AJLP & GS
African Journal on Land Policy and Geospatial Sciences

e-ISSN : 2657-2664, Vol.6, Issue 4, September 2023

https://revues.imist.ma/index.php/AJLP-GS/index

https://doi.org/10.48346/IMIST.PRSM/ajlp-gs.v6i4.41057 *Category of the manuscript : Articles*

Received in: 22 June 2023 **Revised in:** 23 August 2023 ***Accepted in:*** 27 August 2023

Fine-scale mapping of residential land price using machine learning:
An experimental study in the city dominated by informal land markets.

[1]Gideon T. Marandu, [2]Beatrice Tarimo, [3]Vianey Mushi

[1]Department of Geospatial Science and Technology, Ardhi University, gidesep04@gmail.com[1], Dar es Salaam, Tanzania.

[2]Department of Geospatial Science and Technology, Ardhi University, betarimo@gmail.com[2], Dar es Salaam, Tanzania.

[3]Department of Land Management and valuation, Ardhi University, vianeymushi@gmail.com[3], Dar es Salaam, Tanzania.

ABSTRACT

Fine-scale mapping of residential land price (RLP) is essential to the understanding of residential land market dynamics and improving urban planning. However, cartographic resources and experimental studies to map RLP at fine-scale in Sub-Saharan African cities are limited as a result of dominance of informal land markets in shaping the growth and expansion of most of the cities in the region.

Goal and Objectives:

The study seeks to establish an optimized ensemble machine-learning method for mapping RLP at grid-level in Dar-es-Salaam City, Tanzania.

Methodology:

The study utilizes RLPs collected at the sub-ward level via the survey method and uses open data such as Nighttime Lights (NTL) and amenities coordinates points from OpenStreetMap. This paper explores the ability of two (2) ensemble machine learning methods, i.e. Random Forest Regression (RF-R) and XGBoost Regression, for mapping RLP at grid-level.

Results:

Results illustrate that RF-R was slightly superior to XGBoost Regression and thus it was used to map RLP at fine-scale. The relative importance of explanatory variables in the RF-R model demonstrated that NTL was by far the most important determinant for the RLP spatial distribution in Dar-es-Salaam. NTL it is commonly used as a proxy for socioeconomic factors such as Gross Domestic Product (GDP) and population, hence describing typical characteristics of informal land markets. This is contrary to findings in the global-north with a dominance of formal urban land markets whereby factors such as commercial and educational amenities are found to be very important in estimating RLPs. The paper presents a cost-effective methodological approach for mapping land prices at fine-scale in Dar-es-Salaam city and other cities with similar characteristics in the region, hence improving urban decision and policy-making.

Keywords:

Residential land price, land markets, Sub-Saharan Africa, ensemble machine learning, urban, mapping

1. INTRODUCTION

Recently, highest urbanization rates have been observed in African countries as a result of rural-urban migrations, industrialization, and rapid economic growth (Bhanjee & Zhang, 2018) among other factors. Statistically, in a period of 40 years (2010-2050) the urban population of African countries is expected to triple (Güneralp et al., 2018). Most of this urban growth is expected to occur in Sub-Saharan Africa (Buhaug & Urdal, 2013). Consequently, there is a growing demand for residential land by the population in the urban areas (Martínez-Jiménez et al., 2022). It is suggested that effective land resources management and allocation are essentials for sustainable development of cities (Zhang et al., 2021). On the contrary, it is reported that between 50%-75% of new residential development in Sub-Saharan African cities' have occurred on land acquired through informal channels (Andreasen et al., 2020). Thus, most of purchasing and selling of African urban land are performed out of established legal systems (Chimhowu & Woodhouse, 2006). Hence most of the respective governments lack timely and accurate spatial information on residential land prices (RLPs) and their urban land markets (Kironde, 2000).

In urban studies, accurate mapping of RLPs and its distributions is an essential indicator for monitoring and evaluation of land markets (Y. Chen et al., 2016). As well, it might be used quantitatively to simulate horizontal and vertical urban land use changes (S. Hu et al., 2013). Moreover, it facilitates the manifestation of sub-centers development, hence tracking changes of the city structure from monocentric to polycentric (Zhang et al., 2021). It is further emphasized that the knowledge gained from such urban studies highlight declining and vulnerable areas within the city and as a result it may prompt urban planners to emerge with renewal and mitigation plans respectively for those particular areas of the city (Y. Chen et al., 2016; Zhang et al., 2021). Moreover, since early 20th century the land price maps have been key for effective decision making in the mass appraisal for determination of land rent and property tax (Cellmer et al., 2014). This is viable since the land price contains various background information of sociometric, socioeconomic and environmental features which are essential for the adjustment of tax policy and other relevant policies by the government (S. Hu et al., 2013). Generally, spatial distribution of land price surfaces fundamental information of urban land markets which is essential for investors, land managers, urban planners, urban dwellers, and other stakeholders to make effective decision and sound urban policies.

Nonetheless, the described merits are not yet to be realized and grasped effectively by urban studies and land markets in the Sub-Saharan African cities characterized with weak land governance, inadequate human and financial resources, and deficient institutional capacities (Durand-Lasserve et al., 2013). Consequently, the informal land markets have been observed to be dominant in shaping urban expansion of the respective cities (Andreasen et al., 2020). Thus, fine-scale mapping of RLPs in such context becomes of challenge since there is inadequate official information of land parcels sold at particular time. Contrary, the cartographic resources subject to land price mapping are reported to be rich in global north, particularly in the United States of America (USA) and some European countries (Cellmer et al., 2014), as a result of strong institutions and strict formal land markets which enable official release of spatial information regarding land price at tax lot level,

regularly (J. Ma et al., 2020). Successfully making it possible to spatially model the RLPs at parcel level. Nevertheless, efforts to change the course of land management in sub-Saharan African countries have been witnessed recently through the changes and establishment of new urban land policies and land Acts. For instance, The Valuation and Valuers Registration Act (2016) of Tanzania, explicitly directs the Government Chief Valuer to collect and maintain a valuation data bank. Since then the land price data for different land use have been collected by the respective office in an interval of 2-3 years through manual surveys. The method used is daunting, laborious, and time consuming; moreover, the collected data suffers from coarse spatial resolution (Zhang et al., 2021), for instance in Tanzania land price data are aggregated at village or sub-ward administrative level, as a result missing fine-scale attributes (Y. Chen et al., 2016). Furthermore, the existing studies shows that the African valuation system still suffers from lack of quality land market data and one of the reasons is that the collected area-based land price data hinders equity since they do not consider the locational factors (Cheloti & Mooya, 2021; Kang & Kim, 2022). Consequently, affecting urban decisions in various ways, for instance imposing flat land rent and property tax.

Recently, experimental studies that map land price and house price at fine scale in data challenging environment have been observed. The respective studies established machine learning algorithms (MLAs) and locational values as an accurate approach to predict land prices and house prices (Y. Chen et al., 2016; Derdouri & Murayama, 2020; Jeonghyeon Kim et al., 2022; Zhang et al., 2021). Urban economic theories suggest that land buyers acquire two commodities from a piece of land, that is the piece of land itself and its neighborhood characteristics (locational values) such as amenities abundances, accessibility, socioeconomic status, and topography (Alonso, 1960; Brigham, 1965). Moreover, studies have demonstrated that there is complex non-linear relationship between land price and its locational values (Selim, 2009; Wang et al., 2014). Zhang et al., (2021) stated that MLAs are effective in modeling non-linear relationship existing between land price and its explanatory variables (e.g. the locational values) than traditional statistical methods such as Ordinary Least Square (OLS), and hedonic regression (Selim, 2009); also, linear geostatistical methods, e.g. kriging and cokriging methods (Derdouri & Murayama, 2020). Moreover, MLAs do not apply strict assumptions on data distribution as opposed to traditional statistical methods (Berthold et al., 2010; Zhang et al., 2021). As a result, MLAs are viable for prediction of real estate values (e.g. land price) for fine-scale mapping in Sub-Saharan land markets where there is limited availability of data and there might be high unpredictability of interaction of variables.

Ensemble MLAs have demonstrated high capability in predicting real estate values for fine-scale mapping. Y. Chen et al., (2016) applied ensemble learning via the techniques of bagging and stacking for fine-scale mapping of housing rent using the explanatory variables of Nighttime Lights (NTL), Normalized Difference Vegetation Index (NDVI), and several types of amenities in Guangzhou City. In addition, Yang et al., (2021) demonstrated extra-trees regression (ETR) to perform well in estimating and mapping RLPs at fine-scale, mainly using NTL and several types of amenities in Wuhan city. Whilst, the Random Forest (RF) algorithm demonstrated superiority over other geostatistical methods in spatial prediction of land price, mainly using socioeconomic factors, elevation and land use data in Fukushima district, Japan (Derdouri & Murayama, 2020). Moreover, Jeonghyeon Kim et al., (2022) found that RF was also slightly superior than neural network (NN) model in spatial interpolation of house price with missing transaction records in Seoul, South Korea.

In the same city, Jungsun Kim et al., (2021) compared the performance of two ensemble MLAs, that is RF and XGBoost in predicting the land price, with XGBoost yielding slightly higher accuracy than its counterpart RF. The mentioned studies demonstrate the superiority of using ensemble learning algorithms and open data in estimating real estate values. However, their respective study areas are found in the global north, which might share some common characteristics in their land markets, including interaction of variables used in estimating the respective real estate values and formality of the markets. The results might differ in other land markets, particularly in Sub-Saharan African urban areas which are described differently from the reported studies in the global north. There are still limited studies in Sub-Saharan African cities, particularly Dar es Salaam city in Tanzania that demonstrate the utilization of machine learning methods in predicting and mapping RLPs at fine-scales in which field data is lacking.

Thus, this research paper aims at establishing methodological framework for predicting and mapping RLPs at fine-scale using MLAs at Dar-es-Salaam city, Tanzania, whose expansion is largely contributed by informal land markets. Specifically, it aims at (1) building optimized ensemble machine learning regression models; that is Random Forest Regression (RF-R) and XGBoost Regression (XGBoost-R) (2) Comparing the performance of the RF-R and XGBoost-R in estimating RLPs and (3) Establishing the variable importance used in mapping RLPs at fine-scale for Dar-es-Salaam city. The study utilizes data collected by Chief Valuer through survey methods at sub-ward level, which is a relatively coarse-scale. In addition, it utilizes open data such as NTL raster data, along with several types of amenities (e.g. schools, banks, and clinics) from openStreetMap. The remainder of this paper is structured such that: Section 2 presents the details of study area, data and the methodology, Section 3 presents hyper-parameter tuning results of the adopted models, the results of estimated RLPs distribution at fine-scale and variable importance analysis (predictors) of the best model, Section 4 presents the discussion of the results and their implications, and Section 5 gives the conclusions.

2. MATERIALS AND METHODS

2.1 Study Area

Dar-es-Salaam city, Tanzania is the selected area for this study (*figure 1(a)*). It is a coastal city located in Sub-Saharan Africa region, the eastern part of it is bordered by Indian Ocean where the large port that serves the country and other neighboring land-locked countries is located. It has five (5) administrative municipalities (MCs), that is Kigamboni, Temeke, Kinondoni, Ubungo, and Ilala. The city council is situated in Ilala MC with Central Business District (CBD) located at Kivukoni subward. According to year 2022 census, the city had a population of 5,383,728 people within an area of 1393 km^2. That is an average of 3865 people living within 1 km^2. However, the population is not evenly distributed across the city as might be shown by NTL map of year 2020 (*figure 1(b)*), that is some areas might be denser than other areas. It is stated that the physical shape of the city is mainly determined by its transport infrastructure (Peter & Yang, 2019), with main inland entrance and exit from city defined by three (3) trunk roads namely Bagamoyo road, Kilwa road and Morogoro road. According to city's master plan of 2016-2036, it was projected the city to contain population of about 13 million by 2036 (MLHHSD, 2016). Apparently, the city is urbanizing faster than the planning activities (Bhanjee & Zhang, 2021), consequently up to 75% of residence population dwell in informal

settlements (Andreasen et al., 2020). The dominance of informal land markets, small portion of formal land market, rapid growth of the city, and its economic importance around the region makes the study area unique, eligible, and prompt to experiment fine-scale mapping of RLPs using MLAs.

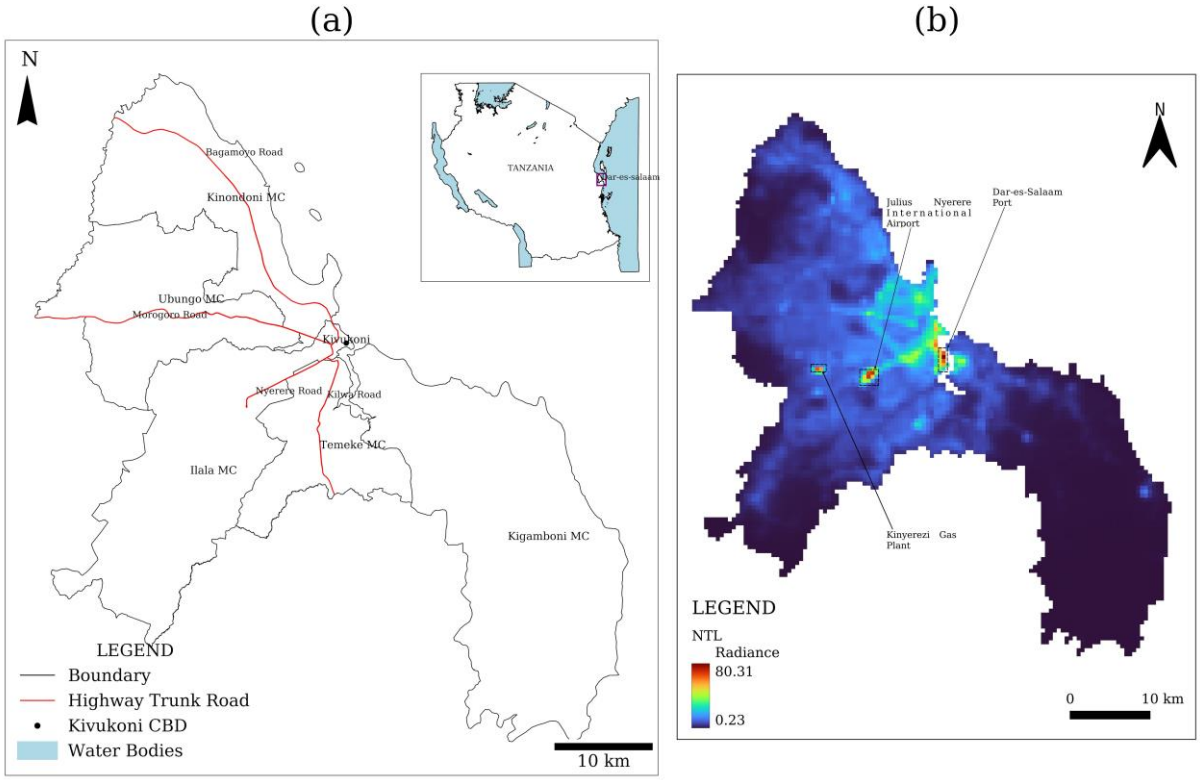

Figure 1: (a) Dar-es-Salaam city and its location in Tanzania and highlighting location of the Central Business District (CBD) of the city (b) The Dar-es-Salaam city NTL distribution for year 2020

2.2 Overall Methodological framework

The overall methodological framework of this study is demonstrated in *figure 2*. It is divided into three (3) steps, (1) Data preparation: the response variable Residential Land Price (RLP) year 2020 data set were gathered from the Ministry of Lands, Housing and Human Settlements Development (MLHHSD), and some explanatory variables from the public accessible repositories. The processed data were aggregated at 500 m X 500 m grid and natural logarithm was applied to all data set. Eventually, the data were randomly split into train set (80%) and test set (20%). (2) Modeling: The machine learning (ML) ensemble regression models that is Random Forest and XGBoost were adopted in this study. The GridSearchCV is used to search for optimal hyper-parameters for each model. Lastly at this stage, the optimal hyper-parameter for each model are used for training and testing the ML models using train set and test set data respectively. (3) The models were evaluated using scientific ML metrics, the better model was selected, variable importance on the best model was investigated, the selected ML model also was used to estimate RLP on the unobserved grids, finally the full fine scale map for RLP of Dar-es-Salaam city was produced to characterize their distribution. Largely, the python scikit-learn ML software (Pedregosa et al., 2011) was used in this study.

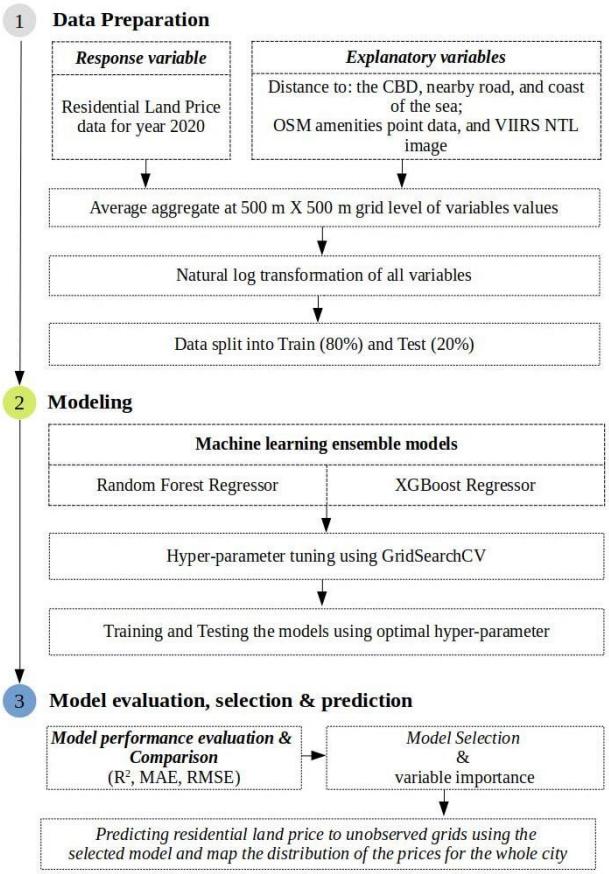

Figure 2: Overall methodological framework

2.3 Data Source and preparation

2.3.1 Residential Land Prices (RLPs)

The RLPs data set for the year 2020 (response variable) were gathered from the Chief Valuer's department placed at Ministry of Lands, Housing and Human Settlements Development (MLHHSD). These RLPs are the result of survey method conducted at the sub-ward (street/village) level after every 2-3 years interval. The RLPs were rendered with defined minimum and maximum price (Tanzania Shillings (TZS)/m2) for each sub-ward in Dar-es-salaam city. Thus, the average RLP for each sub-ward was computed and geocoded using a Dar-es-Salaam administrative sub-ward map from the National Bureau of Statistics (NBS). The map was one of the products of census conducted at year 2012 and had a total 452 sub-wards for the city. Therefore, a total of 452 RLP were geocoded and centroid points for each sub-ward were derived as shown in figure 3(a) with RLP distribution.

Furthermore, a total of 6,941 grids of 500 m X 500 m were generated from the Dar-es-salaam city map. According to the position of a centroid point with the RLP (figure 3(a)), the RLP was assumed and aggregated at that specific grid (figure 3(b)). The mean RLP was calculated for the points which were found in the same grid as a result the total number of RLP mapped into grids were 431 as demonstrated in figure 3(c). Thus, a total of 6,510 grids remain with the unobserved RLP.

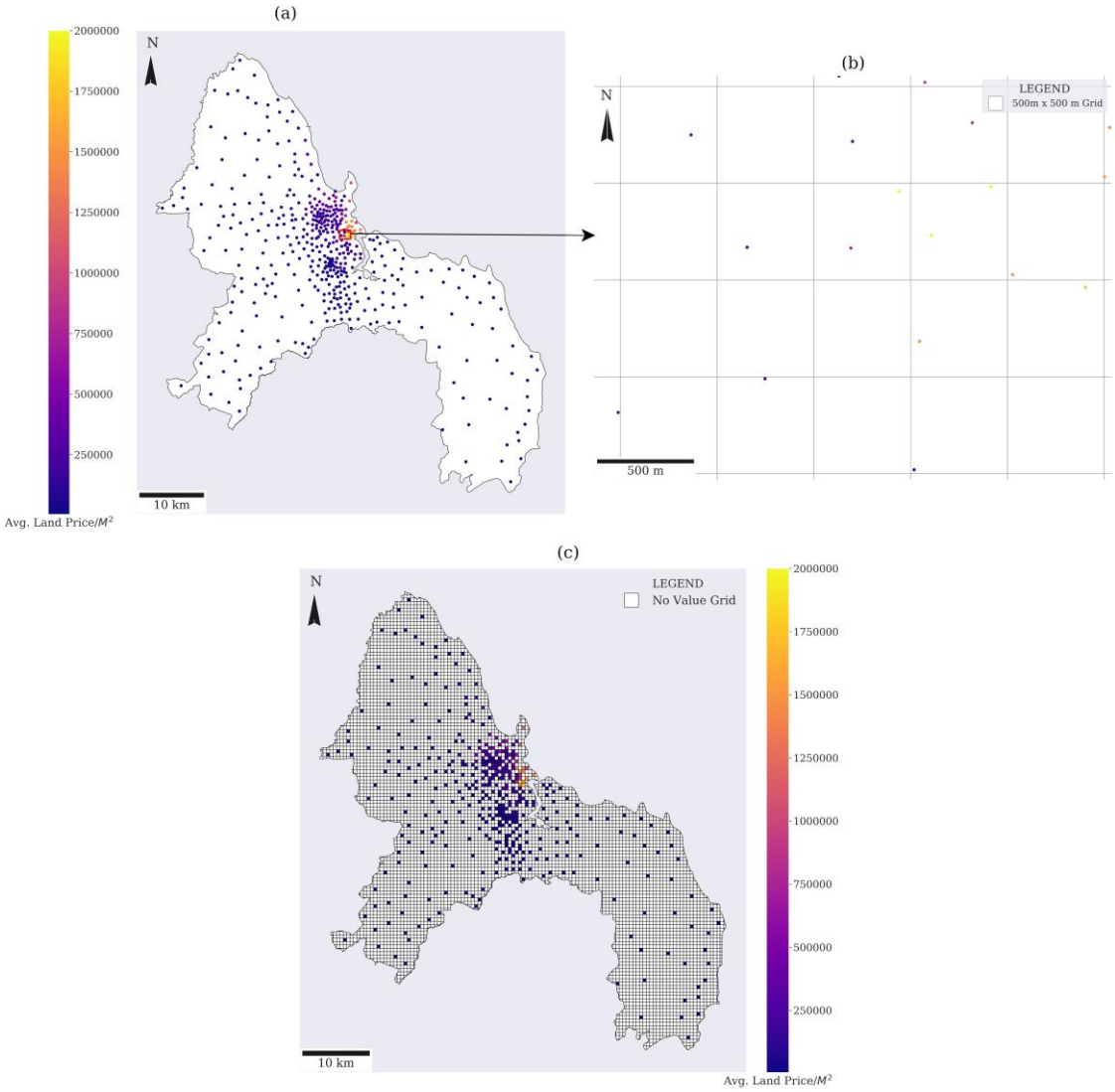

Figure 3: (a) Spatial distribution of RLP for Dar-es-Salaam city for year 2020 (b) sample RLP points as they appear within the grid (c) grids with the average RLP

2.3.2 Explanatory variables

The explanatory variables used in this study were selected according to the relevant existing literature. The description of the explanatory variables, source, and their rationale are presented in Table 1 below.

Table 1: Summary of explanatory variables used in this study and their rationale

Variable	Description	Source	Rationale	Abbreviation
Distance to Central Business District (CBD)	Measured euclidean distance from the center of each grid to the CBD point.		The RLP decrease with distance from the CBD (Alonso, 1960)	log_d_CBD
Night Time Lights (NTL)	NTL monthly satellite images of 500m X 500 m (approx.) spatial resolution for year 2020 were downloaded; one	Academic sector, Colorado School of Mines (https://payneinstitute.mines.edu/eog/)	NTL is spatially useful and consistent as a proxy for socioeconomic factors (e.g population density and Gross	log_NTL

	mean annual composite image was derived from those monthly images and the mean value aggregated to each grid.		Domestic Product (GDP)) which have significant effect on the RLP (Zhang et al., 2021)	
Distance to the sea coast	Measured euclidean distance from the center of each grid to the nearby sea coast		Water body view (e.g sea coast, rivers, and lakes) have significant effect on the RLP (Grimes & Liang, 2009; Jiao & Liu, 2010)	log_d_sea_coast
Distance to the nearby road	Measured euclidean distance from the center of each grid to the nearby road		Proximity to the transport infrastructure such as railways and roads have significant influence on RLP (Derdouri & Murayama, 2020; Morales et al., 2019).	log_d_nearby_road
School amenities Clinic amenities Banks amenities University amenities Public market amenities	Kernel density for amenity point was computed according to Silverman's quartic kernel function (Silverman, 1986) in Quantum GIS (QGIS) environment to represent amenity abundance (Y. Chen et al., 2016; Zhang et al., 2021) and mean density value was aggregated at grid level	All described amenities (school, clinics, banks, university, and markets) coordinate points data were downloaded from the OSM.	Service facilities (school, clinics, etc) availability status within particular area have demonstrated significant influence on RLP (Dong & Wu, 2016; Morales et al., 2019)	log_kschool log_kclinic log_kbank log_kUniv log_kmarket
X_coordinate Y_coordinate	Geographical X and Y coordinates are the Easting and Northing respectively extracted from the centroid points of the grids.		X and Y coordinates are auxiliary variable adopted to account for proximity between grids within the ML ensemble methods adopted which also have considerable influence on RLP (Jeonghyeon Kim et al., 2022)	X_coordinate Y_coordinate

In addition, the amenity data downloaded from OSM were not comprehensive enough to cover the whole administrative area of Dar-es-Salaam city. They were mostly available to the area with relative high price, that is from the CBD to areas few kilometers from the CBD; These are areas which are regarded to be most vibrant areas of the city and contain social-economic data (Bhanjee & Zhang, 2021). Thus, most grids at the out-skirt part of the city had a missing amenity density value. Nevertheless, the k-Nearest Neighbor (kNN) imputer method from the python scikit-learn software (Pedregosa et al., 2011; Troyanskaya et al., 2001), was applied to estimate the missing amenity density values in the respective grids. The number of neighbors (5) and weight points by inverse of their distance were important parameters for the kNN imputer model. The assumption being that the abundance of service amenity tends to decrease from the CBD to outskirt areas of the city in similar way as RLP as shown in *figure 3(a)*.

Furthermore, the explanatory variables which are prefixed with log (see Table 1, abbreviation column) are the ones which natural log was applied to them, including the RLP (response variable). The natural log transformation minimizes the heteroskedasticity problem (Harris et al., 2013), hence ensuring the improvement of prediction power and accuracy of the models.

2.4 Machine learning: Ensemble methods

Ensemble machine learning methods tend to assemble several predictors and combine their outputs to enhance the accuracy of the prediction for classification or regression problem (Y. Chen et al., 2016). Thus, Ensemble methods achieve better accuracy than using individual predictors (Berthold et al., 2010). Moreover, tree-based ensemble regression methods have established a consistent outstanding performance in numerical prediction (Derdouri & Murayama, 2020; Zhang et al., 2021). This study employed and compared proven high-performance tree-based ensemble regression methods, namely Random Forest Regressor (RF-R) and eXtreme gradient boosting Regressor (XGBoost-R) (Jungsun Kim et al., 2021) in predicting and mapping RLP at fine-scale for Dar-es-Salaam city.

RF-R applies multiple decision trees for training and predicting numerical samples (Breiman, 2001; Geurts et al., 2006). Strategically, RF-R used in this study applies bagging method which combines bunch of decisions trees and aggregate the various RLP predictions by averaging, hence controlling the over-fitting and achieving improved accuracy (Jungsun Kim et al., 2021; Zhang et al., 2021). Moreover, during training of RF-R, sampling with replacement (bootstrapping) from original dataset was applied for each decision tree (predictor). In this study, the RF-R algorithm from scikit-learn module was used (Pedregosa et al., 2011).

On the other hand, XGBoost-R applies the boosting method which constructs the decision trees progressively, that is the next predictor is constructed base on the prediction results of the preceded predictor, purposely to correct the error of its predecessor (Berthold et al., 2010; T. Chen & Guestrin, 2016; Jungsun Kim et al., 2021). That is XGBoost-R as a scalable model minimizes over-fitting and bias by applying gradient boosting, regularization, and parallelization when building its decision trees (T. Chen & Guestrin, 2016). As an ensemble model the final prediction is calculated by summing up the results of strong and weak predictors residuals', as a result weak predictor are canceled out by strong predictors to form final accurate estimate of RLP (T. Chen & Guestrin, 2016; Jungsun Kim et al., 2021). In this study, the XGBoost-R was implemented in the scikit-learn python language wrapper module (Pedregosa et al., 2011; Wade, 2020).

2.5 Hyper-parameter tuning

The grid-search cross-validation (GridSearchCV) method from the scikit-learn python module was used to determine and tune the hyper-parameters values for both RF-R and XGBoost-R. The hyper-parameter tuning is of paramount importance to improve the performance of the respective machine learning models (Jungsun Kim et al., 2021; Zhang et al., 2021). The RF-R hyper-parameters tuned in this study were *max_depth* which determine the maximum depth of a tree, *n_estimators* which determines the number of trees (predictors), and the remained hyper-parameters were left with default values. Moreover, the XGBoost-R hyper-parameters tuned were *max_depth*, *n_estimators*, *learning_rate* which prevents the over-fitting of the model by shrinking the feature weights after every boosting step, subsample which will determine the ratio of the original training dataset for preventing over-fitting and subsampling is performed once in every boosting iteration; the remained hyper-parameters of XGBoost-R were left with default values (See *Table 2*).

2.6 Model evaluation

The final dataset with 431 observations (*Section 2.3.1*) were randomly split into training set (80%)

and into testing set (20%). Thereafter, the RF-R and XGBoost-R models were trained using optimized hyper-parameters obtained (*Section 2.5*). Subsequently testing the models using the test set data in order to observe the generalization ability of the models (Berthold et al., 2010). Moreover, the models were evaluated using the following score functions: coefficient of determination (R2), mean absolute error (MAE), and the root mean squared error *subsample = 0.9* (RMSE) as described in the following equations.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

From above equations, n is the total number of grids with observed RLP, y_i and $\hat{y}_i$ are observed and predicted RLPs respectively for the i^{th} grid, and $\bar{y}$ represents the mean value for the observed RLPs.

3. RESULTS

3.1 Results of Hyper-parameter tuning

The optimal hyper-parameters configurations for the Random Forest Regressor (RF-R) and XGBoost Regressor (XGBoost-R) searched by GridSearchCV method are presented in *Table 2*. The combination of optimal hyper-parameter for particular model was found from the defined values and ranges. Comparatively, the RF-R had a deeper tree (*max_depth = 9*) and larger number of trees (*n_estimators = 200*) than XGBoost-R which attained *max_depth = 6* and *n_estimators = 100*, hence RF-R was more complex than XGBoost-R in terms of individual tree structure and number of predictors. Furthermore, XGBoost-R achieved better accuracy with optimal *learning_rate = 0.1* which is crucial for weighting each estimator involved. Moreover, XGBoost-R achieved also an optimal *subsample = 0.9*, which implies 90% of the data were iteratively selected from the original training dataset and fitted in the individual trees (predictors).

Table 2: Hyper-parameter tuning results

Model	Hyper-parameter	Values and ranges	Optimal Hyper-parameter
RF-R	max_depth	[2, 3, 4, 5, 6, 7, 8, 9, 10]	9
	n_estimators	[100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]	200
XGBoost-R	max_depth	[2, 3, 4, 5, 6, 7, 8, 9, 10]	6
	n_estimators	[100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]	100
	learning_rate	[0.01, 0.05, 0.1, 0.2]	0.1
	subsample	[1, 0.9, 0.5, 0.2, 0.1]	0.9

3.2 Performance of Random Forest Regressor and XGBoost Regressor models

The results for RF-R and XGBoost-R for training and testing are presented in *Table 3*. During training the XGBoost-R had a prediction accuracy of 99.82% (R2) which is slight better performance than that of RF-R (98.01%). In addition, XGBoost-R attained lower prediction error, and its MAE and RMSE score are 0.0413 and 0.0537, respectively. However, testing results show that the RF-R had slightly higher accuracy (91.71%) in estimating RLP than XGBoost-R (90.83%). Moreover, the RF-R had lower prediction error than XGBoost-R on testing data set, and its MAE and RMSE score are 0.3093 and 0.4298, respectively. The actual and predicted RLP log values of the testing data-sets for both models are demonstrated in *figure 4*. Apparent, the RF-R (*figure 4(b)*) predicted outcomes were more in line with their actual values than of XGBoost-R (*figure 4(a)*). Generally, testing results show that RF-R controlled the over-fitting of the model, hence more generalization capacity than XGBoost-R. Thus, RF-R was chosen to predict the unobserved grids.

Table 3: Training and Testing results for RF-R and XGBoost-R

Data Sets	Model	R^2	MAE	RMSE
Training	RF-R	0.9801	0.1304	0.1776
	XGBoost-R	0.9982	0.0413	0.0537
Testing	RF-R	0.9171	0.2928	0.4085
	XGBoost-R	0.9083	0.3093	0.4298

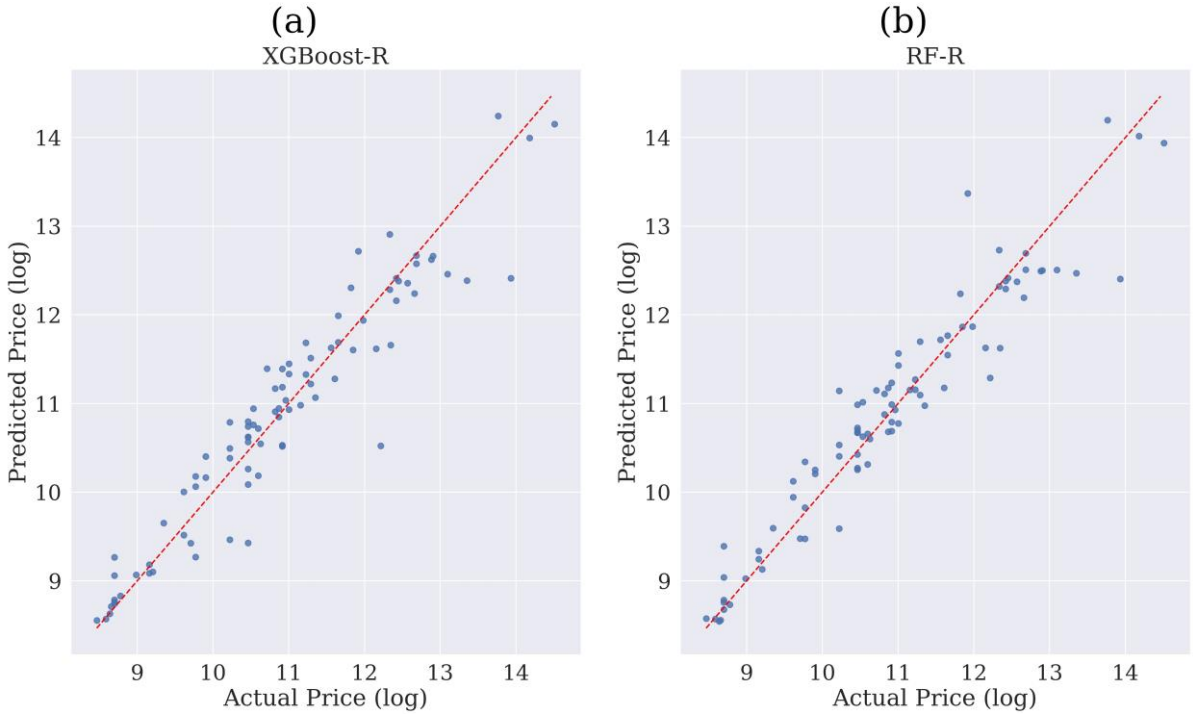

Figure 4: Plot for the results of the XGBoost-R (a) and RF-R (b) for the testing data-sets.

3.3 Variable Importance for the Random Forest Regressor model

Relatively, the importance of predictor variables in RF-R model are demonstrated in *figure 5*. The top five (5) important variables in the model with their score are (1) NightTime lights (log_NTL) – 0.718, (2) Distance to the CBD (log_d_CBD) - 0.076, (3) Y_coordinate – 0.061, (4) Distance to the sea coast (log_d_sea_coast) – 0.044, and (5) Bank amenities (log_kbank) – 0.031. The least important variable

in the respective model was distance to the nearby road (log_d_nearby_road) with the score of 0.008. Apparent, NTL was the most influential variable in estimating the RLP in Dar-es-Salaam city.

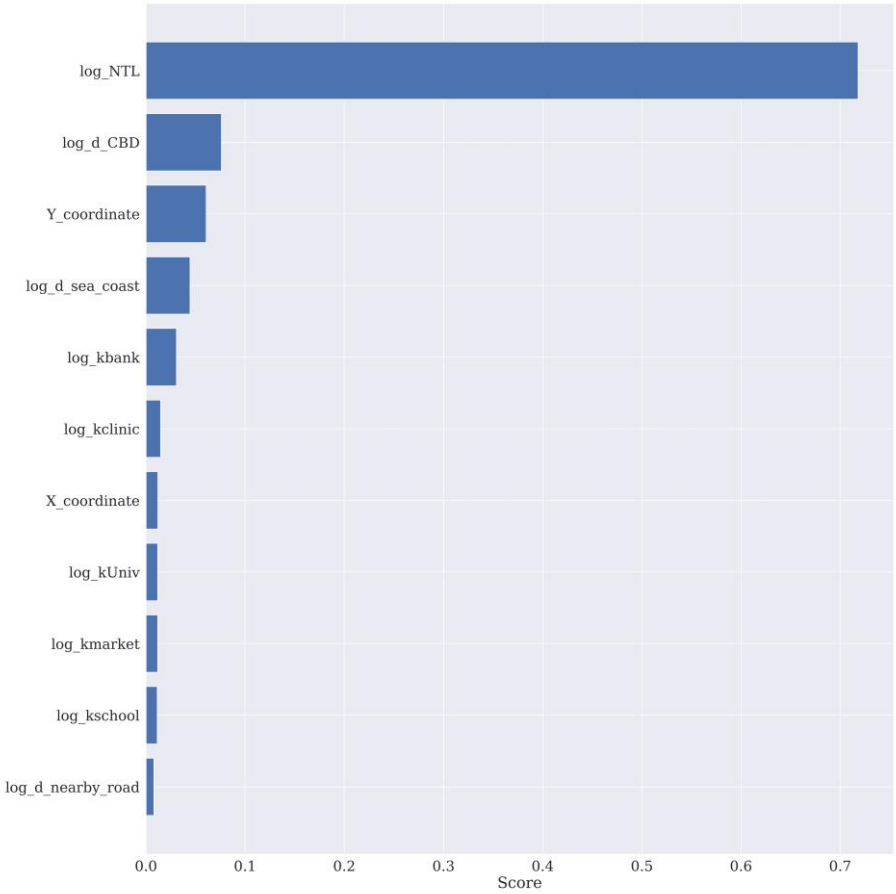

Figure 5: RF-R variable importance

3.4 Mapping estimated value to unobserved grids

Ultimately, the RF-R was used to estimate the RLP of the unobserved grids i.e. grids with no data values. *Figure 6(a)* demonstrate the Dar-es-Salaam city map with the complete RLP mapped grids and its spatial distribution. Apparently, the model was able to estimate the RLPs whilst maintain the diminishing pattern from the city-center to the out-skirt part of the city just like the original data-set, hence achieving spatial consistency. At the CBD and nearby areas where the high RLPs are clustered the model was successfully able to maintain the pattern (*figure 6(b)*). However, moving away from the clustered high prices it is observed that there are some discrete jumps of prices; low and relatively high prices. That could suggest that there is heterogeneity mixture of urban land development, probably with some new emerging centers. Thus, the model was able to capture the non-linear relationship that exist between the RLPs and the explanatory variables involved in the research.

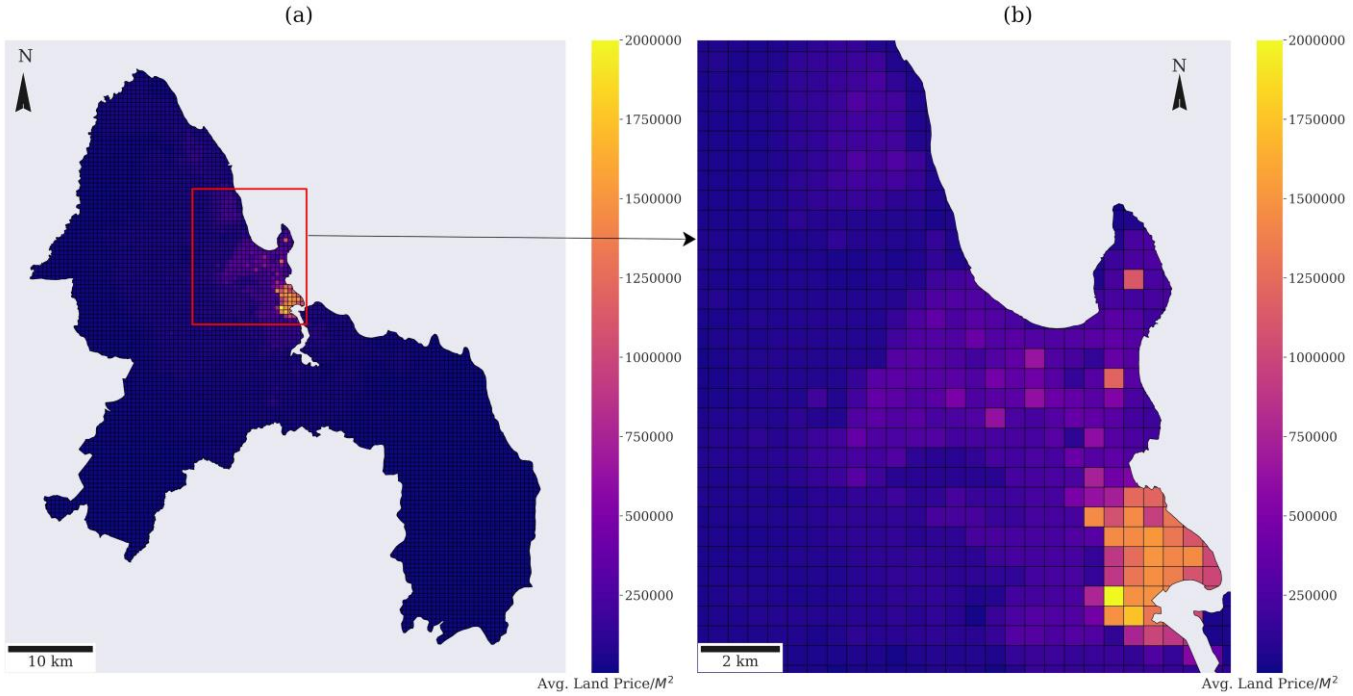

Figure 6: Estimated RLPs of Dar-es-Salaam city at grid level (a) full mapped of the city (b) mapped area at CBD and area around with high economic activities

4. DISCUSSION

The results of XGBoost-R and RF-R have expanded knowledge of application of machine learning ensemble methods in their capability of modeling and estimating urban land price, particularly in the environment that is largely dominated by informal land markets than formal land markets. Generally, the XGBoost-R and RF-R models all performed well in estimating the land price. This is consistent with the similar studies that applied ensemble learning methods in estimating urban land price (Y. Chen et al., 2016; Jungsun Kim et al., 2021; Zhang et al., 2021). However, the RF-R achieved a better result on testing data-set than XGBoost-R, hence good generalization capacity. The reason for XGBoost-R to be outsmarted by RF-R during testing might be due to its sensitivity in dealing with the outliers (Jungsun Kim et al., 2021). The outliers are subject to existence of land sub-markets (L. Hu et al., 2022), largely might be fueled by informal transactions (Andreasen et al., 2020) as a result causing the variations and heterogeneity of the RLP in city.

On the variable importance, vividly, the NTL is the most important variable with the largest score by far than other variables, this might be due to its spatial consistency in explaining the urban dynamics in terms of urban development, that is NTL are consistent and efficient proxy measure for socioeconomic and demographic variables such as GDP and population (T. Ma et al., 2015) which are very important factors in determining urban land price (Kheir & Portnov, 2016). However, its importance in this study does not scale with other studies conducted in global north which describe variables such as education amenities, natural amenities (e.g. water bodies), and public transportation as the most important in determining RLP (Zhang et al., 2021). Contrary, the results of this study have demonstrated services such as schools, clinic, universities, markets, and distance to nearby road (transport) as of least importance. This might be due to the nature of dominating

informal markets of which large population of people prefer to buy fringe unplanned land in out-skirt areas of the city without consideration of availability of mentioned services, and upon request the government formal services follow them in post-settlement era (Andreasen et al., 2020). In fact, the literature report that the public services and amenities, and transport infrastructures supplied are sub-standard, and some places are lack (Peter & Yang, 2019). Some places the land designated for the respective services were encroached, unlawfully turned to other use (W. J. Kombe, 2005). Nonetheless, for the government to supply formal services in post-settlement might imply an intermediate step of transforming informal land markets into formal land markets, particularly to the peri-urban areas (Chimhowu & Woodhouse, 2006). Thus, the model was able to capture the characteristics of the Dar-es-salaam city land market successfully.

In addition, the strength and importance of NTL variable in modeling the RLP is the manifestation that socioeconomic factors such as GDP and population are core drivers in Dar-es-Salaam city land market. Literature suggest that there is growing of inequalities as a result of urban economic growth of African cities (Obeng-Odoom, 2013). This is consistency with the study conducted by Kombe, (1994) where the existence of socioeconomic inequalities within the emergence of informal land markets in Tanzania was observed. Moreover, one of the consequences of the development of informal land markets is that those with high bargain power tend to be in the advantage side than those with lower bargain power, this affects also the amount land that can be afforded, as a result the omission of poorer people in the city land market potentially will continue as the city expands (Chimhowu & Woodhouse, 2006; J. W. Kombe, 1994). Thus, it is recommended for the urban management to come-up with significant pro-poor intervention plan and policies to reduce the dominance of rich and mighty in the land market, consequently mitigating the socioeconomic and political inequality (Obeng-Odoom, 2013).

Moreover, the use of spatial variables has enabled the RF-R to estimate the RLP accurately at fine spatial scale. This technological approach might revolutionize the traditional land appraisal methods (Kang & Kim, 2022) for land taxation and rent estimation, and land compensation for urban development in Tanzania and Sub-Saharan African countries as whole. Land taxation and rent in the city might be estimated correctly, timely, and context-specific at fine spatial scale in lieu of flat rate, as a result the reported problem of government to lack an up-to-date land market information for respective decision making will be alleviated (Kelly, 2004). This is possible since the spatial variables used in this study such as NTL are updated regularly (Zhang et al., 2021). Thus, the machine learning models like RF-R can be adopted to interpolate and extrapolate the land price, accordingly. Hence, improving the revenue collection of the government as a result building and improving transport infrastructures and service facilities to part of the cities with deficiencies (Obeng-Odoom, 2013). Moreover, improving land management and real estate policies.

The chief valuer office can adapt the methodological framework proposed in this study since it has demonstrated fruitful results in estimating RLP at fine scale spatial resolution. Currently, the RLP data are collected at sub-ward level using survey method, only a total of 452 sub-wards had estimated RLP. However, with the proposed approach, the study managed to estimate and map RLP to the 6,941 grids of 500m x 500m. Thus, the generalization of RLP scale was improved from large area (sub-ward) to small area (grids). Hence, reducing the cost, time, and laborious work of collecting

RLP at the particular fine spatial resolution (Zhang et al., 2021). This is a merit not only to Tanzania's cities but also to other developing sub-Saharan Africa countries which are have common problems and challenges in land management and administration described as inadequate human and financial resources, weak governance, and limited institutional capacities (Durand-Lasserve et al., 2013).

The study has several limitations which open the door for future studies and change of practices. First, the assumption of sub-ward centroid points with land price being aggregated at particular grid as the average price of the grid might have some unobserved effects on the results of this study. Thus, it is suggested to the government chief valuer office to include absolute location attribute (x and y coordinates) when collecting the land price data particularly for the transaction done on plots instead of current practice which generalize the land price at sub-ward level (Y. Chen et al., 2016; Zhang et al., 2021). Alternatively, by considering the nature of African land markets which largely is informal, the chief valuer office may adopt the suggested methodology of this study and start collecting land price by sampling the grids. Secondly, previous studies stated that 60%-75% of residential development in Sub-Saharan African cities occurs on land acquired through informal channels (Andreasen et al., 2020; Durand-Lasserve et al., 2013). However, this study has not spatially depicted the areas and grids which were influenced by such nature of land acquisitions. The future study can address this limitation by sampling the grids, hence expanding the knowledge of the Dar-es-Salaam land markets.

5. CONCLUSIONS

This study has established an ensemble machine learning model for fine scale mapping of RLP in Dar-es-Salaam city, one of the cities of Sub-Saharan Africa characterized with informal land markets dominance and fastest sporadic urban growth. Thus, the methodological approach adopted in this study and the results are essential for the land management departments and real estate stakeholders to make informed robust decision, land policy, and resource allocation in an environment faces the challenges of inadequate financial and human resource, lack of timely and local land market information, poor institutional capacities and weak land governance. This study manifested the utility of two (2) ensemble machine learning model, that is XGBoost-R and RF-R to map RLPs at fine scale using open data such as amenity points from openStreetMap and NTL. The RF-R with its unique ability of dealing with noise data demonstrated high accuracy during testing, hence high generalization ability. Thus, RF-R was used to map RLPs at fine scale. In context, the NTL variable which act as a proxy for socio-economic variables such as GDP and population was marked with highest importance than other variables involved. Its importance reflects the characteristics of the land markets for most cities in Sub-Saharan Africa which their markets are informal, that is the existence of socio-economic inequalities which largely affect the choices and demand of land as well the city expansion. Unlike formal land markets in global north which are largely driven by availability and accessibility of commercial and educational amenities. Therefore, the interaction of factors in ensemble machine learning models in the Sub-Saharan African cities land market do differ from those of global north. The intervention is needed from the relevant authorities to improve the urban life of Dar-es-Salaam city by minimizing urban inequality including improving the public services, hence the study has various policy implications.

Generally, this empirical study presents an innovative and cost-effective methodological approach for modeling and mapping land price at fine scale in Dar-es-Salaam city and other cities with similar characteristics in Tanzania, and other Sub-Saharan African countries. The use of open data such as NTL which are effective and are updated regularly present sustainable opportunity to the land managers and other real estate stakeholders to monitor land price and their city land markets for various purposes.

6. ACKNOWLEDGMENT

We thank the Ministry of Lands, Housing and Human Settlements Development (MLHHSD), Tanzania for their provision of land price data, and their cooperation and support during the conduct of this research.

7. FUNDING

This research was supported by DAAD; programme name: In-Country/In-Region Scholarship Programme Tanzania, 2018 (Funding ID: 57429566).

8. AUTHORS' CONTRIBUTIONS

Conceptualization, methodology, data collection, analysis, visualization, and original draft preparation, Gideon T. Marandu ; Supervision, review and editing, Beatrice Tarimo ; Supervision, and proof-reading, Vianey Mushi.

9. REFERENCES

Alonso, W. (1960). A Theory of the Urban Land Market. *Papers and Proceedings of the Regional Science Association*, *6*.

Andreasen, M. H., McGranahan, G., Kyessi, A., & Kombe, W. (2020). Informal land investments and wealth accumulation in the context of regularization: case studies from Dar es Salaam and Mwanza. *Environment and Urbanization*, *32*(1), 89–108. https://doi.org/10.1177/0956247819896265

Berthold, M. R., Borgelt, C., Höppner, F., & Klawonn, F. (2010). *Guide to Intelligent Data Analysis:How to Intelligently Make Sense of Real Data* (D. Gries & F. B. Schneider (eds.)). Springer London.

Bhanjee, S., & Zhang, C. H. (2018). Mapping Latest Patterns of Urban Sprawl in Dar es Salaam, Tanzania. *Papers in Applied Geography*, *4*(3), 292–304. https://doi.org/10.1080/23754931.2018.1471413

Bhanjee, S., & Zhang, S. (2021). Do urban planning and sprawl affect social vulnerability ? An assessment of Dar es Salaam. *Development Southern Africa*, *38*(2), 189–207. https://doi.org/10.1080/0376835X.2020.1818549

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

Brigham, E. F. (1965). The Determinants of Residential Land Values. *Land Economics*, *41*(4), 325–334. https://www.jstor.org/stable/3144665

Buhaug, H., & Urdal, H. (2013). An urbanization bomb ? Population growth and social disorder in cities. *Global Environmental Change*, *23*(1), 1–10. https://doi.org/10.1016/j.gloenvcha.2012.10.016

Cellmer, R., Belej, M., Zrobek, S., & Šubic Kovač, M. (2014). Urban Land Value Maps-A Methodological Approach. *Geodetski Vestnik*, *58*(3), 535–551. https://doi.org/10.15292/geodetski-vestnik.2014.03.535-551

Cheloti, I., & Mooya, M. (2021). Valuation problems in developing countries: A new perspective. *Land*, *10*(12). https://doi.org/10.3390/land10121352

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *ArXiv.Org*. https://doi.org/10.1145/2939672.2939785

Chen, Y., Liu, X., Li, X., Liu, Y., & Xu, X. (2016). Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning. *Applied Geography*, *75*, 200–212. https://doi.org/10.1016/j.apgeog.2016.08.011

Chimhowu, A., & Woodhouse, P. (2006). Customary vs private property rights? Dynamics and trajectories of vernacular land markets in sub-Saharan Africa. *Journal of Agrarian Change*, *6*(3), 346–371. https://doi.org/10.1111/j.1471-0366.2006.00125.x

Derdouri, A., & Murayama, Y. (2020). A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across Fukushima prefecture, Japan. *Journal of Geographical Sciences*, *30*(5), 794–822. https://doi.org/10.1007/s11442-020-1756-1

Dong, G., & Wu, W. (2016). Schools, land markets and spatial effects. *Land Use Policy*, *59*, 366–374. https://doi.org/10.1016/j.landusepol.2016.09.015

Durand-Lasserve, A., Durand-Lasserve, M., & Selod, H. (2013). *A systemic analysis of land markets and land institutions in West African cities. Rules and practices: The case of Bamako, Mali* (Issue November). http://econ.worldbank.org.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees [Article]. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Grimes, A., & Liang, Y. (2009). Spatial determinants of land prices: Does Auckland's Metropolitan urban limit have an effect? *Applied Spatial Analysis and Policy*, *2*(1), 23–45. https://doi.org/10.1007/s12061-008-9010-8

Güneralp, B., Lwasa, S., Masundire, H., Parnell, S., & Seto, K. C. (2018). Urbanization in Africa: Challenges and opportunities for conservation. *Environmental Research Letters*, *13*(1). https://doi.org/10.1088/1748-9326/aa94fe

Harris, R., Dong, G., & Zhang, W. (2013). Using contextualized geographically weighted regression to model the spatial heterogeneity of land prices in beijing, china. *Transactions in GIS*, *17*(6), 901–919. https://doi.org/10.1111/tgis.12020

Hu, L., He, S., & Su, S. (2022). A novel approach to examining urban housing market segmentation: Comparing the dynamics between sales submarkets and rental submarkets. *Computers, Environment and Urban Systems*, *94*. https://doi.org/10.1016/j.compenvurbsys.2022.101775

Hu, S., Cheng, Q., Wang, L., & Xu, D. (2013). Modeling land price distribution using multifractal IDW interpolation and fractal filtering method. *Landscape and Urban Planning*, *110*(1), 25–35. https://doi.org/10.1016/j.landurbplan.2012.09.008

Jiao, L., & Liu, Y. (2010). Geographic Field Model based hedonic valuation of urban open spaces in Wuhan, China. *Landscape and Urban Planning*, *98*(1), 47–55. https://doi.org/10.1016/j.landurbplan.2010.07.009

Kang, S. H., & Kim, B. J. (2022). Designing a Valuation System for Property Tax: The Case of Zanzibar, Tanzania. *Land*, *11*(7). https://doi.org/10.3390/land11070989

Kelly, R. (2004). Property Rates in Tanzania. In *International Handbook of Land and Property Taxation*. https://doi.org/https://doi.org/10.4337/9781845421434.00023

Kheir, N., & Portnov, B. A. (2016). Economic , demographic and environmental factors affecting urban land prices in the Arab sector in Israel. *Land Use Policy*, *50*, 518–527.

Kim, Jeonghyeon, Lee, Y., Lee, M.-H., & Hong, S.-Y. (2022). A Comparative Study of Machine Learning and Spatial Interpolation Methods for Predicting House Prices. *Sustainability (Switzerland)*, *14*(15). https://doi.org/10.3390/su14159056

Kim, Jungsun, Won, J., Kim, H., & Heo, J. (2021). Machine-learning-based prediction of land prices in Seoul, South Korea. *Sustainability (Switzerland)*, *13*(23), 1–14. https://doi.org/10.3390/su132313088

Kironde, L. J. M. (2000). Understanding land markets in African urban areas: The case of Dar es Salaam, Tanzania. *Habitat International*, *24*(2), 151–165. https://doi.org/10.1016/S0197-3975(99)00035-1

Kombe, J. W. . (1994). The Demise of Public Urban Land Management and the Emergence of Informal Land Markets in Tanzania. *Habitat International*, *18*(1), 23–43.

Kombe, W. J. (2005). Land use dynamics in peri-urban areas and their implications on the urban growth and form: The case of Dar es Salaam, Tanzania. *Habitat International*, *29*(1), 113–135. https://doi.org/10.1016/S0197-3975(03)00076-6

Ma, J., Cheng, J. C. P., Jiang, F., Chen, W., & Zhang, J. (2020). Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques. *Land Use Policy*, *94*. https://doi.org/10.1016/j.landusepol.2020.104537

Ma, T., Zhou, Y., Zhou, C., Haynie, S., Pei, T., & Xu, T. (2015). Night-time light derived estimation of spatio-temporal characteristics of urbanization dynamics using DMSP/OLS satellite data. *Remote Sensing of Environment*, *158*, 453–464. https://doi.org/10.1016/j.rse.2014.11.022

Martínez-Jiménez, E. T., Le Gallo, J., Pérez-Campuzano, E., & Aguilar Ibarra, A. (2022). The effects of land price in the peri-urban fringe of Mexico City: Environmental amenities for informal land

parcel purchasers. *Urban Studies*, *59*(1), 222–241. https://doi.org/10.1177/0042098020960968

MLHHSD. (2016). *Dar es Salaam City Master Plan 2016-2036*.

Morales, J. A., Flacke, J., & Zevenbergen, J. (2019). Modelling residential land values using geographic and geometric accessibility in guatemala city. *Environment and Planning B: Urban Analytics and City Science*, *46*(4), 751–776. https://doi.org/10.1177/2399808317726332

Obeng-Odoom, F. (2013). The State of African Cities 2010: Governance, inequality and urban land markets. *Cities*, *31*, 425–429. https://doi.org/10.1016/j.cities.2012.07.007

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peter, L. L., & Yang, Y. (2019). Urban planning historical review of master plans and the way towards a sustainable city: Dar es Salaam, Tanzania. *Frontiers of Architectural Research*, *8*(3), 359–377. https://doi.org/10.1016/j.foar.2019.01.008

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*, 2843–2852. https://doi.org/10.1016/j.eswa.2008.01.044

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* [Book]. Chapman and Hall.

The United Republic of Tanzania. (2016). *The Valuation and Valuers Registration Act*. http://parliament.go.tz/polis/uploads/bills/acts/1480402487-SHERIA 7-THE VALUATION AND VALUERS REGISTRATION ACT.pdf

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Wade, C. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python*. Packt Publishing Ltd.

Wang, X., Wen, J., Zhang, Y., & Wang, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. *Optik*, *125*(3), 1439–1443. https://doi.org/10.1016/J.IJLEO.2013.09.017

Yang, R., Ren, F., Ma, X., Zhang, H., Xu, W., & Jia, P. (2021). Explaining the longevity characteristics in china from a geographical perspective: A multi-scale geographically weighted regression analysis. *Geospatial Health*, *16*(2). https://doi.org/10.4081/gh.2021.1024

Zhang, P., Hu, S., Li, W., Zhang, C., Yang, S., & Qu, S. (2021). Modeling fine-scale residential land price distribution: An experimental study using open data and machine learning. *Applied Geography*, *129*. https://doi.org/10.1016/j.apgeog.2021.102442

10. KEY TERMS AND DEFINITIONS

Ensemble Machine Learning: In the context of this study, the methods involve assembling multiple predictors and combining their outputs to significantly enhance prediction accuracy for a regression problem.

Land Markets: In the context of this study, they exist when and wherever land is treated as a commodity, and rights in land are traded.

Mapping: In the context of this study, is the process of creating an interface for the visualization of geospatial data for the purpose of supporting specific information access and exploratory activities.

Residential Land Price: In the context of this study, it is the average price per square meter at which land rights can be traded for residential purposes.

Urban: In the context of this study, it refers to a city and its suburbs.