The Application of Random Forest Prediction in Developing a Systematic Land Parcel Value in the Urban Area

Deviantari, U. W.,[1,2] Aditya, T.,[1*] Djojomartono, P. N.[1] and Mulyadi[3]
[1]Geomatics Engineering Study Programme, Department of Geodetic Engineering, Universitas Gadjah Mada, Yogyakarta 55284, Indonesia, E-mail: udianawahyudeviantari@mail.ugm.ac.id, triasaditya@ugm.ac.id* prinug@ugm.ac.id
[2]Department of Geomatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia E-mail: udianadeviantari@its.ac.id
[3]Information and Data Center of The Ministry of Agrarian Affairs and Spatial Planning/National Land Agency (ATR/BPN), Jakarta, Indonesia, E-mail: mulyadi.katio@gmail.com
Corresponding Author

DOI: https://doi.org/10.52939/ijg.v21i7.4319

Abstract
Land administration services, especially in urban areas, require a complete and accurate parcel-based land valuation for supporting fair and reliable taxation. For fulfilling that purpose, the utilization of machine learning techniques, instead of full ground survey, to predict land value for each parcel is tested in this work. Although linear regression approach is widely used, the technique is not relevant as not all variables are linearly related to land values. Over the past few years, random forest (RF) models have been applied and seen to be promising for land-administration related analysis. However, as the prediction was typically generating land values for zones or blocks, which are not parcel-based prediction, thus the produced results may not be operational for local land and municipality offices to be used especially for determining land values in property transactions. This study adopted RF to predict parcel-wise land values in 15 administrative districts of Surabaya City and are validated using real transaction data. The independent variables used in this study are as follows: the zoning score, road width, parcels' distance from the central business districts (CBD), schools, markets, arterial roads, urban collector roads, points of interest (POI), and hospitals. The prediction model was trained using assessment values done by Indonesian Society of Land Appraisers. The prediction model produces the mean absolute error (MAE), mean absolute percentage error (MAPE), R^2-score, Coefficient of Variation (COV), and Prediction relative difference (PRD) respectively on the testing data are 39.31 USD.; 6.99%; 0.96; 4.18%; and 1.03. Then, the trained model was validated using official transactional dataset and determined its MAPE and R^2-score. The values were 16.76% and 0.72, respectively. It is concluded that the model performed effectively. Hence, the findings provide a potential solution for delivering completeness and certainty in land valuation for land registration services.

Keywords: Land Value, Machine Learning, Prediction, Random Forest, Urban Area

1. Introduction

Land administration services are essential infrastructure to support a sustainable development in a country [1]. The services include taxation, certification, and valuation of a specific parcel of land which may help the government in gaining optimum revenue [2] and [3]. For instance, [4] explained that the land valuation can contribute to increase national revenues through taxation programs but is hindered by its poor implementation. In addition [5] and [6] stated that land taxes are affecting economic growth of a country. They observed that the taxes can be utilized for expenses to support some government programs in various sectors, including mitigation, emission, and consumption of goods and services. Valuation and taxation are essential functions that make other land administration services work properly. Those indicates the importance of land administration services to a country.

The Indonesian Ministry of Agrarian Affairs and Spatial Plan, well known as ATR/BPN, plays as a main role for supporting land administration services in the country. The ministry's first priority is in providing legal certainty to all land parcels located out of forest areas in the country through land registration programs. The other land administration services that the ministry is being mandated is in managing and regulating valuation of land parcels and spatial plans at the national, province and municipality levels. Meanwhile, the function of taxation is mandated to each and every municipality in the country as sources for local government revenues. The logical framework for generating land and property taxes is that the generated tax map should use maps of land ownerships and land valuation as its reference. Using such an approach, the completeness and logical consistency of the local property (of land and building) taxes can be determined well for each land parcel in the cities and municipalities. However, since 1960s Indonesia experienced mostly sporadic land title adjudication and mapping that slow down the country progress in achieving a complete and reliable cadastral map for the whole country [7]. In addition to that, land management paradigm that requires integration of data and information of land registration, valuation, taxes and plans is not well implemented, as the management of land rights, values and plans are mostly still fragmented among local and national government agencies. As a result, individual land value and tax which is based upon un updated parcel map is still lacking.

By 2024, the ministry has mapped and certified almost 100 million out of targeted 126 million land parcels, under a project namely a Complete Systematic Land Registration in Indonesia (abbreviated as PTSL) [8]. Once the project is completed, an accurate individual land valuation is critical to be produced. A land valuation can be defined as an expert assessment to determine certified land parcels' value, taking into consideration diverse information and characteristics including the market value [5][9] and [10]. Most of the market value data are collected and provided in a typical zone-based uniform value called the mass appraisal technique [11] and [12]. However, due to data generalization in a mass appraisal, the land taxes may not be justified correctly for each land parcel. That implies the urgencies in having a complete land market values for individual parcel, i.e., individual appraisal approach [13].

Direct measurement in field survey for individual land parcel value would be very costly and timely. Therefore, many researchers preferred to adopt several machine learning techniques which require much lesser surveyed land parcels' value data to predict values in the other unsurvey parcels [14]. Current practices in Indonesia have used mostly a linear approach to gather the relationship between the dependent (e.g. land value) and independent variables (e.g. land market value and adjacent properties). The linear approach includes simple linear or ordinary least square (OLS) regression with a locally modified approach such as geographically weighted regression [15] and [16]. However, the output of the linear approach predictions does not represent characteristics of the data which then prone to bias. For example, [17][18] and [19] used a linear regression approach to predict real estate prices with suboptimal accuracy results. Those studies reported that the variables applied had a determination coefficient over the independent variables of less than 40% [18]. Therefore, the prediction results still have a high bias.

Recently, a number of non-linear regression approaches has been utilized and developed for land valuation purpose. The approaches have experienced significant advancements in the field of land valuation. Among many non-linear approaches, support vector machines (SVM) [20], gradient boosting (GB) [21], artificial neural networks (ANN) [22], and random forests (RF) [23], are frequently employed. SVM minimizes the margin and generalization error and can solve small sample, non-linear, and high-dimensional problems. However, the margin of error built into the auxiliary vector can lead to prediction bias [24]. ANN uses neuron connections to extract the relationship characteristics of dependent and independent variables in complex land valuation systems. However, ANNs are highly dependent on the neural network architecture built and are sensitive to the architecture [22]. The GB technique attempts to perform an iterative technique to minimise the residuals in the weak learner, often using a weak learner decision tree (DT) which is then called a gradient boosting decision tree. With its iterative nature, it allows for weak generalization of unseen data [25]. The SVM method and gradient boosting may provide high-quality predictions but are susceptible to outliers. The appearance of outliers in the SVM method reduces model quality [26] and [27]. Meanwhile, outliers in the gradient boosting process produce overfitting due to the outlier values being iterated until the error value is minimal [28]. Whilst, [22] proved that ANN can present an accurate prediction; however, the main disadvantage is its inability to explicitly identify the relationship between dependent and independent variables, hence the term "black box" [29].

RF is considered good enough to build prediction models with structured data and a limited amount of data through a nonlinear approach [30] and [31]. The RF approach is ideal for mass appraisal applications because it has various advantages over other algorithms, such as being resistant to outliers, performing well with missing values, and categorical variables with multiple categories [32]. Based on research conducted by [33] in Ljubljana, Slovenia, RF can better detect variability in apartment values and more effective prediction of land values than multiple linear regression. Results in other research by [23] in Gangnam, Korea, showed that the RF method can overcome complexity or nonlinearity in the property market compared to OLS.

However, using RF models for land value prediction in Indonesia remains uncommon. Systematic land parcel valuations by using random forest algorithms have not been done before in Indonesia, especially applied into the official land parcel map. This research uses RF models for generating land values of land parcels of the city of Surabaya, the second biggest city in Indonesia. This paper aims to validate whether the RF technique provides a reliable result by using inputs of a complete land parcel map of a city and data sampling from professional valuers and validated by official real estate transactions. This study presents an in-depth analysis of the factors that influence land value. The paper is structured as follows: Section 2 describes the research methodology, Section 3 presents the results, and Section 4 provides the discussion and Section 5 draws the conclusion.

2. Research Methodology

2.1 Research Area

Surabaya City, the second largest city in Indonesia, is a capital city of Eastern Java Province [34]. The city, which has an area of approximately 326.36 km^2 is divided into 31 subdistricts and 154 wards. Broadly speaking, it also serves as the service center for Eastern Indonesia's activities. The city is designated as a trade and service city at national and international transportation hubs, providing opportunities to increase its role as a national activity center, as mandated by the central government. As seen in Figure 1, Surabaya City has an urban system development strategy developed through a city spatial structure plan consisting of low-, medium-, and high-density housing, class I and II trade and services, public facility areas, and protected and border areas [35].

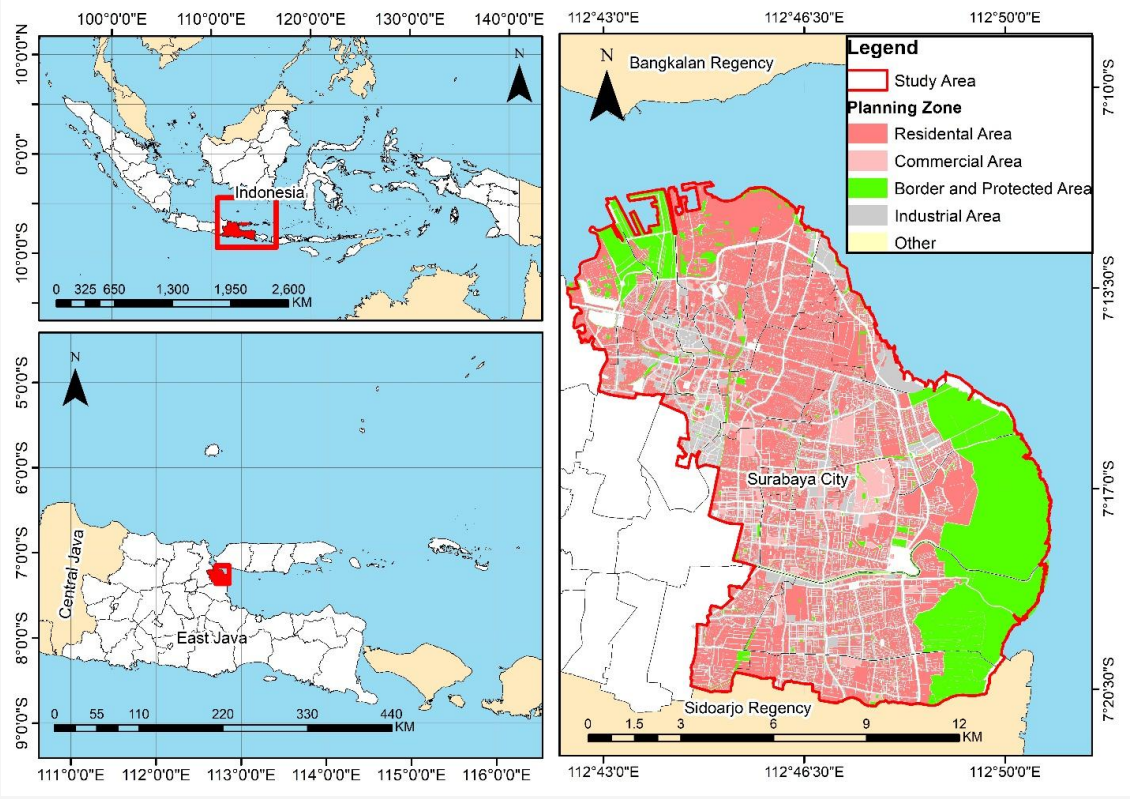


Figure 1: Surabaya city, East Java Province, Indonesia (adapted from https://tanahair.indonesia.go.id and Surabaya City Land and Spatial Planning Office)

2.2 Method

This research implemented the RF model to predict land value from various input variables including zoning score, road-related, and interest-point-related features. The determination of input variables followed the study of [36][37] and [38]. The features of the road are width of the road, distance to arterial road, and distance to collector roads. While for the interest point, the features are distance to markets, distance to central business districts (CBD), distance to point of interests, distance to schools, and distance to hospitals. POI are locations that are the attraction of a particular area, such as tourist attractions, cultural heritage, and landmarks. Easy access to utility services and commercial centers is a great advantage and can lead to high land values [9]. This study encompasses four regions, that are Eastern, Northern, Center, and Southern of Surabaya. That includes 15 subdistricts consisting of Genteng, Bubutan, Simokerto, Semampir, Kenjeran, Bulak, Tambaksari, Gubeng, Rungkut, Tenggilis Mejoyo, Gunung Anyar, Sukolilo, Mulyorejo, Pabean Cantikan, and Krembangan. Figure 2 shows the flowchart of this research which has five main stages: (1) Data collection and preparation; (2) Data pre-processing; (3) RF modelling and hyperparameter adjustment; and (4) Accuracy assessment. The details of each step can be explained in the sub-sections below.

2.3 Data Collection and Preparation

The input of the RF model is zoning score, road-related, and interest-point-related features while the output of the RF model in this study is a parcel-based land valuation (LV^{parcel}). Several sub-sub chapters below describe the collection and pre-processing for the datasets of the input features. Table 1 summarizes the collected data.

2.3.1 Parcel-based land value

The samples of parcel-based land value (locally known as NBT or *Nilai Bidang Tanah*) were obtained from the Indonesian Society of Appraisers (ISA) data totaled 2,836 parcel samples. The land valuation used as training was assessed by an ISA-certified valuer. The technicalities of valuation have been regulated by ISA as the standardisation of property valuation in Indonesia. There are 3 approaches used in the valuation process in accordance with ISA standards, namely the market, income, and cost approaches. The market approach considers sales of similar or replacement properties by comparing market prices. The income approach considers the income and costs associated with the property being valued and estimates the value through capitalization. The cost approach establishes the value of the property by estimating the cost of acquiring the land and the replacement cost of new development on it with comparable utility.

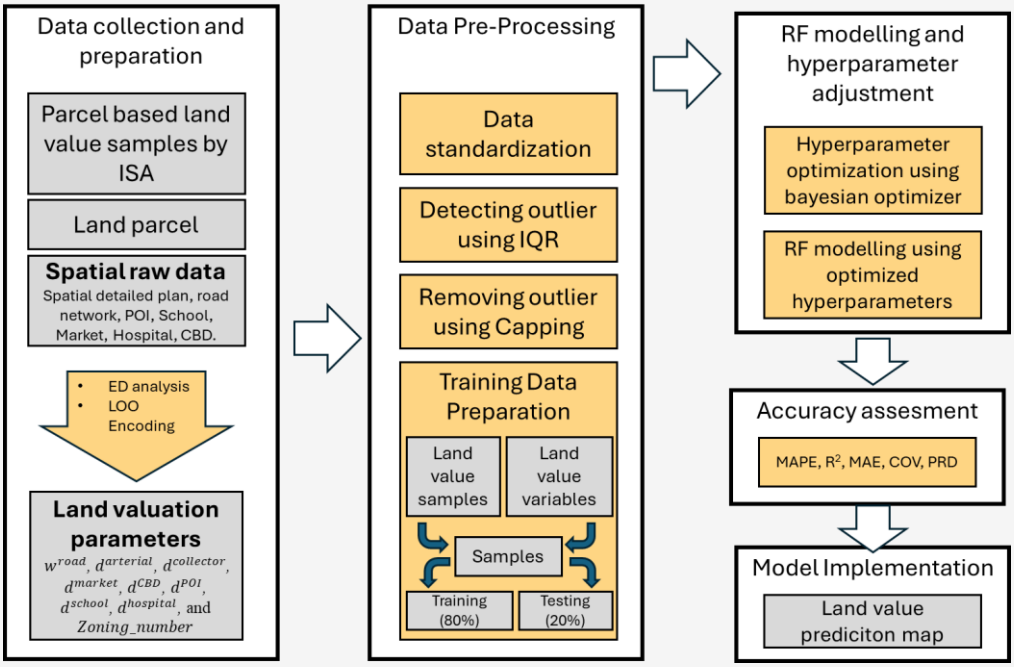
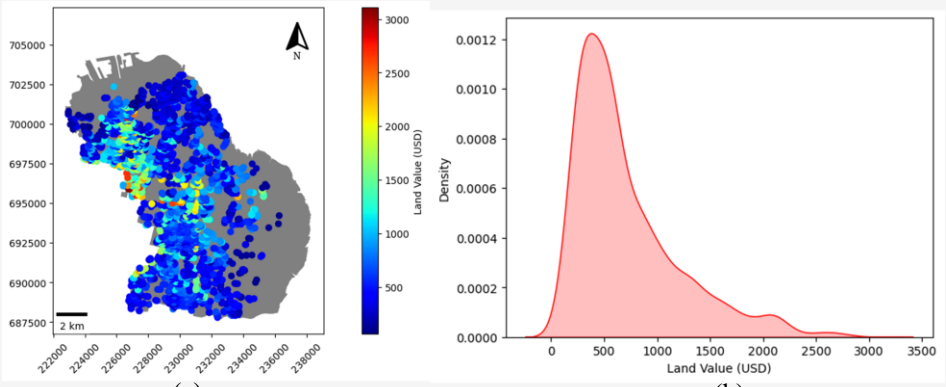


Figure 2: Land parcel valuation prediction study workflow

Table 1: Data sources and research-derived variables

Data	Source of Data	Scale	Variable	Influence on land value
Sample of land value zones	Land Office Surabaya II	1:2,500	Training dataset	-
Land parcel value	Indonesian Society of Appraisers (ISA)	1:2,500	Training dataset	-
Land parcel data	Bhumi.atr.bpn.go.id	1:2,500	Geometric representation of land parcels	-
Spatial city details plan	Surabaya City Land and Spatial Planning Office	1:5,000	Spatial plan zone (*Zone_number*)	Spatial zone regulations represent responsibilities and restrictions on space utilization on a land parcel. Commercial areas have higher land values [39].
Road network	Surabaya City Public Works, Housing, and Settlement Area Office	1:5,000	Distance to the arterial road ($d^{arterial}$) Distance to the urban collector road ($d^{collector}$) Road width (w^{road})	Access to the road network can affect land value. The higher the accessibility of a land parcel, the higher the land value [39].
Point of interest distribution	Surabaya City Land and Spatial Planning Office	1:5,000	Distance to points of interest (d^{POI})	Accessibility of points of interest affects land value because the closer to points of interest, the easier accessibility to entertainment and recreation [40].
School distribution points	Surabaya City Land and Spatial Planning Office	1:5,000	Distance to points of the school (d^{school})	The higher the accessibility to education facilities, the higher the land value [41].
Distribution points of traditional markets	Surabaya City Land and Spatial Planning Office	1:5,000	Distance to the points of traditional market (d^{market})	The market as a place for economic interaction plays a vital role in population mobility, whereas the closer to the market, the higher the land value [41].
Distribution point of hospital location	Surabaya City Land and Spatial Planning Office	1:5,000	Distance to the points of hospital ($d^{hospital}$)	Health facilities are a basic need of the community, therefore the distance to health facilities such as hospitals has an influence on land value. The closer to the hospital, the higher the land value [40].
Distribution point of central business districts	Surabaya City Land and Spatial Planning Office	1:5,000	Distance to the points of CBD (d^{CBD})	The CBD is the driver of community activities. The closer to the CBD, the higher the value of a land parcel [40].

The approach chosen depends on considerations such as the value basis and purpose of the valuation, the availability of information and data, and the methods applied by players in the relevant market [42]. The sample points are spread throughout the study area and already represent the whole area geographically.



(a) (b)

Figure 3: Training point for prediction modelling;
(a) Distribution of land value sample points and (b) density plot of land values

In terms of a representative sample size, the International Association of Assessing Officers (IAAO) recommends the sample calculation by Cochran (1997) as shown in Equations 1 and 2 [43]. Where n_0 is initial sample size, n_s is final sample size, Z is the z-score at the selected confidence level in the normal distribution, p is the proportion of the sample to the population, q is 1 - p, N is population number, and e is the margin of error. In this study, there were 466,393 land parcels, and a confidence level of 95% (z-score = 1.96), a margin of error of 5%, and a sample proportion of 50%. By using Equation 1, the minimum sample size was 384. Meanwhile, the number of samples obtained was 2,836, so the sample size already represented the condition of the land parcels in the study area.

$$n_0 = \frac{Z^2 pq}{e^2}$$

Equation 1

$$n_s = \frac{n_0}{\left(1 + \frac{n_0}{N}\right)}$$

Equation 2

The parcel-based land value point data distribution is integrated into the cadastral land value geometry via spatial joint, this leads to the sample attributes of the parcel-based land value being linked to the cadastral land parcel. The sample distribution can be seen in Figure 3. The samples range from 58.40 USD/m^2 to 3,003.15 USD/m^2; with the average and the standard deviation is 687.57 USD /m^2 and ± 484.56 USD/m^2, respectively. In addition, land parcel data from Land Office Surabaya II which scaled at 1:2,500 is served as complementary data to denote the geometric representation of land parcels.

2.3.2 Road-related features

For road-related features, the dataset is obtained from the Surabaya City Public Works, Housing, and Settlement Area Office which is scaled at 1:5,000. The dataset includes w^{road}, $d^{arterial}$, and $d^{collector}$. These variables indicate the level of accessibility of each land parcel in which the higher the accessibility, the higher the value of a land parcel and vice versa [44]. Proximity analysis is used for the extraction of w^{road}, $d^{arterial}$, and $d^{collector}$. In the analysis, the distance represents the shortest distance from each land parcel to the nearest road feature that is calculated using Euclidean distance defined in Equation 3:

$$d^i_{x,y} = \sqrt{\left(x - x^i_{nearest}\right)^2 + \left(y - y^i_{nearest}\right)^2}$$

Equation 3

Where $d^i_{x,y}$ is the distance of a land parcel located in (x, y) to the nearest features I which located in $(x^i_{nearest}, y^i_{nearest})$. The use of ED for accessibility assessment because it has a relatively low computational cost and is widely used for property valuation systems [45][46] and [47]. However, the use of ED for accessibility assessment has limitations because it does not represent the physical distance which is affected by the impedance of road features [48]. Here, ones shall take a note that i represents not only road-related features (width, distance to arterial, distance to collector), but also interest-point-related features (traditional market, CBD, point of interest, school, hospital). Based on data processing results, the variable distance to arterial roads has a minimum and maximum value of 0 and 2,335 m, respectively, with an average of 338.025 m.

In the distance to urban collector roads variable, the minimum and maximum value is 0 and 4,606.75 m, respectively, with an average of 1,049.91 m. The road width variable is obtained from the measurement of the road width nearest the land parcel. Based on the data obtained, it has minimum, maximum, and average values of 0.5, 16.74, and 4.34 m., respectively.

2.3.3 Interest-point-related features

Further, from the Surabaya City Land and Spatial Planning Office provides the distribution of POI, schools, traditional markets, and the CBD datasets with the scale of 1:5,000. The datasets are thus used to derive point-related features including $d^{amarket}$, d^{CBD}, d^{POI} , d^{school}, and $d^{hospital}$ using Equation (1), which is in line with the study of [49] and [50]. Based on the data processing results, the distance variables to the POI have minimum, maximum, and average values of 57.26, 9.80, 57.26, 9.80, and 2,060.18 m., respectively. The distance variable to the traditional market has a minimum, maximum, and average value of 0, 4,995.33, and 1,010.65 m., respectively. The distance variable to the CBD has a minimum, maximum, and average value of 46.91, 6,963.45, and 2,213.94 m., respectively. Then, the distance variable from the hospital has a minimum, maximum, and average value of 0, 4,790.10, and 1,461.55 m., respectively.

2.3.4 Zoning score feature

Zoning score feature was derived from the "Spatial City Details Plan" (locally abbreviated as RDTR, Rencana Detil Tata Ruang) which represents the zones in the regional spatial plan, is obtained from the Surabaya City Land and Spatial Planning Office with a scale of 1:5,000. The zones consist of 7 different zones; those are protected zones and boundaries, low-density housing, medium-density housing and public facilities, high-density housing, as well as 1st, 2nd, 3rd class trade and service areas. Since were provided in categorical data format, the zones were assigned to certain scores for numerical representation using leave-one-out (LOO) encoding techniques which follow the study of [51]. The assignation into numerical data with LOO was modeled mathematically using Equation 4:

$$Zn_{ip} = Ld_{ip};\ k \neq 1 \qquad \text{Equation 4}$$

Where Zn_{ip} is the assigned score for zone *j:{all zones data from RDTR}* located at k and Ld_{ip} represents the average land value of all parcels for zone j in the location other than k. The result of the LOO encoding technique for the spatial city details plan is delivered in Table 2. The first-class trade and service areas zone had the highest mean LOO value (825.444 USD /m^2), while the protected zones and boundaries had the lowest (597.623 USD/m^2). For summary, Figure 4 shows all input variables for land valuation in this study. Those include w^{road}, $d^{arteriat}$, $d^{collectorial}$, d^{market}, d^{CBD}, d^{POI}, d^{school}, $d^{hospital}$, and Zn_{ip}. Once prepared, all variables were then standardized.

2.4 Data Pre-processing

Variable data preprocessing is undertaken to mitigate errors stemming from uncontrollable factors. Such factors encompass data outliers and nonuniformity in data dimensions. Outliers often arise due to inaccuracies in determining the distribution of random samples, errors in measuring the data, and other factors [52]. Given these challenges, there is a critical demand for a technique to detect and eliminate outliers, ensuring that the data remains uncontaminated. This process allows for the construction of models that perform optimally. The interquartile range (IQR) is a commonly used method for detecting and exercising outliers in machine learning. It serves as a tool to pinpoint outliers in continuously distributed data [53].

Table 2: Result of LOO encoding for spatial city details plan variable

Zone	Minimum value of LOO score (USD/m^2)	Maximum value of LOO score (USD/m^2)	Mean value of LOO score (USD/m^2)	The standard deviation of the LOO score USD/m^2)
Protected zones and boundaries	595.501	598.176	597.624	0.401
Low-density housing	709.198	797.900	776.905	19.893
Medium-density housing	747.905	751.237	750.424	0.669
High-density housing	778.375	782.290	781.416	0.701
1st class trade and service areas	801.540	834.509	825.444	6.553
2nd class trade and service areas	788.527	804.735	800.567	3.436
3rd class trade and service areas	622.856	682.642	661.016	22.531

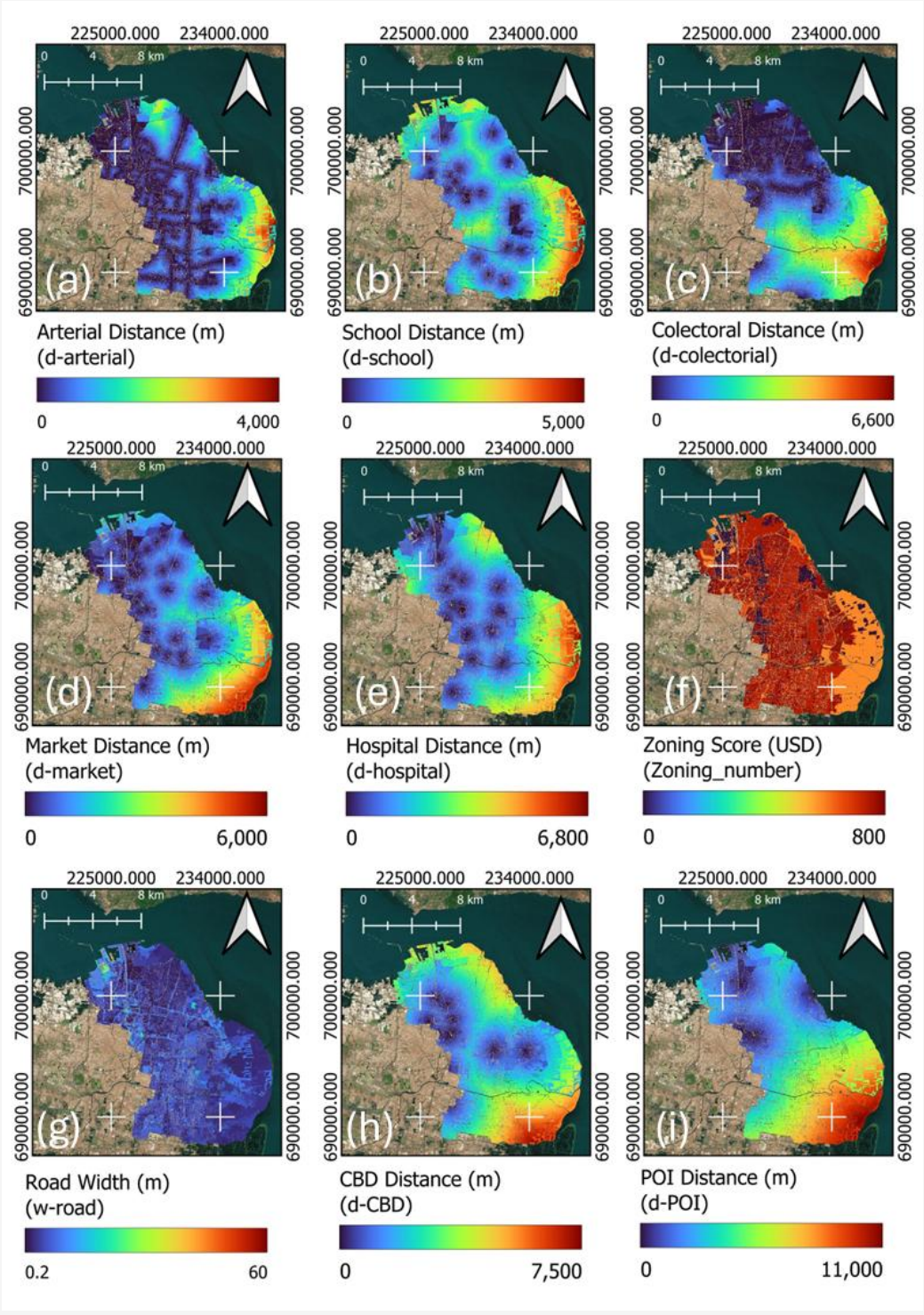


Figure 4: Input variables data for land valuation (coordinate reference system: TM3 49.2 zone); (a) distance from arterial road; (b) distance form school; (c) distance from collector road; (d) distance from market; (e) distance from hospital; (f) zoning score generated from LOO encoding; (g) road width; (h) distance from CBD; (i) distance from POI.

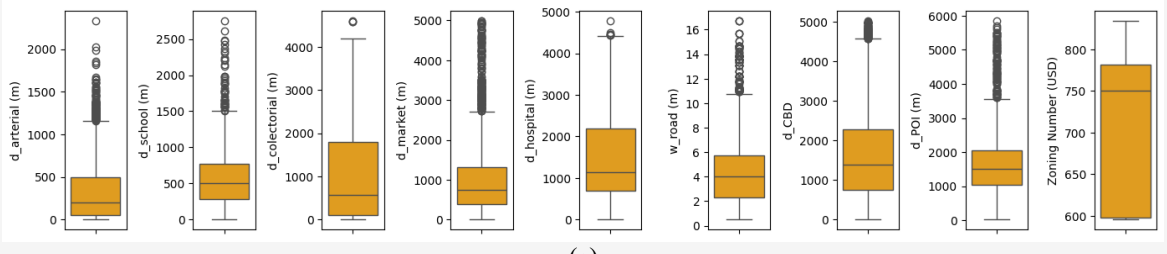
It is also called the midspread or the data distribution covering 50% of the entire dataset, which is technically termed the H-spread. The IQR is determined by calculating the difference between the first and third quartiles, represented as IQR = Q3 − Q1 [54]. Data outside the range of the lower and upper limits is outlier data, so it needs to be removed to make more optimal prediction data. In the preprocessing stage, data standardization is performed so all variable data can have the same scale, thus making it easier for the system to interpret the data [55]. Thus, the model formed will have optimal accuracy. The data normalization technique used in this research is standard-scaler. The method converts the data into a uniform distribution so that all data has a mean of zero and a standard deviation of 1 [56]. The Z score normalization can be formulated in Equation 5 [57]:

$$Z - score = \frac{A_{iv} - A_{iv(avg)}}{\sigma_{A_{iv}}}$$

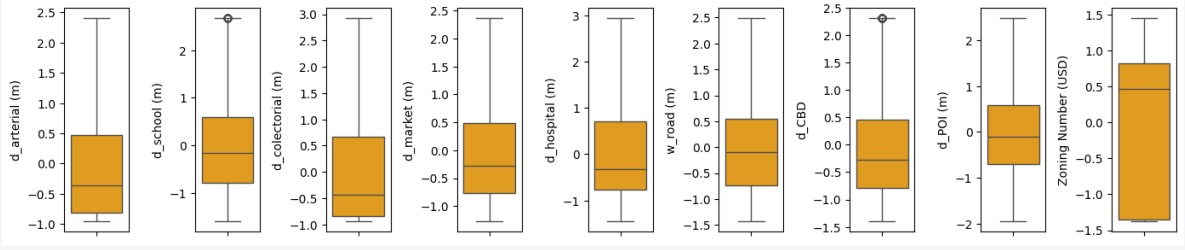
Equation 5

Where A_{iv} and *Z-score* contain the iv^{th} variable before and after standardization with their corresponding mean ($A_{iv(avg)}$) and standard deviation (σ_{Aiv}). In addition, the iv^{th} includes w^{road}, $d^{arteriat}$, $d^{collectorial}$, d^{market}, d^{CBD}, d^{POI}, d^{school}, $d^{hospital}$ and Zn_{ip} in respective order.

Figure 5 visualizes data conditions before and after pre-processing using a box plot involving two key steps: handling outliers and data standardization. In visualisation using boxplots, there are 4 components, namely; (a) inter-quartile range (IQR), which is the value of Q1 and Q3 in the distribution of data, this is represented by a yellow box in Figure 5; (b) lower limit which is the limit of the minimum value range of data included in the normal distribution which is represented by an extension line below the IQR area or often called the lower whisker; (c) upper limit which is the limit of the maximum value range of data that is considered to be included in the normal distribution, represented by an extension line above the IQR area or called the upper whisker; and (d) outlier points which are the values of data that are not included in the upper and lower whisker ranges. Figure 5(a) shows the condition of the data before pre-processing. In this figure, it can be interpreted that almost all data have outliers and different data dimensions. The difference in data dimension will reduce the performance of land value modeling because it is prone to bias. Figure 5(b) is the result after pre-processing, where all data does not have outlier data. In addition, the data dimensions of all variables have the same mean-value of 0 and standard deviation of ±1, a result of data standardization. Thus, the modeling variables are more reliable for further processing. This study uses land value zone data as the dependent variable, while the independent variables use variables that affect the value of land parcels. At the model construction stage, the first thing to do is feature extraction of independent data on dependent data. The stage carried out before modeling is splitting the land parcel value samples. Splitting the land parcel value samples is dividing samples into training and testing data with a ratio of 80:20.



(a)



(b)

Figure 5: Box plot of each variable data; (a) before scaling; (b) after scaling

2.5 RF Modelling and Hyperparameter Adjustment

The algorithm used in this research is RF. RF is an ensemble learning approach from a decision tree that produces output as continuous data for regression problems. Each decision tree divides the input data iteratively into two or more data samples called nodes. This division will continue until certain conditions are met, such that the variation of the data group within a certain variable becomes very small. One node will be subdivided into child nodes in the same way. The data in each child node is used to predict the value of the dependent variable in that node. The results from all child nodes are combined to produce the final prediction [58]. The RF algorithm can be calculated as in Equation 6 [59].

$$LV_{est} = avg\{lv_t(Z-score)\}_{t=1}^{N} \quad \text{Equation 6}$$

Where LV_{est} represents the land value estimation of parcel located at x, y from an RF model which has N number of trees. The LV_{est} is derived by averaging all $lv_t(Z\text{-}score)$ which indicates the estimation of a specific tree number t.

Establishing algorithm architectural models through suitable hyperparameter tuning is an essential component in developing efficient ML models, particularly models based on tree-based algorithms such as RFs. It happens because the RF approach requires optimization of various hyperparameters to achieve optimal model performance. Manual testing using trial and error on hyperparameter settings is still frequent. However, this technique becomes less effective due to several reasons, such as a high number of hyperparameters, complex models, time-consuming model evaluation, and nonlinear interactions between hyperparameter configurations. Consequently, this study proposes to use Bayesian-based hyperparameter tuning to automatically determine the ideal hyperparameter configuration using an approach based on statistics. The land parcel value prediction is trained with the best RF hyperparameter to get the optimal model evaluation value. This study uses the Bayesian search optimizer technique available in the Scikit Learn library to determine the best hyperparameter value from predetermined search spaces as shown in Table 3. Bayesian optimizer is a sequential model-based approach to problem-solving [60].

The parameters optimized in this research are *bootstrap*, *max_depth*, *max_features*, *min_samples_leaf*, *min_sample_split*, and *N_estimators*. The bootstrap parameter identifies whether bootstrap samples are used when building each DT. If bootstrap is false, all datasets are used to build each DT. By default, the value of bootstrap is True. *Max_depth* is a parameter that sets the maximum height of each DT in the RF. By default, the value of *max_depth* is None. *Max_features* is the number of parameters that must be considered to create the best splitting. By default, the *max_features* value is None. The *min_samples_leaf* value is the minimum number of samples a node should have after splitting. By default, the *min_samples_leaf* value is one. The *min_samples_split* value is the minimum number of samples a node must have to split again. The *N_estimator* parameter describes the number of DT built into one RF model. Usually, the more DT, the better. By default, the value of n_estimators is 100.

Figure 6 shows the results of hyperparameter tuning using the Bayesian search optimizer. Table 2 shows the optimal hyperparameter value interpretation results. The optimal input for *bootstrap, max_depth, max_features, min_samples_split, mean_samples_leaf, and N_estimators* is "False," 90, 4, 1, 2, and 1,800, respectively.

Table 3: Search spaces and optimum input of tuning hyperparameter using Bayesian search optimizer

Parameter	Search spaces	Optimum input
bootstrap	"True," "False"	"False"
max_depth	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None	90
max_feature	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	4
min_samples_leaf	1, 2, 4	1
Min_samples_split	2, 5, 10	2
N_estimators	200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000	1800

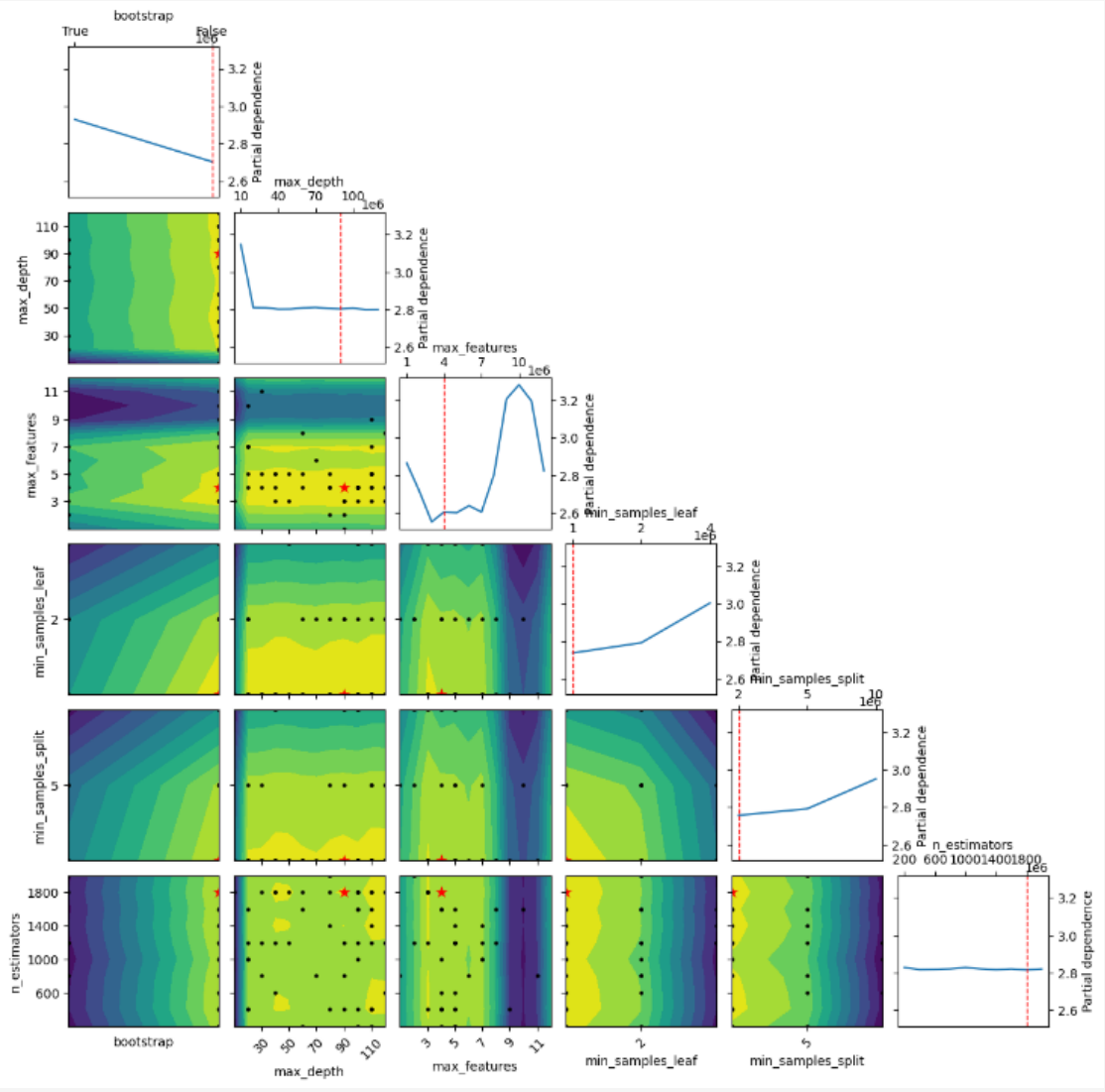


Figure 6: Result of tuning hyperparameter random forest using Bayesian optimizer

2.6 Accuracy Assessment

Evaluation of model performance was conducted with several parameters, such as mean absolute percentage error (*MAPE*), coefficient of determination (R^2), mean absolute error (*MAE*), coefficient of variation (*COV*) land value prediction ratio (*LPR*), and price related differential (*PRD*) as defined in Equations 7 to 12:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{LV_i^{obs} - LV_i^{est}}{LV_i^{obs}}\right| \times 100\%$$

Equation 7

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(LV_i^{obs} - LV_i^{est}\right)^2}{\sum_{i=1}^{n}\left(LV_i^{obs} - LV_{avg}^{est}\right)^2}$$

Equation 8

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|LV_i^{obs} - LV_i^{est}\right|$$

Equation 9

$$LPR_i = \frac{LV_i^{obs}}{LV_i^{est}}$$

Equation 10

$$COV = \frac{\sum_{i=1}^{n}\left(LPR_i - LPR_{avg}\right)^2}{(n-1)LPR_{avg}}$$

Equation 11

$$PRD = \frac{LPR_{avg}\sum_{i=1}^{n} LV_i^{obs}}{\sum_{i=1}^{n} LV_i^{est}}$$

Equation 12

Where LV_i^{obs} is the i^{th} observed land value of parcels, LV_i^{est} LV_i^{est} is the i^{th} estimation land value of parcels, and LV_{avg}^{est} is mean value of estimation land value of parcels.

Further, the relative contribution of each input variable was measured using variable importance. Two variable importance parameters were utilized, that are permutation importance (PI) and mean decreased impurity (MDI). The PI and MDI values have been normalized between 0 and 1. As a result, as the PI and MDI values approximate one, the variable's influence on the prediction model emerges stronger. In contrast, when the MDI or PI value is close to zero, the variable has a lower influence on the prediction model.

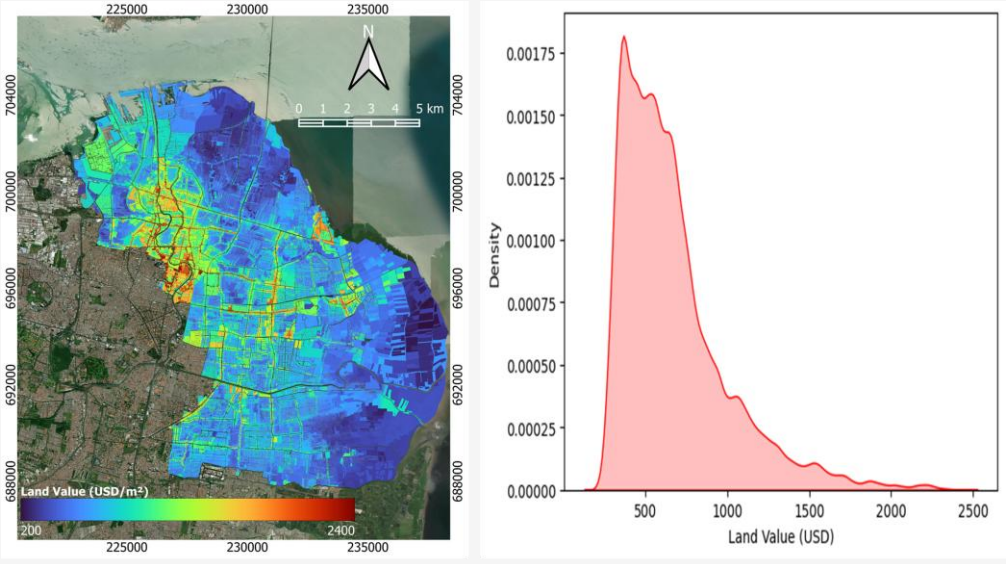
3. Results

3.1 Land Parcel Value Prediction

Once trained, the developed model was applied to predict the $LV_{x,y}^{est}$ for all points. The land parcel value prediction can be seen in Figure 7(a). The predicted land parcel values range from 103.37 USD/m^2 and 2,5083.57 USD/m^2 with the mean of 691.92 USD/m^2 and standard deviation of ± 36.37 USD/m^2. In addition, Figure 7(b) shows the frequency distribution of the prediction results using the prebuilt model in the study area. The x-axis corresponds to the land parcel value in units of USD/m^2 while the y-axis represents the frequency of the range of land parcel values. Visually, the data is asymmetrically distributed. Based on the symmetry–asymmetry of the histogram, the results are included in the positive skew histogram. Modally, the histogram is described as unimodal, with only one peak or mode value. The mode value in the frequency distribution is Rp. 11,507,290.00/m^2.

3.2 Model Performance

The performance of the Random Forest (RF) algorithm in parcel-based land value modeling was evaluated using five parameters: Mean Absolute Error (*MAE*), Mean Absolute Percentage Error (*MAPE*), R-squared (R^2), Coefficient of Variation (*COV*), and Prediction Relative Difference (*PRD*). The evaluation results for these parameters are presented in Table 4. This evaluation was conducted on both training and testing datasets to determine whether the prediction model is overfitting. Overfitting occurs when the performance of the training data is significantly better than that of the testing data. *MAE* represents the average error value of the prediction results compared to the actual. The range of *MAE* values is 0 to ∞, where the closer to 0, the better the prediction model. The *MAE* value on training and testing data is 39.31 USD and 39.25 USD, respectively. Both datasets produce almost the same *MAE* value, but the training data is slightly higher by 0.15% than the testing data. Model evaluation with *MAPE* has the advantage of being easy to interpret compared to *MAE* because the unit is a percentage. *MAPE* is the percentage of the absolute difference between predicted and actual values divided by the actual value.



(a) (b)

Figure 7: Prediction result; (a) Map plot of land parcel value data prediction using random forest (b) KDE plot prediction of land parcel value data

Table 4: Model evaluation result

Metric	Value	
	Training data	**Testing data**
MAE	39.31 USD	39.25 USD
MAPE	6.99%	7.21%
R^2	0.96	0.96
COV	4.18%	4.87%
PRD	1.03	1.02

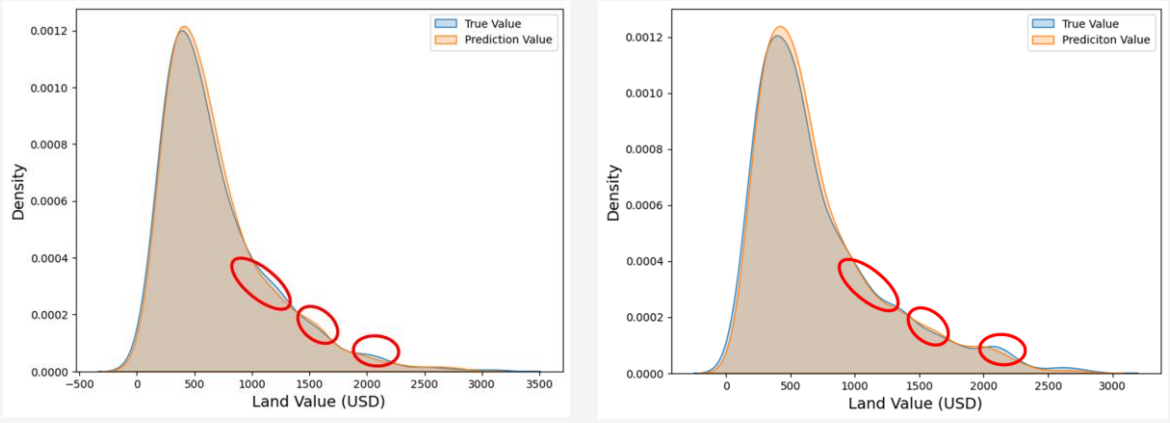
The *MAPE* value has a range of 0-100%; the closer to 0, the better the prediction model. Based on the evaluation results, the *MAPE* values on training and testing data are 6.99% and 7.21%, respectively. Both values have a difference of 3.10%, where the value in the testing data is higher than the training data. The R^2 metric evaluates the relationship of value patterns between predicted and actual. The R^2 value has a range of 0 to 1, where the closer to 1, the higher the performance of the prediction model. Based on the evaluation results, the R^2 value on the same training and testing data is 0.96.

As part of the property valuation context, the International Association of Assessing Officers (IAAO) has a metric to evaluate the valuation result with the actual value. This evaluation is represented by LPR, which is the ratio between the valuation result and the actual value. Then, the metric is defined with *COV* and *PRD*. *COV* is the variance value divided by the average of the *LPR*. Like *MAPE*, *COV* is represented by a percentage where the closer to 0, the better. Based on the evaluation results, the *COV* values on the training and testing data are 4.18% and 4.87%, respectively. The *COV* value of the training data is 15% lower than the testing data. In addition to *COV*, *PRD* is used as a metric to evaluate *LPR* by comparing the average *LPR* with the weighted average (*WA*). *WA* is the ratio between the total number of predicted land values and the total number of actual land values. The *PRD* value has a range of 0 to ∞, where the ideal value of *PRD* is 0.98 to 1.03. This range of values indicates that the predicted land value is not significantly different from the actual value. A value of less than 0.98 indicates that the predicted value is significantly lower than the actual value. Meanwhile, a *PRD* value of more than 1.03 indicates that the predicted value is significantly higher than the actual value [61]. The *PRD* value of the evaluation results on the training and testing data is 1.03 and 1.02, respectively, where the *PRD* value of the training data is 0.98% higher than the testing data. This indicates that the training and testing data prediction results are slightly higher than the actual value but not significantly different.

Figure 8 shows the kernel density estimate (KDE) of predicted values in both the training and testing data. KDE is a visualization technique that estimates the probability density function using all sample point locations and shows multimodality. It compares favorably with 2D histograms as it is smoother and does not require bin orientation [62]. The training and testing data reveal a unimodal frequency distribution model with a positive skewness. The training and testing data demonstrate an almost perfect overlap between the predicted and observed values. However, there are significant differences in frequency distribution between the observed and predicted data in both the training and testing datasets, as seen by the red line in Figure 8. There is a comparatively greater difference between the predicted and observed data in the testing dataset than in the training dataset. This phenomenon is associated with the model being tuned to the training data, resulting in minimal prediction errors on the training data compared to incoming data, specifically the testing data. Despite the close agreement and similar distribution pattern between the observation and prediction data, there are different mode values for both the training and testing parts. The testing dataset shows that the mode value of observation and prediction data is 234.91 USD and 378.82 USD, respectively. The training dataset shows that the mode value of the observation and prediction data is 493.64 USD and 378.82 USD, respectively.

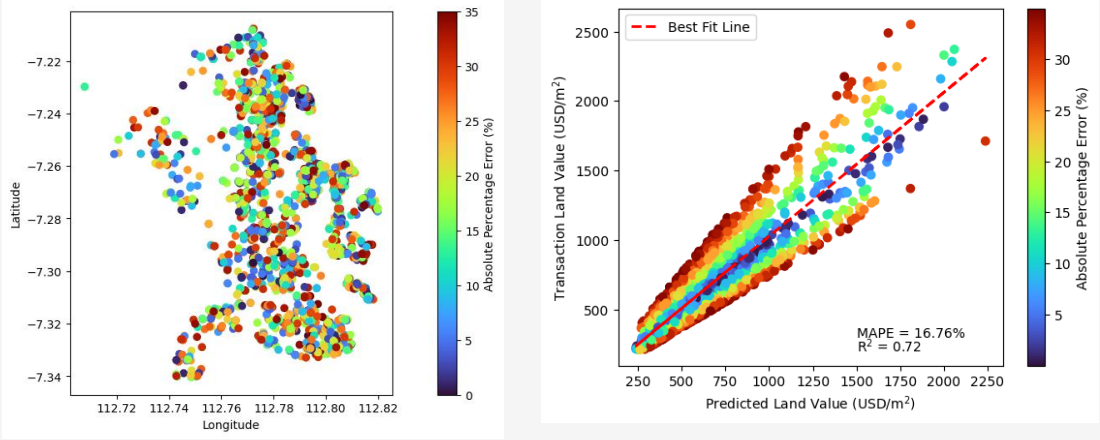
3.3 Validation with Official Transactional Data

In addition to evaluating data quality using data training and testing (notes are considered internal data), this study also tests external data. The external data collected is transactional data from the data and information center of the Indonesian National Land Agency. The data corresponds to official land transactions that occurred and documented by the government in Surabaya City during the year 2023. The data obtained is a point with land value attributes given in rupiah units per square meter. The data is then extracted from the predicted data for further analysis. *MAPE* and R^2 are the evaluation metrics used to compare transaction data to predicted data.



(a) (b)

Figure 8: Training and testing land value KDE-plot: (a) Density distribution of prediction and observation data on training dataset. (b) Density distribution of prediction and observation data on testing dataset



(a) (b)

Figure 9: Results of an evaluation using transaction data for 2023: (a) Residual percentage distribution for each transaction point. (b) Scatter plot demonstrating the relationship between transaction and predicted data

Figure 9(a) shows the distribution of residual percentages for each transaction point. The absolute residual percentage is calculated by dividing the difference between the transaction and predicted land values by the land transaction value and then the absolute value. Figure 9(b) shows the scatter plot relationship among predicted land values and transactions. In Figure 9(b), the x-axis represents the predicted land value, and the y-axis represents the transaction land value. Meanwhile, the red line represents the ideal relationship between the two data sets. The closer to the axis, the lower the difference between the transaction data and the prediction. The color gradation of blue to red represents the absolute percentage error (*APE*) value. According to the scatter plot, the higher the land value, the greater the margin of error between the predicted and transaction land values.

The total amount of test data derived from transaction data is 2,592. The results of the evaluation showed that the absolute percentage error ranged from 0.001% to 34.94%. The *MAPE* between the two datasets is 16.76%. Meanwhile, the coefficient of determination for the two datasets is 0.71. The spatial distribution of the absolute value of the percentage error for each location indicated that the absolute value of the percentage error, both small and large, varied equally over the region, with no centralization in specific areas. A significant difference between the transaction value and the prediction may occur; this is because when conducting a buying and selling transaction, there is a negotiation process that results in the price achieved being more or smaller than the predicted data.

4. Discussion

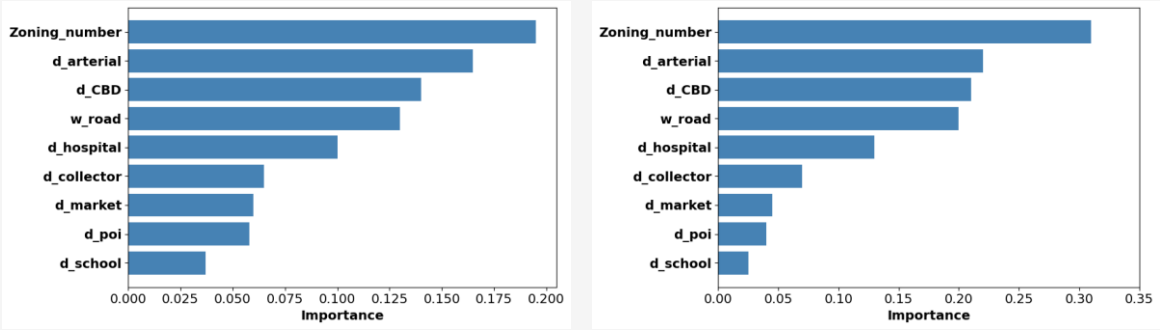
This research endeavors to predict land values using an ML algorithm combined with spatial data. Such explorations are seldom undertaken, especially in Indonesia. This study mainly aims to formulate a land value prediction model for 2023 using the RF algorithm, leveraging the latest variable data. Before deploying the RF modeling, an initial collection of 12 predictor variables is filtered through the important analysis feature space to discard less impactful features. The remaining nine variables undergo analysis using the MDI and PI, which strongly correlate with land value. Employing the RF algorithm, this research successfully crafted a model with high accuracy, R^2 values (exceeding 0.9), and a minimal *MAPE* (less than 10%). This aligns with findings from studies by [23] and [63], which reported a 10% *MAPE* when using the RF algorithm for property value predictions.

RF was chosen as the algorithm used in this research because it has performed very well in previous research related to property valuation [32] and [64]. The good performance of RF compared to other algorithms in solving property valuation problems is because this algorithm combines dataset features by exploring the hierarchical structure of characteristics and the influence of each feature on property values [23]. RF algorithm does not require data complexity assumptions such as normality, as market data characteristic for property valuation is rarely normally distributed [65]. The RF algorithm can capture complex relationships between dependent and independent variables, where the assumption of linearity of the relationship between the two data is not required [66]. This contrasts with the multiple linear regression algorithm, which requires the assumption of the dataset's normality and the linearity of the relationship between the dependent and independent variables [64]. When using multiple linear algorithms, it under-represents the characteristics of property market data. In addition, the RF algorithm can provide an overview of the hierarchy of each DT and information on the level of importance between features to the prediction model, which can increase the transparency of land valuation. Land valuation transparency is needed in systematic land valuation. In contrast, neural network-based algorithms have a 'black box' characteristic, i.e., the user does not know the algorithm's computational complexity. With this characteristic, this research considers the RF algorithm in parcel-based land value assessment.

The evaluation of the relative contribution level of each variable to the model is conducted, among other things, to interpret the built model. Through this evaluation, it is possible to identify which variables most influence the land parcel value prediction in the study area. Evaluation of the relative contribution level of each variable to the model in this study uses two approaches: mean decrease impurity (MDI) and permutation index (PI). Figure 10(a) shows the results of the significance level of each variable in the model with the MDI approach. The three variables that have the most influence on the prediction model are zoning number, distance to arterial roads, and distance to CBD with MDI values of 0.20, 0.18, and 0.15, respectively, while the three variables that have less influence on the prediction model are market distance, distance to point of interest, and distance to schools with MDI values of 0.06, 0.05, and 0.04, respectively. Based on the MDI approach, it can be interpreted that in the study area, zoning score is the variable that most influences the value of land parcels. The distance from the school variable is the variable that has the least influence on the prediction of land parcel value.

Figure 10(b) shows the results of the analysis of the significance level of each variable to the prediction model using the PI approach. The three variables that have the largest relative contribution to the prediction model are zoning number, distance to arterial roads, and distance to CBD with PI scores of 0.32, 0.24, and 0.23, respectively, while the three variables with the lowest relative contribution levels are distance to urban collector roads, distance to point of interests, and distance to schools with PI scores of 0.05, 0.04, and 0.02, respectively. Based on the PI score, it can be interpreted that the zoning score variable has the most influence on the prediction model and the distance from the school variable has the least influence on the prediction model.

In the context of property valuation, there are several terms that define the quality of valuation. One of the terms that can be used is margin of error (MoE). MoE is the concept of acceptable inaccuracy in the property valuation process. MoE can be represented by the percentage variation or difference between the prediction and the actual value. It can also be represented with *MAPE*. According to [67], the threshold for an acceptable property valuation is less than 10%. In addition, the quality of property valuation is also defined by the International Association of Assessing Officers (IAAO) with the coefficient of variance (*COV*) [61]. *COV* is a measure of variance comparison with the average ratio between predicted and actual values. The *COV* value according to IAAO is recommended to be less than 20%. Based on the results of this study, the *MAPE* value of 3.48% means that it does not exceed the MoE limit.



(a) (b)

Figure 10: Feature importance prediction model of land value; (a) mean decrease impurity (MDI); (b) permutation index (PI)

Meanwhile, the *COV* value generated in this study is 4.87%. Thus, it can be interpreted that the land valuation prediction model in this study can meet the IAAO standard criteria and the property valuation margin error threshold.

Despite using the Bagging technique, which can minimize overfitting in the RF algorithm, property valuation cases experience overfitting problems when using RF. Research by [68] found overfitting problems when using several ML algorithms, including RF. The study used ML algorithms for house price valuation in Dar es Salaam, Tanzania. In the study, the *MAPE* value on training and testing data using the RF algorithm was 35.59% and 56.42%, respectively. The percentage difference in *MAPE* values on the two datasets is 45.27%. However, the RF algorithm in the study is the best algorithm with the lowest *MAPE* difference between training and testing compared to linear regression, elastic net, regression tree, boosting, neural network, SVM, and nearest neighbor. There are several causes of overfitting: 1) noise in the training dataset, 2) the algorithm is too complex to learn a simple relationship between dependent and independent variables, and 3) too many iterations for the induction-based algorithm such as boosting [69]. However, in this study, the difference in evaluation values between training and testing data is not too significant, which can be interpreted as free from overfitting problems. The percentage difference in the evaluation value between training and testing in this study ranges from 0% to 15.25%. Where the lowest percentage difference value is R^2 and the highest is *COV*.

This study involved variable selection by examining the correlations between variables and land values and then improving hyperparameter values using Bayesian optimization. Bayesian optimization is frequently used in black box issues through a model to reduce the number of evaluation functions due to its high computational capacity. Consequently, Bayesian optimization can reduce the time and cost of hyperparameter tuning experiments [70]. Despite the absence of postprocessing through regularization, the model used in this study demonstrated accomplished land value prediction.

The accuracy of the predicted outcomes has been validated by comparing them with the transactional data. The validation process involves *MAPE* parameters and an R^2 score. The validation results showed that the *MAPE* was fairly low, with a value of less than 20% [71]. Meanwhile, evaluating the R^2 score provides results with high/strong influence, specifically in the range of 49%–81% [72]. Regional spatial planning zoning is the first variable that most influences the land value. Zoning regulates space utilization, where each zone is allocated by a detailed spatial plan consisting of spatial patterns and structures [73]. This zoning is related to land use for community needs because activity sectors can influence urban space. In scoring, the trade and services area have the highest score because it has a high land value. This is based on the economic capabilities of land, its productivity, and economic strategy [74]. Zoning is also related to the development of an area where, because of greater population activity, public facilities will increase to support the welfare of the residents of Surabaya City. Meanwhile, the variables market distance, distance to point of interest, and distance to schools have less influence on the model.

Besides internal data (testing dataset), this research also uses external data for model validation. Transaction data is the sale and purchase value reported to the local land office, in this case, the Surabaya Region II land office. This administrative report data is very vulnerable to bias, because the bargaining process can still influence the transaction value [75]. Therefore, the transaction value is used as external validation, not to build a prediction model.

Valuation data by ISA is more reliable than transaction data because it is a fixed cost unaffected by the bargaining process. The actual transaction value reported to the land office often differs from the actual sale-purchase value. This is because it is related to the application of sales and purchase tax, which is often lower than the actual sales and purchase value [76]. The number of points with a ratio of value between predictions and transactions of more than 1, which indicates that the transaction value is lower than predicted, is 1,179 points or 52.31% of the total points. With this condition, there is a significant difference at several points, with an APE value of more than 20% when the prediction data is compared to the transaction value. The highest APE value in this external validation process is 34.94%. This exceeds the maximum MoE, as discussed earlier. The number of points with an *APE* of more than 20% is 905 points or 40.15% of the total data. However, in general, when the APE values are averaged, the *MAPE* is less than 20%, 16.76%, to be precise. Thus, it can be interpreted that the prediction model still performs well despite validation with external data, which have various limitations.

Previous studies in various regions have tested the utilization of ML algorithms for property valuation. Research by [25] used geographically weighted regression (GWR), ordinary least square (OLS), Lasso regression (LaR), Ridge Regression (RR), ElasticNet, SVR, DT, RF, GB, XGBoost, and Multi-Layer Perceptron (MLP) algorithms for house and apartment valuation in Sydney, Australia. The GB algorithm has the best performance of these algorithms, with a *MAPE* of 7.38%. Meanwhile, the *MAP*E for RF was 7.81%. Another study by [41] used SVR, RF, XGBoost, and deep neural network (DNN) algorithms for land valuation in Victoria, Melbourne, Australia. In that study, RF had the best performance with a *MAPE* value of 16.6%, and DNN had the lowest performance with a *MAPE* of 57.3%. House price valuation in Rotterdam using OLS and RF algorithms was conducted by [77]. RF is the best algorithm in the study, with an R^2 value of 0.74, while OLS has an R^2 of 0.61. Meanwhile, in this study, using the RF algorithm for parcel-based land value assessment in the case study of Surabaya City, Indonesia, resulted in a *MAPE* value of 7.21% and R^2 0.96. Although the model in this study performs better than the three previous studies, it does not necessarily have the same performance when applied to the three cases; it could be lower or higher, even the same. This is due to the limitations of ML prediction models for generalization to other case studies, which have different characteristics from the training data in a particular case study. Therefore, other prediction models or RF testing should be used in other case study locations, especially in Indonesia.

This research selects samples by reconstructing the entire sample and selecting training and testing sets at random. Utilizing the sample land value data from 2022, which could be trained for valuing land by 2023. However, our access to information is often limited when evaluating land valuation practices. To resolve this issue, transparent and efficient land value data modeling performance is required. Despite the limitations of the land value due to the incomplete PTSL, this study can generate a model for an accurate prediction of the land value. This research takes a case study in Surabaya City, representing the characteristics of a reasonably complex city in Indonesia. Applying the RF algorithm for land valuation in Surabaya City produces reasonably good performance.

However, modeling using the ML algorithm is highly dependent on the conditions and characteristics of the training data. An RF algorithm that performs well for land valuation in Surabaya City may not necessarily perform well for other case study applications. There are several limitations of the RF algorithm as consideration for using the same algorithm for land valuation in other case studies, namely: 1) RF only minimizes variance, not bias, so it is susceptible to data containing bias[78]; 2) Algorithm performance is highly dependent on hyperparameter settings [79]; 3) Requires dimensionality reduction settings to minimize overfitting problems [80]; 4) Heavy computational cost when using large datasets or large number of DT and DT depth hyperparameter settings[81]; and 5) reduced performance when dealing with high-cardinal categorical data [82]. In addition, land values are highly dependent on temporal variations. Parcel-based land value in this study uses training data in 2022, and the value needs to be updated continuously and annually, considering dynamic urban development.

5. Conclusions

This study aims to ascertain the variables that exert an influence on land value and develop a predictive model for urban land value in Surabaya City. This chapter has presented the significance and influence of variables in ascertaining land value. Various factors have demonstrated efficacy in regulating the variability of land value. Three of the nine variables considered are road width, zoning, and distance to the CBD. The land value model with the RF algorithm demonstrates promising predictive capabilities that can be effectively used in Surabaya City.

The values of *MAE*, *MAPE*, R^2, *COV*, and *PRD* for each test dataset have been calculated as 39.25 USD, 7.21%, 0.96, 4.87%, and 1.02 respectively. Based on the validation process with transaction data, it was discovered that the prediction model continues to perform well regarding *MAPE* and R^2 scores, with values of 16.76%. and 0.72, respectively. Although certain aspects were not considered in this study, such as bargaining in the purchase and sale process, the model can still predict the actual land value effectively.

The most interesting finding of this research is that it can predict the value of land parcels. To address a practical question, the RF algorithm can serve as a tool for individualized valuation that is both relevant and efficient. This means that it can be applied to locations with characteristics like the area studied in this research case. Several land parcels still need to undergo the PTSL process, but land valuation must be conducted immediately. Overall, this study systematically highlights the advantages of the RF algorithm in predicting urban land valuation. The limitation of this research is that it only uses a single algorithm for land parcel value modelling, namely RF. RF has shown excellent performance in urban case studies, especially Surabaya City, Indonesia. Experimental research to evaluate other algorithms is recommended because a model that performs well in this research is not necessarily suitable for other case studies. This is because modelling with ML algorithms is very sensitive to the characteristics of the training data. Another limitation of the study is the availability of variable data for land value prediction. Therefore, further research is needed to add variables related to the internal characteristics of land parcels, such as land tenure.

Acknowledgements

The authors would like to thank Geomatics Engineering, Universitas Gadjah Mada (UGM), for facilitating this research. The authors are also grateful the Directorate of Research and Community Service at Institut Teknologi Sepuluh Nopember for its financial support. The authors also appreciate data access from Information and Data Center/National Land Agency (ATR/BPN).

References

[1] Dawidowicz, A. and Źróbek, R., (2017). Land Administration System for Sustainable Development - Case Study of Poland. *Real Estate Management and Valuation*, Vol. 25(1), 112–122. https://doi.org/10.1515/remav-2017-0008.

[2] Williamson, I., Enemark, S., Wallace, J. and Rajabifard, A., (2010). *Land Administration for Sustainable Development. Redlands: ESRI.* Esri Press.

[3] Subedi, G., (2016). *Land Administration and Its Impact on Economic Development*. Doctoral Thesis. Real Estate and Planning, Henley Business School. University of Reading.

[4] Schwerhoff, G., Edenhofer, O. and Fleurbaey, M., (2022). *Equity and Efficiency Effects of Land Value Taxatio*n. International Monetary Fund, Vol. 2022(263). https://doi.org/10.5089/9798400227943.001.

[5] Che, S., Kumar, R. R. and Stauvermann, P. J., (2021). Taxation of Land and Economic Growth. *Economies,* Vol. 9(2), 1–20. https://doi.org/10.3390/economies9020061.

[6] Yang, Z., (2018). Differential Effects of Land Value Taxation. *Journal of Housing Economics*, Vol. 39, 33–39. https://doi.org/10.1016/j.jhe.2017.11.002.

[7] Aditya, T., Santosa, P. B., Yulaikhah, Y., Widjajanti, N., Atunggal, D. and Sulistyawati, M., (2021). Land Use Policy Title Validation and Collaborative Mapping to Accelerate Quality Assurance of Land Registration. *Land Use Policy*, Vol. 109. https://doi.org/10.1016/j.landusepol.2021.105689.

[8] Kementerian ATR/BPN [The Indonesian Ministry of Agrarian Affairs and Spatial Plan], (2022). PTSL Tingkatkan Efektivitas Proses Legalisasi Aset [Complete Systematic Land Registration in Indonesia Improves the Effectiveness of the Asset Legalisation Process]. 2022.

[9] Li, L., Prussella, P. G. R. N. I., Gunathilake, M. D. E. K., Munasinghe, D. S. and Karadana, C. A., (2015). Land Valuation Systems Using GIS Technology Case of Matara Urban Council Area, Sri Lanka. *Bhumi, The Planning Research Journal*, Vol. 4(2). https://doi.org/10.4038/bhumi.v4i2.6.

[10] Janoušková, J. and Sobotovičová, Š., (2019). Fiscal Autonomy of Municipalities in the Context of Land Taxation in the Czech Republic. *Land Use Policy*, Vol. 82(2018), 30–36. https://doi.org/10.1016/j.landusepol.2018.11.048.

[11] Kontrimas, V. and Verikas, A., (2011). The Mass Appraisal of the Real Estate by Computational Intelligence. *Applied Soft Computing Journal*, Vol. 11(1). 443–448. https://doi.org/10.1016/j.asoc.2009.12.003.

[12] Gnat, S., (2021). Property Mass Valuation on Small Markets. *Land*, Vol. 10(4). https://doi.org/10.3390/land10040388.

[13] Tezcan, A., Büyüktaş, K. and Aslan, Ş. T. A., (2020). A Multi-Criteria Model for Land Valuation in the Land Consolidation. *Land Use Policy*, Vol. 95. https://doi.org/10.1016/j.landusepol.2020.104572.

[14] Lin, C. C. and Mohan, S. B., (2011). Effectiveness Comparison of the Residential Property Mass Appraisal Methodologies in the USA. *International Journal of Housing Markets and Analysis*, Vol. 4(3). 224–243. https://doi.org/10.1108/17538271111153013.

[15] Kisworini, I. D., (2022). Land Value Prediction Model in the Urban Fringe. *Journal of Marine-Earth Science Technology*, Vol. 2(3), 28–34. https://doi.org/10.12962/j27745449.v2i3.442

[16] Dziauddin, M. F. and Idris, Z., (2017). Use of Geographically Weighted Regression (GWR) Method to Estimate the Effects of Location Attributes on the Residential Property Values. *Indonesian Journal of Geography*, Vol. 49(1), 97. https://doi.org/10.22146/ijg.27036.

[17] Ghosalkar, N. N. and Dhage, S. N., (2018). Real Estate Value Prediction Using Linear Regression. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*; 1–5. https://doi.org/10.1109/ICCUBEA.2018.8697639.

[18] Kusumawardhani, R. and Budisusanto, Y., (2016). Kajian Nilai Tanah Berdasarkan Harga Pasar Menggunakan Metode Regresi Linier Berganda (Studi Kasus: Kecamatan Gunung Anyar, Surabaya) [Assessment of Land Value Based on Market Price Using Multiple Linear Regression Method (Case Study: Gunung Anyar Sub-District, Surabaya)]. *Jurnal Teknik ITS*, Vol. 5(2), 2–5. https://doi.org/10.12962/j23373539.v5i2.17183.

[19] Zhang, Q., (2021). Housing Price Prediction Based on Multiple Linear Regression. *Scientific Programming*, Vol. 2021. https://doi.org/10.1155/2021/7678931.

[20] Wang, X., Wen, J., Zhang, Y. and Wang, Y., (2014). Real Estate Price Forecasting Based on SVM Optimized by PSO. *Optik*, Vol. 125, 1439–1443. https://doi.org/10.1016/j.ijleo.2013.09.017.

[21] Hjort, A., Pensar, J., Scheel, I. and Sommervoll, D. E., (2022). House Price Prediction with Gradient Boosted Trees under Different Loss Functions. *Journal of Property Research*, Vol. 39(4), 338–364. https://doi.org/10.1080/09599916.2022.2070525.

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

[23] Hong, J., Choi, H. and Kim, W. S., (2020). A House Price Valuation Based on the Random Forest Approach: The Mass Appraisal of Residential Property in South Korea. *International Journal of Strategic Property Management*, Vol. 24(3), 140–152. https://doi.org/10.3846/ijspm.2020.11544.

[24] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. and Lopez, A., (2020). A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends. *Neurocomputing*, Vol. 408, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118.

[25] Gao, Q., Shi, V., Pettit, C. and Han, H., (2022). Property Valuation Using Machine Learning Algorithms on Statistical Areas in Greater Sydney, Australia. *Land Use Policy*, Vol. 123. https://doi.org/10.1016/j.landusepol.2022.106409.

[26] Mountrakis, G., Im, J. and Ogole, C., (2011). Support Vector Machines in Remote Sensing: A Review. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 66(3). 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.01.

[27] Baldomero-Naranjo, M., Martínez-Merino, L. I. and Rodríguez-Chía, A. M., (2021). A Robust SVM-Based Approach with Feature Selection and Outliers Detection for Classification Problems. *Expert Systems with Applications*, Vol. 178. https://doi.org/10.1016/j.eswa.2021.115017.

[28] UC Business Analyst, (2018). Gradient Boosting Machines, UC Business Analytics R Programming Guide. [Online]. Available: http://uc-r.github.io/gbm_regression. [Accessed: Jan. 04, 2023]

[29] Tu, J. V., (1996). Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. *Journal of Clinical Epidemiology*, Vol. 49(11), 1225–1231. https://doi.org/10.1016/S0895-4356(96)00002-9.

[30] Luan, J., Zhang, C., Xu, B., Xue, Y. and Ren, Y., (2020). The Predictive Performances of Random Forest Models with Limited Sample

Size and Different Species Traits. *Fisheries Research*, Vol. 227. https://doi.org/10.1016/j.fishres.2020.105534.

[31] Kontschieder, P., Bulò, S. R., Bischof, H. and Pelillo, M., (2011). Structured Class-Labels in Random Forests for Semantic Image Labelling. *Proceedings of the IEEE International Conference on Computer Vision*; 2190–2197. https://doi.org/10.1109/ICCV.2011.6126496.

[32] Antipov, E. A. and Pokryshevskaya, E. B., (2012). Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a CART-Based Approach for Model Diagnostics. *Expert Systems with Applications*, Vol. 39(2), 1772–1778. https://doi.org/10.1016/j.eswa.2011.08.077.

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

[34] Badan Informasi Geospasial [Indonesia Geospatial Information Agency], (2018). Peta RBI Per Wilayah [Indonesia Topographic Map by Region]. [Online]. Available: https://tanahair.indonesia.go.id/portal-web/. [Accessed: Dec. 15, 2023]

[35] Mayor of Surabaya, (2014). *Rencana Tata Ruang Wilayah Kota Surabaya 2014-2034 [Surabaya City Regional Spatial Plan 2014-2034]*. [Online]. Available: https://jdih.surabaya.go.id/uploads/peraturan/perda_731.pdf. [Accessed: May 12, 2023].

[36] M.Nzau, B., (2003). *Modelling the Influence of Urban Sub - Centres on Spatial and Temporal Urban Land Value Pattern: Case Study of Nairobi, Kenya*. The Netherlands: International Institute for Aerial Survey and Earth Sciences (ITC).

[37] Han, S. S. and Basuki, A., (2001). The Spatial Pattern of Land Values in Jakarta. *Urban Studies*, Vol. 38(10), 1841–1857. https://doi.org/10.1080/00420980120084886.

[38] Waddell, P., Berry, B. J. L. and Hoch, I., (1993). Residential Property Values in a Multinodal Urban Area: New Evidence on the Implicit Price of Location. *The Journal of Real Estate Finance and Economics*, Vol. 7(2), 117–141. https://doi.org/10.1007/BF01258322.

[39] Abdulla, H. M., Ibrahim, M. A. and Al-Hinkawi, W. S., (2023). The Impact of Urban Street Network on Land Value: Correlate Syntactical Premises to the Land Price. *Buildings*, Vol. 13(7). https://doi.org/10.3390/buildings13071610.

[40] Yin, Z., Li, W., Li, C. and Zheng, Y., (2025). The Relationship between Accessibility and Land Prices: A Focus on Accessibility to Transit in the 15-Min City. *Travel Behaviour and Society*, Vol. 38. https://doi.org/10.1016/j.tbs.2024.100914.

[41] Jafary, P., Shojaei, D., Rajabifard, A. and Ngo, T., (2024). Automated Land Valuation Models: A Comparative Study of Four Machine Learning and Deep Learning Methods Based on a Comprehensive Range of Influential Factors. *Cities*, Vol. 151. https://doi.org/10.1016/j.cities.2024.105115.

[42] Standard Appraisal Drafting Committee of Indonesia, (2022). Standar Penilaian Indonesia (SPI) 204 [Indonesian Valuation Standards 204], 204, Jakarta., Oct. 01, 2022.

[43] Cochran, W. G., (1977). *Sampling Techniques*. New York. John Wiley & Sons.

[44] Son, N. P., Van Manh, L., Thuy, N. T., Vu, T. N. and Hanh, N. T., (2020). Factors that Affect Land Values and the Development of Land Value Maps for Strengthening Policy Making in Vietnam: The Case Study of Non-Agricultural Land in Quang Ninh Province, Vietnam. *EQA-International Journal of Environmental Quality*, Vol. 36, 23–35.

[45] Salonen, M., Toivonen, T., Cohalan, J. M. and Coomes, O. T., (2012). Critical Distances: Comparing Measures of Spatial Accessibility in the Riverine Landscapes of Peruvian Amazonia. *Applied Geography*, Vol. 32(2), 501–513. https://doi.org/10.1016/j.apgeog.2011.06.017.

[46] Kara, A., van Oosterom, P., Çağdaş, V., Işıkdağ, Ü. and Lemmen, C., (2020). 3 Dimensional Data Research for Property Valuation in the Context of the LADM Valuation Information Model. *3D Land Administration for 3D Land Uses*, Vol. 98. https://doi.org/10.1016/j.landusepol.2019.104179.

[47] Crompton, J. L. and and Nicholls, S., (2020). Impact on Property Values of Distance to Parks and Open Spaces: An Update of U.S. Studies in the New Millennium. *Journal of Leisure Research*, Vol. 51(2), 127–146. https://doi.org/10.1080/00222216.2019.1637704.

[48] Crosby, H., Damoulas, T., Caton, A., Davis, P., Porto De Albuquerque, J., and Jarvis, S. A., (2018). Road Distance and Travel Time for an Improved House Price Kriging Predictor. *Geospatial Information Science*, Vol. 21(3), 185–

194. https://doi.org/10.1080/10095020.2018.1503775.

[49] Bykowa, E., Hełdak, M., and Sishchuk, J., (2020). Cadastral Land Value Modelling Based on Zoning by Prestige: A Case Study of a Resort Town. *Sustainability (Switzerland)*, Vol. 12(19). https://doi.org/10.3390/SU12197904.

[50] Fitrianingsih, D. and Ghozali, A., (2019). Modeling of Land Prices in Karang Joang, Balikpapan City. *IOP Conference Series: Earth and Environmental Science*, Vol. 313 (1). https://doi.org/10.1088/1755-1315/313/1/012001.

[51] Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F. and Cho, Y. I., (2024). Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. *Mathematics*, Vol. 12(16). https://doi.org/10.3390/math12162553.

[52] Barra, I., Haefele, S. M., Sakrabani, R. and Kebede, F., (2021). Soil Spectroscopy with the Use of Chemometrics, Machine Learning and Pre-Processing Techniques in Soil Diagnosis: Recent Advances–A Review. *TrAC Trends in Analytical Chemistry*, Vol. 135. https://doi.org/10.1016/j.trac.2020.116166.

[53] Yang, J., Rahardja, S. and Fränti, P., (2019). Outlier Detection. *Proceedings of Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, Dec. 19, 2019. ACM. https://doi.org/10.1145/3371425.3371427.

[54] Clark-Carter, D., (2005). Interquartile Range, in *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013192.bsa311.

[55] Abdulkareem, K. H., Mohammed, M. A., Salim, A., Arif, M., Geman, O., Gupta, D. and Khanna, A., (2021). Realizing an Effective COVID-19 Diagnosis System Based on Machine Learning and IoT in Smart Hospital Environment. *IEEE Internet of Things Journal*, Vol. 8(21), 15919–15928. https://doi.org/10.1109/JIOT.2021.3050775.

[56] Hajdu, G., Minoso, Y., Lopez, R., Acosta, M., and Elleithy, A., (2019). Use of Artificial Neural Networks to Identify Fake Profiles. *Proceedings of 2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*,May 2019. IEEE. https://doi.org/10.1109/LISAT.2019.8817330.

[57] Fei, N., Gao, Y., Lu, Z., and Xiang, T., (2021). Z-Score Normalization, Hubness, and Few-Shot Learning. *Proceedings of the IEEE International Conference on Computer Vision*; 142–151. https://doi.org/10.1109/ICCV48922.2021.00021.

[58] Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., Naidan, G., Ochir, C., Legtseg, B., and Byambaa, T., (2019). Evaluation of Random Forest Regression and Multiple Linear Regression for Predicting Indoor Fine Particulate Matter Concentrations in a Highly Polluted City. *Environmental pollution*, Vol. 245; 746–753.

[59] Breiman, L., (2001). Random Forests. *Machine learning*, Vol. 45 (1); 5–32.

[60] Royset, J. O. and Wets, R. J. B., (2021). An Optimization Primer. *Springer Series in Operations Research and Financial Engineering*; 1–668. https://doi.org/10.1007/978-1-4684-9388-7.

[61] IAAO, (2013). Standard on Ratio Studies, Kansas City, Missouri., Apr. 2013.

[62] Węglarczyk, S., (2018). Kernel Density Estimation and Its Application. *ITM Web of Conferences*, Vol. 23; 00037. https://doi.org/10.1051/itmconf/20182300037.

[63] Choy, L. H. and Ho, W. K., (2023). The Use of Machine Learning in Real Estate Research. *Land*, Vol. 12 (4); 740.

[64] Dimopoulos, T., Tyralis, H., Bakas, N. P., and Hadjimitsis, D., (2018). Accuracy Measurement of Random Forests and Linear Regression for Mass Appraisal Models That Estimate the Prices of Residential Apartments in Nicosia, Cyprus. *Advances in Geosciences*, Vol. 45; 377–382. https://doi.org/10.5194/adgeo-45-377-2018.

[65] Alburshaid, E. A. and Ksantini, R., (2024). Real Estate Predictive Analysis for the Kingdom of Bahrain Using Web Scraping and Machine Learning. *Proceedings of 2024 Arab ICT Conference (AICTC)*,*2024 Arab ICT Conference (AICTC)*,*Manama, Bahrain*. Feb. 27, 2024. IEEE. https://doi.org/10.1109/AICTC58357.2024.10735042.

[66] Gala, D. M., Pawar, B., Band, G., Dalal, A., Chakraborty, A., Das, U., and Patil, B. V., (2024). Evaluating the Effectiveness of Random Forest versus Decision Tree Algorithms in Predictive Analytics for Enhancing Sustainable Development Strategies in Real Estate, Infrastructure, and Construction: A Machine Learning Perspective. *Proceedings of 2024 International Conference on Intelligent Systems and Advanced Applications (ICISAA)*,*2024 International Conference on*

Intelligent Systems and Advanced Applications (ICISAA),*Pune, India*. Oct. 25, 2024. IEEE. https://doi.org/10.1109/ICISAA62385.2024.10828645.

[67] Rossini, P. and Kershaw, P., (2008). Automated Valuation Model Accuracy: Some Empirical Testing. *Proceedings of 14th Pacific Rim Real Estate Society Conference*,*Kuala Lumpur, Malaysia*. Jan. 2008. Pacific Rim Real Estate Society.

[68] Nyanda, F., Muyingo, H., and Wilhelmsson, M., (2024). Machine Learning Valuation in Dual Market Dynamics: A Case Study of the Formal and Informal Real Estate Market in Dar Es Salaam. *Buildings*, Vol. 14 (10). https://doi.org/10.3390/buildings14103172.

[69] Ying, X., (2019). An Overview of Overfitting and Its Solutions. *Journal of Physics: Conference Series*, Vol. 1168; 022022. https://doi.org/10.1088/1742-6596/1168/2/022022.

[70] Chuaqui, T. R. C., Rhead, A. T., Butler, R., and Scarth, C., (2021). A Data-Driven Bayesian Optimisation Framework for the Design and Stacking Sequence Selection of Increased Notched Strength Laminates. *Composites Part B*, Vol. 226 (September); 109347. https://doi.org/10.1016/j.compositesb.2021.109347.

[71] Montaño, J., Palmer, A., Sesé, A., and Cajal, B., (2013). Using the R-MAPE Index as a Resistant Measure of Forecast Accuracy. *Psicothema*, Vol. 25; 500–506. https://doi.org/10.7334/psicothema2013.23.

[72] Nurunnisha, G., Sinaga, O., Rohmattulah, A., and maulansyah, risky, (2020). Analysis Of Consumer Acceptance Factors Against Fintech At Bandung SMEs. *PalArch's Journal of Archaeology of Egypt/ Egyptology*, Vol. 17; 841–855.

[73] Mayor of Surabaya, (2021). *Peraturan Daerah (perda) Kota Surabaya Nomor 4 Tahun 2021 Tentang Rencana Pembangunan Jangka Menengah Daerah Kota Surabaya Tahun 2021-2026 [Surabaya City Regional Regulation (perda) Number 4 of 2021 Concerning the Medium Term Development Plan of the City of Surabaya Year 2021-2026]*. [Online]. Available: https://peraturan.bpk.go.id/Details/193589/perda-kota-surabaya-no-4-tahun-2021. [Accessed: Dec. 29, 2023].

[74] Darin-Drabkin, H., (2013). *Land Policy and Urban Growth: Pergamon International Library of Science, Technology, Engineering and Social Studies*. Elsevier.

[75] Harnita, H., Muazzin, M., and Idami, Z., (2019). Tanggung Jawab PPAT dalam Penetapan Nilai Transaksi Jual Beli Tanah dan Bangunan di Kota Banda Aceh [Responsibilities of Land Deed Officials in Determining the Value of Land and Building Sale and Purchase Transactions in Banda Aceh City]. *Jurnal Magister Hukum Udayana (Udayana Master Law Journal)*, Vol. 8 (3); 354–370. https://doi.org/10.24843/JMHU.2019.v08.i03. p05.

[76] Fitriady, E., Effendy, M., and Buana, M. S., (2023). Harga Jual Beli dalam Akta Jual Beli (Ajb) Dikaitkan dengan Pajak Pemungutan Bea Perolehan Hak Atas Tanah Dan Bangunan (BPHTB) [Sale and Purchase Price in the Deed of Sale and Purchase in Relation to the Collection of Tax on Acquisition of Land and Building Rights]. *Notary Law Journal*, Vol. 2 (3), 203–215. https://doi.org/10.32801/nolaj.v2i3.44.

[77] Potrawa, T. and Tetereva, A., (2022). How Much Is the View from the Window Worth? Machine Learning-Driven Hedonic Pricing Model of the Real Estate Market. *Journal of Business Research*, Vol. 144, 50–65. https://doi.org/10.1016/j.jbusres.2022.01.027.

[78] Zhang, G. and Lu, Y., (2012). Bias-Corrected Random Forests in Regression. *Journal of Applied Statistics*, Vol. 39(1), 151–160. https://doi.org/10.1080/02664763.2011.578621.

[79] Contreras, P., Orellana-Alvear, J., Muñoz, P., Bendix, J., and Célleri, R., (2021). Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment. *Atmosphere*, Vol. 12 (2). https://doi.org/10.3390/atmos12020238.

[80] Borup, D., Christensen, B. J., Mühlbach, N. S. and Nielsen, M. S., (2023). Targeting Predictors in Random Forest Regression. *International Journal of Forecasting*, Vol. 39 (2), 841–868. https://doi.org/10.1016/j.ijforecast.2022.02.010.

[81] Wu, Z., Yao, F., Zhang, J. and Liu, H., (2024). Estimating Forest Aboveground Biomass Using a Combination of Geographical Random Forest and Empirical Bayesian Kriging Models. *Remote Sensing*, Vol. 16(11). https://doi.org/10.3390/rs16111859.

[82] Zhu, T., (2020). Analysis on the Applicability of the Random Forest. *Journal of Physics: Conference Series*, Vol. 1607(1). https://doi.org/10.1088/1742-6596/1607/1/012123.