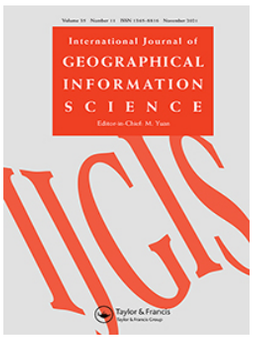

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area

Myung-Jin Jun

To cite this article: Myung-Jin Jun (2021) A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area, International Journal of Geographical Information Science, 35:11, 2149-2167, DOI: 10.1080/13658816.2021.1887490

To link to this article: https://doi.org/10.1080/13658816.2021.1887490

Published online: 01 Mar 2021.

Submit your article to this journal

Article views: 995

View related articles

View Crossmark data

Citing articles: 23 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tgis20

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2021, VOL. 35, NO. 11, 2149–2167
https://doi.org/10.1080/13658816.2021.1887490

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Check for updates

A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area

Myung-Jin Jun

Department of Urban Planning and Real Estate, Chung-Ang University, Seoul, South Korea

ABSTRACT

This study compares the performance of gradient boosting decision tree (GBDT), artificial neural networks (ANNs), and random forests (RF) methods in LUC modeling in the Seoul metropolitan area. The results of this study showed that GBDT and RF have higher predictive power than ANN, indicating that tree-based ensemble methods are an effective technique for LUC prediction. Along with the outstanding predictive performance, the DT-based ensemble models provide insights for understanding which factors drive LUCs in complex urban dynamics with the relative importance and nonlinear marginal effects of predictor variables. The GBDT results indicate that distance to the existing residential site has the highest contribution to urban land use conversion (30.4% of the relative importance), while other significant predictor variables were proximity to industrial and public sites (combined 32.3% of relative importance). New residential development is likely to be adjacent to existing residential sites, but nonresidential development occurs at a distance (about 600 m) from such sites. The distance to the central business district (CBD) had increasing marginal effects on residential land use conversion, while no significant pattern was found for nonresidential land use conversion, indicating that Seoul has experienced more population suburbanization than employment decentralization.

ARTICLE HISTORY
Received 16 January 2020
Accepted 4 February 2021

KEYWORDS
Land use change; gradient boosting decision tree; random forests; artificial neural networks; Seoul

1. Introduction

Accurate and detailed spatial prediction about urban land use change (LUC) is essential for achieving smart and sustainable urban development. The understanding of the LUC process and its precise prediction is challenging due to the complexity of the problems that arise from dynamic and nonlinear LUC processes by various factors, such as demographic and socioeconomic conditions, topographic and physical factors, and policies and regulations. A wide variety of LUC models have been developed alongside rapid advances in information and computing technology, including logistic regression (Hu and Lo 2007), cellular automata (Li and Yeh 2002, Van Vliet *et al*. 2009, Yao *et al*. 2017), Markov models (Lopez *et al*. 2001), and activity-based models (Groeneveld *et al*. 2017). However, these

CONTACT Myung-Jin Jun mjjun1@cau.ac.kr
© 2021 Informa UK Limited, trading as Taylor & Francis Group

models do not perform well in predicting LUC when applied to large-scale complex problems (Li and Yeh 2002, Aburas *et al*. 2016).

Significant progress has been made in the past few decades on the application of machine learning (ML) techniques to LUC prediction (Pijanowski *et al*. 2002, 2005). The ML model is advantageous for quantifying the interactions between LUC and its drivers as it has no assumptions on data distribution and no a priori understanding of variable relationships, and it is able to account for nonlinear association between those variables (Pijanowski *et al*. 2002, Dai *et al*. 2005). Numerous empirical applications of ML algorithms to LUC analysis are found in the literature. The Land Transformation Model (LTM) developed by Pijanowski *et al*. (2002) is a popular example of the ANN model applied to LUC. LTM combines ANN and Geographic Information Systems to simulate the potential effects of LUC in diverse fields and to forecast future LUCs. Empirical applications of the LTM include (1) investigating how the driving forces of LUCs, such as roads, highways, recreational facilities, and quality of views can affect urbanization patterns in Michigan's Grand Traverse Bay Watershed (Pijanowski *et al*. 2002) (2) forecasting future vacant lands and compare vacancy patterns of a shrinking city (Chicago, USA) and a growing city (Fort Worth, USA) (Lee *et al*. 2018); and (3) predicting infill development patterns of Tabriz, Iran, and examine any increase in built area or loss of agricultural land and wasteland area by 2021 (Rahimi 2016).

A few comparative studies exist in the literature in terms of the predictive powers of different algorithms in LUC modeling. Tayyebi and Pijanowski (2014) simulated multiple land use classes by building three ML techniques – ANN, Classification and Regression Trees, and Multivariate Adaptive Regression Splines – and concluded that the ANN technique provided the best accuracy in both areas for all three land use classes. Park *et al*. (2011) built frequency ratio, analytical hierarchy process, logistic regression, and ANN approaches to forecast urban LUCs in Korea and found that the ANN approach produced the highest overall accuracy at 92.3%.

However, Kamusoko and Gamba (2015) compared a random forest-cellular automata model with support vector machine cellular automata and logistic regression cellular automata models for simulating urban growth. They found that the random forest-cellular automata model outperformed the support vector machine cellular automata and logistic regression cellular automata models. Georganos *et al*. (2018) compared extreme gradient boosting algorithm, RF, and support vector machines for object-based urban land use–land cover classification. The results demonstrate that extreme gradient boosting parameterized with a Bayesian procedure systematically outperformed other algorithms, mainly in larger sample sizes. Sun *et al*. (2020) compared a GBDT with classifiers based on decision tree, distance weighted k-nearest neighbor, and the adaptive network-based fuzzy inference system for GPS signal reception classification in urban environments. Test results show that the GBDT-based algorithm is superior to the other three algorithms in terms of model accuracy. Du *et al*. (2018) explored the potential of tree-based methods by comparing the predictive power of the ANN model and four tree-based models (bagged trees, RF, extremely randomized trees, and bagged gradient boosting decision trees) to simulate the LUC in the Greater Tokyo Area. They found that tree-based models generally outperform ANN and extremely randomized trees perform better than the other tree-based models.

Though the predictive power of the model is an important criterion for model selection, the interpretive ability of the model is also critical for understanding the driving factors of land use change in the case of the LUC model. ANNs have been widely applied in LUC modeling due to their high adaptability and predictability, but they are barely interpretable as the approach operates like a black-box model (Pijanowski *et al*. 2002, Kavzoglu and Mather 2003, Guan *et al*. 2005, Etemad-Shahidi and Mahjoobi 2009, Rahimi 2016, Lee *et al*. 2018). On the other hand, decision trees are a popular type of ML algorithm due to their simplicity, fast training, and high interpretability, but they have poor predictability when encountering complex problems and tend to overfit data with exhaustive branches and deep trees (Solomatine and Xue 2004, Etemad-Shahidi and Mahjoobi 2009, Li *et al*. 2014, 2015, Kamusoko and Gamba 2015). The ensemble method has been suggested to improve the predictability of DT by training multiple DT models, and it is regarded as an alternative to achieve both interpretability and predictability (James *et al*. 2013, Du *et al*. 2018). Two ensemble techniques are widely used: bagging and boosting. Bagging generates several subsets of the original data for training by randomly sampling with replacements, while boosting is a sequential ensemble technique in which observations are selected based on errors in previous predictors.

This study compares the performance of ANNs and DT-based ensemble models to simulate the LUC in the Seoul metropolitan Area from 1995 to 2005[1]. ANNs and the DT-based ensemble models were selected for the comparison because these methods were recommended as appropriate algorithms by machine learning experts and developers in the literature in terms of accuracy and interpretation. In the case of the DT-based model, RF and GBDT models were selected because they are typical methods that use bagging and boosting techniques, respectively, to achieve high levels of accuracy and interpretability as well as overcome the overfitting problem (James *et al*. 2013). The predictive performances of three methods are compared by using accuracy, precision, f1-scores, and kappa scores, while the effects of different driving factors are interpreted from the results of the DT-based ensemble method with the relative importance and nonlinear marginal effects of predictor variables. The results of this study provide empirical evidence for selecting ML models that have not only high predictive power but also interpretability of the effects of different driving factors. Moreover, this study applies the Bayesian optimization hyperparameters tuning technique for finding optimal hyperparameters of the selected ML models.

2. Methodology and data

2.1. Artificial neural network

ANNs are a type of ML algorithm that is designed to mimic the learning patterns of the neurons of the brain. The most common neural network is a multilayer perceptron, a feed-forward ANN model (Krenker *et al*. 2011). A multilayer perceptron typically consists of three types of layers as the input, hidden, and output layers, in which the learning processing is performed via a system of weighted 'connections' among nodes (neurons), while no interconnections are between nodes within the same layer (Pijanowski *et al*. 2002, Kavzoglu and Mather 2003).

Since multilayer perceptron is one type of supervised learning algorithms, a learning algorithm is essential for the ANN application. Among various learning strategies, the backpropagation training algorithm, developed by Rumelhart *et al*. (1986), is the most popular learning technique in multilayer perceptron (Reed and Marks 1998, Kavzoglu and Mather 2003). In this algorithm, the input data are forward processed with randomly initialized network weights to produce training output, and the training output is compared with the target values to obtain the difference (error). The error then goes back through the network to adjust the weights and biases in such a way that the error is reduced. After repeating this process through iterative training cycles, the network usually converges to some state where the error of the calculations is minimized.

A challenging task in the application of multilayer perceptrons is the selection of appropriate hyper-parameter values (or network parameters) since they have a significant impact on the model performance. There are two types of hyperparameters in the ANN: optimizer and model-specific hyperparameters. Optimizer hyperparameters are related to the optimization in training process, including learning rate, batch size, and number of epochs. Learning rate controls the step size at each iteration in searching for a minimum of a loss function. Choosing the appropriate learning rate is challenging since a value too small or too large may result in a long training time to reach an ideal state or no algorithm convergence, respectively. A large batch size allows computational boost at the cost of more memory, while a small batch size is useful to prevent the training process from stopping at local minimum but induces more noise in error calculations. It is necessary to set the appropriate epoch values to prevent underfitting and overfitting. An epoch value that is too small or too large may result in underfitting or overfitting, respectively. Early stopping is a technique to avoid overfitting by determining when to stop training once the validation error has not improved.

Model-specific hyperparameters are numbers of hidden layers and neurons in each layer, dropout, and activation function. The numbers of hidden layers and neurons in each layer determines the network structure and the representational capacity of the network. Generally, there is no rule of thumb to determine the optimal numbers of layers and neurons. The use of cross-validation has been suggested to test the accuracy of the test set. Dropout is a method to randomly remove neurons out of the network during training, making the network less sensitive to the specific weights of neurons. Through this technique, a network is more likely to be generalized and less likely to overfit the training data.

2.2. Tree-based ensemble algorithms

2.2.1. Random forest

RF is an ensemble learning model made of many decision trees using bootstrapping (random sampling with replacement), random selection of subsets of features, and average voting to make predictions (Breiman 2001). RF is an enhanced version of the standard DT method as the performance of many weak learners (i.e. a single decision tree) can be improved via a voting scheme. The training procedure of RF is first to create a new training dataset using bootstrapping sampling for each tree in the forest, then to grow a tree for the optimal split by searching a random subset of input features at each node, and predictions resulting from every tree are aggregated to make a final prediction by

averaging the values of all trees (regression) or the majority vote (classification) (Svetnik *et al*. 2003).

RFs overcome overfitting problems by obtaining the final prediction aggregated from the individual prediction of each decision tree, and it has high accuracy through variance reduction (Breiman 1996). RFs reduce the variance of a single decision tree, leading to better predictions on new data. In addition, RFs are very flexible, as they do not require data preprocessing, such as data scaling and the assumptions of the model or linearity in the dataset (Caruana and Niculescu-Mizil 2006). The main disadvantage of RFs, however, is complexity that requires large computational resources and hinders intuitive interpretation of the relationship between the response and the predictor variables. RF involves the tuning of several hyperparameters to improve the predictive accuracy and control overfitting, including the number of random features to sample at each split point (*max_features*), the number of trees in the forest (*n_estimators*), the maximum depth of the tree (max_depth), the minimum number of samples required to be at a leaf node (min_samples_leaf), and the minimum number of samples required to split an internal node (min_samples_split).

2.2.2. Gradient boosting decision tree

Like RFs, GBDT, developed by Friedman (2001), is also a DT-based ensemble model. However, there are two major differences between RF and GBDT. First, unlike RFs, which build individual trees independently by a bagging technique to improve their accuracy through variance reduction, GBDTs sequentially build an ensemble to improve accuracy through bias reductions. They use a boosting process in which weighted resampling is carried out to put more weight on samples with lower prediction accuracy at the end of each iteration; thus, the samples with low accuracy would have higher chances of selection at the next iteration (Quinlan *et al*. 1996, Zhang and Haghani 2015). Second, unlike RFs, which combine results at the end of the process by averaging or using majority votes, GBDTs use sequential learning procedures that fits single DT models to minimize the current pseudo-residuals by least squares at each iteration utilizing the negative gradient of the loss function (Friedman 2002).

GBDT has the advantage of having high predictive power, but it is likely to result in overfitting if the dataset has a lot of noise (Opitz and Maclin 1999). A few regularization methods to avoid overfitting can be utilized to control the fitting process for balancing model fit and predictive power (Hastie *et al*. 2009). Regularization is essential for GBDT because its sequential learning procedures allow trees to be added until the model is overfitted (Elith *et al*. 2008). Model regularization can be achieved by jointly optimizing several hyperparameters such as the number of trees, learning rate (shrinkage), and tree depth (Elith *et al*. 2008). The number of trees refers to the number of iterations of gradient boosting, and more trees can reduce training error, so a large number usually results in better performance, as gradient boosting is fairly robust to overfitting. The learning rate is used to shrink the influence a single tree has on the overall prediction, and a lower learning rate is recommended to avoid overfitting. Subsample is the fraction of samples to be used for fitting the individual base learners. A subsample less than 1 leads to a reduction of variance and an increase in bias. The maximum depth limits the number of nodes in the tree, while the maximum features limit the number of features to consider when looking for the best split. In the GBDT model, there is a tradeoff between the

learning rate and the number of trees, as a smaller value of learning rate is likely to minimize loss function while requiring a large number of trees (Zhang and Haghani 2015). The tree complexity is another parameter affecting the performance of the GBDT algorithm. Tree complexity refers to the number of splits (or nodes) in each decision tree. Higher tree complexity (larger number of splits) tends to capture complex interactions among variables. Since the number of trees, learning rate, and tree complexity collectively influence the performance of the GBDT, it is important to find the best combination of these parameters for the optimal model.

2.3. Feature importance

Since the predictor variables have, in general, different impacts on the target, it is useful to measure the relative importance of each predictor variable in predicting the response variable. DTs can identify the importance of features by calculating information gain across over the splits using Gini Impurity or Entropy measures for classification trees. Breiman *et al*. (1984) proposed the following measure as an approximation of the relative importance of the predictor variable in the response for the single decision tree model:

$$I_j^2 = \frac{1}{M}\sum_{m=1}^{M} I_j^2(T_m) \tag{3}$$

where $I_j^2(T)$ refers to the importance of a predictor variable, which is based on the number of times the variable is chosen for splitting and weighted by the squared improvement to the model as a result of each split (Friedman and Meulman 2003). After averaging the importance measures over all trees, the relative influence is standardized to make sure that they add up to 100%, with higher numbers indicating stronger influence on the response.

2.4. Hyperparameter tuning

Hyperparameter tuning is a crucial step in the process of applied ML because the model performance is highly dependent on hyperparameter values. Two methods, grid and random search, have been frequently used for algorithm tuning over the years. However, both approaches are criticized for being highly time-consuming, as the number of parameter combinations increases exponentially with the addition of more hyperparameters and may thus evaluate unpromising areas of the search space. Recently, hyperparameter optimization has been conducted by automated methods such as Bayesian optimization that carries a more comprehensive search process while at the same time fully reflecting 'prior knowledge' when conducting a survey of the new hyperparameter values each time (see Brochu *et al*. 2010).

Bayesian optimization aims to find the optimal solution (x^*) of an unknown objective function f that takes an input value x and maximizes the function value $f(x)$. It is usually assumed that the expression of the objective function is not known explicitly (i.e. black-box function), but it takes a long time to calculate a function $f(x)$. In this situation, the main goal is to quickly and effectively find x^*, the optimal solution that $f(x)$ is maximized by sequentially examining the function value for as few input candidates as possible.

There are two essential elements to Bayesian optimization. First is the surrogate model, which performs stochastic estimation of the approximate form of an unknown objective function based on the input value-function points. The most frequently used probabilistic model as a surrogate model is the Gaussian process (GP). Unlike the ordinary probability model, which expresses the probability distribution for a specific variable, GP is a probability model for representing the probability distribution over the kinds of functions and assumes that the joint distribution between its components follows the Gaussian distribution. Second, the acquisition function refers to a function that recommends the next input candidate that is 'most useful in finding the optimal input (x^*)' based on the probabilistic estimation results of the objective function. Expected improvement (EI) is widely used as an acquisition function (Brochu *et al*. 2010). This study uses Bayesian optimization to find the optimal hyperparameters of three ML algorithms using the Scikit-Optimize library (skopt).

2.5. Evaluation of model performance

Several methods are used to evaluate an ML model's performance, including a confusion matrix, precision and recall, f1-score, and Kappa coefficients. A confusion matrix is a summary table showing the number of correct and incorrect predictions by class. It provides insight into the accuracy as well as the size and type of errors made by the ML model. Accuracy is a simple fraction of correctly predicted observation to the total incidents. Precision, also known as percent correct metric, is the fraction of correctly predicted positive observations among the total predicted positive observations $\left(\frac{\text{TruePositives(TP)}}{\text{(TruePositives(TP)+FalsePositives(FP))}}\right)$, while recall is the fraction of correctly predicted positive observations among all observations in the actual class $\left(\frac{\text{TruePositives(TP)}}{\text{(TruePositives(TP)+FalseNegatives(FN))}}\right)$. In contrast, f1-score is widely used when the data have an uneven class distribution. The f1-score considers false positives and false negatives together and computes the weighted average of precision and recall $\left(\frac{2*(\text{recall}*\text{precision})}{(\text{recall}+\text{precision})}\right)$. The Kappa coefficient is used to measure the degree of accuracy and reliability in a classification model. It compares the agreement presented from an actual transition map with the agreement expected from a predicted transition map with the following equation (Cohen 1960):

$$k = \frac{p_o - p_e}{1 - p_e} \tag{4}$$

where p_o is the relative observed agreement (accuracy) and p_e is the hypothetical probability of chance agreement. The general guidelines for interpreting the Kappa coefficient characterize the value of 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–-0.80 as substantial, and 0.81–1.0 as almost perfect agreement (Landis and Koch 1977).

2.6. Data and model variables

It is necessary to obtain land use information during the study period to simulate urban land use change using ML models. Building input data for the ML models require several steps. First, in order to detect land cover changes of the SMA between 1995 and 2005, two Landsat thematic mapper (TM) 5 satellite images were retrieved from the United States

Geological Survey. Acquisition dates were 19 February 1995, and 13 January 2005, for both satellite images. These images were selected for the analysis because of low cloud cover on land, with 4% in 1995 and 7% in 2005. Second, supervised land cover classifications of Landsat 5 images were conducted using 'Support Vector Machine Classifier' in ArcGIS Pro (Esri Inc., Redlands, California, USA). Figure 1 illustrates the land cover classifications of 1995 and 2005 for the SMA with five land use types. It shows that substantial LUCs took place in the southern part of the city of Seoul. Third, since land cover data include information on the natural environment, such as forests, wetlands, and non-built-up lands to make up the land surface, it provides limited information on land use type that is determined by human activity. This study combines the land cover classification maps with parcel-based Land Management Information System (LMIS)[2] to identify land use conversion for each grid cell during the study period. Fourth, we eliminated the physical factors hindering urban development by removing grid cells in which the slope is greater than 15° and higher than 200 m above sea level. We also eliminated grid cells regarding institutional factors such as national parks, freshwater protection areas, and military camps. After the removal of these grid cells, 155,662 developable grid cells were selected for the analysis. Fifth, we classified the grid cells into three types of land use conversion during the last decade: from vacant to residential uses or to nonresidential uses or remaining vacant. We found that 12,277 grid cells have been converted from vacant to residential and 5,484 cells to nonresidential use, while 88.6% of total developable land in 1995 remained vacant in 2005.

Data on nine predictor variables were collected for predicting urban LUCs, as these variables are considered to affect urban land use conversions. Two variables on physical conditions (average slope and altitude) and four distance variables representing proximity to neighboring sites (distances to residential, commercial, industrial, and public sites) were included. We also considered three location and accessibility variables: distances to roads, to

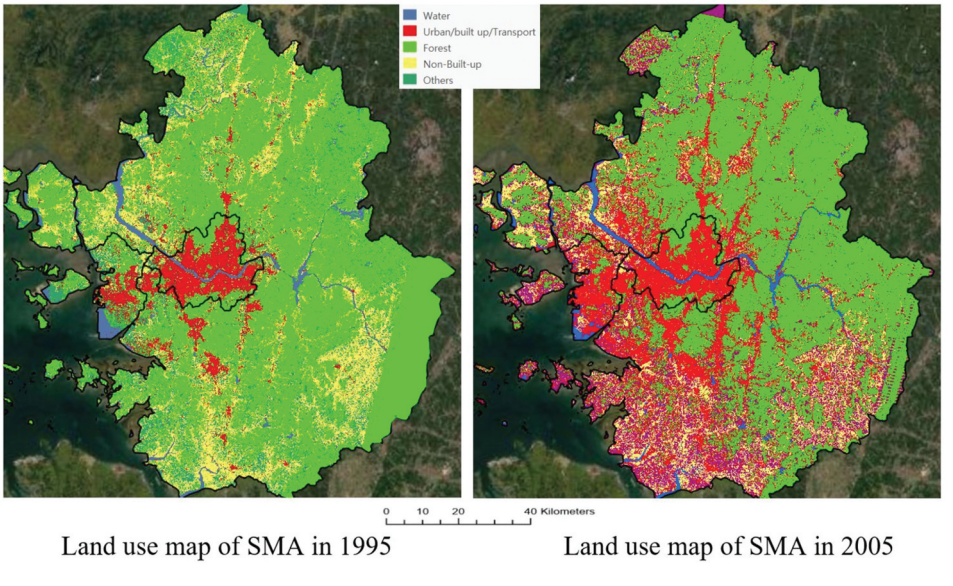

Figure 1. Results of supervised land cover classification of Landsat 5 images.

Table 1. Basic statistics of the predictor variables.

Description	Mean	SD	Min	Max
Average slope (degrees)	3.04	2.91	0.00	15.00
Altitude (meters)	36.50	26.33	0.00	200.00
Distance to a residential site (km)	0.79	1.43	0.07	16.12
Distance to a commercial site (km)	4.60	4.71	0.07	31.50
Distance to an industrial site (km)	4.31	6.65	0.06	38.21
Distance to a public site (km)	2.77	3.14	0.07	23.42
Distance to a road (km)	1.07	2.22	0.00	22.82
Distance to Seoul's CBD (km)	44.68	16.17	1.56	81.39
Distance to the nearest subway station (km)	11.79	9.25	0.01	49.70
N	155,662			

the nearest subway station, and to the city center. Table 1 presents the descriptive statistics for the variables used in the ML models, while Figure 2 illustrates raster images for the nine predictor variables. Since the dataset in this study was highly imbalanced, with only 11.4% of vacant cells developed for residential and nonresidential uses between 1995 and 2005 while 88.6% remained vacant, we employed the Synthetic Minority Over-sampling Technique (SMOTE) to adjust the class distribution of our dataset (Chawla *et al.* 2002).

3. Results

3.1. Comparison of model performance

Land use changes were predicted using the three ML models with optimal hyperparameters by the Bayesian optimization process,[3] and the accuracy of the model was evaluated by comparing model-predicted data with the actual test dataset. Table 2 presents model performances of the three models. The GBDT model was found to have the highest accuracy (91.8% vs. 89.7% for the ANN and 89.9% for the RF), precision, and f1 values compared to those of the ANN and RF models. Precision values of the GBDT on the transited sites are higher at 0.64 (0.40 for the ANN and 0.55 for the RF) and at 0.67 (0.38 for the ANN and 0.50 for the RF) than those of other models, respectively, for residential and nonresidential conversions. The f1-score, the weighted average of precision and recall, were 0.58 and 0.59 for residential and nonresidential conversions, respectively. The DT-based ensemble models had higher Kappa coefficients (0.574 and 0.539 for the GBDT and RF) than the ANN model (0.438), indicating that the DT-based models have a relatively higher model performance with a moderate and substantial predictability.

Along with the superior predictive performance, the DT-based models have an interpretive advantage regarding the driving factors of land use change through the analysis of the relative importance and nonlinear effects of predictor variables. Since GBDT and RF generate similar results, this study discusses the results of the GBDT model, which has relatively higher predictive power than the RF model. In addition, the comparisons between the actual LUC and model prediction were conducted with the GBDT model results.

3.2. Relative importance of predictor variables from the GBDT model

The relative contributions of predictor variables for land use conversion were measured based on the GBDT model. A higher value of a predictor variable's relative importance

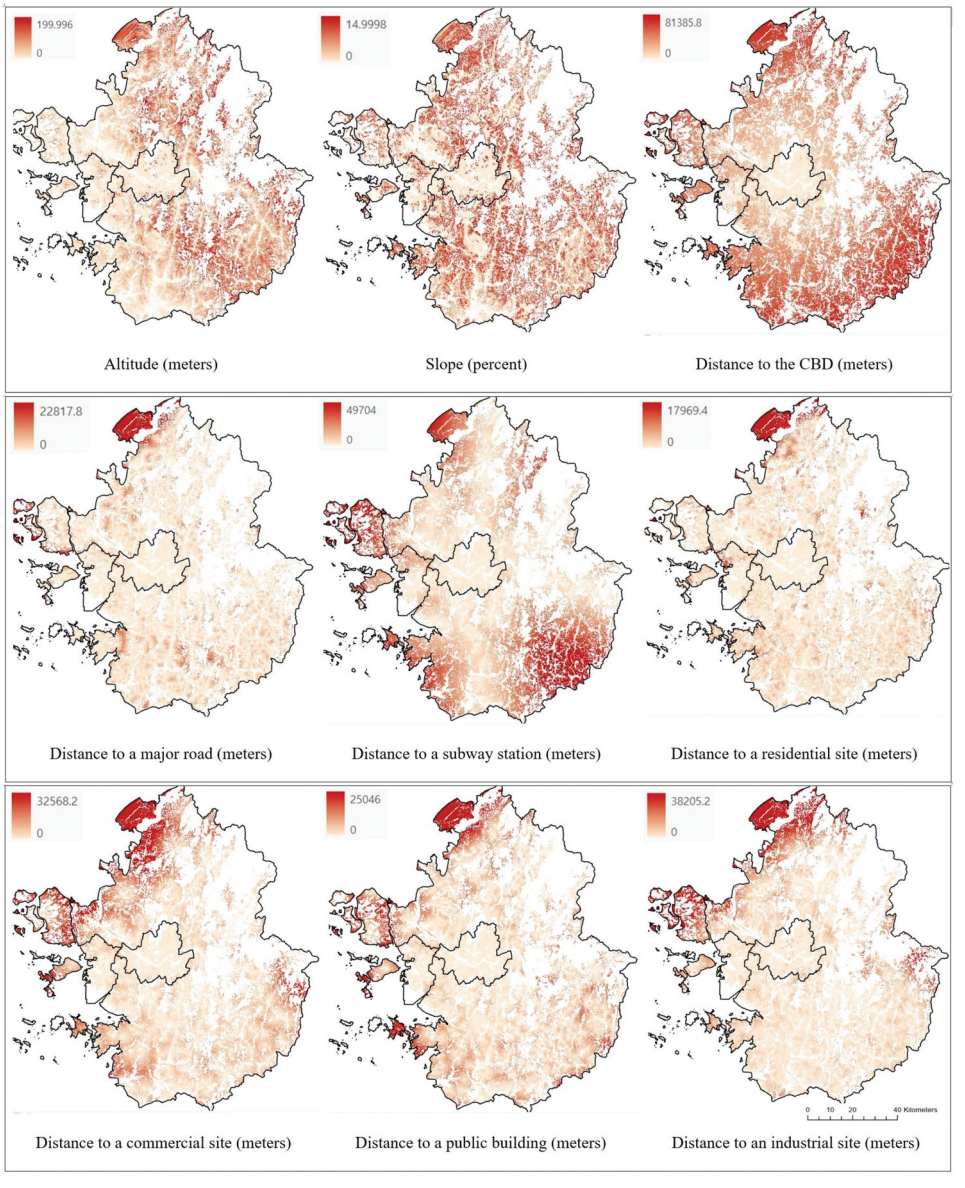

Figure 2. Raster maps for the input variables.

implies its stronger influence in predicting land use conversion in Seoul. Unlike a sensitivity analysis that evaluates changes in an outcome variable responding to the change in a predictor variable, the GBDT model estimates the contribution of each predictor variable on land use conversion simultaneously while taking the possible relationships among predictor variables into account (Zhang and Haghani 2015, Ding *et al*. 2016).

Figure 3 illustrates the relative contribution of predictor variables and its ranking in predicting land use conversion. Distance to the residential site is the most important predictor variable affecting land use conversion to either residential or nonresidential use with a 30.4% contribution. Distances to the industrial and public site have significant

Table 2. Performance of three models (test dataset).

	Observed \ Predicted	Vacant	Residential	Nonresidential	Recall
Gradient Boosting Decision Tree	Vacant	**39,996**	1037	304	0.97
	Residential	1587	**1990**	137	0.54
	Nonresidential	665	103	**880**	0.53
	Precision	0.95	0.64	0.67	
	f1-score	0.96	0.58	0.59	-
	Accuracy		91.79%		
	Kappa Score		0.574		
Artificial Neural Network	Vacant	**36,772**	3162	1403	0.89
	Residential	1309	**2194**	211	0.59
	Nonresidential	502	161	**985**	0.6
	Precision	0.95	0.40	0.38	
	f1-score	0.92	0.48	0.46	-
	Accuracy		89.71%		
	Kappa Score		0.438		
Random Forest	Vacant	**38,889**	1625	823	0.94
	Residential	1413	**2101**	200	0.57
	Nonresidential	525	110	**1013**	0.61
	Precision	0.95	0.55	0.50	
	f1-score	0.95	0.56	0.55	-
	Accuracy		89.94%		
	Kappa Score		0.539		

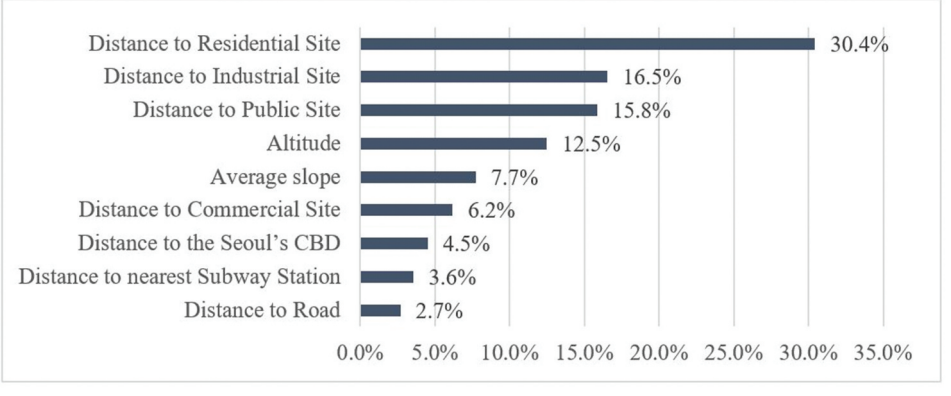

Figure 3. Relative contribution of predictor variables.

impacts on land use conversion with 16.5–15.8% contributions, respectively, while altitude has a relatively modest impact with 12.5%. It is also interesting to find that slope, distance to a commercial site and to the CBD, as well as distance to a subway station and a road, have relatively minor influences. The reason for the relatively low importance of distance variables from subway stations and major roads seems to be because these variables have a high correlation with existing residential and industrial sites, so the feature importance of these variables is likely to be shared with the correlated features.

3.3. Nonlinear effects of predictor variables from the GBDT model

Another interpretive advantage of the DT-based ensemble method is its capability to measure nonlinear effects of predictor variables on the target variable. Unlike the traditional sensitivity analysis based on linear regression, the DT-based ensemble method can

illustrate the nonlinear association between predictor variables and the target variable using partial dependence plots since there is no linearity assumption (Hastie *et al*. 2009, Saha *et al*. 2015, Ding *et al*. 2016). A partial dependence plot is a useful tool to illustrate the marginal effect of a predictor variable on the target variable given the average effects of all other predictor variables.

Figure 4 shows the marginal effects of predictor variables on residential land use conversion in the GBDT model. Distance to the residential site has diminishing marginal impacts on residential land use conversion, indicating that new residential development sites are likely to be located closer to the existing residential area, but the development likelihood decreases as the distance to the residential site increases. Distances to the industrial site and the CBD have positive marginal effects, indicating that residential development is likely to take place farther away from industrial sites and the CBD. This implies population suburbanization and new residential development far from industrial sites. However, distances to a public site and a road have negative marginal impacts, implying that new residential developments are likely to be located near public sites and roads.

Figure 5 shows the marginal effects of predictor variables on nonresidential land use conversion. The nonresidential conversion has increasing marginal effects within 600 m of the existing residential site, implying that nonresidential conversions are likely to take

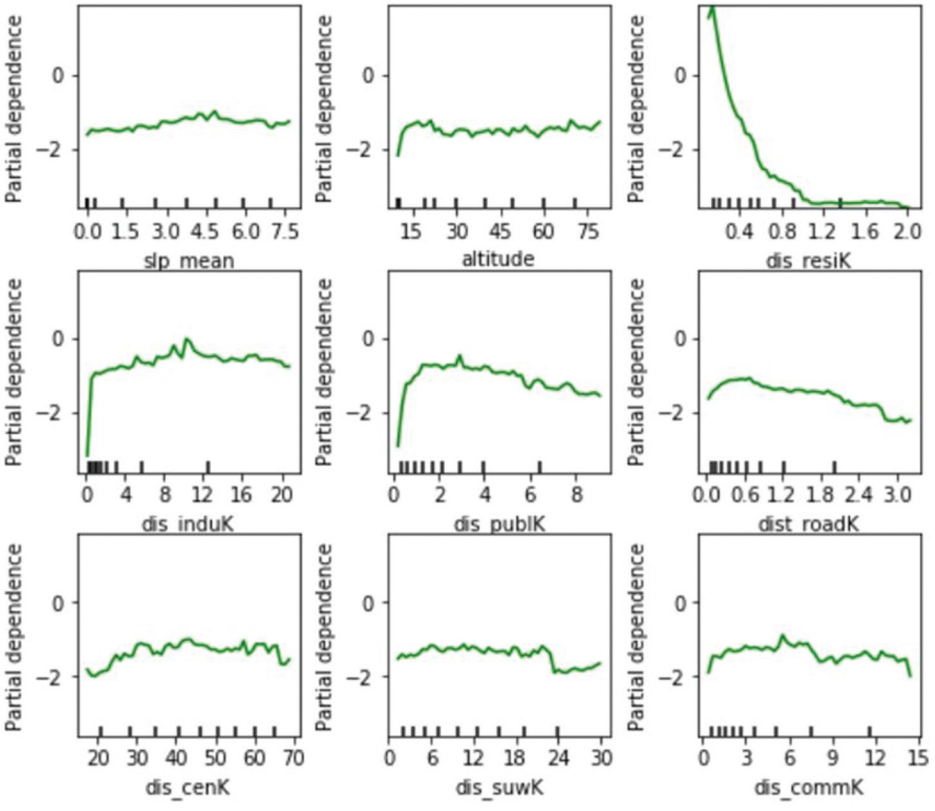

Figure 4. Marginal effects of predictor variables on residential land use conversion.

place within walking distance from the residential sites. On the other hand, distances to industrial, public, and commercial sites have diminishing marginal impacts on nonresidential land use conversion, implying that the likelihood for nonresidential development increases near these sites.

3.4. Comparisons between actual land use change and GBDT model prediction

Figure 6 presents the comparisons between actual LUC and GBDT model prediction for residential and nonresidential land use conversion. The probability matrix derived from the GBDT model is used for the prediction of urban LUCs. True positives indicate actual residential or nonresidential change and correct prediction by the model, while false positives refer to land use conversions that are incorrectly predicted by the model. The number of true and false positives are presented in the confusion matrix shown in Table 2. The GBDT model accurately predicted the residential development in the northeast (Gimpo), north (Yangju and Pocheon), and south (Yongin and Anseong) regions, while false positives were concentrated in the southeast (Pyungteak) and southwest (Icheon and Yeoju) regions. True positives for nonresidential development were located in the north (Yangju and Pocheon) and mid-south (Yongin) regions, whereas false positives were

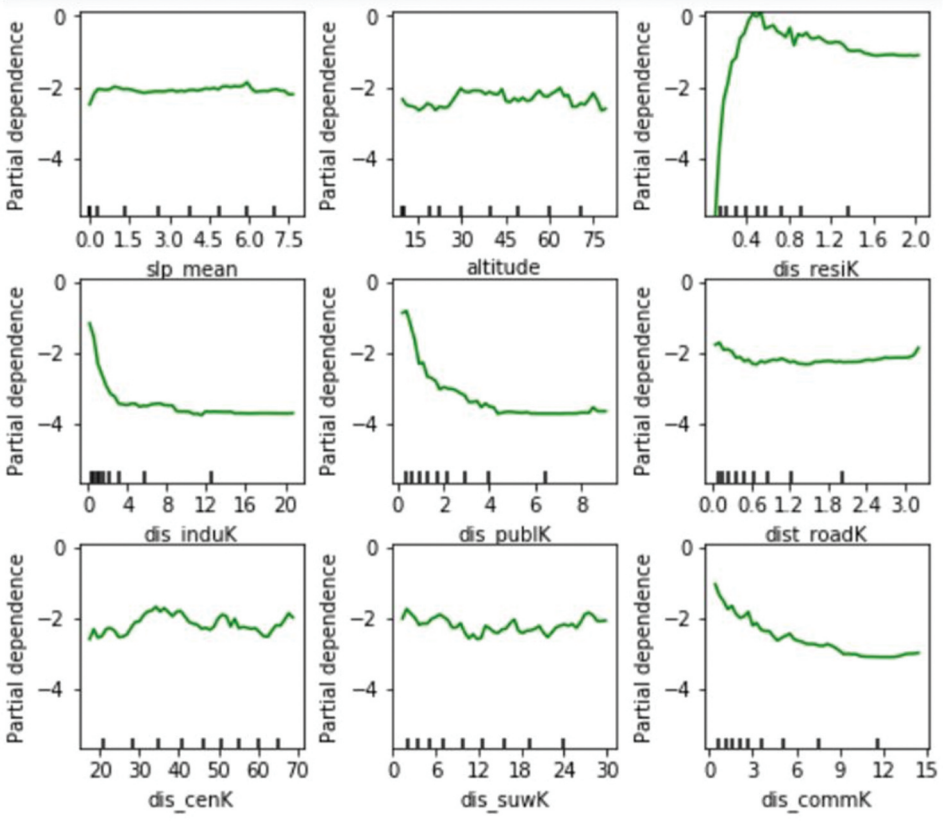

Figure 5. Marginal effects of predictor variables on nonresidential land use conversion.

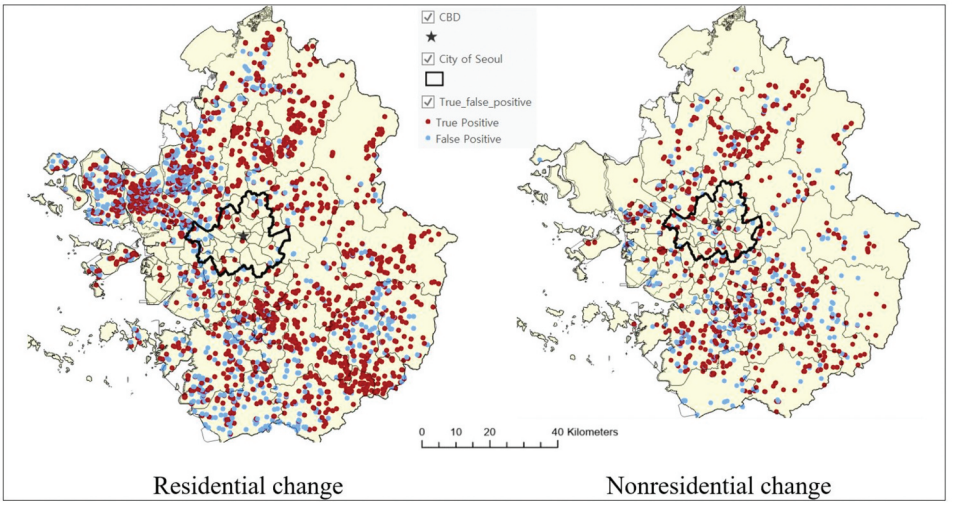

Figure 6. Comparison between observed LUCs and GBDT model prediction.

concentrated in the northeast (Incheon and Goyang) and south (Hwasung, Pyungteak, Anseong, and Icheon) regions. The concentration of false positives in the suburban areas seems to be highly related to the urbanization process that is happening outside of the central city (City of Seoul), which is weakly explained by predictor variables used in this study such as through proximity to existing development sites, accessibility to subway and major roads, and distance from the CBD.

4. Conclusions

In this study, the predictive and interpretive performance of three ML methods (GBDT, ANNs, and RF) was compared in modeling urban LUCs in the SMA. The results of this study showed that GBDT and RF have higher predictive power than ANN, indicating that DT-based ensemble methods are an effective technique for LUC prediction. This finding is consistent with those of Du *et al*. (2018), which reported the outperformance of a tree-based model over ANN for LUC modeling in the Greater Tokyo Area.

Along with the outstanding predictive performance, the DT-based ensemble models provide insights for understanding what factors drive LUCs in complex urban dynamics. The results of the GBDT model indicate that proximity to a residential site has the highest contribution to LUC, with 30.4% of the relative importance. Other significant predictor variables were proximity to industrial and public sites, having 32.3% of relative importance together. This finding is similar to that of other previous studies that reported the impact of the land use neighborhood characteristics on LUC (Van Vliet *et al*. 2013, Du *et al*. 2018).

GBDT's capability to measure the nonlinear marginal influence of a single predictor variable, while controlling for the effects of other predictor variables, further explains the relationship between driving factors and target variables in detail. The findings of this study indicate that marginal impacts of driving factors vary depending on land use conversion type. For example, proximity to the residential site had

diminishing marginal impacts on residential land use conversion but increasing marginal impacts on nonresidential land use conversion. This finding implies that new residential development was likely to be adjacent to the existing residential sites, but nonresidential development takes place at a distance (about 600 m) from existing residential sites. Furthermore, new residential development was likely to take place far from existing industrial sites while new nonresidential development was located closer to the industrial, public, and commercial sites.

Looking at the land use conversion from the perspective of urban spatial structure, the distance to the CBD had increasing marginal effects on residential land use conversion while no significant pattern was found for nonresidential land use conversion, indicating that Seoul has experienced more population suburbanization than employment dispersion. This argument can be supported by two land use policies initiated by the Korean government: 1) large-scale suburban new town development projects to construct approximately 300,000 new housing units in five new towns during the study period, located approximately 20–30 km from the CBD, and 2) the greenbelt regulation, a 10- to 15-km-wide donut-shaped green space surrounding the City of Seoul beginning 15 to 20 km from the CBD, thereby prohibiting new development. The new town projects have contributed to suburban leapfrog residential development, while the greenbelt policy plays a role in containing nonresidential development within the central city, pursuing agglomeration economies within the central city.

Due to its relatively high predictive performance and detailed interpretive power, the approach proposed in this study can be used as a comprehensive decision support tool for establishing effective zoning legislation and land use policy in metropolitan areas. Unlike results from land cover modeling, which analyzes the physical and environmental changes of land surface, the results of this study can be utilized to mitigate the negative consequences of urban sprawl; this is characterized as unlimited suburban expansion of low-density development and large-scale conversion of open space and environmentally sensitive lands for human use. In addition to the utility of the model in the public sector, it can be used to support specific time–space investment decisions by private developers and investors. The ML methods proposed in this study can be applied to other cities with two requirements. First, it is necessary to obtain land use information at two time points for tracking land use changes and information on predictor variables affecting land use changes. Second, once the input dataset is established, hyperparameter tuning should be performed to find the optimal ML models for the study area, because the hyperparameter values are data dependent.

Several limitations of the current approach can be improved in future studies. The predictive power of the model could be improved by including more predictor variables, such as socioeconomic and individual-related variables than the current approach, by more detailed land use classification for the target variable, and by obtaining data about long-term LUCs. Model reliability can be strengthened by more empirical applications of the model in other countries and metropolitan areas to elucidate the relationship between predictor and target variables in the LUC modeling.

Notes

1. The SMA has been selected as the case study area because the SMA, the capital region of South Korea, has experienced massive complex LUCs in the past few decades, making it a good test bed for building ML-based LUC modeling. The SMA has experienced rapid LUC over the last several decades via suburbanization and decentralization of population and employment. The SMA is one of the largest and densest cities in the world in terms of its population, which increased five-fold from 5.2 million in 1960 to 25.3 million in 2015 (Korean Statistical Information Service: http://kosis.kr).
2. LIMS is the parcel boundary map provided by National Spatial Data Infrastructure Portal. The data are downloadable at http://data.nsdi.go.kr/dataset/12771.
3. This study conducted a Bayesian optimization process for the selected ML models with the train dataset after the dataset was randomly split 70%:30% into train and test datasets, respectively. The optimal hyperparameters for the GBDT were learning rate = 0.1, number of iterations = 1415, maximum depth = 8, minimum sample split = 10, minimum sample leaf = 10, and subsample = 0.9, while those for the ANN were learning rate = 0.000392, num_dense_layers = 4, num_dense_nodes = 512, dropout = 0.2, epoch = 200, batch_-size = 126, and activation = 'relu.' The optimal hyperparameters for the RF were number of iterations = 2000, maximum depth = 40, minimum sample split = 10, minimum sample leaf = 10, and max_features = 8.

Data and codes availability statement

The codes and data that support the findings of the present study are available on Figshare at (10.6084/m9.figshare.12749813).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Myung-Jin Jun is professor in the Department of Urban Planning and Real Estate at Chung-Ang University, South Korea. His research interests include urban modelling, urban big data analysis, and machine learning, and their applications in urban sciences.

References

Aburas, M.M., *et al*., 2016. The simulation and prediction of spatio-temporal urban growth trends using cellular automata models: a review. *International Journal of Applied Earth Observation and Geoinformation*, 52, 380–389. doi:10.1016/j.jag.2016.07.007

Ben-Gal, I., *et al*., 2014. Efficient construction of decision trees by the dual information distance method. *Quality Technology & Quantitative Management*, 11 (1), 133–147. doi:10.1080/16843703.2014.11673330

Breiman, L., *et al*., 1984. *Classification and regression trees*. Boca Raton, FL: CRC Press.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24 (2), 123–140. doi:10.1007/BF00058655

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

Brochu, E., Cora, V.M., and Freitas, N.D. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599.

Caruana, R. and Niculescu-Mizil, A., 2006. *An empirical comparison of supervised learning algorithms | machine learning | support vector machine*. Available from: https://www.scribd.com/document/113006633/2006-An-Empirical-Comparison-of-Supervised-Learning-Algorithms# .

Chawla, N.V., *et al*., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 341–378. doi:10.1613/jair.953

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), 37–46. doi:10.1177/001316446002000104

Dai, E., *et al*., 2005. Modeling change-pattern-value dynamics on land-use: an integrated GIS and artificial neural networks approach. *Environmental Assessment*, 36 (4), 576–591.

Ding, C., *et al*., 2016. A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data. *Transportation Research Part C: Emerging Technologies*, 72, 225–238. doi:10.1016/j.trc.2016.09.016

Du, G., *et al*., 2018. A comparative approach to modelling multiple urban land use changes using tree-based methods and cellular automata: the case of Greater Tokyo Area. *International Journal of Geographical Information Science*, 32 (4), 757–782. doi:10.1080/13658816.2017.1410550

Elith, J., Leathwick, J.R., and Hastie, T., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77 (4), 802–813. doi:10.1111/j.1365-2656.2008.01390.x

Etemad-Shahidi, A. and Mahjoobi, J., 2009. Comparison between M5′ model tree and neural networks for prediction of significant wave height in Lake Superior. *Ocean Engineering*, 36 (15–16), 1175–1181. doi:10.1016/j.oceaneng.2009.08.008

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29 (5), 1189–1232. doi:10.1214/aos/1013203451

Friedman, J.H., 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38 (4), 367–378. doi:10.1016/S0167-9473(01)00065-2

Friedman, J.H. and Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 22 (9), 1365–1381. doi:10.1002/sim.1501

Georganos, S., *et al*., 2018. Very high resolution object-based land use–land cover urban classification using extreme gradient boosting. *IEEE Geoscience And Remote Sensing Letters*, 15 (4), 607–611. doi:10.1109/LGRS.2018.2803259

Groeneveld, J., *et al*., 2017. Theoretical foundations of human decision-making in agent-based land use models – a review. *Environmental Modelling and Software*, 87, 39–48. doi:10.1016/j.envsoft.2016.10.008

Guan, Q., Wang, L., and Clarke, K.C., 2005. An artificial-neural-network-based, constrained CA model for simulating urban growth. *Cartography and Geographic Information Science*, 32 (4), 369–380. doi:10.1559/152304005775194746

Hastie, T., Tibshirani, R., and Friedman, J., 2009. *The elements of statistical learning*. 2nd. New York: Springer.

Hu, Z. and Lo, C.P., 2007. Modeling urban growth in Atlanta using logistic regression. *Computers, Environment and Urban Systems*, 31 (6), 667–688. doi:10.1016/j.compenvurbsys.2006.11.001

James, G., *et al*., 2013. *An introduction to statistical learning with applications in R*. New York: Springer.

Kamusoko, C. and Gamba, J., 2015. Simulating urban growth using a Random Forest-Cellular Automata (RF-CA) model. *ISPRS International Journal of Geo-Information*, 4 (2), 447–470. doi:10.3390/ijgi4020447

Kavzoglu, T. and Mather, P.M., 2003. The use of backpropagating artificial neural networks in land cover classification. *International Journal of Remote Sensing*, 24 (23), 4907–4938. doi:10.1080/0143116031000114851

Krenker, A., Bester, J., and Kos, A., 2011. Introduction to the artificial neural networks. *In*: K. Suzuki, ed. *Artificial neural networks: methodological advances and biomedical applications*. PLACE: Publisher, 1–18.

Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159–174. doi:10.2307/2529310

Larivière, B. and den Poel, D.V., 2005. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29 (2), 472–484. doi:10.1016/j.eswa.2005.04.043

Lee, J., Newman, G., and Park, T., 2018. A comparison of vacancy dynamics between growing and shrinking cities using land transformation model. *Sustainability*, 10 (5), 1–17.

Li, X., Liu, X., and Gong, P., 2015. Integrating ensemble-urban cellular automata model with an uncertainty map to improve the performance of a single model. *International Journal of Geographical Information Science*, 29 (5), 762–785. doi:10.1080/13658816.2014.997237

Li, X., Liu, X., and Yu, L., 2014. A systematic sensitivity analysis of constrained cellular automata model for urban growth simulation based on different transition rules. *International Journal of Geographical Information Science*, 28 (7), 1317–1335. doi:10.1080/13658816.2014.883079

Li, X. and Yeh, A.G.O., 2002. Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science*, 16 (4), 323–343. doi:10.1080/13658810210137004

Lopez, E., *et al*., 2001. Predicting land-cover and land-use change in the urban fringe: a case in Morelia city, Mexico. *Landscape and Urban Planning*, 55 (4), 271–285. doi:10.1016/S0169-2046(01)00160-8

Opitz, D.W. and Maclin, R., 1999. Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research*, 11, 169–198. doi:10.1613/jair.614

Park, S., *et al*., 2011. Prediction and comparison of urban growth by land suitability index mapping using GIS and RS in South Korea. *Landscape and Urban Planning*, 99 (2), 104–114. doi:10.1016/j.landurbplan.2010.09.001

Pijanowski, B.C., *et al.*, 2000. A land transformation model for the Saginaw Bay Watershed. *In*: J. Sanderson and L. Harris, eds. *Landscape ecology: a top down approach*. Boca Raton: Lewis Publishers, 183–198.

Pijanowski, B.C., *et al*., 2002. Using neural networks and GIS to forecast land use changes: a land transformation model. *Computers, Environment and Urban Systems*, 26 (6), 553–575. doi:10.1016/S0198-9715(01)00015-1

Pijanowski, B.C., *et al*., 2005. 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*, 19 (2), 197–215. doi:10.1080/13658810410001713416

Quinlan, J.R., *et al*., 1996. Bagging, boosting, and C4. 5. *In*: *Proceedings of the thirteenth national conference on artificial intelligence*. Cambridge, MA: AAAI Press/MIT Press, 725–730.

Rahimi, A., 2016. A methodological approach to urban land-use change modeling using infill development pattern—a case study in Tabriz, Iran. *Ecological Processes*, 5 (1), 1–15. doi:10.1186/s13717-016-0044-6

Reed, R.D. and Marks, R.J., 1998. *Neural smithing: supervised learning in feedforward artificial neural networks*. Cambridge, MA: MIT Press.

Rumelhart, D.E., Mcclelland, J.L., and Group, P.D.P., 1986. *Parallel distributed processing: explorations in the microstructure of cognition: foundations*. London: The MIT Press.

Saha, D., Alluri, P., and Gan, A., 2015. Prioritizing highway safety manual's crash prediction variables using boosted regression trees. *Accident Analysis and Prevention*, 79, 133–144. doi:10.1016/j.aap.2015.03.011

Solomatine, D.P. and Xue, Y., 2004. M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai river in China. *Journal of Hydrologic Engineering*, 9 (6), 491–501.

Sun, R., *et al*., 2020. A gradient boosting decision tree based GPS signal reception classification algorithm. *Applied Soft Computing Journal*, 86 (105942), 1–12. doi:10.1016/j.asoc.2019.105942

Svetnik, V., *et al*., 2003. Random Forest: a Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences*, 43 (6), 1947–1958. doi:10.1021/ci034160g

Tayyebi, A. and Pijanowski, B.C., 2014. 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*, 28, 102–116. doi:10.1016/j.jag.2013.11.008

Van Vliet, J., *et al*., 2013. Measuring the neighbourhood effect to calibrate land use models. *Computers, Environment and Urban Systems*, 41, 55–64. doi:10.1016/j.compenvurbsys.2013.03.006

Van Vliet, J., White, R., and Dragicevic, S., 2009. Modeling urban growth using a variable grid cellular automaton. *Computers, Environment and Urban Systems*, 33 (1), 35–43. doi:10.1016/j.compenvurbsys.2008.06.006

Yao, Y., *et al*., 2017. Investigation on the expansion of urban construction land use based on the CART-CA model. *ISPRS International Journal of Geo-Information*, 6 (5), 149. doi:10.3390/ijgi6050149

Zhang, Y. and Haghani, A., 2015. A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308–324. doi:10.1016/j.trc.2015.02.019