Computers, Environment and Urban Systems 94 (2022) 101801

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Machine learning application to spatio-temporal modeling of urban growth

Yuna Kim[a], Abolfazl Safikhani[b], Emre Tepe[c,*]

[a] *Department of Statistics, University of Florida, 106D Griffin-Floyd Hall, P.O. Box 118545, Gainesville, FL 32611, USA*
[b] *Department of Statistics, University of Florida, 203 Griffin-Floyd Hall, P.O. Box 118545, Gainesville, FL 32611, USA*
[c] *Department of Urban and Regional Planning, University of Florida, 444 Architectural Building, P.O. Box 115706, Gainesville, FL 32611, USA*

ARTICLE INFO

Keywords:
Urban growth
Machine learning method
Spatio-temporal modeling
Random forest
Prediction

ABSTRACT

Understanding the dynamics of urban growth is among the most important tasks in urban planning due to their influence on policy decision-making. Specifically, prediction of urban growth at regional levels is crucial for regional policy makers. Making such predictions is difficult because of the existence of complex topological structures and the high-dimensional nature of data sets related to urban growth. Spatial and temporal autocorrelation and cross-correlations, together with regional social and physical covariates, need to be properly accounted for improving the forecasting power of any statistical or machine learning method. To that end, we develop novel machine learning methodologies to perform predictions of urban growth at regional levels by incorporating lead-lag *non-linear* relationships among past urban changes in each region and its neighbors. Based on this analysis, machine learning algorithms outperform more classical methods, such as a logistic regression, in terms of classifying low/high urban growth regions, and the random forest algorithm seems to have the best prediction accuracy among the selected machine learning methods. Moreover, the random forest method *without* any external covariates has still a high prediction accuracy which not only confirms that most of variability of urban growth can be described by past observations of self and neighboring changes, but also makes it possible to perform real forecasting of urban growth without accessing any external covariates. The latter makes this modeling framework useful for local policy makers in allocating budget and directing resources appropriately based on such predictions.

1. Introduction

Accurate prediction of future urban growth and land development is one of the fundamental goals of urban modeling. Urban growth dynamics depend on the multidimensional aspects of the physical, social and economic environments. The land development potential on a given site is determined based on human behavior and physical and institutional limitations. Increased complexity in the dynamics requires equivalent mathematical representations in quantitative models, providing additional challenges in formulating suitable modeling methods, in data requirements, and computational power.

The main goal in modeling land-use change is to mimic the human activities which characterize urban development. Investment decisions that result in changes in existing land conditions depend on the expected utility from land conversion (Irwin & Geoghegan, 2001). However, estimating utility expectation for a potential land-use change on a given parcel is not a simple task, due to the unavailability of the necessary information. In many studies, proxy information is used to approximate the utility function in land-use change models. As mentioned in Verburg, Ritsema van Eck, de Nijs, and Dijst (2004), land-use changes depend on the complex interactions between human activities and the physical environment. In land-use change models, researchers incorporate multiple explanatory variables to approximate people's utility maximization behaviours. Tepe and Guldmann (2017) highlight the importance of working with disaggregated data to achieve robust model results. Finally, there is no consensus about which information should be used as proxy. Therefore, there are significant data challenges in land-use change modeling.

Statistical models of urban systems can successfully represent actual system dynamics if relationships in real life are precisely formulated. However, there are many limitations in building such complex models. The first challenge is introducing a successful methodology to account for spatial and temporal correlations as well as cross-correlations of urban systems in different regions. Recent studies in the field highlight the importance of dynamic historical and contemporaneous neighborhood relations (Bhat, Dubey, Alam, & Khushefati, 2015; Huang, Zhang,

* Corresponding author at: Department of Urban and Regional Planning, University of Florida, 1480 Inner Road, Gainesville, FL 32611, USA.
E-mail addresses: ynkim732@ufl.edu (Y. Kim), a.safikhani@ufl.edu (A. Safikhani), emretepe@ufl.edu (E. Tepe).

https://doi.org/10.1016/j.compenvurbsys.2022.101801
Received 4 September 2021; Received in revised form 20 March 2022; Accepted 21 March 2022

& Wu, 2009; Irwin, Bell, & Geoghegan, 2003; Tepe & Guldmann, 2017; Tepe & Guldmann, 2020). The second challenge is data availability. Theoretically, land-use change models are based on the utility-maximization behaviours of consumers, and land investment decisions are affected by institutional and physical limitations. These factors must be represented in statistical and machine learning models in order to achieve robust results. While some of these factors can be represented using direct information, the remainder can only be incorporated using proxy data. In parallel to advances in geo-coded data collection, modeling methods at spatially disaggregated levels provide heterogeneous information that can help eliminate some data requirements limitations. The third challenge is computational bottlenecks in running complex quantitative models with increasing data sizes.

Machine learning methods (such as random forest, neural networks, gradient boosting, etc.) are among the powerful prediction tools available to scientists and forecasters. They are computationally feasible methods which can be applied to large-scale data sets and can capture complex relationships between the dependent variables and the predictors due to the highly non-linear nature of their estimation. Predicting future land-use changes is a complicated task due to the multiple parameters affecting urban dynamics and the existence of spatial and temporal correlations among urban changes in different regions. The main idea of this paper is to tackle the problem of forecasting future urban growth using machine learning methods. The growth rates of residential, commercial, and occupied parcels at the block group level are investigated using the Auditor's geo-coded tax database for the State of Florida. The available information on when construction took place on each parcel is used to derive measures of land-use dynamics. Different spatio-temporal models, incorporating space and time and their interactions, are used to investigate land development dynamics. The developed machine learning methods are successfully able to capture the non-linear dynamics of urban growth in this rich data set (see more details in Section 1 of the supplementary document) and achieve satisfactory prediction accuracy (Section 5.1). Another interesting outcome of the analysis is the existence of strong non-linear spatial and temporal signals in urban growth rate data, which can be utilized to perform real out-of-sample forecasts *without* the use of any external covariates (see more details in Section 5.3). Note that there is no access to external variables in the future; thus, any method which uses external variables cannot make real out-of-sample predictions. This is an important observation which makes it possible for regional policy makers to have access to highly accurate prediction of urban changes and modify local budgets accordingly to reach certain policy goals. A brief review of existing works in the literature is provided next.

2. Related works

In land-use change modeling, discrete-response models are commonly preferred due to the categorical nature of land uses as dependent variables. Chomitz and Gray (1996) implement a multinomial logit to model relationships between new roads and deforestation as a result of land conversion from agriculture to commercial uses in Belize. Semisubsistence farming, commercial farming, and natural soil nitrogen, slope, distance to Belize, etc. vegetation are used as the response variables while a set of soil and locational characteristics are incorporated as explanatory variables. The potential bias from road endogeneity is mostly eliminated by incorporating soil quality into the model. Verburg et al. (2004) implement a stepwise logistic regression model to investigate temporal dynamics in land development between 1989 and 1996 in the Netherlands. The determinants of changes in a set of land use categories are investigated using detailed location of features and accessibility measures. They indicate that accessibility, neighborhood interactions and spatial policies play an important role in recent years, as compared to historical land developments.

In addition to temporal dynamics, spatial dependencies are important components in land-use change modeling, and accounting for such dynamics introduces additional complexity in the models. Nahuelhual, Carmona, Lara, Echeverria, and Gonzalez (2012) use an autologistic regression to analyze timber plantation expansion in south-central Chile over two separate periods (1975–1990, 1990–2007), accounting for land characteristics and accessibility measures. They conclude that the spatiotemporal dynamics formed as a result of the interactions of natural and socio-economic drivers are an important factor in timber plantation expansion. Yu and Srinivasan (2016) also implement a binary autologistic regression to investigate rural-to-urban land-use change over 2000–2010 in Beijing. Their explanatory variables are grouped into proximity, neighborhood, physical, jurisdictional, and socio-economic categories. Their findings indicate a positive association between existing land and vacant land in close proximity. Both Nahuelhual et al. (2012) and Yu and Srinivasan (2016)) account for spatial dependencies in their modeling approaches, using the autocorrelation of neighboring spatial units through a spatial weight matrix. Moreover, Carrion-Flores and Irwin (2004) utilize a probit model to investigate the land development potential at the parcel level in the rural areas of the Cleveland metropolitan area, where spatial dependencies are assumed to exist within the error term. Bhat et al. (2015) model land development patterns for Austin's CBD and surrounding areas using a spatial discrete-continuous probit model accounting for the spatial lag of the dependent variable.

Incorporating temporal dynamics is critical in land use change modeling. Irwin et al. (2003) implement a duration model of land-use changes at the parcel level by controlling for spatial dependency without an explicit temporal lag. Incorporating spatial dependence and temporal dynamics separately is not sufficient to achieve robust model results. To that end, Huang et al. (2009) model the spatial and temporal dynamics of conversions from rural to urban land uses in New Castle County, Delaware, over three separate periods (1984–1992, 1992–1997, 1997–2002), while Ferdous and Bhat (2013) introduce a spatial panel ordered-response probit model controlling for both spatial interactions and temporal lags. Gao et al. (2020) compare methods used to control for spatial heterogeneity in land development and conclude that the spatial lag and localized modeling approaches (such as GWR) provide better modeling results. Finally, Tepe and Guldmann (2017, 2020) introduce a novel approach for spatio-temporal modeling of land-use changes. Both binary and multinomial spatio-temporal autologistic regression models are developed for estimating land-use conversions at the parcel level in Delaware county, Ohio. Their findings show that land developments in neighboring parcels attract the same land use and historical land development trends are also positively associated with contemporaneous parcel development.

Spatial components in land-use change modeling introduce computational challenges. When a spatial weight matrix is incorporated using spatial lag or spatial error approaches, the computation of the inverse matrix is required (Anselin, 1988; Ord, 1975). Alternatively, simulation methods, such as Gibbs sampler and EM algorithms, can be considered to solve complex log-likelihood functions of discrete-response models with spatial lags (Fleming, 2004). However, simulation-based optimization procedures do not guarantee convergence of the maximum likelihood function during parameter estimation. Most proposed land-use change models with such explicit spatial components have less than 3000 sample observations because of these computational challenges (Bhat et al., 2015; Ferdous & Bhat, 2013; Huang et al., 2009; Nahuelhual et al., 2012; Yu & Srinivasan, 2016). Tepe and Guldmann (2017, 2020) substantially improve the computational feasibility of discrete response models by using simulation-based approaches. However, other methods are required to achieve robust results when model complexity increase. The computational advantages of the Random Forest (RF) and Artificial Neural Network (ANN) methods can be considered for such modeling.

Recent years have witnessed efforts to integrate Machine Learning (ML) methods into land-use change models, based on the Cellular Automata (CA) approach. Gounaridis, Chorianopoulos, Symeonakis, and Koukoulas (2019) apply a Random Forest approach to classify detailed land-use categories, accounting for environmental, physical, accessibility, and socio-economic indicators in Attica, Greece. They introduce a hybrid modeling approach for land-use change, using both CA and RF models. Karimi, Sultana, Babakan, and Suthaharan (2019) implement a Support Vector Machine (SVM) model for urban expansion in Guilford County, analyzing land-use changes over 2001–2006. Their model classifies a given piece of land as vacant or built-up, based on a set of predictors grouped under site-specific, proximity and neighborhood categories. Their model provides highly accurate results. Also, their findings show the importance of the spatial clustering pattern of land use / land category types. Xing, Qian, Guan, Yang, and Huayi (2020) integrate spatial and temporal dynamics in a Deep Learning (DL) model used as the transition function in CA models. Landsat images and road networks are used to derive spatiotemporal dynamics in land cover proportions, site-specific measures (implicit spatial features, elevation, slope), and distances to a set of point of interests (river, railway, highway, first class road, minor road, city road, railway station, bus station, main POIs, and central city). This model successfully captures the neighborhood dynamics that are vital to land-use change models, with an overall model accuracy of almost 94%. Liang, Dang, Sun, and Wang (2020) propose a CA approach combining Markov Chain and RF methods to model land use changes in Shanghai, using site-specific, proximity, socio-economic characteristics, and planning guidelines. This is the only study accounting for institutional factors in land development dynamics. Similarly, Okwuashi and Ndehedehe (2021) integrate SVM and Markov chain approaches into cellular automata for modeling urban changes. Lv et al. (2021) introduce a gravity-based approach to account for spatial interactions between cities in their RF-CA model. Their model classifies a binary choice of urban and non-urban land use types, based on a set of predictor variables covering economic, social, educational characteristics and infrastructure and environment conditions. Use of a gravitational model in RF-CA advances traditional CA modeling by accounting for travel cost in the system. Shafizadeh-Moghadam, Minaei, Jr, Asghari, and Dadashpoor (2021) apply a Forward Feature Selection algorithm for RF models used as transition rules in CA-based land-use change modeling. In RF models, urban growth and non-urban persistence are classified based on grids' characteristics: slope, altitude and distances from roads, crop, greenery, urban, and barren. This study shows the effectiveness of accounting for proximity factors in the absence of socio-economic factors. Finally, Yu, Hagen-Zanker, Santitissadeekorn, and Hughes (2021) discuss the lack of sufficient historical information to calibrate Cellular Automata land-use change models. They introduce a Markov Chain Monte Carlo approximation based on Bayesian computation to calibrate CA models.

There are also a few ML applications to land-use change modeling. Zhai et al. (2020) implement a Convolution Neural Network (CNN) approach to Vector-based CA modeling. CNN effectively classifies a given parcel's land-use category based on parcel site-specific and proximity characteristics. This novel approach effectively mimics local neighborhood dynamics, using the convolution kernel and local connectors. Ron-Ferguson, Chin, and Kwon (2021) investigate land development dynamics by analyzing the actions taken on vacant lands and existing constructions and accounting for a wide range of explanatory, including socio-economic, built environment characteristics, and landscape metrics. They show the importance of the RF method to account for complex non-linear relationships in the data. Talukdar et al. (2021) introduce a spatiotemporal analysis of land-cover changes using Machine Learning algorithms such as Bagging and RF, where water bodies, agricultural land, vegetation, sand bar, bare land and built-up area categories are used as response variable and a set of landscape metrics is used as the explanatory variables. The bagging model produces more accurate predictions, due to higher levels of tree depths as compared to the RF model. The model successfully captures land cover fragmentation in the study area. These methods provide highly accurate predictions due to their incorporating non-linear relations (Bahadori, Yu, & Liu, 2014; Delasalles, Ziat, Denoyer, & Gallinari, 2019). Basse, Omrani, Charif, Gerber, and Bódis (2014) highlight the use of a Cellular Automata (CA) based approach using Artificial Neural Networks (ANNs) in order to increase model accuracy. Soares-Filho, Rodrigues, and Follador (2013) introduce a heuristic modeling approach based on the Genetic Algorithm (GA) to improve the accuracy of land-use change models.

The remainder of the paper is organized as follows. A brief introduction to the Auditor's geo-coded tax database for the State of Florida is provided in Section 3 while four machine learning methods applied to this data set are summarized in Section 4. The main results of the paper, including the prediction performance of the developed machine learning algorithms, are presented in Section 5. Comparisons of the results with similar results in the other literature are provided in Section 6. Finally, some concluding remarks and future research directions are stated in Section 7.

3. Data set

The state of Florida is selected as the study area. The University of Florida GeoPlan Center provides statewide Auditor's Parcel Databases. The publicly available 2019 database comprises nearly 9 million parcels, with data on parcel geometry, year-built, land use, and the two most recent sales. Using the information on when constructions took place, historical land development conditions at the parcel level are generated for all years between 1900 and 2019. These parcel histories provide an opportunity to compute the average distances from any parcel to a set of points of interest (POI)(recreation, stores, supermarket, etc.; see Table 3 for the full list of POIs) within a range of 2 miles. Fig. 1 illustrates the procedure to compute the average distance to a certain POI from a given parcel in a single year. Once parcel-level computations are completed, parcel-level data are aggregated at the block group level. The aggregated data characterize all Florida 11,394 block groups, with no missing data issues.

We focused on the most recent 5 years, when Florida has experienced rapid land development. Table 1 presents descriptive statistics for the numbers of single-family, commercial, and occupied parcels in a block group in 2015 and 2019. In 2019, almost 59% of all parcels (8,995,663) in Florida were single-family residential parcels, a 4.5% increase over 5 years. Commercial parcels account for approximately 3% of all parcels, with a 2.1% raise over 5 years. In 2019, occupied parcels constitute almost 73% of the total with a 4% increase over 5 years. The number of other land-use parcels increased by 1.9%. Fig. 2 presents parcel maps of occupied parcels in Florida in 2015 and 2019. Bigger circles indicate larger occupancy.

4. Methodology

Statistical and machine learning methods used to model land-use changes in Florida are briefly introduced, including Logistic Regression (LR), Random Forest (RF), Artificial Neural Network (ANN), and Extreme Gradient Boosting (EGB). The data set is divided into a training set (72% of the data set), validation set (8% of the data set) and a test set (the remaining 20% of the data set). Vabalas, Gowen, Poliakoff, and Casson (2019) and Hansen et al. (2013) used 90% of the data set as a training set and 10% of the data set as a validation set, but they used the validation set as a test set. We differentiate validation and test sets for parameter tuning. We conduct sensitivity analysis for multiple splits, but the difference between the result from a single split and multiple splits is less than 2%. Before discussing these methods, the dependent variables and predictors are presented.

4.1. Dependent variables

The data set includes the numbers of parcels for each land use at the block group level (single-family residential, vacant, commercial, other residential, open spaces, and services). The numbers of single-family

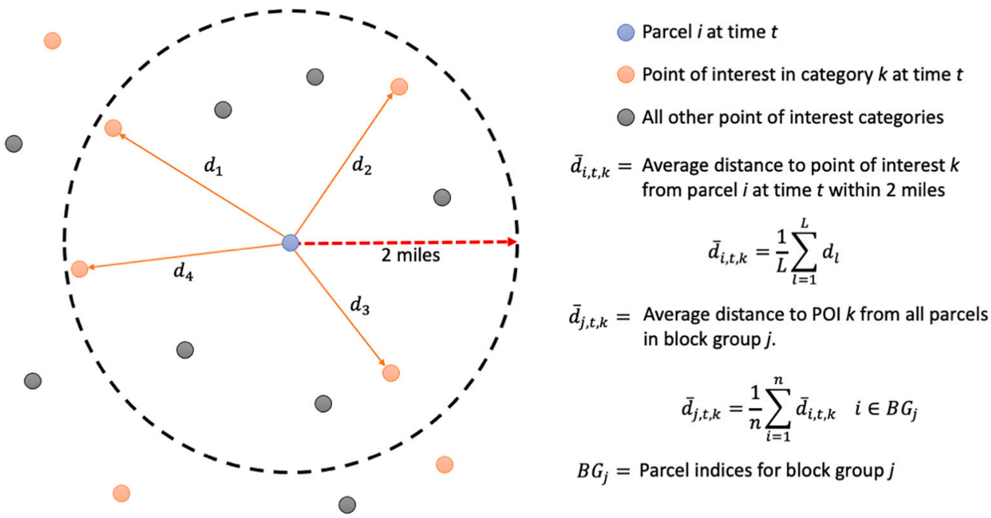

d_1
d_2
d_3
d_4
2 miles
Parcel i at time t
Point of interest in category k at time t
All other point of interest categories
$\bar{d}_{i,t,k}$ = Average distance to point of interest k from parcel i at time t within 2 miles
$\bar{d}_{i,t,k} = \frac{1}{L}\sum_{l=1}^{L} d_l$
$\bar{d}_{j,t,k}$ = Average distance to POI k from all parcels in block group j.
$\bar{d}_{j,t,k} = \frac{1}{n}\sum_{i=1}^{n} \bar{d}_{i,t,k} \quad i \in BG_j$
BG_j = Parcel indices for block group j

Fig. 1. Illustration of the procedure to compute the mean of average distances between a given parcel i to POI k within a 2-mile buffer at time t.

Table 1
Descriptive statistics for the numbers of single-family residential, commercial, and occupied parcels in block groups in 2015 and 2019.

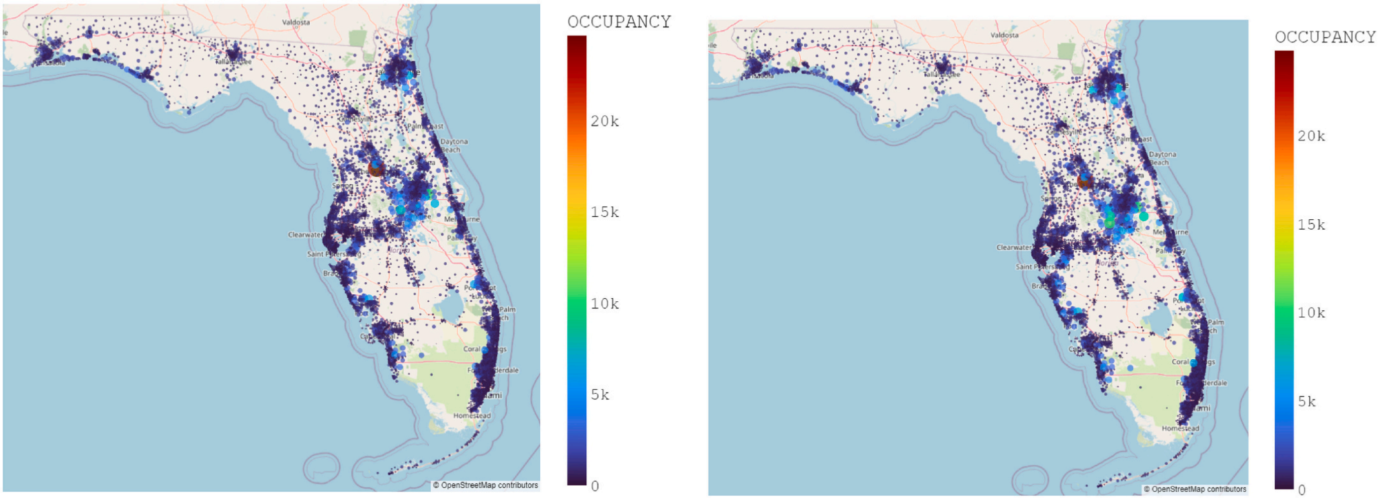
Year	Land use category	Min	Mean	Median	Max	Std. Dev.
2015	single-family residential	0	448.1	345	24,269	525.1
2015	commercial	0	21.5	456	613	35.1
2015	occupied	0	557.4	58	24,638	549.9
2019	single-family residential	0	468.4	352	24,389	577.7
2019	commercial	0	22.0	11	616	35.7
2019	occupied	0	579.8	464	24,772	605.1

residential, commercial, and vacant parcels are the dominant, and other land-use categories are excluded from further modeling due to significant imbalances among the different land uses. The growth rates of single-family residential ($Y_1(t,k)$), occupied ($Y_2(t,k)$), and commercial ($Y_3(t,k)$) parcels are used as the dependent variables, where t is the index of year and k is the number of years in the past from which growth rates are computed. All the dependent variables are normalized using log-transformation and standardization. A small fraction of 1 (0.001) is added to the denominator in order to avoid infinite values in the calculation of growth rates. Table 2 further describes the dependent variables, $Y_j(t,k)$, $(j = 1,2,3)$. Table 2 in Section 4 of the supplementary file provides descriptive statistics of these variables.

Binary versions of the above continuous dependent variables have been created and will serve as the main dependent variables in the analysis. They are defined as follows:

Table 2
Description of the dependent variables.

Name	Description
$Y_1(t, k)$	Transformed growth rate of the number of single-family residential parcels from year $(t - k)$ to t
$Y_2(t, k)$	Transformed growth rate of the number of occupied parcels from year $(t - k)$ to t
$Y_3(t, k)$	Transformed growth rate of the number of commercial parcels from year $(t - k)$ to t



(a) Year 2015

(b) Year 2019

Fig. 2. Occupancy map of Florida in 2015 and 2019.

$$Z_{1i}(t,k) = \begin{cases} 0\ (\text{Less developed}), & \text{if } Y_{1i}(t,k) < m_{Y_1(k)}, \\ 1\ (\text{Mor edeveloped}), & \text{otherwise}, \end{cases}$$
$$Z_{2i}(t,k) = \begin{cases} 0\ (\text{Less developed}), & \text{if } Y_{2i}(t,k) < m_{Y_2(k)}, \\ 1\ (\text{More developed}), & \text{otherwise}, \end{cases} \tag{1}$$
$$Z_{3i}(t,k) = \begin{cases} 0\ (\text{Less developed}), & \text{if } Y_{3i}(t,k) \leq m_{Y_3(k)}, \\ 1\ (\text{More developed}), & \text{otherwise}, \end{cases}$$

where i is the index of a block $\{1,\ldots,11394\}$, t is the index of a year $\{2015, 2016, 2017, 2018, 2019\}$, k is the time lag, set to 5 and 10, $m_{Y1(k)}$, $m_{Y2(k)}$, $m_{Y3(k)}$ are the median values of $Y_1(\cdot,k)$, $Y_2(\cdot,k)$, and $Y_3(\cdot,k)$, respectively, where $Y_j(\cdot,k) = \underset{t\in T}{\cup} Y_j(t,k)$, $T = \{2015,\ 2016,\ 2017,\ 2018,\ 2019\}$, for $j = 1,2,3$. The median growth rate is used as a threshold to create balanced data. Table 3 in Section 4 of the supplementary file provides the frequencies for these binary variables.

4.2. Predictors

In land-use change models, proximity to POIs is used as proxy information. Accessibility to services provided within a close proximity affects investors' decisions. In the absence of direct factors affecting the utility function, proximity factors are important approximations (Shafizadeh-Moghadam et al., 2021). Twenty eight accessibility measures are calculated as means of average distances from parcels to certain POIs within 2 miles for each block group. The Table 3 lists and describes these variables. The descriptive statistics of these 28 accessibility measures are provided in Table 4 of Section 4 in the supplementary file.

In addition to these twenty eight variables, we use four temporal variables and four spatio-temporal variables. The temporal variables, $Y_{ji}(t-k,k)$, $Y_{ji}(t-2k,k)$, $Y_{ji}(t-3k,k)$, $Y_{ji}(t-4k,k)$, are the k-lag, $2k$-lag, $3k$-lag, and $4k$-lag growth rates. $N_{ji}(t,k,k)$, $N_{ji}(t,2k,k)$, $N_{ji}(t,3k,k)$, and $N_{ji}(t,4k,k)$ are the spatio-temporal variables for $j = 1,2,3$, with:

$$N_{ji}(t,l,k) = \frac{1}{10}\sum_{p=1}^{10} log\left(Y_{jip}(t-l,k)+1\right), \tag{2}$$

where $Y_{jip}(t-l,k)$ is the $Y_{ji}(t-l,k)$ value of the p^{th} closest neighborhood block of block i, for $l = k$, $2k$, $3k$, $4k$. The reason for selecting 10 neighbors is as follows. We use the K-nearest neighbor spatial conceptualization approach. Global Moran's I tests (Moran, 1950) are conducted for the dependent variables at various K-degree between 1 and 50. The Global Moran's I test results indicate that there are statistically significant spatial dependencies in the dependent variables, while the index values reach their highest levels between 5 and 10 nearest neighbors.

All predictors are normalized to enhance the prediction performance of statistical and machine learning models.

Table 3
Names and descriptions of the accessibility measures.

Name	Description
Distance to Recreation	recreational parcels
Distance to Stores	department stores parcels
Distance to Supermarket	supermarkets parcels
Distance to Regional Shopping Center	regional shopping centers
Distance to Community Mall	community shopping centers
Distance to One-story Office	one-story office buildings, non-professional service buildings
Distance to Multi-Office	multi-story office buildings, non-professional service buildings
Distance to Pro-Service	professional service buildings
Distance to Transport	airports, bus terminals, marine terminals, piers, marinas
Distance to Restaurant	restaurants, cafeterias
Distance to Driven-In Restaurant	drive-in restaurants
Distance to Financial	financial institutions
Distance to Insurance	insurance company offices
Distance to Other Commercial	other commercial parcels
Distance to Other Service	other service parcels
Distance to Wholesale	wholesale outlets, produce houses, manufacturing outlets
Distance to Entertainment	entertainment parcels
Distance to Hotel	hotels and motels
Distance to Light Industry	light industrial parcels
Distance to Heavy Industry	heavy industrial parcels
Distance to Industry	industrial parcels
Distance to Agricultural	agricultural parcels
Distance to Institution	institutional parcels
Distance to Education	educational parcels
Distance to Military	military parcels
Distance to Open Space	open spaces
Distance to Hospital	hospitals
Distance to Government	government parcels

Note: all distances are means of average distances from parcels to POIs within a 2 miles range in each block group.

4.3. Machine learning methods

Machine-learning algorithms, including LR, RF, ANN, and EGB, are trained and tested using 36 input variables. These models help to predict urban development trends for several land-use categories, such as single-family residential, occupied, and commercial uses.

Each method is briefly described next. We use the randomforest package (Liaw & Wiener, 2002) for RF, TensorFlow (Allaire & Tang, 2020) and Keras (Allaire & Chollet, 2020) for ANN, and xgboost (Chen et al., 2020) for EGB, all in R.

LR is a regression model in which a binary dependent variable, with values of 0 and 1, is associated with a set of independent variables. More details on estimation and inference are provided in Hastie, Tibshirani, and Friedman (2009).

RF is an ensemble learning method for classification (and regression) consisting of many decision trees, in which bootstrap samples build each decision tree (Breiman, 2001). Specifically, one creates bootstrap samples from the training data set and grows a tree from each bootstrap sample. At each node of a tree, a set of features is randomly selected to build the next node. RF makes a prediction by aggregating results from all grown decision trees. When training RF models, we use 1000 trees and set up the size of a random set of features (mtry) to 4 (the number of randomly selected variables as a candidate set at each split is 4). The node size is set to 3 (the minimum node size is 3).

The randomforest package provides variable importance measures such as Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) (Han, Guo, & Hua, 2016). Considering both MDA and MDG is more robust than using any single one of them. Han et al. (2016) suggest a new method called MDAMDG, combining MDA and MDG to measure variable importance. A large MDA, MDG, and MDAMDG value indicate that the corresponding variable is important.

ANN is a mathematical model inspired by biological neural networks (Hastie et al., 2009). ANN is an interconnected group of simple processing elements called nodes. ANN consists of multiple layers, including input layer, hidden layers, and output layer. Each layer involves interconnected nodes containing activation functions and determines the output of the node given input data. In each layer, the output from the previous layer becomes an input for the current layer. As an input value enters into the node, it gets multiplied by a weight value and added on a bias. All weights and biases are estimated by backpropagation (Fan, Ma, & Zhong, 2019).

In our analysis, three layers are used, each with 50, 20, and 2 as dimensions of the output spaces. The rectified linear unit (relu) and softmax activation functions (Nair & Hinton, 2010; Nwankpa, Ijomah, Gachagan, & Marshall, 2018) are used for the hidden layers and output layer, respectively. 60 epochs are used to train the model (the learning algorithm passes through the entire training data set 60 times) and batch

Table 4
Prediction accuracy of four ML methods over different model variable sets.

	Dependent	Model Variable Set	LR	RF	ANN	EGB
over 5 years	Single-family growth	temporal	0.6323 ± 0.0089	0.7707 ± 0.0077	0.7943 ± 0.0074	0.7857 ± 0.0075
		spatio-temporal	0.6348 ± 0.0088	0.7925 ± 0.0074	0.7869 ± 0.0075	0.791 ± 0.0075
		dist. Based covariates	0.6517 ± 0.0087	0.7068 ± 0.0084	0.6594 ± 0.0087	0.6853 ± 0.0085
		spatio-temporal + dist. Cov.	0.668 ± 0.0086	0.8066 ± 0.0073	0.7338 ± 0.0081	0.8024 ± 0.0073
	Occupancy growth	temporal	0.645 ± 0.0088	0.6814 ± 0.0086	0.7015 ± 0.0084	0.6991 ± 0.0084
		spatio-temporal	0.6624 ± 0.0087	0.7068 ± 0.0084	0.7066 ± 0.0084	0.7133 ± 0.0083
		dist. Based covariates	0.5775 ± 0.0091	0.6418 ± 0.0088	0.6134 ± 0.0089	0.6147 ± 0.0089
		spatio-temporal + dist. Cov.	0.6773 ± 0.0086	0.7275 ± 0.0082	0.6781 ± 0.0086	0.7207 ± 0.0082
	Commercial growth	temporal	0.7223 ± 0.0082	0.751 ± 0.0079	0.7679 ± 0.0078	0.7654 ± 0.0078
		spatio-temporal	0.7216 ± 0.0082	0.7644 ± 0.0078	0.757 ± 0.0079	0.7668 ± 0.0078
		dist. Based covariates	0.7219 ± 0.0082	0.7219 ± 0.0082	0.6852 ± 0.0085	0.7167 ± 0.0083
		spatio-temporal + dist. Cov.	0.7198 ± 0.0082	0.7769 ± 0.0077	0.7062 ± 0.0084	0.758 ± 0.0079
over 10 years	Single-family growth	temporal	0.551 ± 0.0091	0.7732 ± 0.0077	0.7888 ± 0.0075	0.7807 ± 0.0076
		spatio-temporal	0.6453 ± 0.0088	0.7872 ± 0.0075	0.7909 ± 0.0075	0.7862 ± 0.0075
		dist. Based covariates	0.6642 ± 0.0087	0.7171 ± 0.0083	0.6773 ± 0.0086	0.6924 ± 0.0085
		spatio-temporal + dist. Cov.	0.6725 ± 0.0086	0.7976 ± 0.0074	0.7378 ± 0.0081	0.7924 ± 0.0074
	Occupancy growth	temporal	0.6296 ± 0.0089	0.6594 ± 0.0087	0.679 ± 0.0086	0.6801 ± 0.0086
		spatio-temporal	0.6366 ± 0.0088	0.6853 ± 0.0085	0.6746 ± 0.0086	0.6816 ± 0.0086
		dist. Based covariates	0.5711 ± 0.0091	0.645 ± 0.0088	0.5996 ± 0.009	0.6076 ± 0.009
		spatio-temporal + dist. Cov.	0.6463 ± 0.0088	0.7134 ± 0.0083	0.6499 ± 0.0088	0.7041 ± 0.0084
	Commercial growth	temporal	0.5353 ± 0.0092	0.7098 ± 0.0083	0.736 ± 0.0081	0.7268 ± 0.0082
		spatio-temporal	0.5382 ± 0.0092	0.7198 ± 0.0082	0.7203 ± 0.0082	0.7228 ± 0.0082
		dist. Based covariates	0.5759 ± 0.0091	0.6439 ± 0.0088	0.6217 ± 0.0089	0.6347 ± 0.0088
		spatio-temporal + dist. Cov.	0.5747 ± 0.0091	0.7349 ± 0.0081	0.661 ± 0.0087	0.7164 ± 0.0083

size is set at 512. The categorical cross-entropy is used as the loss function while the selected optimizer is Adam (Kingma & Ba, 2014) which is an algorithm for gradient-based optimization of stochastic objective functions. ANNs were trained using a higher number of layers (4 and 5), different epoch sizes (80 and 100), and different batch sizes (256 and 1024). However, the prediction accuracy results did not change significantly (less than 0.01). Hence, the original settings were retained throughout the analysis.

EGB, which is an efficient implementation of gradient boosting, is an ensemble learning method combining the predictive power of multiple models to provide an optimal solution (Chen & Guestrin, 2016). Gradient boosting builds models sequentially, so that each consecutive model tries to correct the errors present in the previous model.

In this analysis, we use a tree classifier. The maximum depth of a tree is set as 15; γ, a parameter to control minimum loss reduction, is set as 3; λ, a tuning parameter to control the regularization term, is set at 0; and the evaluation metrics is a binary classification error rate. Using different values of λ, such as 1, 10, and 50, did not improve the prediction accuracy. Therefore, the default value of 0 is used.

Additional details about these machine learning methods are provided in Section 2 of the supplementary file.

5. Main results

We present the main results from the analysis of the Florida urban data set, using the machine learning (ML) methods described in the previous section. 11394 block groups are used: 9116 block groups for model training and 2278 for model testing. Different sets of the 28 distance-related covariates and 8 spatio-temporal covariates are used as input features: models with (1) temporal variables (4, the number of covariates), (2) spatio-temporal variables (8), (3) distance-related variables (28), and (4) spatio-temporal and distance-related variables (36).

5.1. Prediction performance

Prediction accuracy is an important measure to evaluate binary classification models. Various models with different binary dependent variables, $Z_1(t,5)$, $Z_1(t,10)$, $Z_2(t,5)$, $Z_2(t,10)$, $Z_3(t,5)$, and $Z_3(t,10)$, are evaluated in terms of prediction accuracy at year $t \in \{2015, 2016, 2017, 2018, 2019\}$. Table 4 provides the prediction accuracy rates of the four methods over different models variable sets with a 95% confidence interval (CI). The best prediction accuracy rates are highlighted using a bold font.

Table 4 shows that the highest prediction accuracy is achieved using the RF model with spatio-temporal and accessibility covariates. A prediction accuracy rate of around 80% is achieved for single-family growth prediction over 5 years. It is interesting that ML methods outperform statistical models (LR), which confirms the usefulness of such methods in urban growth prediction. Comparing the prediction accuracy of the accessibility covariates models (28 variables) with the prediction accuracy of the spatio-temporal and distance covariates models, the latter models have a 10% higher accuracy rate on average. This suggests that the 8 spatio-temporal covariates play significant roles in the models. The 28 accessibility features are useful but less informative when compared to the spatio-temporal variables, since the temporal models from RF, ANN, and EGB have all a higher prediction accuracy than the model incorporating only distance-based covariates. Models with smaller time intervals display better performances than those with larger time intervals. This is expected because prediction over 10 years in future seems to be a harder task than prediction over 5 years.

5.2. Variable selection

Another important measure to evaluate models is variable selection, which can be done using several variable importance measures. Since RF has the best prediction performance, we focus these on this method. We choose MDAMDG (Han et al., 2016) to evaluate variable importance since MDAMDG does variable selection as well as compute variable importance. Table 5 provides MDAMDG values for the optimal model. The highlighted numbers with an asterisk are the top ten most important variables. A dash mark indicates a variable not selected in the optimal model. Several temporal and spatial predictors are among the most important variables. This justifies the use of these predictors in our model. For the single-family growth model, Distance to Education is the most important variable among the 28 accessibility variables, followed by Distance to Government, Distance to Agricultural, and Distance to Institution. For the occupancy growth rate model, Distance to Institution is the most important variable, followed by Distance to Government, Distance to Industry, Distance to Education, and Distance to Other Commercial. For the commercial model, Distance to Government is the most important variable among the 28 distance-based variables, followed by Distance to Institution, Distance to Education, Distance to

Table 5
Variable importance score (MDAMDG) from RF.

Variable	$Z_1(t,k)$		$Z_2(t,k)$		$Z_3(t,k)$	
	$k=$ 5	$k=$ 10	$k=$ 5	$k=$ 10	$k=$ 5	$k=$ 10
k-lag temporal	__36*__	__18*__	__36*__	__36*__	__34*__	__36*__
2*k*-lag temporal	__34*__	__13*__	__34*__	__18*__	__36*__	__32*__
3*k*-lag temporal	__31*__	–	__18*__	__32*__	__32*__	__26*__
4*k*-lag temporal	__19*__	–	14	__20*__	__27*__	__22*__
k-lag neighborhood effect	__16*__	__8*__	14	__31*__	–	__20*__
2*k*-lag neighborhood effect	–	–	__21*__	13	–	–
3*k*-lag neighborhood effect	14	__8*__	__21*__	9	–	6
4*k*-lag neighborhood effect	__16*__	__9*__	–	11	–	10
Distance to Driven-In Restaurant	–	–	–	–	–	–
Distance to Financial	14	–	16	13	12	–
Distance to Education	__21*__	__14*__	__18*__	__23*__	__19*__	__23*__
Distance to Pro-Service	15	–	–	–	14	–
Distance to Restaurant	14	–	–	–	14	–
Distance to Multi-Office	–	–	–	–	–	–
Distance to Agricultural	__21*__	__3*__	11	14	–	–
Distance to Community Mall	–	–	10	–	–	14
Distance to Other Service	–	–	–	–	13	–
Distance to Other Commercial	13	–	__18*__	__21*__	__17*__	__21*__
Distance to One-story Office	__16*__	–	14	17	12	12
Distance to Government	__20*__	__8*__	__23*__	__20*__	__24*__	__25*__
Distance to Institution	__16*__	__9*__	__26*__	__24*__	__23*__	__22*__
Distance to Industry	14	–	__24*__	__18*__	__15*__	14
Distance to Entertainment	–	–	13	–	__16*__	17
Distance to Open Space	–	–	–	–	–	–
Distance to Hotel	–	–	11	11	10	13
Distance to Light Industry	–	–	–	11	9	9
Distance to Recreation	12	–	–	–	__15*__	__20*__
Distance to Stores	–	–	–	–	–	–
Distance to Supermarket	–	–	–	–	–	–
Distance to Transport	–	–	–	–	–	–
Distance to Heavy Industry	–	–	–	–	–	–
Distance to Wholesale	–	–	–	–	–	–
Distance to Regional Shopping Center	–	–	–	–	–	–
Distance to Hospital	–	–	–	–	–	–
Distance to Insurance	–	–	–	–	–	–
Distance to Military	–	–	–	–	–	–

Other Commercial, and Distance to Recreation.

Model results highlight important dynamics in single-family residential developments. Historical conditions in block groups are the most important factor in new residential developments. Similarly, historical conditions of neighboring block groups play a significant role in land development dynamics. In addition, distances to educational institutions, agricultural lands and government facilities are more important than historical conditions of neighboring block groups. Finally, distances to financial institutions, professional services, restaurants, commercial lands, one-story office buildings, institutional buildings, and industrial units are important factors in single-family residential developments.

Similar to the single-family residential model, the model results for commercial land development provide insights about dynamics in commercial land development in Florida. Historical conditions in block groups are also the most important factors in contemporaneous commercial land development. However, historical conditions of neighboring block groups do not play a significant role. Average distances to institutional facilities and government facilities are the second important factors in commercial land development. Distances to educational institutions, other commercial lands, industrial lands, recreational areas, and industrial facilities are also significant determinants behind the commercial land development in Florida.

5.3. Out-of-sample forecast of land development

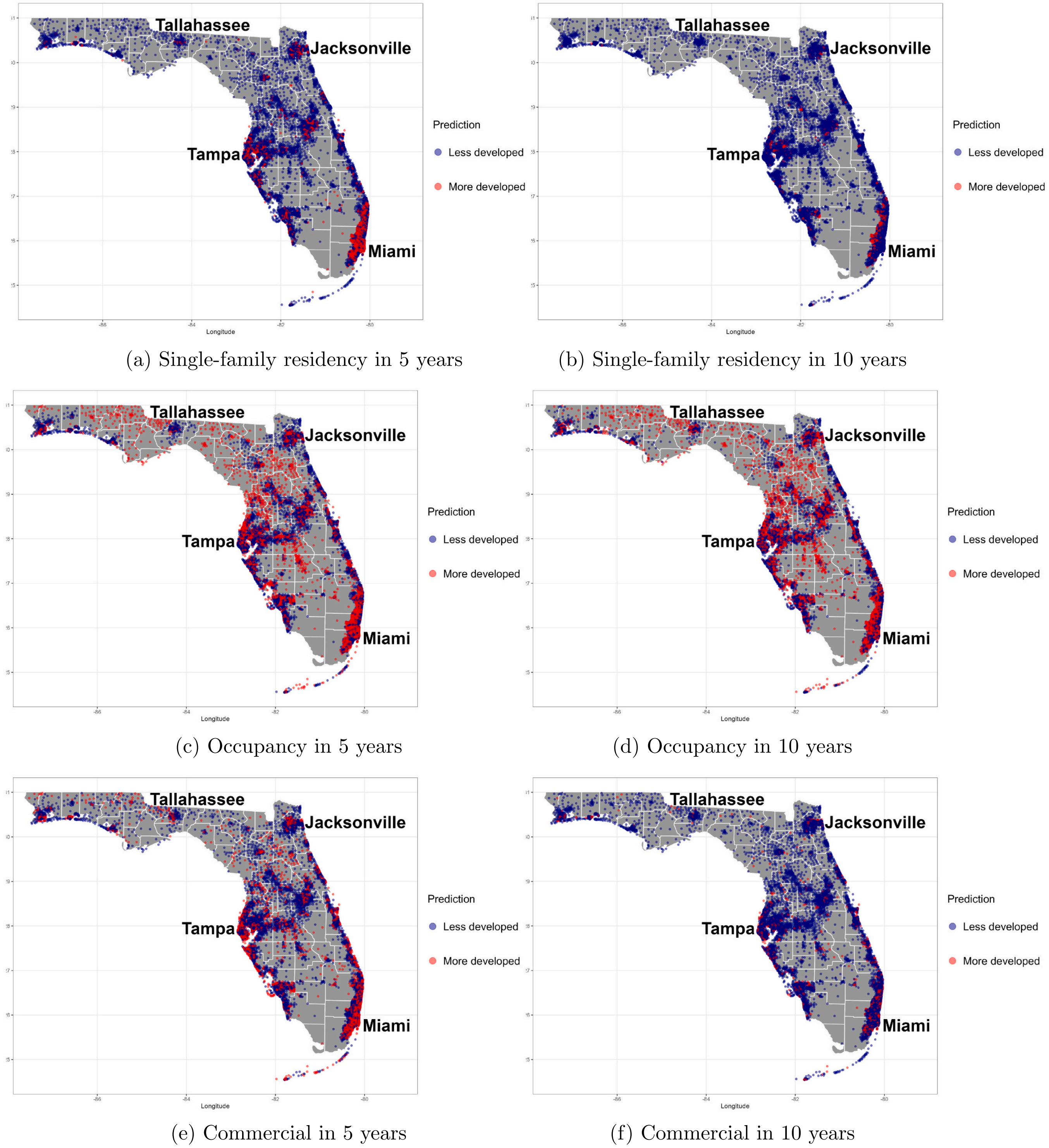
Land development in the future can be predicted by the developed machine learning models. The RF prediction model incorporating only spatio-temporal variables provides accurate results. When distance-based covariates are included, the model accuracy increases by approximately 1%. Distance-based variables can be included in predictions of future land developments if the covariates are updated based on future scenarios. For example, potential future land development can be estimated based on planned new hospitals, schools, shopping centers, etc. However, here, we introduce an accurate prediction modeling approach using minimum information. We train a RF prediction model with 8 spatio-temporal variables using a data set with $t = 2019$ and $k = 5$, where RF displays the best performance Section 5.1. The predicted values of the growth rates of the numbers of single-family residential parcels, occupied parcels, and commercial parcels in 2024 and 2029, are depicted in Fig. 3. Blue dots present areas where the growth rate is above the median ($\widehat{Z_j(t,k)} = 1$), whereas red dots present areas in which the growth rate is below the median ($\widehat{Z_j(t,k)} = 0$). These maps suggest that the central areas of Orlando, Tampa, and Miami may be less developed in the future. Since the central areas of these major cities are mostly developed, they provide fewer opportunities for new land developments.

All the machine learning models are developed independently based on the dependent variable, thus inference can only be made from the corresponding model. However, it is worth to cross interpret the results from different models to get meaningful insights. Fig. 4(a) presents areas where the growth rate of the number of single-family residential parcels is below the median, but the growth rate of occupied parcels is above the median. The areas where the growth rate of the number of single-family residential parcels is above the median but the growth rate of occupied parcels is below the median are shown in Fig. 4(b). Fig. 4(c) presents areas where the growth rate of the number of single-family residential parcels is below the median but the growth rate of commercial parcels is above the median. The areas where the growth rate of the number of single-family residential parcels is above the median but the growth rate of commercial parcels is below the median is shown in Fig. 4(d). The predictions presented in these two figures indicate critical land developments in Florida. When we look at single-family residential developments, we clearly expect fewer developments within built-up areas. According to our future predictions, Florida will experience new land developments in the peripheral areas of cities, with single-family residential being the major driver of these land developments. Such conditions could result in threatening environmentally-sensitive areas and increasing public infrastructure costs. Such an urban growth pattern is expected based on our theoretical understanding of location choice of certain land uses. Single-family residential land use cannot compete with non-residential and existing residential land uses in developed areas, therefore peripheral locations are desirable choices (O'Sullivan, 2012).

Similar to occupancy and single-family growth rate, the occupancy and commercial results also provide reasonable predictions. Commercial parcels are predominantly expected in densely urbanized and coastal areas. More commercial developments near coastal areas require further investigations in order to evaluate potential risks under sea level rises. There are also some expected commercial areas in rural areas, but these locations are close to major transportation networks. Similar to our discussions on single-family residential parcels, we also see meaningful predictions based on our theoretical understanding. Finally, model results successfully capture profit maximization behaviours of commercial activities. (O'Sullivan, 2012).

6. Discussion

In this section, the best-fitted model results are further discussed by



(a) Single-family residency in 5 years

(b) Single-family residency in 10 years

(c) Occupancy in 5 years

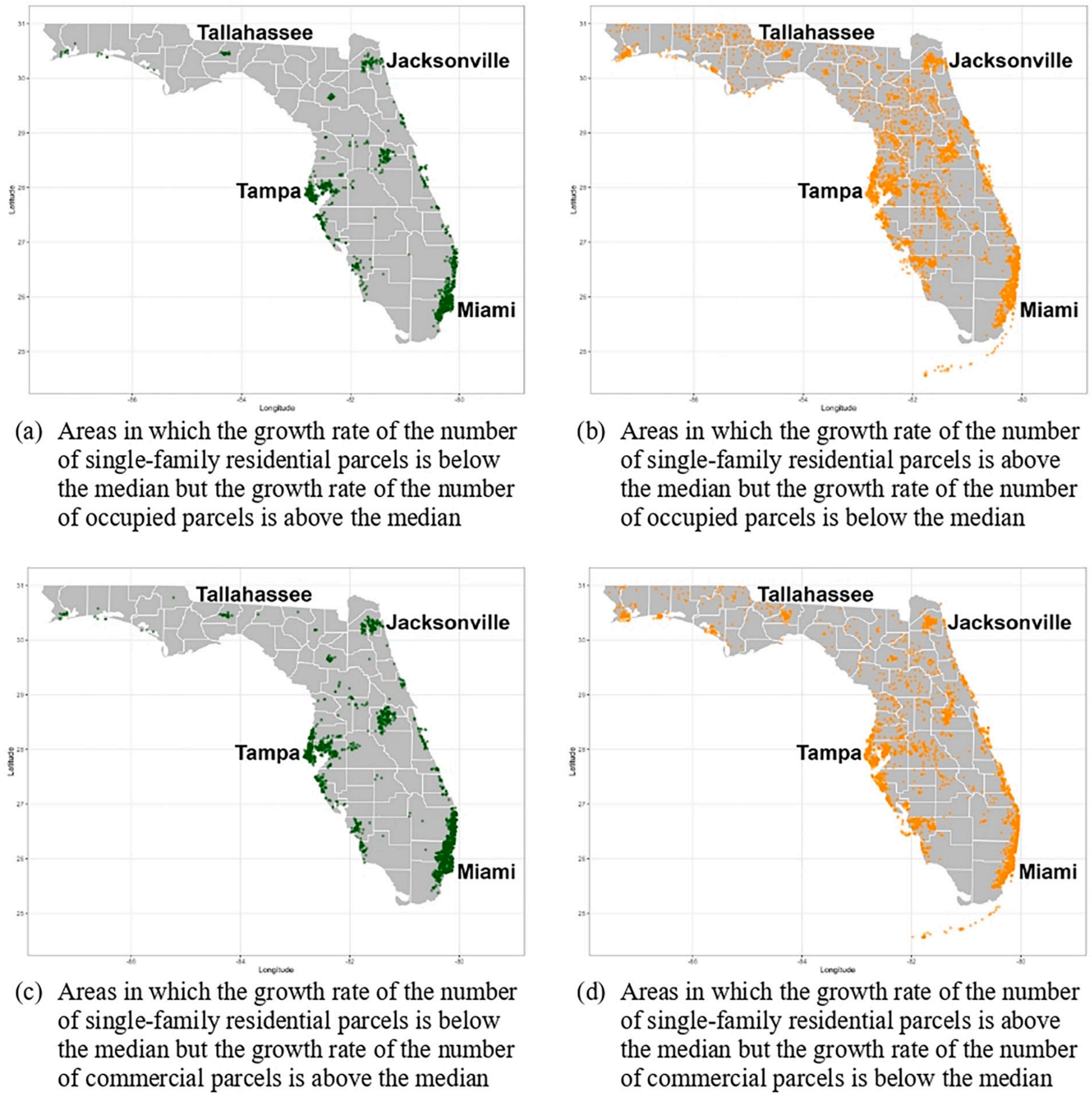
(d) Occupancy in 10 years

(e) Commercial in 5 years

(f) Commercial in 10 years

Fig. 3. Predicted values of the growth rate of single-family residency, occupancy, and commercial parcels in 5 years and 10 years using RF.

comparing them with results from similar studies in the literature. The implemented RF model allows us to rank the importance of the effects of the explanatory variables on land development dynamics. Our results show the clustering patterns of the same land uses increase future land development potentials of this land use. Karimi et al. (2019) obtained a similar result in their land-use change modeling. Gounaridis et al. (2019) found that the most important factors in land developments are road and enterprise densities. In this study, we did not explicitly control for these densities. However, being close to certain POIs in dense urban areas, such as commercial and institutional services, is a significant factor in commercial land developments in Florida. Lv et al. (2021) report that densities of restaurants, hospitals and markets play significant roles in land development dynamics. Further, Shafizadeh-Moghadam et al. (2021)'s RF model results indicate that distance to greenery is the most significant factor in land development, followed by altitude, distances from crops, urban roads and barren land. Our results for single-family residential developments, which are the main drivers in Florida's land developments, are significantly affected by distances to agricultural



(a) Areas in which the growth rate of the number of single-family residential parcels is below the median but the growth rate of the number of occupied parcels is above the median

(b) Areas in which the growth rate of the number of single-family residential parcels is above the median but the growth rate of the number of occupied parcels is below the median

(c) Areas in which the growth rate of the number of single-family residential parcels is below the median but the growth rate of the number of commercial parcels is above the median

(d) Areas in which the growth rate of the number of single-family residential parcels is above the median but the growth rate of the number of commercial parcels is below the median

Fig. 4. Numbers of occupied parcels and numbers of single-family residential parcels in 2024 (RF).

and large government facilities. In the study conducted by Liang et al. (2020), the proposed RF model for land-use changes in Shanghai shows that population is the most important factor, followed by GDP, distance to subways, distance to airports, and other transportation infrastructures.

We also compare our model prediction accuracy with the bagging approach introduced by Talukdar et al. (2021) and Support Vector Machine by Karimi et al. (2019). Our test results are presented in Table 6. Our best RF models with spatio-temporal and distance based covariates outperform Bagging and SVM models. The prediction accuracy of Bagging is higher than that of SVM in all cases and our best RF models are about 2.5% more accurate than the Bagging model.

7. Concluding remarks

Accurate predictions of future land development depend on the successful representations of actual complex dynamics. A robust land-use change model should account for non-linear relationships using proxy information about a wide range of aspects of land development. Previously introduced models focused on using site-specific, socio-economic, neighborhood and accessibility components of land development. Accessing this information is a major limitation to build such models. There are also computational challenges in statistical models when non-linear relationships are incorporated. In this paper, machine learning methods, including Random Forest (RF), Artificial Neural Networks (ANNs), and Extreme Gradient Boosting (EGB), are tested using only accessibility to certain points of interest while also including spatial and temporal components.

We have used the Auditor's geo-coded tax database for Florida to derive historical land developments at the block-group level, based on actual year-built information. Since the database includes land use categories, accessibilities to a set of services and infrastructures are computed. We examine relationships between a set of explanatory variables and growth rates of residential, commercial, and occupied parcels at the block group level. The RF model provides the most accurate predictions and its results are in line with urban growth theories. Also,

Table 6
Prediction accuracy of SVM and Bagging over different model variable sets.

Dependent variables		Model	Temporal	Spatio-temporal	Dist. based covariates	Spatio-temporal + Dist. cov.
over 5 years	Single-family growth	Bagging	0.7614 (0.0078)	0.7817 (0.0076)	0.6743 (0.0086)	0.7837 (0.0076)
		SVM(linear) - cost = 0.1	0.5857 (0.009)	0.5924 (0.009)	0.6537 (0.0087)	0.672 (0.0086)
		SVM(linear) - cost = 10	0.592 (0.009)	0.5972 (0.009)	0.6549 (0.0087)	0.6729 (0.0086)
		SVM(radial basis) - cost = 10, gamma = 1	0.7674 (0.0078)	0.7573 (0.0079)	0.6398 (0.0088)	0.6382 (0.0088)
		SVM(polynomial) - cost = 10, gamma = 1	0.4807 (0.0092)	0.531 (0.0092)	0.5788 (0.0091)	0.5969 (0.009)
		SVM(polynomial) - cost = 10, gamma = 2	0.5203 (0.0092)	0.5274 (0.0092)	0.578 (0.0091)	0.5953 (0.009)
		SVM(polynomial) - cost = 10, gamma = 3	0.4798 (0.0092)	0.5253 (0.0092)	0.551 (0.0091)	0.5899 (0.009)
	Occupancy growth	Bagging	0.6737 (0.0086)	0.6883 (0.0085)	0.6068 (0.009)	0.691 (0.0085)
		SVM(linear) - cost = 0.1	0.6379 (0.0088)	0.6601 (0.0087)	0.5838 (0.0091)	0.6788 (0.0086)
		SVM(linear) - cost = 10	0.6379 (0.0088)	0.6602 (0.0087)	0.582 (0.0091)	0.6737 (0.0086)
		SVM(radial basis) - cost = 10, gamma = 1	0.6983 (0.0084)	0.6687 (0.0086)	0.5805 (0.0091)	0.5773 (0.0091)
		SVM(polynomial) - cost = 10, gamma = 1	0.6187 (0.0089)	0.5293 (0.0092)	0.5439 (0.0091)	0.5798 (0.0091)
		SVM(polynomial) - cost = 10, gamma = 2	0.5383 (0.0092)	0.5205 (0.0092)	0.5359 (0.0092)	0.5793 (0.0091)
		SVM(polynomial) - cost = 10, gamma = 3	0.5207 (0.0092)	0.5343 (0.0092)	0.5213 (0.0092)	0.5772 (0.0091)
	Commercial growth	Bagging	0.748 (0.008)	0.7504 (0.0079)	0.7032 (0.0084)	0.753 (0.0079)
		SVM(linear) - cost = 0.1	0.7219 (0.0082)	0.7219 (0.0082)	0.7219 (0.0082)	0.7219 (0.0082)
		SVM(linear) - cost = 10	0.7219 (0.0082)	0.7219 (0.0082)	0.7219 (0.0082)	0.7219 (0.0082)
		SVM(radial basis) - cost = 10, gamma = 1	0.7368 (0.0081)	0.7274 (0.0082)	0.7141 (0.0083)	0.7163 (0.0083)
		SVM(polynomial) - cost = 10, gamma = 1	0.2822 (0.0083)	0.3069 (0.0085)	0.5923 (0.009)	0.6024 (0.009)
		SVM(polynomial) - cost = 10, gamma = 2	0.2887 (0.0083)	0.3089 (0.0085)	0.5715 (0.0091)	0.6087 (0.009)
		SVM(polynomial) - cost = 10, gamma = 3	0.7159 (0.0083)	0.3137 (0.0085)	0.3041 (0.0084)	0.6043 (0.009)
over 10 years	Single-family growth	Bagging	0.7691 (0.0077)	0.7782 (0.0076)	0.6792 (0.0086)	0.7825 (0.0076)
		SVM(linear) - cost = 0.1	0.572 (0.0091)	0.5552 (0.0091)	0.661 (0.0087)	0.6723 (0.0086)
		SVM(linear) - cost = 10	0.5696 (0.0091)	0.555 (0.0091)	0.6597 (0.0087)	0.6727 (0.0086)
		SVM(radial basis) - cost = 10, gamma = 1	0.7784 (0.0076)	0.7561 (0.0079)	0.6655 (0.0087)	0.5337 (0.0092)
		SVM(polynomial) - cost = 10, gamma = 1	0.6916 (0.0085)	0.5718 (0.0091)	0.5632 (0.0091)	0.6223 (0.0089)
		SVM(polynomial) - cost = 10, gamma = 2	0.6491 (0.0088)	0.548 (0.0091)	0.5478 (0.0091)	0.6064 (0.009)
		SVM(polynomial) - cost = 10, gamma = 3	0.5864 (0.009)	0.5558 (0.0091)	0.5506 (0.0091)	0.6076 (0.009)
	Occupancy growth	Bagging	0.6525 (0.0087)	0.6681 (0.0086)	0.6035 (0.009)	0.6777 (0.0086)
		SVM(linear) - cost = 0.1	0.6231 (0.0089)	0.6422 (0.0088)	0.5769 (0.0091)	0.6491 (0.0088)
		SVM(linear) - cost = 10	0.623 (0.0089)	0.6422 (0.0088)	0.5751 (0.0091)	0.6469 (0.0088)
		SVM(radial basis) - cost = 10, gamma = 1	0.6773 (0.0086)	0.6392 (0.0088)	0.583 (0.0091)	0.5716 (0.0091)
		SVM(polynomial) - cost = 10, gamma = 1	0.6017 (0.009)	0.5679 (0.0091)	0.5174 (0.0092)	0.5613 (0.0091)
		SVM(polynomial) - cost = 10, gamma = 2	0.5896 (0.009)	0.5665 (0.0091)	0.5478 (0.0091)	0.5647 (0.0091)
		SVM(polynomial) - cost = 10, gamma = 3	0.58 (0.0091)	0.5555 (0.0091)	0.5264 (0.0092)	0.568 (0.0091)
	Commercial growth	Bagging	0.7096 (0.0083)	0.7029 (0.0084)	0.6191 (0.0089)	0.7059 (0.0084)
		SVM(linear) - cost = 0.1	0.532 (0.0092)	0.5322 (0.0092)	0.5448 (0.0091)	0.56 (0.0091)
		SVM(linear) - cost = 10			0.5544 (0.0091)	0.5596 (0.0091)

(*continued on next page*)

Table 6 (*continued*)

Dependent variables	Model	Temporal	Spatio-temporal	Dist. based covariates	Spatio-temporal + Dist. cov.
		0.5322 (0.0092)	0.5322 (0.0092)		
	SVM(radial basis) - cost = 10, gamma = 1	0.7219 (0.0082)	0.6904 (0.0085)	0.5442 (0.0091)	0.5546 (0.0091)
	SVM(polynomial) - cost = 10, gamma = 1	0.6738 (0.0086)	0.4573 (0.0091)	0.5645 (0.0091)	0.5423 (0.0091)
	SVM(polynomial) - cost = 10, gamma = 2	0.5751 (0.0091)	0.4488 (0.0091)	0.4943 (0.0092)	0.5402 (0.0092)
	SVM(polynomial) - cost = 10, gamma = 3	0.5452 (0.0091)	0.4488 (0.0091)	0.5227 (0.0092)	0.5603 (0.0091)

the RF model results reveal that spatio-temporal lags are major factors in capturing the variability of the dependent variable. This is an important finding, because spatial and temporal lags can be easily incorporated in land-use change models and can be utilized to make appropriate policy decisions.

Model results provide important insights into single-family residential and commercial land developments in Florida. Historical conditions in block groups are the most important factor in both new single-family residential and commercial developments. Distances to educational institutions, agricultural lands and government facilities are important factors for single-family residential land development, while, average distances to institutional and government facilities play significant roles in commercial land development.

We outline several areas of further research. In this paper, we separately model land use categories. Multivariate approaches may reach better prediction accuracy and are an interesting topic of future research. Further, prediction at finer geographical levels may yield more accurate predictions, but will require handling additional computational issues. In addition, distance-based covariates cannot be utilized in out-of-sample predictions without designing future scenarios. Designing future paths of distance-based covariates may be an interesting future research direction. Finally, ranking variables based on their importance provides some insight about explanatory variables; however, our current methodological approach should be improved to obtain causal relationships.

CRediT authorship contribution statement

Yuna Kim: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Abolfazl Safikhani:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Emre Tepe:** Conceptualization, Data curation, Visualization, Investigation, Writing – original draft, Writing – review & editing, Project administration, Funding acquisition.

Acknowledgements

This work was supported by the University of Florida Research [grant number AGR DTD 12-02-20, 2020]; the National Science Foundation [grant number 2124507, 2021]. The authors would like to thank anonymous reviewers for helpful comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compenvurbsys.2022.101801.

References

Allaire, J. J., & Chollet, F. (2020). keras: R Interface to 'Keras'. https://CRAN.R-project.org/package=keras.

Allaire, J. J., & Tang, Y. (2020). tensorflow: R Interface to 'TensorFlow'. https://CRAN.R-project.org/package=tensorflow.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Springer Netherlands.

Bahadori, M. T., Yu, Q. R., & Liu, Y. (2014). Fast multivariate spatio-temporal analysis via low rank tensor learning. In *Advances in neural information processing systems* (pp. 3491–3499).

Basse, R. M., Omrani, H., Charif, O., Gerber, P., & Bódis, K. (2014). Land use changes modelling using advanced methods: Cellular automata and artificial neural networks. the spatial and explicit representation of land cover dynamics at the cross-border region scale. *Applied Geography, 53*, 160–171. https://doi.org/10.1016/j.apgeog.2014.06.016. ISSN 0143-6228 https://www.sciencedirect.com/science/article/pii/S0143622814001325.

Bhat, C. R., Dubey, S. K., Alam, M. J. B., & Khushefati, W. H. (2015). A new spatial multiple discrete-continuous modeling approach to land use change analysis. *Journal of Regional Science, 55*(5), 801–841.

Breiman, L. (2001). Random forests. *Machine Learning, 450*(1), 5–32.

Carrion-Flores, C., & Irwin, E. G. (2004). Determinants of residential land-use conversion and sprawl at the rural-urban fringe. *American Journal of Agricultural Economics, 86* (4), 889–904.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. (2020). xgboost: Extreme gradient boosting. https://CRAN.R-project.org/package=xgboost.

Chomitz, K. M., & Gray, D. A. (1996). Roads, land use, and deforestation: A spatial model applied to Belize. *The World Bank Economic Review, 10*(3), 487–512.

Delasalles, E., Ziat, A., Denoyer, L., & Gallinari, P. (2019). Spatio-temporal neural networks for space-time data modeling and relation discovery. *Knowledge and Information Systems, 61*, 1241–1267.

Fan, J., Ma, C., & Zhong, Y. *A selective overview of deep learning*. (2019). *arXiv preprint arXiv:1904.05526*.

Ferdous, N., & Bhat, C. R. (2013). A spatial panel ordered-response model with application to the analysis of urban land-use development intensity patterns. *Journal of Geographical Systems, 15*, 1–29.

Fleming, M. M. (2004). Techniques for estimating spatially dependent discrete choice models. In L. Anselin, M. M. Fischer, G. J. D. Hewings, P. Nijkamp, & F. Snickars (Eds.), *Advances in spatial econometrics* (pp. 145–166). Springer-Verlag.

Gao, C., Feng, Y., Tong, X., Lei, Z., Chen, S., & Zhai, S. (2020). Modeling urban growth using spatially heterogeneous cellular automata models: Comparison of spatial lag, spatial error and gwr. *Computers, Environment and Urban Systems, 81*, 101459. https://doi.org/10.1016/j.compenvurbsys.2020.101459. ISSN 0198-9715 https://www.sciencedirect.com/science/article/pii/S0198971519303928.

Gounaridis, D., Chorianopoulos, I., Symeonakis, E., & Koukoulas, S. (2019). A random forest-cellular automata modelling approach to explore future land use/cover change in Attica (Greece), under different socio-economic realities and scales. *Science of the Total Environment, 646*, 320–335. https://doi.org/10.1016/j.scitotenv.2018.07.302. ISSN 0048-9697 https://www.sciencedirect.com/science/article/pii/S0048969718328006.

Han, H., Guo, X., & Hua, Y. (2016). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th ieee international conference on software engineering and service science (icsess)* (pp. 219–224). IEEE.

Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., … Muller, K.-R. (2013). Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of Chemical Theory and Computation, 9*(8), 3404–3419.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Huang, B., Zhang, L., & Wu, B. (2009). Spatiotemporal analysis of rural-urban land conversion. *International Journal of Geographical Information Science, 23*(3), 379–398.

Irwin, E. G., Bell, K. P., & Geoghegan, J. (2003). Modeling and managing urban growth at the rural-urban fringe: A parcel-level model of residential land use change. *Agricultural and Resource Economics Review, 32*(1), 83–102.

Irwin, E. G., & Geoghegan, J. (2001). Theory, data, methods: Developing spatially explicit economic models of land use change. *Agriculture, Ecosystems and Environment, 85*, 7–23.

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.

ISSN 0198-9715 https://www.sciencedirect.com/science/article/pii/S0198971518304332.

Kingma, D. P., & Ba, J. *Adam: A method for stochastic optimization*. (2014). *arXiv preprint arXiv:1412.6980*.

Liang, Z., Dang, X., Sun, Q., & Wang, S. (2020). Multi-scenario simulation of urban land change in shanghai by random forest and ca-markov model. *Sustainable Cities and Society, 55*, 102045. https://doi.org/10.1016/j.scs.2020.102045. ISSN 2210-6707 https://www.sciencedirect.com/science/article/pii/S2210670720300329.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News, 2* (3), 18–22. https://CRAN.R-project.org/doc/Rnews/.

Lv, J., Wang, Y., Liang, X., Yao, Y., Ma, T., & Guan, Q. (2021). Simulating urban expansion by incorporating an integrated gravitational field model into a demand-driven random forest-cellular automata model. *Cities, 109*, 103044. https://doi.org/10.1016/j.cities.2020.103044. ISSN 0264-2751 https://www.sciencedirect.com/science/article/pii/S0264275120313925.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1/2), 17–23.

Nahuelhual, L., Carmona, A., Lara, A., Echeverria, C., & Gonzalez, M. E. (2012). Land-cover change to forest plantations: Proximate causes and implications for the landscape in south-Central Chile. *Landscape and Urban Planning, 107*, 12–20.

Nair, V., & Hinton, G. E. (2010). *Rectified linear units improve restricted boltzmann machines*. Icml.

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. *Activation functions: Comparison of trends in practice and research for deep learning*. (2018). *arXiv preprint arXiv:1811.03378*.

Okwuashi, O., & Ndehedehe, C. E. (2021). Integrating machine learning with markov chain and cellular automata models for modelling urban land use change. *Remote Sensing Applications: Society and Environment, 21*, 100461. https://doi.org/10.1016/j.rsase.2020.100461. ISSN 2352-9385 https://www.sciencedirect.com/science/article/pii/S2352938520306364.

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association, 20*(349), 120–126.

O'Sullivan, A. (2012). *Urban economics* (Eight ed.). McGraw-Hill/Irwin.

Ron-Ferguson, N., Chin, J. T., & Kwon, Y. (2021). Leveraging machine learning to understand urban change with net construction. *Landscape and Urban Planning, 216*, 104239. https://doi.org/10.1016/j.landurbplan.2021.104239. ISSN 0169-2046 https://www.sciencedirect.com/science/article/pii/S0169204621002024.

Shafizadeh-Moghadam, H., Minaei, M., Jr, R. G. P., Asghari, A., & Dadashpoor, H. (2021). Integrating a forward feature selection algorithm, random forest, and cellular automata to extrapolate urban growth in the tehran-karaj region of iran. *Computers, Environment and Urban Systems, 87*, 101595. https://doi.org/10.1016/j.compenvurbsys.2021.101595. ISSN 0198-9715 https://www.sciencedirect.com/science/article/pii/S0198971521000028.

Soares-Filho, B., Rodrigues, H., & Follador, M. (2013). A hybrid analytical-heuristic method for calibrating land-use change models. *Environmental Modelling & Software, 43*, 80–87. https://doi.org/10.1016/j.envsoft.2013.01.010. ISSN 1364-8152 https://www.sciencedirect.com/science/article/pii/S1364815213000236.

Talukdar, S., Eibek, K. U., Akhter, S., Ziaul, S., Islam, A. R. M. T., & Mallick, J. (2021). Modeling fragmentation probability of land-use and land-cover using the bagging, random forest and random subspace in the teesta river basin, bangladesh. *Ecological Indicators, 126*, 107612. https://doi.org/10.1016/j.ecolind.2021.107612. ISSN 1470-160X https://www.sciencedirect.com/science/article/pii/S1470160X21002776.

Tepe, E., & Guldmann, J.-M. (2017). Spatial and temporal modeling of parcel-level land dynamics. *Computers, Environment and Urban Systems, 64*, 204–214. https://doi.org/10.1016/j.compenvurbsys.2017.02.005. ISSN 0198-9715 https://www.sciencedirect.com/science/article/pii/S0198971516301880.

Tepe, E., & Guldmann, J.-M. (2020). Spatio-temporal multinomial autologistic modeling of land-use change: A parcel-level approach. *Environment and Planning B: Urban Analytics and City Science, 47*(3), 473–488. https://doi.org/10.1177/2399808318786511

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS One, 14*(11). e0224365.

Verburg, P. H., Ritsema van Eck, J. R., de Nijs, T. C. M., & Dijst, P. S. M. J. (2004). Determinants of land-use change patterns in The Netherlands. *Environment and Planning. B, Planning & Design, 31*, 125–150.

Xing, W., Qian, Y., Guan, X., Yang, T., & Huayi, W. (2020). A novel cellular automata model integrated with deep learning for dynamic spatio-temporal land use change simulation. *Computers & Geosciences, 137*, 104430. https://doi.org/10.1016/j.cageo.2020.104430. ISSN 0098-3004 https://www.sciencedirect.com/science/article/pii/S0098300419307708.

Yu, D., & Srinivasan, S. (2016). Urban land use change and regional access: A case study in Beijing, China. *Habitat International, 51*, 103–113.

Yu, J., Hagen-Zanker, A., Santitissadeekorn, N., & Hughes, S. (2021). Calibration of cellular automata urban growth models from urban genesis onwards - a novel application of markov chain monte carlo approximate bayesian computation. *Computers, Environment and Urban Systems, 90*, 101689. https://doi.org/10.1016/j.compenvurbsys.2021.101689. ISSN 0198-9715 https://www.sciencedirect.com/science/article/pii/S019897152100096X.

Zhai, Y., Yao, Y., Guan, Q., Liang, X., Li, X., Pan, Y., … Zhou, J. (2020). Simulating urban land use change by integrating a convolutional neural network with vector-based cellular automata. *International Journal of Geographical Information Science, 34*(7), 1475–1499. https://doi.org/10.1080/13658816.2020.1711915