance. Figure 3.4 and Figure 3.5 show the importance of the predictor variables using these methods. Both CART and RF models demonstrated that proximity to highways, city centers, and uptown are the most important factors for urban growth, while proximity to greenspaces and population density are the least important in this case study. The CART, however, identifies ‘proximity to the urban area’ as an additional factor for urban expansion. However, these two models exhibit almost the same most and least essential predictor variables, they identify different magnitudes for the variables in this estimation.

Table 3.5: Comparison of the Methods Concerning the Number of FN and the Number of FP Predicted Land Cells for 2011 and 2016

Models	2011				2016			
	TN	TP	FN	FP	TN	TP	FN	FP
CART	566805	971144	25519	5762	487640	1008914	71336	1340
RF	589329	969557	2995	7349	553370	1001870	5606	8384
SVM	566787	959277	26737	17958	514718	980273	45382	30386
LR	542556	957461	49768	19445	499067	1008411	59909	1843
NN	580783	957543	11541	19363	545133	988927	13843	21327

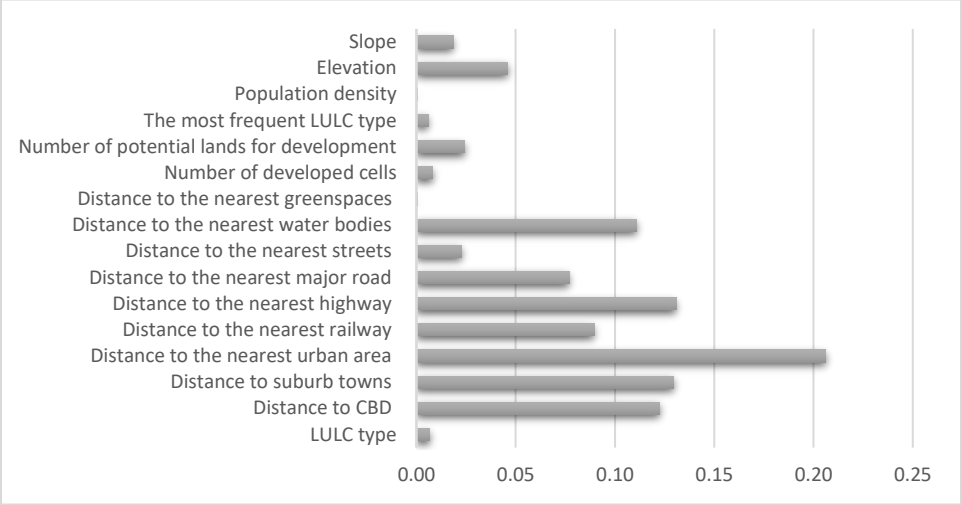


Figure 3.4: Predictor Variables Importance Based on CART Model

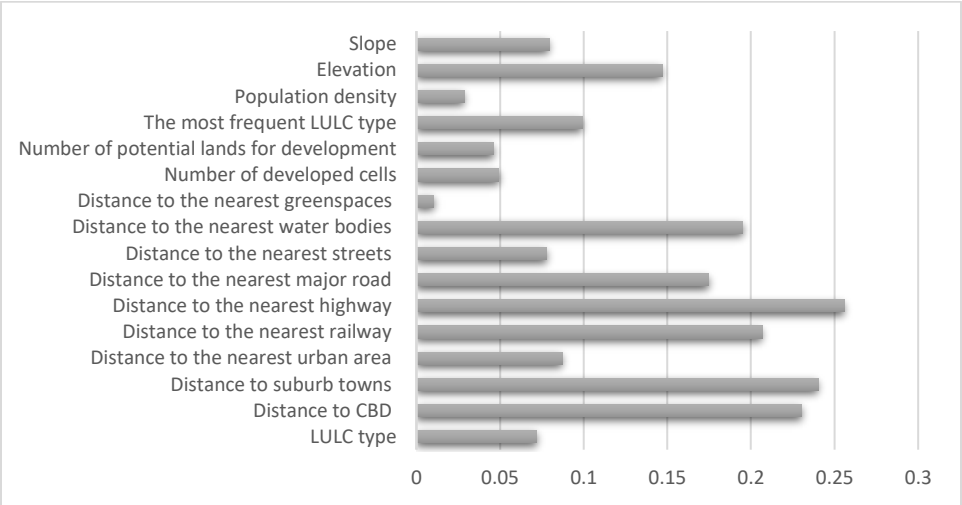


Figure 3.5: Predictor Variables Importance Based on RF Model

3.5 Conclusion

In this research, five machine learning models, including RF, CART, SVM, LR, and ANN were assessed and compared for urban growth modeling and prediction in Mecklenburg County, NC, USA over the 2001-2016 period. 16 predictor variables were extracted for including distance to CBD, suburb towns, the nearest highway, the nearest major road, the nearest streets, the nearest railway, the nearest urban area. the nearest greenspaces, and the nearest water bodies (km), number of developed cells in a 5×5 neighborhood, number of potential lands for development in a 5×5 neighborhood, the most frequent LULC type in a 5×5 neighborhood, LULC type, slope, elevation, and population density. The importance of predictor variables was analyzed by CART and RF methods and as the result, proximity to urban areas, highways, city centers, and uptown are the most critical variables and proximity to green spaces and population density are the least important variables.

The models were trained, and the performances were evaluated using training accuracy, testing accuracy, precision, NPR, sensitivity, specificity, F-score, MCC, Kappa statistics, and AUC. According to this case study, all five models exhibit reasonably good performances based on the evaluation metrices; however, the RF model showed the highest predictive capability compared with other models due to the lower number of FN and FP. The RF model, with an accuracy equal to 0.99, is a promising method for urban expansion modeling and prediction. In addition, the models were compared concerning the number of hyperparameters that they require, their run time, and the need for data preparation in the modeling process. It is found that RF requires low training time and

requires a low number of hyperparameters to be regularized compared to other ML models. RF does not need data preparation like scaling and hot encoding. As well, RF resulted in a lower number of FT and FN in the study area. The number of FP and FN land cells for each model shows the effectiveness of them for predicting developed and undeveloped land cells.

However, in the cases that the dataset is enormous, and time is of the essence, DT method is suggested. In confronting large datasets SVM needs a long time to deliver the results and requires a computer system with large memory. Because of the difficulties of ANN with generalization and overfitting the data, its results may not be reliable for predicting future patterns. LR is not suggested for urban expansion modeling as it is not as accurate as other methods, and it has difficulty with non-linear datasets and non-normal distributions and is less effective in modeling spatial-temporal data.

The results of this study may be useful for decision-makers and planners in urban growth studies as it is always a need for new ways to enhance urban development predictions for effective planning and determination of future policies of urban development.

CHAPTER IV

A SCENARIO-BASED SIMULATION OF URBAN GROWTH BY COUPLING RANDOM FOREST AND CELLULAR AUTOMATA[3]

4.1 Introduction

Worldwide the urban systems are expanding at a faster rate than population growth (Angel et al.; Sultana "Land Use and Transportation"). Such rapid urban development patterns are often subject to central debates of land-use/land-cover (LULC) change in urban, suburban, and surrounding rural areas and the LULC researchers have criticized this inefficient use of land resources and energy which leads to environmental costs (Camagni et al.; Sultana et al.). Land-cover is an important factor of environmental activities, particularly in terms of hydrological processes (Lindh; J. D. Wickham et al. "Ageography of Ecosystem Vulnerability"; Williams), natural resources (Masri; Yankson and Gough), and regional and global warming (Alcoforado and Andrade; Stone Jr The City and the Coming Climate: Climate Change in the Places We Live). With the expansion of urban systems, land-covers making up the natural environment such as forests, wetlands, and farmlands have been replaced by urban land-uses. As a result, impervious surfaces such as roads, sidewalks, parking lots, and airports, that are covered by water-resistant materials, change the hydrological system and deteriorate water quality

[3] Karimi, F., Sultana, S. (2021). A Scenario-Based Simulation of Urban Growth by Coupling Random Forest and Cellular Automata. *Cities*. In review.

(Arnold Jr and Gibbons; Tong and Chen), as well this conversion causes an increase in land surface temperature (Brovkin et al.), pollution (Shukla and Parikh), and deforestation and agricultural land loss (Zhou and Wang). While the compact urban form has been adopted for sustainable planning practices in the developed world, including in the United States, since the 1990s, the empirical evidence suggests that there are more low-density and decentralized development than intensification and efficient expansion (Jantz et al.; Sultana and Weber "Journey-to-Work Patterns in the Age of Sprawl: Evidence from Two Midsize Southern Metropolitan Areas"; Weber and Sultana; Zhao et al.)(Jantz et al., 2004; Sultana and Weber, 2007; Weber and Sultana, 2008; Zhao et al. 2020). Hence, predicting future environmental outcomes for supporting sustainable development requires being able to develop simulation models to analyze where and how the conversion has happened and predict the alternative spatial pattern of urban expansion (Hersperger et al.; Kamusoko and Gamba). The visualization and quantification of the simulations can enable planners and decision-makers to explore various plans, predict possible environmental impacts, and seek optimal land-use patterns (Karimi et al. "An Enhanced Support Vector Machine Model for Urban Expansion Prediction").

In recent years, urban growth models integrated with geographic information system (GIS) and remote sensing have emerged to generate spatially explicit simulations. Such simulations present applicable information about locations, types, range, quantity, and land conversion density that will probably occur (Jiang and Yao). Cellular automata (CA), a relatively simple model, is a spatial dynamic modeling method that has been widely applied to simulate convoluted urban dynamics and predicting spatial patterns of

urban development (Batty and Xie; Batty et al.; Berberoğlu et al.; Clarke et al.; Deep and Saklani; Li and Yeh "Calibration of Cellular Automata by Using Neural Networks for the Simulation of Complex Urban Systems"; Santé et al.). The advantage of CA is that it models very complex behaviors and global structures from some simple rules. CA requires a grid of cells as a window that changes the state of the center cell as the model iterates (Clarke et al.). The changes are determined by the rules that define a set of neighborhood conditions to be fulfilled (O'Sullivan). CA model not only can predict future urban development, but it can also explore development alternatives by integrating different sustainable elements and policies to the model for planning sustainable urban areas and forecasting the consequences of plans and policies (Li and Yeh "Calibration of Cellular Automata by Using Neural Networks for the Simulation of Complex Urban Systems"). However, CA has been criticized for its seeming inability to define transition rules for producing a realistic simulation of urban areas which are open and non-linear complex systems involving spatial and sectoral interactions (Batty et al.). Moreover, CA concentrates on the simulation of spatial patterns rather than on the spatiotemporal urban growth analysis and has deficiency of employing socio-economic and demographic variables (Z. Hu and C.P. Lo).

Definition of proper transition rules can well model the temporal and spatial complexities of urban systems and urban growth by incorporating various models such as artificial neural network (ANN) (Li and Yeh "Calibration of Cellular Automata by Using Neural Networks for the Simulation of Complex Urban Systems"; Li and Yeh "Neural-Network-Based Cellular Automata for Simulating Multiple Land Use Changes Using

Gis"), agent-based model (Tian et al.), logistic regression (J.J. Arsanjani et al.) and support vector machine (SVM) (Feng et al.; Yang et al.) in CA model. However, logistic regression is considered a generalized linear model (Yang et al.), in which it fails to model the non-linear, complicated, and self-organized change patterns and processes that are often characterized by (Liu et al.). In addition to linearity, logistic equations cannot provide explicit rules (Wu and Martin). Artificial neural networks (ANN) can handle nonlinear relationships, but its black-box approach makes it crucial to understand the meanings of its parameter values; also, the difficulty of regulating ANN models, leads to overfitting the dataset (Li and Yeh "Neural-Network-Based Cellular Automata for Simulating Multiple Land Use Changes Using Gis"). SVM is non-linear, but it generally requires more training time, especially for large datasets (Resler et al.). Moreover, it is sometimes challenging to handle complex relationships.

Random forest (RF), a powerful machine learning algorithm, provides levels of performance superior to conventional and other machine learning methods. It retrieves explicit rules for easier understanding and implementation and can recognize patterns through the training process and simulate development plans (Kamusoko and Gamba). RF method can handle the uncertainties of spatial data, runs efficiently on large datasets of both continuous and categorical variables, gives estimates of the importance of variables, requires less training time compared to other machine learning methods, and requires a low number of hyperparameters to be regularized (Breiman "Random Forests"; Rodriguez-Galiano et al.). Also, the RF method handles spatiotemporal data with non-normal distributions and non-linear relationships, prevents overfitting, produces an

internal unbiased estimate of the generalization error, and ensures good generalization performance, is almost robust to outliers and noise, and is computationally light and fast (Breiman "Random Forests"; Rodriguez-Galiano et al.). RF can deal with the difficulties and uncertainties in defining the transition rules for CA as it can estimate development probability at each iteration of the CA simulation. RF may be the best way to reveal complex processes of urban systems. RF can simultaneously be calibrating with CA during the rule-induction process. Nevertheless, the existing studies (Xu et al.) which used RF to define transition rules for CA did not regularize the hyperparameters simultaneously with various configurations and did not simulate developmental alternatives.

As the simulation of development alternatives is advantageous for urban and regional planning to prevent existing problems from happening again in the future, this study develops a RF-CA model to simulate three different urban development scenarios for the planning of sustainable urban development. The simulated urban development patterns are evaluated by using a cost indicator to find which type of development scenario can better fulfill the sustainability criteria. Besides, still several problems need to be addressed to improve the effectiveness of urban growth modeling. While most of the urban growth models assume that the rate of expansion is constant for all periods (Brown et al.) and their transition function retains similar properties over the whole study period, the rate of development is not constant in some study areas in long term due to the dynamics of economic, social, and political driving forces (Li and Yeh "Data Mining of Cellular Automata’s Transition Rules"). Consequently, this assumption may result in

poor and unreliable predictions in these cases. The reliability and performance of the proposed model are tested to see if it can discover the knowledge to model and predict urban growth in such a case study. Also, despite the previous studies (e.g., (Li and Yeh "Data Mining of Cellular Automata's Transition Rules")) that used the constant value of iterations for simulation using the CA model, in this study, the value of iterations is calculated in the process of training the RF-CA model. This helps the process to be fast and saves processing time.

The proposed model incorporates different data sampling strategies, predictor variables, various configurations of RF-CA, and constraints in Mecklenburg County, North Carolina (NC), USA in which the current urban growth trend is not acceptable for sustainable development (Sustain Charlotte). For this purpose, first, the most effective sampling method is selected to create an appropriate training dataset by testing several sampling strategies. Second, the significance of the predictor variables is investigated using RF to extract the most and the least significant predictor variables. Third, the most efficient configuration of the RF-CA-based model is developed by regulating RF and CA parameters simultaneously. Fourth, the accuracy, reliability, and predictability of the model are assessed by pertinent evaluation metrics. Fifth, the model is used to simulate the urban expansion in 2021 and 2026. Finally, different urban development scenarios are incorporated into the model using constraints and stochastic variables, and the simulated patterns are evaluated to find which type of development scenario can better fulfill the criteria of sustainability. The paper is structured as follows; section 2 briefly introduces RF-CA urban growth modeling approach. Section 3 presents a case study. Section 4

presents the implementation. Section 5 discusses the outcomes of the implemented approach, and finally, the paper concludes with a summary and some suggestions for future works.

4.2 Random Forest-Cellular Automata (RF-CA) Model

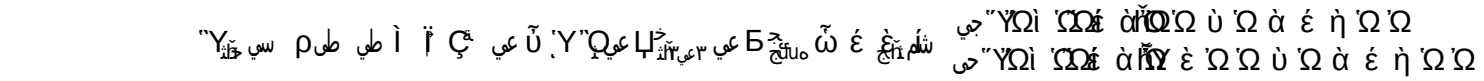
CA is a bottom-up and discrete dynamic model developed to examine the logical nature of self-reproducing systems (White and Engelen). This model has the ability to simulate intricate global patterns by applying local interactions and transition rules which decide how a cell change under some conditions (Clarke et al.). In CA, space is tessellated into regular cells, and the state of each cell is specified by the state of the neighboring cells and some transition rules. The state of each cell is updated in discrete time steps (Liu et al.). The transition rules for classic and simple CA are neighborhood-based as the transition potential of a central cell is specified by the number of developed land cells in the neighborhood (Batty). However, the neighborhood-based factor cannot address the complex dynamics of urban development. More factors should be integrated into the CA model to improve simulation performance using transition probabilities of mathematical and machine learning models (Shafizadeh-Moghadam et al.). Defining random variables is essential for addressing the stochastic nature and uncertainties in the convoluted process of urban expansion which makes the simulation more realistic (White and Engelen) and produces fractals that are common in actual land-use patterns. Also, various planning objectives as constraints based on study area features can be incorporated to regulate development patterns in line with urban policies (Li and Yeh "Neural-Network-Based Cellular Automata for Simulating Multiple Land Use Changes

Using Gis"). Thus, in the CA model, discrete steps repeat, and in each iteration, all the cells alter their state as a function of their state, the state of the neighboring cells, defined transition rules, stochastic variables, and constraints simultaneously, and finally by comparing the state of each cell with a threshold (Batty and Xie; Wu and Martin).

RF is a tree-based ensemble model and uses bootstrap aggregated sampling to construct plenty of decision trees (DTs) for modeling and prediction (Breiman "Random Forests"). Thus, a RF model has greater stability than a DT model, which leads to higher accuracy and being robust facing little variations in input data (Breiman "Random Forests"). The algorithm split the input dataset into homogenous subsets by using a random subset of predictor variables and choose a training sample to build the model. Then, to evaluate the performance of the model, the RF model uses the subsets that are not in the training sample named out-of-bag (OOB) sample dataset (Breiman "Bagging Predictors"; Dietterich). Tree-based methods need to select suitable variables that maximize dissimilarity between classes. A CART-based RF model uses the Gini Index for this purpose. RF estimates the importance of predictor variables during the model construction. To evaluate the significance of each variable, RF changes one of the input random variables while keeping the others fixed, and it measures the OOB error increase and Gini Index decrease (Breiman "Random Forests"), which is an implication of the significance of that variable. The RF model only requires defining two parameters for building a prediction model, the number of DT in the model (k) and the number of predictor variables (m) randomly selected at each node to make the tree grow. Thus, the

most frequent value for the total number of k trees will be the value allocated to each example.

In this study, the RF model is proposed as the decision function to construct non-linear transition rules for the CA simulations. Thus, in each iteration of the simulation, the state of a cell is estimated based on a random variable, the output from the RF model, neighborhood effect, constraints, and a threshold value (equation 1) (Yang et al.):



(1)

In this equation, the first term is the random variable which helps to generate a more realistic pattern by adding uncertainty. Various values of parameter a can be used to explore possible urban forms due to uncertainties, and parameter γ is a random variable between 0 and 1(White and Engelen). Higher values of parameter a can lead to more dispersed urban areas in the simulation and setting it to 1 presents a small amount of uncertainty in the simulation (Li and Yeh "Calibration of Cellular Automata by Using Neural Networks for the Simulation of Complex Urban Systems"). This term is always greater or equal to 1 and in the case of an equal to 1, the uncertainty term will be between 1 and 4 (Mustafa et al.). The second term is the urban development probability of cell i, which is retrieved from the RF model and is a value between 0 and 1. The third term is the influence of neighborhood and is the total number of urban cells within the 5×5 Moore’s neighborhood of cell i over 25 in iteration t (Shafizadeh-Moghadam et al.). Implementing CA with varying neighborhood sizes Shafizadeh-Moghadam et al. found

that a 5×5 window had the best accuracy among other neighborhood sizes. The fourth term indicates the constraint in the modeling process. Some layers which values of the land cells are between 0 and 1 can be used to outline constraints to urban land development using GIS spatial data such as excluding water bodies from the process. The neighborhood variable is dynamically updated during the simulation. The updated variable is the inputs of the model at each loop. At each iteration, the result of equation 1 for each cell is compared with a predefined threshold to determine if the development has occurred or not. If the state of a cell is greater than the threshold value, it will be converted to a developed land cell (Li and Yeh "Calibration of Cellular Automata by Using Neural Networks for the Simulation of Complex Urban Systems"). The simulation of urban development is conducted by running the model iteratively until the accumulated mismatch between the actual and simulated urban development reaches the minimum amount.

4.3 Case Study

North Carolina has been the fastest-growing state in the United States since 2010, and a large segment (66%) of its population resides in urban areas (North Carolina State Data Center). The proposed model has been applied and tested in Mecklenburg County, home of the city of Charlotte—the third fastest-growing city in the United States, located in southwestern NC, USA (Figure 4.1). According to the U.S. Census Bureau "The Census Bureau’s Population Estimates Program ", the population of this county was 514,831, 700,802, and 923,202 in 1990, 2000, and 2010, respectively, and the estimated population of it was 1,100,000 in 2019, making it the most populous county in North

Carolina and the first county in the Carolinas exceeding 1 million in population. The total area of this county is about 141400 ha, of which 56% was developed urban area and 44% was the natural area in 2001 (MRLC). The developed urban area has been grown to 60% in 2006, 62% in 2011, and 64% in 2016. Additionally, the city of Charlotte in Mecklenburg County, the major city and commercial hub of North Carolina and the American South (Graves and Smith; Sultana "Edge Cities in the Era of Megaprojects"), is characterized by low density and dispersed urban growth patterns (Shoemaker et al.).

Population growth (U.S. Census Bureau), the development of the transportation network specially, the construction of the interstate highways such as I-485 around the city of Charlotte (U.S. Census Bureau "Tiger/Line Shapefiles and Tiger/Line Files"), together with the economic development (City of Charlotte) have had a tremendous effect on urban growth (UNCCharlotte) in this urban area and the county has become exemplary of the sprawl debate. As a result, the low-density development has endangered the natural environment of Mecklenburg County by loss and fragmentation of the natural resource and reduced the quality of life (Shoemaker et al.; Yang). While the continuous population growth and urban expansion in this county highlight the importance of urban growth modeling to aim at growth management and natural resource protection, the different rate of development in the long term in this county (Figure 4.2, 4.3) makes serious challenges for prediction and requires specific considerations to develop an appropriate and reliable model.

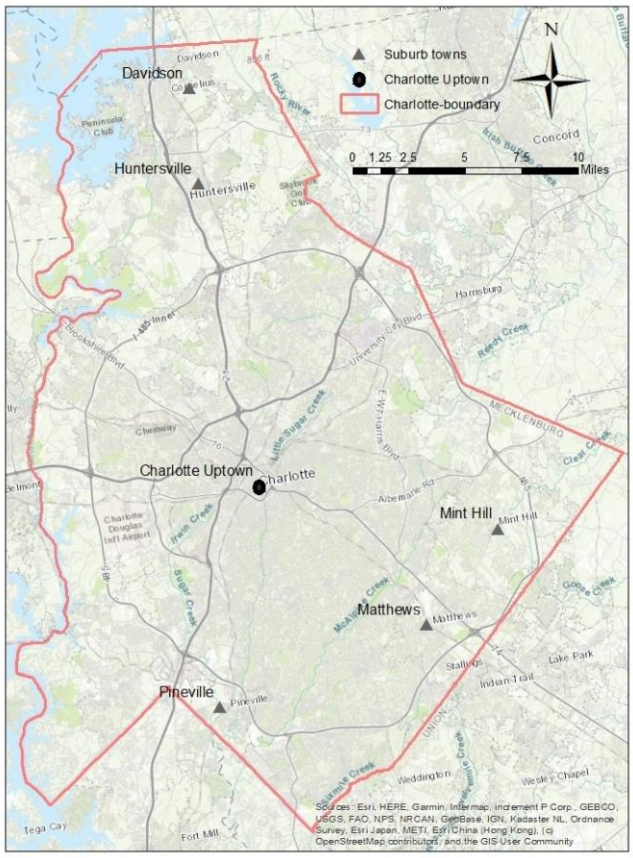


Figure 4.1: The Location Map of the Study Area, Mecklenburg County

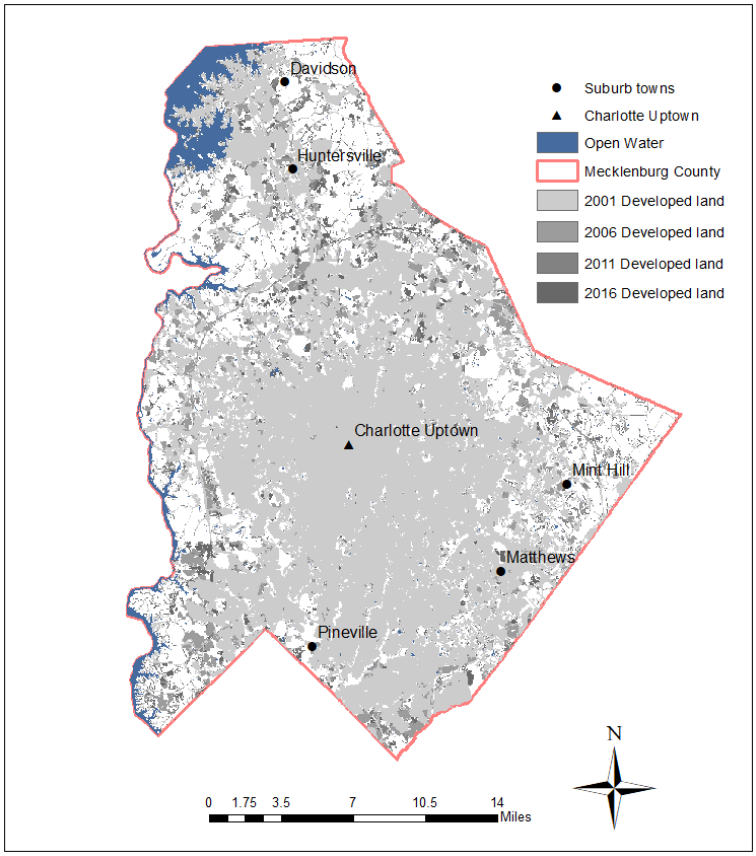


Figure 4.2: The Urban Development in Mecklenburg County in the Study Period (2001-2016)

The proposed model is integrated with a GIS for using the spatial data. The spatial data of LULC types were collected from the National Land Cover Database (USGS "The National Map") at the spatial resolution of 30 meters for the years 2001, 2006, 2011, and 2016 to constitute regular time steps (5-year period) for temporal mapping and simulation. These maps provide beneficial information about the trend of LULC changes in the study area. In these LULC maps, the developed urban area and the natural environment are classified into land types that are presented in Table 4.1. Vector data of transportation networks were gathered from (U.S. Census Bureau "Tiger/Line Shapefiles and Tiger/Line Files") and prepared for the study area over the study period. The vector

data of urban areas, city centers, green spaces, and water bodies were extracted from LULC maps and used to produce the proximity raster maps. Population data were collected from (U.S. Census Bureau "The Census Bureau's Population Estimates Program ") at the scale of census tracts for 2000 and 2010. As the census data is not available for the study period, a population estimation model is used to approximate population density in the study period years from the census data in 2000 and 2010.

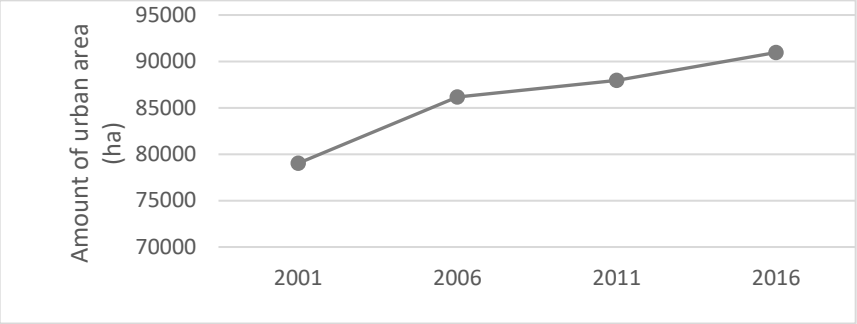


Figure 4.3: Rate of Development in the Urban Area in Mecklenburg County for 2001-2006, 2006-2011, and 2011-2016

Table 4.1: The Classification of the Developed Urban Area and the Natural Environment

LULC class	Description
Developed urban area	developed open spaces, low-intensity, medium-intensity, and high-intensity developed
Natural environment	open water, barren land, deciduous forest, evergreen forest, mixed forest, shrub and scrub, herbaceous, and hay and pasture

In the urban expansion model, the development probability of each cell is determined by the predictor variables of that land cell acquired by applying GIS spatial analyses. Table 4.2 lists the predictor variables used for urban expansion modeling, considering the most common factors used in previous studies (Aburas et al.; Bhatta; Li and Yeh "Calibration of Cellular Automata by Using Neural Networks for the Simulation

of Complex Urban Systems"; White and Engelen). The target variable is the urban development in the study period, obtained from change detection using 2001, 2006, 2011, and 2016 LULC maps. Proximity variables were achieved by calculating the Euclidian distances to the central business district (CBD), suburb towns, highways, major roads, railways, and urban areas. The role of proximity variables in urban development is that a closer distance to the major transportation networks and urban areas leads to higher development probability and expanding urban areas generate more infrastructure to maintain future urban expansion. Neighborhood variables are essential spatial elements that emphasize the dynamics of different conversion events. The neighboring land type surrounding a land cell affects the process of change. The neighboring variables were calculated in the Moore neighborhood of 5×5. Furthermore, most land change replaces agricultural lands and open lands with urban areas, which are potential lands for development (Karimi et al. "Land Suitability Evaluation for Organic Agriculture of Wheat Using Gis and Multicriteria Analysis"). The physical variables affect the development probability of land cells through the type of LULC of each land cell and terrain features such as slope and elevation that pose constraints to urban development. The layer of the slope was generated from the digital elevation model (DEM) which represents the elevation. Finally, a population density map outlining the number of residents per cell is created as a social variable. All these spatial data were prepared and converted to raster format and then to ASCII files using ESRI ArcGIS 10.3 software to facilitate the calculation and simulation. The resolution was fixed at a ground resolution of 30m2 to match the resolution of the National Land Cover Database maps.

4.4 Implementation

In this study, an integrated model of RF and CA methods and GIS is developed to enhance the reliability of urban expansion modeling and simulation of development alternatives in an urban area with varying development rates. In most urban expansion studies the assumption is that the rate of urban expansion is constant, thus, the simulation of urban development is conducted based on the past development trend and the projection from two previous periods. While in some urban areas the rate of urban expansion in sequence periods is not constant. In this study, although the study area has experienced a varying rate of development, the proposed model assumes that the relationship between spatial variables and land change does not change. To train the model, 16 data layers of predictor variables for 2001 and the binary classification of the LULC map for the target variable in 2006 are prepared. The trained model is evaluated using 16 data layers of predictor variables for 2006 and the binary classification of LULC map for the target variable in 2011 and also the data layer of predictor variables for 2011 and the binary classification of LULC map for target variable in 2016. Then, the simulated map is compared with the actual map of 2011 using the confusion matrix to evaluate the performance of the configured model.

For the RF model configuration, first, a random sampling strategy is applied to produce a proper training dataset. The sampling strategy helps to increases the computational performance and improve the prediction accuracy of the model (Karimi et al. "An Enhanced Support Vector Machine Model for Urban Expansion Prediction"). Comparison of the LULC map of 2001 and 2006 shows that there are enough recently

developed cells in this period to create a reliable training dataset by a combination with a random sampling of undeveloped land cells. The second step is the configuration of the RF model, in which the model hyperparameters are regularized on the training dataset (sample dataset of 2001-2006) and the performance is evaluated on the testing dataset (whole dataset of 2006-2011 and 2011-2016). In this study, the RF method uses the CART algorithm for discovering transition rules. The RF algorithm generates initial large and complex trees by splitting the predictor variables into independent groupings based on binary decisions. As larger trees tend to overfit the training dataset and results in lower performance, the RF model parameters including the number of input variables selected at each node split from 1 to 16 and the total number of trees in the model from 1 to 1000 are adjusted.

Table 4.2: Target and Predictor Variables for Urban Expansion Modeling

Type	Variables
Target variable	Developed urban area (1) and undeveloped area (0)
Proximity predictor variables	Distance to CBD (km) Distance to suburb towns (km) Distance to the nearest highway (km) Distance to the nearest major road (km) Distance to the nearest streets (km) Distance to the nearest railway (km) Distance to the nearest urban area (km) Distance to the nearest greenspaces (km) Distance to the nearest water bodies (km)
Neighboring predictor variables	Number of developed cells in a 5×5 neighborhood Number of potential lands for development in a 5×5 neighborhood The most frequent LULC type in a 5×5 neighborhood
Physical predictor Variables	LULC type Slope (%) Elevation (m)
Social variable	Population density

By enhancing the number of trees, the generalization error decreases, and overfitting is not a problem. Reducing the number of predictor variables decreases the robustness of each individual tree of the model, but also reduces the computational complexity of the algorithm and the correlation between trees, which enhances the accuracy of the model. Therefore, optimization of the hyperparameters k and m leads to generalization error minimization. The performance of the RF model for producing the transition potential map is tested using Kappa statistics. In addition, the RF model is able to determine the most significant predictor variables in the process of urban expansion. The whole process of calibrating the RF-CA model and simulating the future patterns is presented in Figure 4.4.

The CA model uses the explicit transition rules derived from RF for the year 2001 together with the influence of neighborhood, a stochastic disturbance term, constraint information, and a threshold value to evaluate how land cells are converted from 2001 status to 2006 status. For the neighborhood effect, the proportion of neighboring developed land cells in 5×5 Moore's neighborhood is considered. For the stochastic variable $(1+(-\log \gamma)\, a)$, the parameter γ is a random number between 0 and 1, and the parameter a, the random disturbance, is set to 1 to present a small amount of uncertainty in the simulation. The constraint layer (Figure 4.5) determines areas that are entirely or partially protected for development. The simulation process is conducted by running the model iteratively and updating the spatial variables dynamically until the simulated map of 2006 corresponds to the actual map of 2006. The CA method uses discrete-time to renew the status of each cell step by step and there are multiple repetitions for obtaining

the results in urban simulation. Indeed, in each iteration of the RF-CA model configuration, an updated neighboring layer map is applied. In previous CA-based urban growth studies, the number of iterations was determined by a constant number between 100 and 200 (Li and Yeh "Data Mining of Cellular Automata's Transition Rules"), or the configuration is stopped by the total land consumption in a given period (Yang et al.).

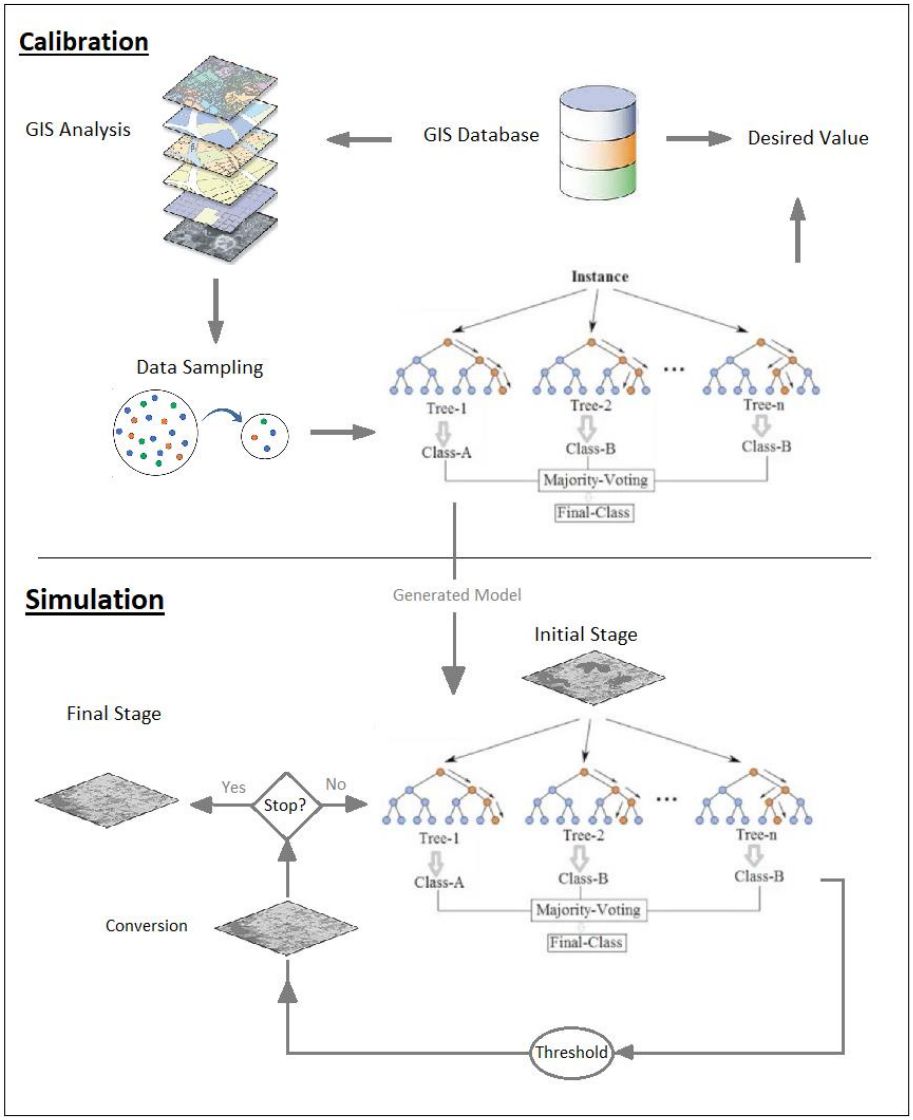


Figure 4.4: The Flowchart of the RF-CA Model for Simulating Urban Development

However, in this study, the simulation of urban development is conducted by different values for threshold from 50% to 95% and running the model iteratively until the accumulated mismatch between the actual and simulated urban development reaches the minimum amount. The value of this measure shows the sum of false positive and false negative through the confusion matrix (Suthaharan), which should be minimum with a balance between these two values. Also, to achieve an efficient model, the hyperparameters of the RF-CA model are adjusted by comparing the evaluation of various combinations of them. Training accuracy is used to evaluate the model performance over the training dataset, and testing accuracy, testing precision, testing sensitivity, testing specificity, F-score, Kappa statistics, and Area Under Curve (AUC) are used to evaluate the reliability and predictability of the adjusted model over the testing dataset for 2011 and 2016. A high value of these metrices demonstrates a higher performance and validity (Karimi et al. "Urban Expansion Modeling Using an Enhanced Decision Tree Algorithm").

The final step is to use the configured model to predict urban expansion for 2021 and 2026 by incorporating environmental constraints retrieved from GIS in the RF-CA model to formulate alternative urban development patterns. The environmental constraint is used to protect the natural environment and agricultural lands as urban development endangers these valuable lands. The model uses environmental constraints to indicate whether a land cell can be converted to a developed urban area or not regarding the environmental considerations. The environmental constraints can be defined based on the exclusion of environmentally sensitive areas and buffer distances to these areas. The

development alternatives are current trends, controlled urban development, and environmentally sustainable urban development scenarios for both years. In the current trend scenario, water bodies are fully protected from development as these land cells have not been decreased from 2011 to 2016. The controlled urban development scenario ensures a more intense commitment to spatially focused growth and preserving agricultural lands and resource conservation. In addition to exclusion of water bodies, agricultural lands and forests are partially excluded with the exclusion values of 30% and 50% respectively and the development is allowed just in a buffer of 2 0 0 meters in the developed urban area. The third scenario, environmentally sustainable urban development, shows a stronger constraint for limited growth and natural resource protection. The constraint layer same as the two previous scenarios specifies waterbodies as wholly unavailable for development, forests and agricultural lands are partially excluded with the exclusion values of 50% and 70% respectively, and the development is allowed just in a smaller buffer of 100 meters of the developed urban area.

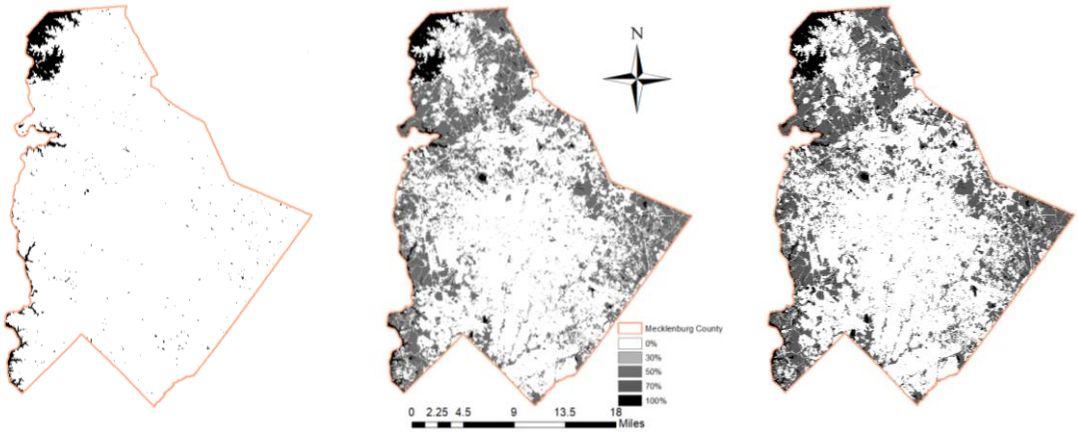


Figure 4.5: Constraint Layer for Exclusions of (a) Current Trends, (b) Controlled Urban Development, and (c) Environmentally Sustainable Urban Development Scenarios for Mecklenburg County, NC.

The conversion of natural or agricultural land into urban land-use imposes environmental costs. The constraints in the CA model help the simulation of developing land cells to be shifted to suitable sites to decrease the environmental costs. But the amount of the environmental cost for simulations should be indicated to assure the capability of the models for environment preservation. Thus, a potential environmental cost indicator is estimated to assess the alternative development simulations for environmentally sustainable development. By considering the natural environment and the land conversion from the natural environment to urban land uses, the environmental cost index can be written as:

$$EC = \frac{D \times A}{A}$$

where EC is the amount of the environmental cost; D is the size of the developed sites and A is the size of the natural environment. The index value is normalized by the

total area, in which a higher value shows that urban development leads to higher environmental costs, such as the loss of considerable proportions of suitable agricultural land.

4.5 Results

The RF- CA model is applied to Mecklenburg County, NC, USA for the simulation of development alternatives. The current development trend is not acceptable for sustainable development because of the environmental problems and costs. The positive aspect is that the simulation of possible development alternatives is helpful for urban planners and practitioners which can support preventing current environmental problems from arising again in the future through urban and regional planning. This study aims to simulate developmental alternatives under three scenarios.

First, a sampling approach is implemented to recognize the best training dataset, and then the model is assessed on the testing dataset. In this study, all the recently developed land cells of the 2001-2006 period and different numbers of undeveloped land cells are combined and evaluated for choosing the training dataset. The results of employing this sampling approach and the efficiency of different combinations for urban expansion modeling in Mecklenburg County for the 2001-2006 period when k equals to100 and m equals 1 are shown in Figure 4.6. The figure displays the relationships between the sampling points and the prediction error. It reveals the reduction in prediction error by enhancing the number of undeveloped land cells in the training dataset. The prediction error is 38.2% by using 10% of the recently developed land cells from undeveloped lands, and it is reduced to 35.0% and 32.5% by using 20% and 30% of

that data, respectively. The improvement rates are insignificant after the first 30% of the data. Therefore, this study used all the recently developed land cells of 2001-2006 and 30% of undeveloped land cells for further analysis of urban expansion modeling in Mecklenburg County and deriving the transition rules.

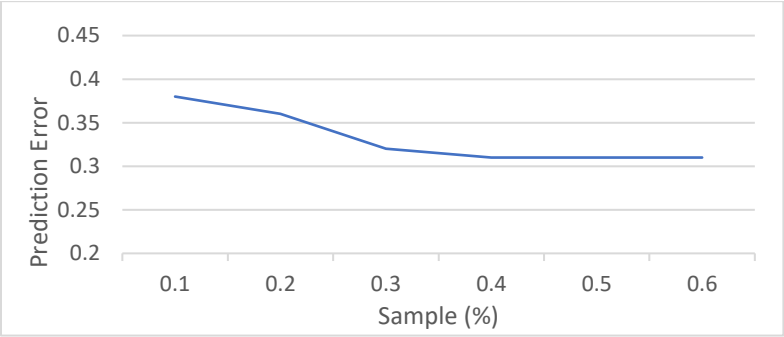


Figure 4.6: Random Sampling Rate and Prediction Error

In order to obtain optimum model performance, different amounts of k and m are tested, thus k ranges from 1 to 1000, and m ranges from 1 to 16, using intervals of 1. This way a total number of 16000 different RF models are generated. The resulting models are assessed using Kappa statistics and the most accurate model is the one with the highest Kappa value, the highest k, and the lowest m. Based on the findings presented in Figure 4.7, it can be realized that a higher amount of k and a lower amount of m leads to better results. To achieve the highest Kappa a RF model made up of 206 DTs with 1 split variable is adopted. Using the calibrated RF model, the transition potential map of urban development occurrence (Figure 4.8) is computed. The probability map displays a particular amount of probability between 0 and 1 for every single land cell that will be developed.

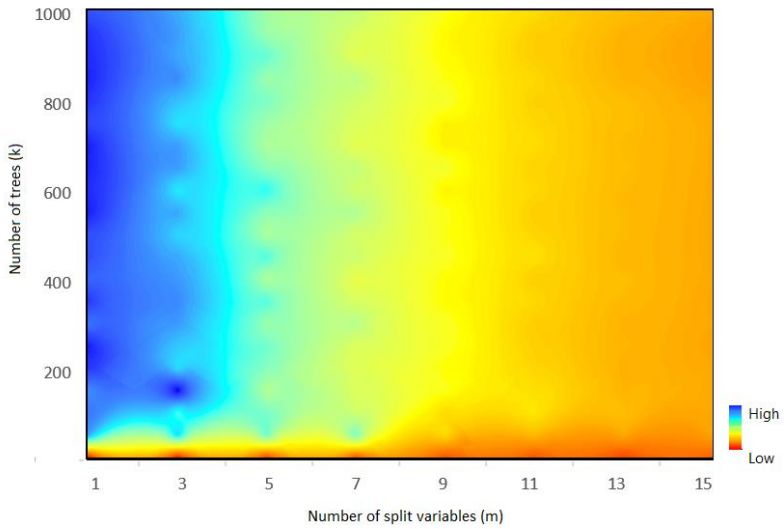


Figure 4.7: The Value of Kappa Statistics with Respect to the Number of Trees (k) and the Number of Random Split Variables (m)

This method cannot designate the quantity and location of development but integrating with the CA model can resolve the problem. Thereafter, the obtained probability of development is allocated in the entire map and can be quantified. The CA model used the transition potential map derived from RF for the year 2006 together with the proportion neighboring developed land cells in 5×5 Moore's neighborhood, a stochastic disturbance term in which γ is a random number between 0 and 1 and the parameter a is set to 1, constraint information from Figure 4.5a, and a threshold value to evaluate how land cells were converted from 2001 status to 2006 status. The simulation process is conducted by different values for threshold (0.6, 0.7, 0.8, and 0.9) and running the model iteratively and updating the neighborhood variable dynamically until the accumulated mismatch between the simulated map and the actual map reaches the minimum amount. As more iterations enhance the danger of adding False Negative land cells, the number of iterations was specified precisely. The threshold of 0.8 and 98 and

106 iterations led to the least mismatch between actual and simulated map for 2011 and 2016 maps, respectively. Thus, 100 iterations would be proper for predicting the next periods.

Table 4.3 shows testing accuracy, testing precision, testing sensitivity, testing specificity, F-score, MCC, Kappa, and AUC values which are used to evaluate the reliability and predictability of the adjusted model over the testing dataset by comparing the observed and predicted map of 2011 and 2016 through confusion matrix. The training accuracy for simulating the urban growth in 2001-2006 is 100% and the best accuracy is 99% and 98% in 2006-2011 and 2011-2016 respectively. The values of the evaluation metrices are high due to the higher amount of consistent land cells to converted land cells, therefore it is necessary to investigate the number of False Negative and False Positive land cells. Table 4.4 shows the number of True Negative, True Positive, False Negative, and False Positive land cells for 2011 and 2016 for the RF-CA and RF model. It shows that integrating the RF model with the CA model reduces the number of False Negative land cells in addition to allowing developing different alternative urban expansion simulations. Figure 4.9 shows the predicted map of the study area over the testing dataset. As it is shown, the mismatched land cells are dispered clusters near previous urban environment.

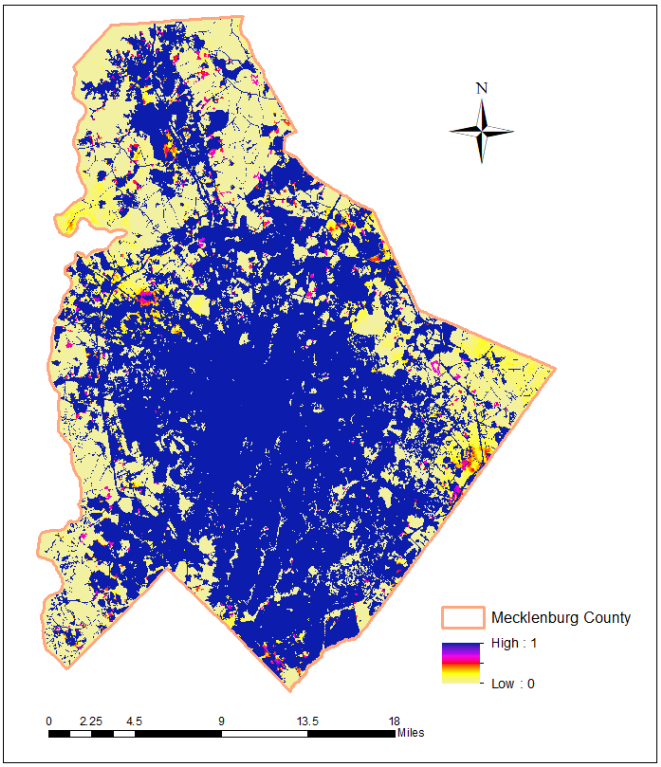


Figure 4.8: Transition Potential Map of Mecklenburg County, NC for 2011

Table 4.3: The Performance Achieved from the Cross-tabulation of the Overlay Analysis of Actual and Predicted Urban Development Map of Mecklenburg County, NC for 2011 and 2016

Period	Accuracy	Precision	NPR	Sensitivity	Specificity	F_S	MCC	Kappa	AUC
2006-2011	0.99	0.99	0.98	0.99	0.99	0.99	0.98	0.98	0.99
2011-2016	0.98	0.99	0.97	0.98	0.98	0.99	0.96	0.96	0.98

Table 4.4: The Number of TN, TP, FN, and FP Land Cells for 2011 and 2016 for the RF and RF-CA Model

Models	2011				2016			
	TN	TP	FP	FN	TN	TP	FP	FN
RF	589329	969557	2995	7349	553370	1001870	5606	8384
RF-CA	589278	969983	3046	6923	553258	1002518	5718	7736

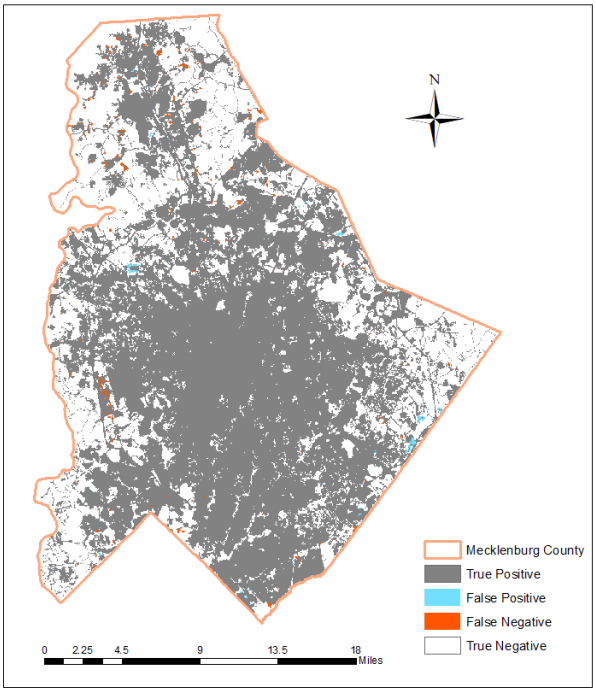


Figure 4.9: The Predicted Map of Mecklenburg County, NC for 2011 Consideing True Positive, True Negative, False Positive and False Negative Land Cells

As shown in Figure 4.3 the amount of urban development in the 2001-2006 interval is more than 2006-2011 and 2011-2016 intervals and it seems that if the

simulation of urban development is only based on the data from 2001-2006 period, the simulated urban areas will be larger than the actual urban areas for the next periods with a 5-year interval. In addition, the geographical setting of this study area is relatively complex, with different geomorphologic characteristics (e.g. forests, wetlands, and lakes) and some suburban centers (5 towns). However, using the transition rules derived from the 2001-2006 dataset, high performance in the simulation of urban development in 2006-2011 and 2011-2016 was obtained (Table 4.3). This is because of the ability of the RF model to realize complex relationships.

As mentioned before, RF is able to give estimates of the importance of predictor variables. Based on the regularized RF model, distance to highways, distance to suburb towns and distance to the city center are the most important predictor variables for simulation development in this urban area. On the other hand, distance to forests, population, and neighboring potential cells are the least important (Figure 4.10).

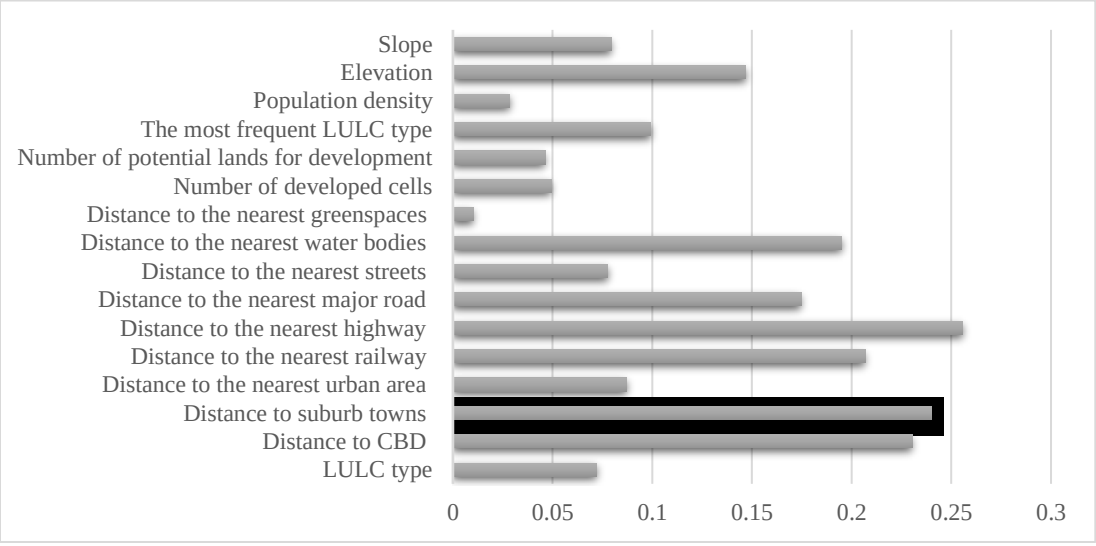


Figure 4.10: The Importance of Predictor Variables

The RF-CA model is approximately accurate at detecting development as reflected by components of the agreement. The reliability and performance of the RF-CA model are because of the comparatively accurate RF transition potential maps. Thus, this study highlights the potential of the RF-CA model for urban development simulation. Figure 4.11 shows the actual urban development under the current trend scenario in Mecklenburg County for 2001, 2006, 2011, and 2016 and the predicted urban development for 2021 and 2026. The figure shows that urban development will appear near current or newly built urban clusters or adjacent to the major roads.

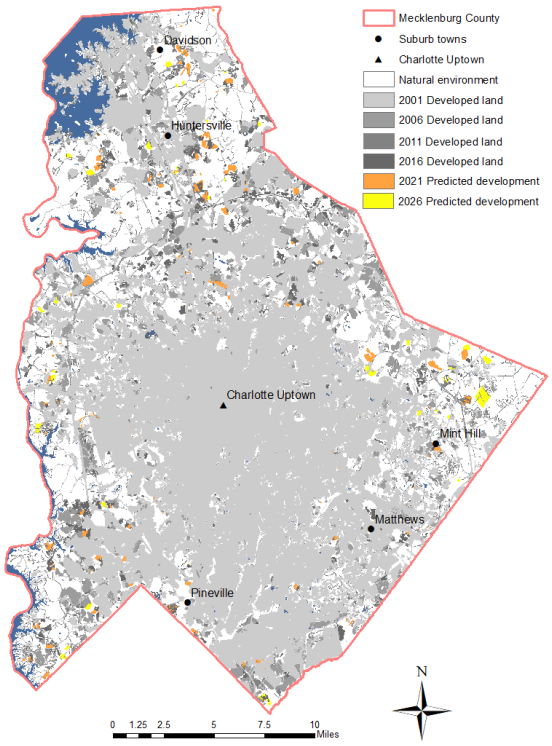


Figure 4.11: The Predicted Urban Development Map of Mecklenburg County, NC for 2021 and 2026

The configured model is used to predict urban expansion for 2021 and 2026 under controlled and environmentally sustainable urban development scenarios by incorporating environmental constraints retrieved from GIS in the RF-CA model (Figure 4.11). The model uses environmental constraints to indicate whether a land cell can be converted to a developed urban area or not regarding environmental considerations. It is obvious that the controlled and environmentally sustainable urban development scenarios decrease the rate of urban expansion into the natural environment in this county. Figure 4.12 shows the simulated urban development map for 2021 and 2026 under controlled development and environmentally sustainable urban development scenarios. The transportation network and the urban extent exhibit assurance to focused growth and no new major planned developments appear in these scenarios. This means that there is no more space in this county for development by considering controlled and environmentally sustainable urban development in decision-making.

The environmental cost index is used to evaluate the performance of simulated urban developments under the three scenarios in relation to sustainable development. Table 4.4 shows the environmental cost index for simulated urban developments under current trends, controlled urban development, and environmentally sustainable urban development scenarios for 2021 and 2026. It can easily be seen that the current trends in 2021 and 2026 are dispersed patterns which reinforces that current trend development in Mecklenburg County endangers the natural environment. Table 4.4 shows that the controlled and environmentally sustainable urban development scenarios for 2021 and 2026 are much higher compact. It can also be seen from these maps that merely achieving

controlled development will not minimize environmental costs unless more intense environmental constraints are also taken into consideration. The third scenario shows the effects of more intense environmental constraints. This scenario leads to highly compact and less horizontally development which produces low environmental costs.

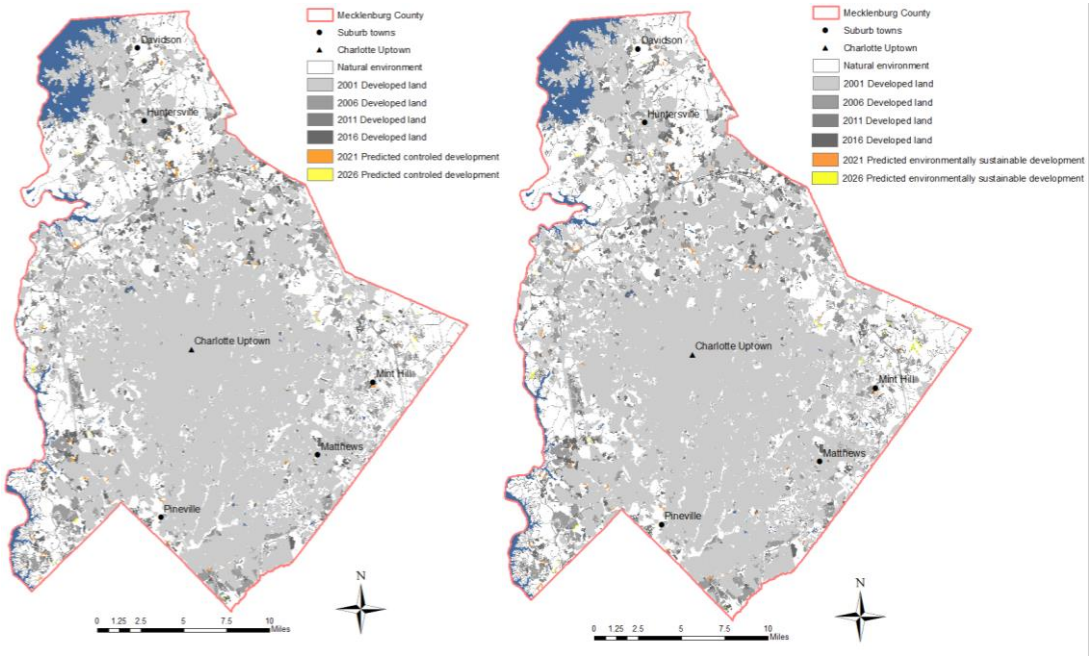


Figure 4.12: Simulated Urban Development Map for (a) 2021 and 2026 Under Controlled Development Scenario, (b) 2021 and 2026 Under Environmentally Sustainable Urban Development Scenario

Table 4.5: The Environmental Cost Index for Simulated Urban Developments Under Current Trends, Controlled Urban Development, and Environmentally Sustainable Urban Development Scenarios for 2021 and 2026

	Predicted year	Environmental Cost
Current trends	2021	4%
Controlled urban development	2021	2%
Environmentally sustainable development	2021	0.8%
Current trends	2026	2.8%
Controlled urban development	2026	1%
Environmentally sustainable development	2026	0.6%

4.6 Conclusion

The prediction of future urban development patterns presents helpful information for urban planning and environmental management and gives an estimation of the potential impacts of urban development. Sustainable urban development plans can be expressed to alleviate the negative impacts and control the amount and extent of the development before it actually occurs. The RF model is spatially explicit and most importantly, allows a much deeper understanding of the predictor variables and the formation of the urban spatial pattern. However, this approach can be integrated with the CA model to quantify and allocate the development. While CA is unable to define transition rules for producing a realistic simulation of urban areas. The integration of RF and CA methods substantially overcomes the shortcomings of each method, effectively models spatiotemporal urban development patterns, and explicitly describes urban growth dynamics in the urban areas with varying rates of development.

Applying the proposed RF-CA model in Mecklenburg County, NC, USA in 2001-2016, the results demonstrate that the model can achieve reliable and high accurate results and significantly mitigate the adverse effects of the varying rate of development. The results show that the precise regulation of the RF-CA model's parameters and a training dataset containing all the recently developed cells and 30% of them from undeveloped cells considerably enhances the accuracy and performance of the model. The RF analysis of predictor variables shows that the proximity to highways, proximity to suburb towns, and proximity to city center are the most significant factors for simulation development in this urban area. The current urban development in

Mecklenburg County for 2021 and 2026 will appear near current or newly built urban clusters or adjacent to the major roads, the controlled and environmentally sustainable urban development scenarios for 2021 and 2026 are much higher compact and minimize environmental costs.

CHAPTER V

CONCLUSION

The uncontrolled, rapid and low-density urban growth affects the environment, economy, and society significantly. Thus, an appropriate insight into how and where urban growth occurs and its causes, and consequences is required for urban planners, geographers, and practitioners for managing future urban plans, and sustainable development. Urban growth models facilitate understanding the dynamic and convoluted process of urban expansion, assessing causal factors, and analyzing the consequences of policies. This dissertation studied urban growth modeling and prediction using machine learning algorithms to model urban expansion patterns effectively and efficiently. An introduction to the research problem was provided in Chapter I, where I reviewed the existing literature and the concepts.

Chapter II explored the capabilities of the SVM method, as a powerful machine learning algorithm, using three different data sampling methods, regulating, and evaluating the model with the emphasis on feature selection. The implementation of the developed model in Guilford County, NC, throughout 2001-2011, as a case study, demonstrated highly accurate and reliable results. According to the summary of the most widely used factors in previous studies conducted by Musa et al., the historical spatiotemporal LULC change patterns in the study area, and data availability, 19 predictor variables were considered including population density, current LULC type of a cell, proximity to the city center, urban built area, the nearest highways, major

roads,streets, railways, water bodies, green spaces, the number of potential cells for expansion, water body cells, forest cells, wetlands cells, barren land cells, open-space developed cells, low-intensity developed cells, medium-intensity developed cells,high-intensity developed cells in a 3*3 Moore's neighborhood of each cell. As irrelevant and redundant predictor variables affect the modeling results through overfitting and poor generalization feature selection process was conducted to determine the most significant predictor variables using information gain measure. The result showed, the current LULC type was the most significant predictor, following with distance to highways, neighboring with medium-intensity developed areas, neighboring with potential lands for urban expansion, and distance to water bodies as the next high-ranked predictors. Neighboring with forests, neighboring with developed open spaces, neighboring with water bodies, neighboring with wetlands, and neighboring with barren lands were the least significant predictors.

A comparative study in which the models were investigated in the same case study was addressed in Chapter III, which will enable decision-makers to thoroughly understand the performance of the models and their strengths and limitations. This study compared five machine learning methods, including DT, RF, SVM, LR, and ANN. The models were trained and validated to simulate urban expansion in Mecklenburg County, NC, USA over the 2001-2016 period. The results showed that RF is superior to other models concerning evaluation metrices, the number of hyperparameters, run time, the need for data preparation, and the number of FP and FN land cells in the prediction in this case study. The importance of predictor variables was analyzed by CART and RF

methods and the analysis recognized proximity to urban areas, highways, city centers, and uptown as the most critical variables and proximity to green-spaces and population density as the least important variables.

Developing different urban growth scenarios with the goal of sustainable urban growth management were investigated in Chapter IV. Innovative integration of RF and CA (RF-CA) is suggested to simulate urban development under three urban growth scenarios, including current trends, controlled urban development, and environmentally sustainable urban development. While current trends allow the urban fringe to be uncontrollably developed, the controlled and environmentally sustainable urban development scenarios constrain future developments and reduce the environmental implications. A variety of data sampling strategies, predictor variables, and model configurations were explored to enhance the accuracy and predictability of the proposed model. The model was calibrated using spatiotemporal data of 2001-2016 and was applied to simulate future urban developments in 2021 and 2026 for rapidly urbanizing Mecklenburg County, NC, USA. The accuracy and reliability of the model were evaluated by apposite evaluation metrics, and the simulated urban development patterns were examined using a cost indicator from the perspective of sustainable development. The results demonstrated that the proposed model is a fast, high-performance, and accurate model with low uncertainty; therefore, it can be effectively utilized to evaluate a wide range of urban development policies and scenarios and support decision-making to achieve the goal of establishing sustainable development in Mecklenburg County.

The results from three studies demonstrated that the sampling strategy, the regulation of models, and the importance of the predictor variables depend on the case study, the study period, and predictor variables. More precisely, the results depend on the data. This conclusion raises the data challenge. Past studies mostly have used environmental and physical data, and the use of social and economic data due to their availability was limited. Collecting urban local and regional planning policies such as restricted areas (Alsharif and Pradhan; J. J. Arsanjani et al.), social factors such as income and affluence (Echenique et al.; Haase et al.), and economic factors such as housing/land prices and rent (Waddell), job availability and job growth (Z. Hu and C. P. Lo), if possible, is a future direction. Nowadays social media affects various aspects of human life. What people say or how they act in social media may affect urban development, as these are people's thoughts and decisions and indirectly affect the economy and society. This data can include in modeling and its importance in the modeling can be another future direction. The size of the study area is an issue in urban growth modeling and prediction, due to lack of memory mostly small study areas were chosen. Selecting a large case study such as a state to investigate the models reliability, which needs a supercomputer with a large memory, should be addressed in future studies.

REFERENCES

Aburas, M. M. et al. "Spatio-Temporal Simulation and Prediction of Land-Use Change Using Conventional and Machine Learning Models: A Review." *Environmental monitoring and assessment*, vol. 191, no. 4, 2019, p. 205.

Aburasa, M.M. et al. "The Simulation and Prediction of Spatio-Temporal Urban Growth Trends Using Cellular Automata Models: A Review." *International Journal of Applied Earth Observation and Geoinformation*, vol. 52, 2016, pp. 380-389.

Adams, J. S. "Residential Structure of Midwestern Cities." *Annals of the Association of American Geographers*, vol. 60, no. 1, 1970, pp. 37–62.

Agarwal, C. et al. "A Review and Assessment of Land Use Change Models: Dynamics of Space, Time, and Human Choice." *Gen. Tech. Rep. NE-297. Newton Square, PA: US Department of Agriculture, Forest Service, Northeastern Research Station*, vol. 61, 2002, p. 297.

Alcoforado, M. J. and H. Andrade. "Global Warming and the Urban Heat Island." *Urban Ecology*, Springer, 2008, pp. 249-262.

Alsharif, A.A. and B. Pradhan. "Urban Sprawl Analysis of Tripoli Metropolitan City(Libya) Using Remote Sensing Data and Multivariate Logistic Regression Model." *J.Indian Soc. Remote Sens*, vol. 42, no. 1, 2014, pp. 149-163.

Amato, F. et al. "Using Spatiotemporal Analysis in Urban Sprawl Assessment and Prediction Computational Science and Its Applications." *ICCSA 2014. Springer*, 2014, pp. 758-773.

Angel, S. et al. "The Dimensions of Global Urban Expansion: Estimates and Projections for All Countries, 2000-2050." *Progress in Planning*, vol. 75, no. 2, 2011, pp. 53-107.

Armesto, J. J. et al. "Old-Growth Temperate Rainforests of South America: Conservation, Plant-Animal Interactions, and Baseline Biogeochemical Processes." *Old-Growth Forests*, edited by C. Wirth et al., Springer, 2009, pp. 367-390.

Arnold Jr, C. L. and C. J. Gibbons. "Impervious Surface Coverage: The Emergence of a Key Environmental Indicator." *Journal of the American planning association*, vol. 62, no. 2, 1996, pp. 243-258.

Arsanjani, J. J. et al. "Spatiotemporal Simulation of Urban Growth Patterns Using Agent-Based Modeling: The Case of Tehran." *Cities*, vol. 32, 2013, pp. 33-42.

Arsanjani, J.J. et al. "Integration of Logistic Regression, Markov Chain and Cellular Automata Models to Simulate Urban Expansion." *Int. J. Appl. Earth Obs. Geoinf.*, vol. 21, 2013, pp. 265-275.

Azhagusundari, B. and A. S. Thanamani. "Feature Selection Based on Information Gain." *International Journal of Innovative Technology and Exploring Engineering(IJITEE)*, vol. 2, no. 2, 2013, pp. 18-21.

Babakan, A. S. and M. Taleai. "Impacts of Transport Development on Residence Choice of Renter Households: An Agent-Based Evaluation." *Habitat International*, vol. 49, 2015, pp. 275-285.

Baja, S. and S. Arif. "Gis-Based Modelling of Land Use Dynamics Using Cellular Automata and Markov Chain." *J Environ Earth Sci*, vol. 4, 2014, pp. 61-66.

Balk, Gene. "Census: Seattle Is the Fastest-Growing Big City in the U.S." *Seattle Times*, 2014.

Basheera, I.A. and M. Hajmeer. "Artificial Neural Networks: Fundamentals, Computing, Design, and Application." *Journal of Microbiological Methods*, vol. 42, 2000, pp. 3-31.
Batty, M. "Cellular Automata and Urban Form: A Primer." *Journal of the American planning association*, vol. 63, 1997, pp. 266-274.
Batty, M. and Y. Xie. "From Cells to Cities." *Environment and Planning B: Planning and Design*, vol. 21, no. 7, 1994, pp. S31-S48.Batty, M. et al. "Modeling Urban Dynamics through Gis-Based Cellular Automata." *Computers, Environment and Urban Systems*, vol. 23, 1999, pp. 205-233.
Belgiu, M. and L. Drăguţ. "Randomforest in Remote Sensing: A Reviewof Applications and Future Directions." *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, no. 24-31, 2016.
Belousov, A. I. et al. "Application Aspects of Support Vector Machines." *Journal of Chemometrics*, vol. 16, 2002, pp. 482-489.
Bengston, D. N. et al. "Public Policies for Managing Urban Growth and Protecting Open Space: Policy Instruments and Lessons Learned in the United States." *Landscape and urban planning*, vol. 69, no. 2-3, 2004, pp. 271-286.
Berberoğlu, S. et al. "Cellular Automata Modeling Approaches to Forecast Urban Growth for Adana, Turkey: A Comparative Approach." *Landscape and urban planning*, vol. 153, 2016, pp. 11-27.
Bhatta, B. "Causes and Consequences of Urban Growth and Sprawl." *Analysis of Urban Growth and Sprawl from Remote Sensing Data*, Springer, 2010, pp. 17-36.
Black, D. and V. Henderson. "A Theory of Urban Growth." *Journal of political economy*, vol. 107, no. 2, 1999, pp. 252-284.
Blumenfeld, H. "The Tidal Wave of Metropolitan Expansion." *Journal of the American Institute of Planners*, vol. 20, 1954, pp. 3–14.
Boser, B.E. et al. "A Training Algorithm for Optimal Margin Classifiers." *Proceedings of the fifth annual workshop on Computational learning theory*, ACM, 1992, pp. 144-152.
Bramley, G. and S. Power. "Urban Form and Social Sustainability: The Role of Density and Housing Type." *Environment and Planning B: Planning and Design*, vol. 36, no. 1, 2009, pp. 30-48.
Breiman, L. "Bagging Predictors." *Machine learning*, vol. 24, no. 2, 1996, pp. 123-140.
---. "Random Forests." *Mach. Learn.*, vol. 45, 2001, pp. 5-32.
Breiman, L. et al. *Classification and Regression Trees.* Chapman & Hall, 1984.
Brovkin, V. et al. "Role of Land Cover Changes for Atmospheric Co2 Increase and Climate Change During the Last 150 Years." *Global Change Biology*, vol. 10, no. 8, 2004, pp. 1253-1266.
Brown, D.G. et al. *Advancing Land Change Modeling: Opportunities and Research Requirements.* National Academies Press, 2014.
Bruegmann, R. *Sprawl: A Compact History.* The University of Chicago Press, 2005.
Burton, E. et al. *The Compact City: A Sustainable Urban Form? .* Routledge, 2003.
Burton, E. et al. "The Compact City and Urban Sustainability: Conflicts and Complexities." *The Compact City: A Sustainable Urban Form*, 1996, pp. 231-247.
Camagni, R. et al. "Urban Mobility and Urban Form: The Social and Environmental Costs of Different Patterns of Urban Expansion." *Ecological economics*, vol. 40, no. 2, 2002, pp. 199-216.
Carolina Population Center. "Urbanization Trends." University of North Carolina https://demography.cpc.unc.edu/2015/01/05/urbanization-trends/.

Cervero, R. and J. Landis. "Twenty Years of the Bay Area Rapid Transit System: Land Use and Development Impacts." *Transportation Research Part A: Policy and Practice*, vol. 31, no. 4, 1997, pp. 309-333.
Chandrashekar, G. and F. Sahin. "A Survey on Feature Selection Methods." *Computers & Electrical Engineering*, vol. 40, no. 1, 2014, pp. 16-28.
Chapelle, O. et al. "Support Vector Machines for Histogram-Based Image Classification." *IEEE transactions on Neural Networks*, vol. 10, no. 5, 1999, pp. 1055-1064.
Chen, S. et al. "Remote Sensing and Gis for Urban Growth Analysis in China." *Photogrammetric Engineering and Remote Sensing*, vol. 66, no. 5, 2000, pp. 593-598.
Cheng, J. and I. Masser. "Urban Growth Pattern Modeling: A Case Study of Wuhan City, Pr China." *Landscape and urban planning*, vol. 62, no. 4, 2003, pp. 199-217.
Cheng, J. et al. "Understanding Urban Growth System: Theories and Methods." *8th international conference on computers in urban planning and urban management*, 2003.
City of Charlotte. "Economic Development " https://charlottenc.gov/CityCouncil/focus-areas/Pages/EconomicDevelopmentFocusArea.aspx.
Clarke, K.C. et al. "A Self-Modifying Cellular Automaton Model of Historical Urbanization in the San Francisco Bay Area." *Environment and Planning B: Planning and Design*, vol. 24, 1997, pp. 247–261.
Cohen, B. "Urbanization in Developing Countries: Current Trends, Future Projections, and Key Challenges for Sustainability." *Technology in Society*, vol. 28, 2006, pp. 63-80.
Cohen, J. "A Coefficient for Agreement for Nominal Scales." *Educational and Psychological Measurement*, vol. 20, no. 37-46, 1960.
Coseo, P. and L. Larsen. "How Factors of Land Use/Land Cover, Building Configuration, and Adjacent Heat Sources and Sinks Explain Urban Heat Islands in Chicago." *Landscape and urban planning*, vol. 125, 2014, pp. 117–129.
Cox, D.R. "The Regression Analysis of Binary Sequences " *Journal of the Royal Statistical Society*, vol. Series B, no. 20, 1958, pp. 215-242.
Cristianini, N. and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, 2000.
Dadvand, P. et al. "Green Spaces and General Health: Roles of Mental Health Status, Social Support, and Physical Activity." *Environment international*, vol. 91, 2016, pp. 161-167.
Darin-Drabkin, H. *Land Policy and Urban Growth* Pergamon Press, 1977.
de Noronha Vaz, E. et al. "A Multi-Scenario Forecast of Urban Change: A Study on Urban Growth in the Algarve." *Landsc Urban Plan*, vol. 104, 2012, pp. 201-211.
Debeljak, M. and S. Dzeroski. "Decision Trees in Ecological Modeling." *Modeling Complex Ecological Dynamics*, Springer, 2011, pp. 197-209.
Deep, S. and A. Saklani. "Urban Sprawl Modeling Using Cellular Automata." *Egypt J Remote Sens Space Sci*, vol. 17, 2014, pp. 179-187.
del Mar López, T. et al. "Urban Expansion and the Loss of Prime Agricultural Lands in Puerto Rico." *Ambio: a Journal of the Human environment*, vol. 30, no. 1, 2001, pp. 49-54.
Delen, D. et al. "Measuring Firm Performance Using Financial Ratios: A Decision Tree Approach." *Expert Systems with Applications*, vol. 40, no. 10, 2013, pp. 3970-3983.
Deng, Y. and S. Srinivasan. "Habitat International." *Urban land use change and regional access: A case study in Beijing, China*, vol. 51, 2016, pp. 103-113.
Dietterich, T. G. "Ensemble Methods in Machine Learning." *International workshop on multiple classifier systems*, vol. 1-15, Springer, 2000.
Duranton, G. and M. A. Turner. "Urban Growth and Transportation." *Review of Economic Studies*, vol. 79, no. 4, 2012, pp. 1407-1440.

Echenique, M. H. et al. "Growing Cities Sustainably: Does Urban Form Really Matter?" *Journal of the American planning association*, vol. 78, no. 2, 2012, pp. 121-137.
Ekins, P. *Economic Growth and Environmental Sustainability: The Prospects for Green Growth.* Routledge, 2002.
Evans, J. S. et al. "Modeling Species Distribution and Change Using Random Forest." *Predictive Species and Habitat Modeling in Landscape Ecology*, Springer, 2011, pp. 139-159.
Feng, Y. et al. "Modeling Urban Growth with Gis Based Cellular Automata and Least Squares Svm Rules: A Case Study in Qingpu–Songjiang Area of Shanghai, China." *Stochastic Environmental Research and Risk Assessment*, vol. 30, no. 5, 2016, pp. 1387–1400.
Friedl, M. A. and C. E. Brodley. "Decision Tree Classification of Land Cover from Remotely Sensed Data." *Remote Sensing of Environment*, vol. 61, no. 3, 1997, pp. 399-409.
Fry, J. et al. "Completion of the 2006 National Land Cover Database for the Conterminous United States." *Photogrammetric Engineering and Remote Sensing*, vol. 77, no. 9, 2011, pp. 858-864.
Gislason, P. O. et al. "Random Forests for Land Cover Classification." *Pattern Recognition Letters*, vol. 27, no. 4, 2006, pp. 294-300.
Gober, P. and E. K. Burns. "The Size and Shape of Phoenix's Urban Fringe." *Journal of Planning Education and Research*, vol. 21, 2002, pp. 379–390.
Graves, W. and H. A. (Eds.) Smith. "Charlotte, Nc: The Global Evolution of a New South City." *University of Georgia Press*, 2010.
Greene, R. P. "The Farmland Conversion Process in a Polynucleated Metropolis." *Landscape and Urban Planning*, vol. 36, 1997, pp. 291–300.
Haase, D. et al. "Endless Urban Growth? On the Mismatch of Population, Household and Urban Land Area Growth and Its Effects on the Urban Debate." *PLoS One*, vol. 8, no. 6, 2013, p. e66531.
Hart, J. F. "The Perimetropolitan Bow Wave." *Geographical Review*, vol. 81, 1991, pp. 35–51.
Haughton, G. and C. Hunter. *Sustainable Cities.* Jessica Kingsley, 1994. *Regional Studies Association, Regional Policy and Development Series 7.*
Hersperger, A. M. et al. "Urban Land-Use Change: The Role of Strategic Spatial Planning." *Global Environmental Change*, vol. 51, 2018, pp. 32-42.
Homer, C. et al. "Completion of the 2001 National Land Cover Database for the Conterminous United States." *Photogrammetric Engineering and Remote Sensing*, vol. 73 (4), 2007, pp. 337-341.
Homer, C.G. et al. "Completion of the 2011 National Land Cover Database for the Conterminous United States-Representing a Decade of Land Cover Change Information." *Photogrammetric Engineering and Remote Sensing*, vol. 81, no. 5, 2015, pp. 345-354.
Horner, M.W. and D. Schleith. "Analyzing Temporal Changes in Land-Use–Transportation Relationships: A Lehd-Based Approach." *Applied Geography*, 2012, pp. 491-498.
Hosseinali, F. et al. "Agent-Based Modeling of Urban Land-Use Development, Case Study: Simulating Future Scenarios of Qazvin City." *Cities*, vol. 31, 2013, pp. 105–113.
Hu, Z. and C. P. Lo. "Modeling Urban Growth in Atlanta Using Logistic Regression." *Computers, Environment and Urban Systems*, vol. 31, no. 6, 2007, pp. 667-688.
Hu, Z. and C.P. Lo. "Modeling Urban Growth in Atlanta Using Logistic Regression." *Computers, Environment and Urban Systems*, vol. 31, 2007, pp. 667-688.
Huang, B. et al. "Support Vector Machines for Urban Growth Modeling." *Geoinformatica*, vol. 14, no. 1, 2010, p. 83.
Huang, B. et al. "Land-Use-Change Modeling Using Unbalanced Support-Vector Machines." *Environment and Planning B: Planning and Design*, vol. 36, 2009, pp. 398-416.

Huang, C. et al. "An Assessment of Support Vector Machines for Land Cover Classification." *International Journal of Remote Sensing*, vol. 23, no. 4, 2002, pp. 725-749.
Huang, X. and L. Zhang. "An Svm Ensemble Approach Combining Spectral, Structural, and Semantic Features for the Classification of High-Resolution Remotely Sensed Imagery." *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 1, 2013, pp. 257-272.
IPUMS-USA. University of Minnesota www.ipums.org. 2019.
Jantz, C. A. et al. "Using the Sleuth Urban Growth Model to Simulate the Impacts of Future Policy Scenarios on Urban Land Use in the Baltimore-Washington Metropolitan Area." *Environment and Planning B: Planning and Design*, vol. 31, no. 2, 2004, pp. 251-271.
Jiang, B. and X. Yao. "Geospatial Analysis and Modelling of Urban Structure and Dynamics." *Berlin: Springer Science & Business Media.*, vol. 99, 2010.
Jiang, F. et al. "Measuring Urban Sprawl in Beijing with Geo-Spatial Indices." *Journal of Geographical Sciences*, vol. 17, no. 4, 2007, pp. 469-478.
Kamusoko, C. and J. Gamba. "Simulating Urban Growth Using a Random Forest-Cellular Automata (Rf-Ca) Model." *ISPRS International Journal of Geo-Information*, vol. 4, no. 2, 2015, pp. 447-470.
Karimi, F. et al. "An Enhanced Support Vector Machine Model for Urban Expansion Prediction." *Computers, Environment and Urban Systems*, vol. 75, no. 61-75, 2019.
---. "Urban Expansion Modeling Using an Enhanced Decision Tree Algorithm." *Geoinformatica*, 2019, pp. 1-16.
Karimi, F. et al. "Land Suitability Evaluation for Organic Agriculture of Wheat Using Gis and Multicriteria Analysis." *Papers in Applied Geography*, vol. 4, no. 3, 2018, pp. 326-342.
Kelley, A. C. and J. G. Williamson. "Population Growth, Industrial Revolutions, and the Urban Transition." *Population and Development Review*, 1984, pp. 419-441.
Kim, H. et al. "A Geographic Assessment of the Economic Development Impact of Korean High-Speed Rail Stations." *Transport Policy*, vol. 66, 2018, pp. 127-137.
Kim, J.P. and J.M. Guldmann. "Land-Use Planning and the Urban Heat Island." *Environment and Planning B: Planning and Design*, vol. 41, 2014, pp. 1077 – 1099.
Kim, Y. et al. "Feature Selection in Data Mining." *Data Mining: Opportunities and Challenges*, edited by J. Wang, Idea Group, 2003, pp. 80–105.
Kucsicsa, G. and I. Grigorescu. "Urban Growth in the Bucharest Metropolitan Area: Spatial and Temporal Assessment Using Logistic Regression." *Journal of Urban Planning and Development*, vol. 144, no. 1, 2018, p. 05017013.
Landis, J. R. and G. G. Koch. "An Application of Hierarchical Kappa-Type Statistics in the Assessment of Majority Agreement among Multiple Observers." *Biometrics*, 1977, pp. 363-374.
Lee, C. and G. G. Lee. "Information Gain and Divergence-Based Feature Selection for Machine Learning-Based Text Categorization." *Information processing & management*, vol. 42, no. 1, 2006, pp. 155-165.
Li, H. et al. "Devel-Oping Alternative Forest Cutting Patterns: A Simulation Approach." *Lands Ecology*, vol. 8, 1993, pp. 63-75.
Li, J. et al. "An Examination of Historical and Future Land Use Changes in Uganda Using Change Detection Methods and Agent-Based Modelling." *African Geographical Review*, vol. 35, no. 3, 2016, pp. 247–271.
Li, X. and A. G. O. Yeh. "Calibration of Cellular Automata by Using Neural Networks for the Simulation of Complex Urban Systems." *Environment and Planning A*, vol. 33, no. 8, 2001, pp. 1445-1462.

---. "Data Mining of Cellular Automata's Transition Rules." *International Journal of Geographical Information Science*, vol. 18, 2004, pp. 723-744.
---. "Neural-Network-Based Cellular Automata for Simulating Multiple Land Use Changes Using Gis." *International Journal of Geographical Information Science*, vol. 16, no. 4, 2002, pp. 323-343.
Li, X. et al. "Projecting Global Urban Area Growth through 2100 Based on Historical Time Series Data and Future Shared Socioeconomic Pathways." *Earth's Future*, vol. 7, no. 4, 2019, pp. 351-362.
Liao, FH. and YD. Wei. "Modeling Determinants of Urban Growth in Dongguan, China: A Spatial Logistic Approach." *Stoch Env Res Risk Assess*, vol. 28, 2014, pp. 801-816.
Ließ, M. et al. "Uncertainty in the Spatial Prediction of Soil Texture: Comparison of Regression Tree and Random Forest Models." *Geoderma*, vol. 170, 2012, pp. 70-79.
Lin, J. and A. Yang. "Does the Compact-City Paradigm Foster Sustainability? An Empirical Study in Taiwan." *Environment and Planning B: Planning and Design*, vol. 33, no. 3, 2006, p. 365.
Lindh, G. "Urbanization: A Hydrological Headache." *Ambio: a Journal of the Human environment*, 1972, pp. 185-201.
Lippmann, R. . "An Introduction to Computing with Neural Nets." *IEEE ASSP Magazine*, vol. 4, no. 2, 1987, pp. 4 - 22.
Liu, X. et al. "Simulating Complex Urban Development Using Kernel-Based Non-Linear Cellular Automata." *Ecological Modeling*, vol. 211, 2008, pp. 169-181.
Liu, Y. and S. Phinn. "Mapping the Urban Development of Sydney (1971–1996)with Cellular Automata in a Gis Environment." *J. Spat. Sci*, vol. 49, no. 2, 2004, pp. 57-74.
Marjanović, M. et al. "Landslide Susceptibility Assessment Using Svm Machine Learning Algorithm." *Engineering Geology*, vol. 123, no. 3, 2011, pp. 225-234.
Martellozzo, F. et al. "Modelling the Impact of Urban Growth on Agriculture and Natural Land in Italy to 2030." *Applied Geography*, vol. 91, 2018, pp. 156-167.
Marzluff, J. M. *Worldwide Urbanization and Its Effects on Birds. In Avian Ecology and Conservation in an Urbanizing World.* Springer, 2001.
Masri, R. "Environmental Challenges in Lebanon." *INTERNATIONAL STUDIES IN SOCIOLOGY AND SOCIAL ANTHROPOLOGY*, 1997, pp. 73-115.
Matthews, B. W. "Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme." *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, 1975, pp. 442-451.
McCauley, L. A. et al. "Isolated Wetland Loss and Degradation over Two Decades in an Increasingly Urbanized Landscape." *Wetlands*, vol. 33, no. 1, 2013, pp. 117-127.
McDonald, R. I. et al. "Global Urban Growth and the Geography of Water Availability, Quality, and Delivery." *Ambio: a Journal of the Human environment*, vol. 40, no. 5, 2011, pp. 437-446.
McKinney, M. L. "Urbanization, Biodiversity, and Conservation:The Impacts of Urbanization on Native Species Are Poorly Studied, but Educating a Highly Urbanized Human Population About These Impacts Can Greatly Improve Species Conservation in All Ecosystems." *Bioscience*, vol. 52, no. 10, 2002, pp. 883-890.
Memon, R.A. et al. "A Review on the Generation, Determination and Mitigation of Urban Heat Island." *Environmental Sciences*, vol. 20, 2008, pp. 120-128.
Menard, S. "Coefficients of Determination for Multiple Logistic Regression Analysis." *The American Statistician*, vol. 54, no. 1, 2000, pp. 17-24.

Meyer, W. B. and B. L. Turner. "Human Population Growth and Global Land-Use/Cover Change." *Annual review of ecology and systematics*, vol. 23, no. 1, 1992, pp. 39-61.
Michalak, W. Z. "Gis in Land Use Change Analysis: Integration of Remotely Sensed Data into Gis." *Applied Geography*, vol. 13, no. 1, 1993, pp. 28-44.
Mireles, F. et al. "Assessing Urban Soil Pollution in the Cities of Zacatecas and Guadalupe, Mexico by Instrumental Neutron Activation Analysis." *Microchemical Journal*, vol. 103, 2012, pp. 158-164.
Mirzaei, P.A. and F. Haghighat. "Approaches to Study Urban Heat Island - Abilities and Limitations." *Building and Environment*, vol. 45, 2010, pp. 2192-2201.
Mohammady, S. and M. R. Delavar. "Urban Sprawl Assessment and Modeling Using Landsat Images and Gis." *Modeling Earth Systems and Environment*, vol. 2, no. 3, 2016, pp. 155–169.
Mom, K. and S. Ongsomwang. "Urban Growth Modeling of Phnom Penh, Cambodia Using Satellite Imageries and a Logistic Regression Model." *Suranaree J. Sci. Technol.*, vol. 23, no. 4, 2016, pp. 481-500.
MRLC. "The National Land Cover Database (Nlcd) " https://www.mrlc.gov/data?f%5B0%5D=year%3A2001&f%5B1%5D=category%3Aland%20cover.
Muñoz-Marí, J. et al. "Semisupervised One-Class Support Vector Machines for Classification of Remote Sensing Data." *IEEE transactions on geoscience and remote sensing*, vol. 48, no. 8, 2010, pp. 3188-3197.
Murray-Rust, D. et al. "Agent-Based Modelling of Land Use Dynamics and Residential Quality of Life for Future Scenarios." *Environ Model Softw*, vol. 46, 2013, pp. 75-89.
Musa, S. I. et al. "A Review of Geospatial-Based Urban Growth Models and Modelling Initiatives." *Geocarto International*, vol. 32, no. 8, 2017, pp. 813-833.
Mustafa, A. et al. "Measuring the Effect of Stochastic Perturbation Component in Cellular Automata Urban Growth Model." *Procedia Environmental Sciences*, vol. 22, 2014, pp. 156-168.
Myint, S. W. and L. Wang. "Multicriteria Decision Approach for Land Use Land Cover Change Using Markov Chain Analysis and a Cellular Automata Approach." *Canadian Journal of Remote Sensing*, vol. 32, no. 6, 2006, pp. 390-404.
Nagendra, H. et al. "From Pattern to Process: Landscape Fragmentation and the Analysis of Land Use/Land Cover Change." *Agriculture, Ecosystems & Environment*, vol. 101, no. 2-3, 2004, pp. 111-115.
Newman, P. and J. Kenworthy. *Sustainability and Cities: Overcoming Automobile Dependence.* Island press, 1999.
News and Records. https://www.greensboro.com/uploaded_pdfs/w-nws-census-p/pdf_35e11ba5-4daf-5ed0-8be6-bfcd8bc729ee.html.
North Carolina State Data Center. "The Urban and Rural Faces of North Carolina." 2018. https://files.nc.gov/ncosbm/documents/files/2018ACSNC.pdf.
Nurul, W. M. R. W. "Compact Urban Form for Sociability in Urban Neighbourhoods." *International Journal of Social Science and Humanity*, vol. 5, no. 10, 2015, p. 822.
Nuruzzaman, Md. . "Urban Heat Island: Causes, Effects and Mitigation Measures." *International Journal of Environmental Monitoring and Analysis*, vol. 3(2), 2015, pp. 67-73.
O'Sullivan, D. "Exploring Spatial Process Dynamics Using Irregular Cellular Automaton Models." *Geographical analysis*, vol. 33, no. 1, 2001, pp. 1-18.
Onishia, A. et al. "Evaluating the Potential for Urban Heat-Island Mitigation by Greening Parking Lots." *Urban Forestry & Urban Greening*, vol. 9, 2010, pp. 323-332.

Pal, M. "Random Forest Classifier for Remote Sensing Classification." *International Journal of Remote Sensing*, vol. 26, no. 1, 2005, pp. 217-222.

Pal, M. and P. M. Mather. "An Assessment of the Effectiveness of Decision Tree Methods for Land Cover Classification." *Remote Sensing of Environment*, vol. 86, no. 4, 2003, pp. 554-565.

Peng, C. Y. J. et al. "Modeling Categorical Variables by Logistic Regression." *American Journal of Health Behavior*, vol. 25, no. 3, 2001, pp. 278-284.

Peng, H. et al. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min Redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, pp. 1226-1238.

Pijanowski, B.C. et al. "Using Neural Networks and Gis to Forecast Land Use Changes: A Land Transformation Model." *Computers, Environment and Urban Systems*, vol. 26, no. 6, 2002, pp. 553-575.

Pijanowski, B.C. et al. "A Land Transformation Model: Integrating Policy, Socioeconomics and Environmental Drivers Using a Geographic Information System." *Landscape ecology: a top down approach. Lewis Publishers, Boca Raton*, 2000.

Pijanowski, B.C. et al. "Calibrating a Neural Network-Based Urban Change Model for Two Metropolitan Areas of the Upper Midwest of the United States." *International Journal of Geographical Information Science*, vol. 19, no. 2, 2005, pp. 197–215.

Pourebrahim, N. et al. "Enhancing Trip Distribution Prediction with Twitter Data: Comparison of Neural Network and Gravity Models." *2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* 2018.

Pradhan, B. *Spatial Modeling and Assessment of Urban Form.* Springer, 2017.

Puertas, O. L. et al. "Assessing Spatial Dynamics of Urban Growth Using an Integrated Land Use Model: Application in Santiago Metropolitan Area , 2010–2045." *Land use policy*, vol. 38, 2014, pp. 415-425.

Qin, B. et al. "Dtu: A Decision Tree for Uncertain Data." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2009, pp. 4-15.

Quinlan, J.R. *C4. 5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

Resler, L. et al. "Predicting Functional Role and Occurrence of Whitebark Pine (Pinus Albicaulis) at Alpine Treelines: Model Accuracy and Variable Importance." *Annals of the American Association of Geographers*, vol. 104, no. 1-20, 2014.

Richards, Peter and Leah VanWey. "Where Deforestation Leads to Urbanization: How Resource Extraction Is Leading to Urban Growth in the Brazilian Amazon." *Annals of the Association of American Geographers*, vol. 105, no. 4, 2015, pp. 806-823, doi:10.1080/00045608.2015.1052337.

Rienow, A. and R. Goetzke. "Supporting Sleuth–Enhancing a Cellular Automaton with Support Vector Machines for Urban Growth Modeling." *Computers, Environment and Urban Systems*, vol. 49, 2014, pp. 66-81.

Rodriguez-Galiano, V. F. et al. "An Assessment of the Effectiveness of a Random Forest Classifier for Land-Cover Classification." *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, 2012, pp. 93-104.

Romero, H. et al. "Rapid Urban Growth, Land-Use Changes and Air Pollution in Santiago, Chile." *Atmospheric Environment*, vol. 33, no. 24-25, 1999, pp. 4039-4047.

Rosenblatt, F. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review*, vol. 65, 1958, pp. 386-408.

Samard zic´ -Petrovic´, M. et al. "Modeling Urban Land Use Changes Using Support Vector Machines." *Transactions in GIS*, vol. 20, no. 5, 2016, pp. 718-734.

Samard zic´ -Petrovic´, M. et al. "Modeling Urban Land Use Changes Using Support Vector Machines." *Transactions in GIS*, vol. 20, no. 5, 2016, pp. 718–734.
Samardži´c-Petrovi´c, M. et al. "Machine Learning Techniques for Modelling Short Term Land-Use Change." *ISPRS Int. J. Geo-Inf.*, vol. 6, 2017, p. 387.
Samardžić-Petrović, M. et al. "Exploring the Decision Tree Method for Modelling Urban Land Use Change." *Geomatica*, vol. 69, no. 3, 2015, pp. 313-325.
Samardžić-Petrović, M. et al. "Machine Learning Techniques for Modelling Short Term Land-Use Change." *ISPRS International Journal of Geo-Information*, vol. 6, no. 12, 2017, p. 387.
Santé, I. et al. "Cellular Automata Models Forthe Simulation of Real-World Urban Processes: A Review and Analysis." *Landsc.Urban Plan*, vol. 96, no. 2, 2010, pp. 108-122.
Shafizadeh-Moghadam, H. et al. "Coupling Machine Learning, Tree-Based and Statistical Models with Cellular Automata to Simulate Urban Growth." *Computers, Environment and Urban Systems*, vol. 64, 2017, pp. 297-308.
Shannon, C.E. "A Mathematical Theory of Communication." *Bell Syst. Tech. J.*, vol. 27, 1948, pp. 379–423.
Shirzadi, A. et al. "Public Transportation Mode Selection in an Urban Corridor: Application of Multi-Criteria Decision Making Methods." *Urban-Regional Studies and Research Journal*, vol. 5, no. 18, 2013, pp. 1-6.
Shirzadi Babakan, A. and A. Alimohammadi. "An Agent-Based Simulation of Residential Location Choice of Tenants in Tehran, Iran." *Transactions in GIS*, vol. 20, no. 1, 2016, pp. 101-125.
Shirzadi Babakan, A. et al. "An Agent-Based Evaluation of Impacts of Transport Developments on the Modal Shift in Tehran, Iran." *Journal of Development Effectiveness*, vol. 7, no. 2, 2015, pp. 230-251.
Shoemaker, D. A. et al. "Anticipating Trade-Offs between Urban Patterns and Ecosystem Service Production: Scenario Analyses of Sprawl Alternatives for a Rapidly Urbanizing Region." *Computers, Environment and Urban Systems*, vol. 74, 2019, pp. 114-125.
Shukla, V. and K. Parikh. "The Environmental Consequences of Urban Growth: Cross-National Perspectives on Economic Development, Air Pollution, and City Size." *Urban Geography*, vol. 13, no. 5, 1992, pp. 422-449.
Singh, S. and P. Gupta. "Comparative Study Id3, Cart and C4. 5 Decision Tree Algorithm: A Survey." *International Journal of Advanced Information Science and Technology (IJAIST)*, vol. 27, no. 27, 2014, pp. 97-103.
Sisodia, P. S. et al. "Prediction of Urban Sprawl Using Remote Sensing, Gis and Multilayer Perceptron for the City Jaipur." *Intelligent Systems Technologies and Applications*, 2016, pp. 403-410.
Skapura, D. "Building Neural Networks." 1996.
Sokolova, M. et al. " Beyond Accuracy, F-Score and Roc: A Family of Discriminant Measures for Performance Evaluation." *Australasian joint conference on artificial intelligence* Springer, Berlin, Heidelberg, 2006, pp. 1015-1021.
Song, Y. Y. and L. U. Ying. "Decision Tree Methods: Applications for Classification and Prediction." *Shanghai archives of psychiatry*, vol. 27, no. 2, 2015, p. 130.
Statnikov, A. *A Gentle Introduction to Support Vector Machines in Biomedicine: Theory and Methods* vol. 1, world scientific, 2011.
Stone Jr, B. *The City and the Coming Climate: Climate Change in the Places We Live.* Cambridge University Press, 2012.

---. "Urban Heat and Air Pollution: An Emerging Role for Planners in the Climate Change Debate." *Journal of the American planning association*, vol. 71, no. 1, 2005, pp. 13-25.
Sultana, S. "Edge Cities in the Era of Megaprojects." *Engineering Earth: The Impacts of Mega-Engineering Projects*, edited by Stan Brunn, Springer, 2011, pp. 1071-1088.
---. "Land Use and Transportation." *The International Encyclopedia of Geography: People, the Earth, Environment and Technology (IEG)*, John Wiley & Sons, Ltd, 2017, pp. 1-11. doi:DOI: 10.1002/9781118786352.wbieg0697.
Sultana, S. et al. "Household Energy Expenditures in North Carolina: A Geographically Weighted Regression Approach." *Sustainability*, vol. 10, no. 5, 2018, p. 1511.
Sultana, S. and J. Weber. "Journey-to-Work Patterns in the Age of Sprawl: Evidence from Two Midsize Southern Metropolitan Areas." *The Professional Geographer*, vol. 59, no. 2, 2007, pp. 193-208.
---. "The Nature of Urban Growth and the Commuting Transition: Endless Sprawl or a Growth Wave?" *Urban studies*, vol. 51, no. 3, 2014, pp. 544-576.
Sustain Charlotte. https://www.sustaincharlotte.org/.
Suthaharan, S. *Machine Learning Models and Algorithms for Big Data Classifi Cation.* vol. 36, Springer, 2016.
Swain, J. "Charlotte's Transit Plans Must Match Our Growth." https://www.charlotteobserver.com/opinion/article248377225.html.
Swenson, J. J. and J. Franklin. "The Effects of Future Urban Development on Habitat Fragmentation in the Santa Monica Mountains." *Landscape Ecology*, vol. 15, no. 8, 2000, pp. 713-730.
Tahami, H. et al. "Virtual Spatial Diversity Antenna for Gnss Based Mobile Positioning in the Harsh Environments." *The 31st International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2018)*, 2018, pp. 3186-3198.
Tahami, H. et al. "The Preliminary Study on the Prediction of a Hurricane Path by Gnss Derived Pwv Analysis." *Proceedings of the ION 2017 Pacific PNT Meeting*, 2017.
Tan, P. N. et al. "Classification: Basic Concepts, Decision Trees, and Model Evaluation." *Introduction to data mining*, vol. 1, 2006, pp. 145-205.
Taravat, A. et al. "Urbanization Dynamics of Tehran City (1975–2015) Using Artificial Neural Networks." *Journal of Maps*, vol. 13, 2017, pp. 24-30.
Tayyebi, A. et al. "A Spatial Logistic Regression Model for Simulating Land Use Patterns: A Case Study of the Shiraz Metropolitan Area of Iran." *In: Chuvieco E., Li J., Yang X. (eds) Advances in Earth Observation of Global Change. Springer, Dordrecht*, 2010.
Tayyebi, A. et al. "Predicting the Expansion of an Urban Boundary Using Spatial Logistic Regression and Hybrid Raster-Vector Routines with Remote Sensing and Gis." *International Journal of Geographical Information Science*, vol. 28, no. 4, 2014, pp. 639–659.
Tayyebi, A. and B. C. Pijanowski. "Modeling Multiple Land Use Changes Using Ann, Cart and Mars: Comparing Tradeoffs in Goodness of Fit and Explanatory Power of Data Mining Tools." *International Journal of Applied Earth Observation and Geoinformation*, vol. 28, 2014, pp. 102-116.
Tayyebi, A. et al. "An Urban Growth Boundary Model Using Neural Networks, Gis and Radial Parameterization: An Application to Tehran, Iran." *Landsc Urban Plan*, vol. 100, 2011, pp. 35-44.
Theobald, D. M. "Land-Use Dynamics Beyond the American Urban Fringe." *Geographical Review*, vol. 91, 2001, pp. 544–564.

Tian, G. et al. "Simulation of Urban Expansion and Encroachment Using Cellularautomata and Multi-Agent System Model—a Case Study of Tianjinmetropolitan Region, China." *Ecological Indicators*, vol. 70, 2016, pp. 439-450.
Timofeev, R. *Classification and Regression Trees (Cart) Theory and Applications.* Humboldt University, 2004.
Tomlinson, C.J. et al. "Including the Urban Heat Island in Spatial Heat Health Risk Assessment Strategies: A Case Study for Birmingham, Uk." *International Journal of Health Geographics*, vol. 10:42, 2011.
Tong, S. T. and W. Chen. "Modeling the Relationship between Land Use and Surface Water Quality." *Journal of environmental management*, vol. 66, no. 4, 2002, pp. 377-393.
Torrens, P. M. and D. O'Sullivan. "Cellular Automata and Urban Simulation: Where Do We Go from Here?" *Environment and Planning B: Planning and Design*, vol. 28, 2001, pp. 163-168.
Triantakonstantis, D. and G. Mountrakis. "Urban Growth Prediction: A Review of Computational Models and Human Perceptions." *J Geogr Inform Syst.*, vol. 4, 2012, pp. 555-587.
Turner, B. L. et al. "The Emergence of Land Change Science for Global Environmental Change and Sustainability." *Proceedings of the National Academy of Sciences*, 2007, pp. 20666-20671.
U.S. Census Bureau. https://factfinder.census.gov.
---. https://factfinder.census.gov.
---. "The Census Bureau's Population Estimates Program " https://www.census.gov/glossary/#term_Populationestimates. 2020.
---. "Tiger/Line Shapefiles and Tiger/Line Files." https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html. 2019.
UNCCharlotte. "Introducing 'Future Charlotte,' a Podcast About Our City's Growth." https://ui.uncc.edu/story/introducing-future-charlotte-podcast-about-our-citys-growth.
United Nations. "World Urbanization Prospects: The 2018 Revision (St/Esa/Ser.A/420)." Department of Economic and Social Affairs, Population Division, 2019.
USGS. "Elevation Products (3dep)." https://viewer.nationalmap.gov/basic/#/.
---. "The National Map." https://nationalmap.gov/landcover.html.
---. "The National Map." https://nationalmap.gov/landcover.html. 2018.
Vapnik, V. "The Support Vector Method of Function Estimation." *Nonlinear Modeling*, Springer, 1998, pp. 55-85.
Vapnik, V. and A. Lerner. "Pattern Recognition Using Generalized Portrait Method." *Automation and Remote Control*, vol. 24, 1963, pp. 774-780.
Verburg, P. H. et al. "A Method to Analyse Neighbourhood Characteristics of Land Use Patterns." *Computers, Environment and Urban Systems*, vol. 28, no. 6, 2004, pp. 667-690.
Verburg, P.H. et al. "Modeling the Spatial Dynamics of Regional Land Use: The Clue S Model." *Environmental Management*, vol. 30, no. 2, 2002, pp. 391-405.
Waddell, P. "Urbansim: Modeling Urban Development for Land Use, Transportation, and Environmental Planning." *Journal of the American planning association*, vol. 68, no. 3, 2002, pp. 297-314.
Weber, C. and A. Puissant. "Urbanization Pressure and Modeling of Urban Growth: Example of the Tunis Metropolitan Area." *Remote Sensing of Environment*, vol. 86, no. 3, 2003, pp. 341-352.
Weber, J. and S. Sultana. "Employment Sprawl, Race and the Journey to Work in Birmingham, Alabama." *southeastern geographer*, vol. 48, no. 1, 2008, pp. 53-74.

Wehrwein, G. S. "The Rural–Urban Fringe." *Economic Geography*, vol. 18, 1942, pp. 217–228.
Westreich, D. et al. "Propensity Score Estimation: Neural Networks, Support Vector Machines, Decision Trees (Cart), and Meta-Classifiers as Alternatives to Logistic Regression." *Journal of clinical epidemiology*, vol. 63, no. 8, 2010, pp. 826-833.
White, R. and G. Engelen. "Cellular Automata and Fractal Urban Form: A Cellular Modelling Approach to the Evolution of Urban Land-Use Patterns." *Environment and Planning A*, vol. 25, no. 8, 1993, pp. 1175-1199.
Wickham, J. D. et al. "Ageography of Ecosystem Vulnerability." *Landscape Ecology*, vol. 15, 2000, pp. 495-504.
Wickham, J. D. et al. "Thematic Accuracy of the Nlcd 2001 Land Cover for the Conterminous United States." *Remote Sensing of Environment*, vol. 114, no. 6, 2010, pp. 1286-1296.
Wickham, J. D. et al. "Accuracy Assessment of Nlcd 2006 Land Cover and Impervious Surface." *Remote Sensing of Environment*, vol. 130, no. 294-304, 2013.
Wickham, J. et al. "Thematic Accuracy Assessment of the 2011 National Land Cover Database (Nlcd)." *Remote Sensing of Environment*, vol. 191, 2017, pp. 328-341.
Wiesmeier, M. et al. "Digital Mapping of Soil Organic Matter Stocks Using Random Forest Modeling in a Semi-Arid Steppe Ecosystem." *Plant and soil*, vol. 340, no. 1-2, 2011, pp. 7-24.
Williams, P. W. "Impact of Urbanization on the Hydrology of Wairau Creek, North Shore, Auckland." *Journal of Hydrology (New Zealand)*, 1976, pp. 81-99.
Wu, F. and D. Martin. "Urban Expansion Simulation of Southeast England Using Population Surface Modelling and Cellular Automata." *Environment and Planning A*, vol. 34, 2002, pp. 1855-1876.
Xu, L. et al. "Modelling Urban Expansion Guided by Land Ecological Suitability: A Case Study of Changzhou City, China." *Habitat International*, vol. 75, 2018, pp. 12-24.
Yang, Q. et al. "Cellular Automata for Simulating Land Use Changes Based on Support Vector Machines." *Computers & geosciences*, vol. 34, no. 6, 2008, pp. 592-602.
Yang, Y. "A Tale of Two Cities: Physical Form and Neighborhood Satisfaction in Metropolitan Portland and Charlotte." *Journal of the American planning association*, vol. 74, no. 3, 2008, pp. 307-323.
Yang, Y. and J. O. Pedersen. "A Comparative Study on Feature Selection in Text Categorization." *the 14th International Conference on Machine Learning*, 1997, pp. 412-420.
Yankson, P. W. K. and K. V. Gough. "The Environmental Impact of Rapid Urbanization in the Peri-Urban Area of Accra, Ghana." *Geografisk Tidsskrift-Danish Journal of Geography*, vol. 99, no. 1, 1999, pp. 89-100.
Yao, F. et al. "Simulating Urban Growth Processes by Integrating Cellular Automata Model and Artificial Optimization in Binhai New Area Oftianjin, China." *Geocarto Int*, 2015, pp. 1-16.
Zhang, J. P. et al. "A Parallel Svm Training Algorithm on Large-Scale Classification Problems." *International Conference on Machine Learning and Cybernetics*, vol. 3, IEEE, 2005, pp. 1637-1641.
Zhao, C. et al. "Characterizing the 3-D Urban Morphology Transformation to Understand Urban-Form Dynamics: A Case Study of Austin, Texas, USA." *Landscape and urban planning*, vol. 203, 2020, p. 103881, doi:https://doi.org/10.1016/j.landurbplan.2020.103881. .
Zhou, X. and Y. C. Wang. "Spatial-Temporal Dynamics of Urban Green Space in Response to Rapid Urbanization and Greening Policies." *Landscape and urban planning*, vol. 100, no. 3, 2011, pp. 268-277.