



CORRIGENDUM TO:

Research Article

Real Estate Appraisal Performance Improvement by Adapting a Hybrid Model: Geographically Weighted Regression and Extreme Gradient Boosting in Al Bireh, Palestine

Jamal A.A. Numan¹, Izham Mohamad Yusoff^{2,*}

¹*Institute of Postgraduate Studies, Universiti Sains Malaysia (USM), 11800 USM Penang, Malaysia*

²*Geography Section, Transdisciplinary Research on Environmental Science, Occupational Safety and Health, Universiti Sains Malaysia, 11800 USM Penang, Malaysia*

*Corresponding author Jamal A.A. Numan: jamalnuman@student.usm.my; Tel.: +6046532271; Fax: 6046576000
Should be written *Corresponding author Izham Mohamad Yusoff: izham@usm.my; Tel.: +604-6532271; Fax: 604 - 6576000



Research Article

Real Estate Appraisal Performance Improvement by Adapting a Hybrid Model: Geographically Weighted Regression and Extreme Gradient Boosting in Al Bireh, Palestine

Jamal A.A. Numan ^{1,*}, Izham Mohamad Yusoff ²

¹Institute of Postgraduate Studies, Universiti Sains Malaysia (USM), 11800 USM Penang, Malaysia

²Geography Section, Transdisciplinary Research on Environmental Science, Occupational Safety and Health, Universiti Sains Malaysia, 11800 USM Penang, Malaysia

*Corresponding author: jamalnuman@student.usm.my; Tel.: +6046532271; Fax: 6046576000

Abstract: Real estate appraisal functions as a decision support system in pivotal financial, economic, and business processes, encompassing buying and selling houses, property tax, bank lending, and insurance companies. Therefore, improving appraisal performance, among other daunting challenges such as interpretability, stability, generalizability, data availability, and evaluation metrics, has been a consistent and substantial demand, attracting attention from academia and stakeholders over the years. The original contribution of this research to knowledge is the adaptation of a hybrid model that combines Geographically Weighted Regression (GWR) with Extreme Gradient Boosting (XGBoost)—termed the hybrid GWR-XGBoost model—using Cook's Distance (Cook's D), distinguishing this research from other studies aimed at improving performance. The rationale behind this hybrid model is to simultaneously address the nonstationarity and nonlinearity aspects inherent in the real estate industry to improve performance. This aim is achieved by identifying features influencing real estate appraisal, particularly apartments within residential buildings, in the context of Al Bireh city, Palestine, designing the hybrid GWR XGBoost model, and evaluating its performance. The findings indicate that, across five commonly used statistical performance evaluation metrics—namely MSE, RMSE, MAE, MAPE, and R^2 —the hybrid GWR-XGBoost model significantly outperforms both GWR and XGBoost, achieving an 11% improvement in R^2 . The implications of this research are threefold: first, features influencing residential real estate appraisal are identified, serving as a reference checklist for stakeholders, including buyers, sellers, developers, appraisers, and investors. Second, enhanced appraisal performance aids in making well-informed financial, economic, and business decisions, ultimately affecting various aspects of people's lives. Lastly, within Palestine, this study lays the groundwork for exploring advanced model-based methods as an alternative to traditional manual methods for real estate appraisal. Overall, improving appraisal performance facilitates informed, evidence-based decision-making, maximizes benefits in economic and financial transactions, and impacts both individuals and their environments policymakers.

Keywords: GWR; Hybrid GWR-XGBoost model; Linear regression; Nonstationarity; Nonlinearity; Real estate appraisal; XGBoost

1. Introduction

Real estate appraisal, also termed property valuation or house pricing, serves as a crucial decision-support tool used in a wide range of important financial, economic, and business transactions, including property selling and buying, bank loans, insurance companies, property taxation, property transfers, partnership dissolutions, expropriations, settlements, auctions, and other related activities (Droj et al., 2024, Jin et al., 2024, Oust et al., 2023, Sisman and Aydinoglu, 2022, Mankad, 2022, Alzain et al., 2022, Steurer et al., 2021, Xu and Zhang, 2021). At the individual level, for instance, purchasing an apartment often represents one's largest lifetime investment, underscoring the importance of appraisal tools to ensure accurate valuations and informed decision-making (Cheng and Tsai, 2022; Bogin and Shui, 2020; Oust et al., 2020).

Typically, methods used for real estate appraisal, whether individual or mass, are divided into non-model-based and model-based categories (Guliker et al., 2022). The non-model-based methods encompass approaches such as comparative, cost, income, profit, and development (Sisman and Aydinoglu, 2022; Gabrielli and French, 2021). The model-based methods are divided into three subcategories: supervised, unsupervised, and reinforcement, which can be parametric, semi-parametric, or non-parametric (Potrawa and Tetereva, 2022). These are further categorized into traditional or modern models. Linear Regression (LR) and Geographically Weighted Regression (GWR) are examples of traditional model-based approaches. Modern models primarily refer to artificial intelligence (AI) techniques, encompassing two branches: machine learning and deep learning. Examples of machine learning models include Decision Trees (DT), Random Forest (RF), Gradient Boosting (GB), Adaptive Boosting (AdaBoost), Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), as well as methods like K-nearest Neighbor (KNN) and Support Vector Machine (SVM), while Artificial Neural Networks (ANN) serve as examples of deep learning models.

The focus on model-based methods, aimed at enhancing their performance, has garnered considerable attention from both academic circles and relevant stakeholders. This emphasis encompasses addressing challenges such as interpretability, stability, reliability, scalability, flexibility, simplicity, adaptability, applicability, generalizability, comprehensibility, data availability, and evaluation metrics. (Chen et al., 2024b; Elnaeem Balila and Shabri, 2024; Hoxha, 2024; Hurley and Sweeney, 2024; Mathotaarachchi et al., 2024; Song and Ma, 2024; Geerts and De Weerd, 2023; Oust et al., 2023; Stang et al., 2023; Hoang and Wiegatz, 2023). In response, and given the absence of hybrid models in the field of real estate appraisal that combine Geographically Weighted Regression (GWR), a nonstationarity-aware linear regression model, with Extreme Gradient Boosting (XGBoost), a nonlinearity-aware model, this research aims to fill this gap.

The rationale behind implementing the hybrid GWR-XGBoost model, which combines GWR, a nonstationarity-aware model, with XGBoost, a nonlinearity-aware model, is to address both nonlinearity and nonstationarity inherent in the real estate sector simultaneously. Nonstationarity, a form of spatial heterogeneity, is addressed explicitly by incorporating GWR into the design of the hybrid model. It arises from the fact that the relationship between real estate features and their appraisals varies from one location to another, in alignment with Tobler's First Law of Geography, which states that nearby events are more correlated than distant ones (Deppner and Cajias, 2024; Li and Niu, 2022; Koohpayma and Argany, 2021). Meanwhile, nonlinearity, addressed by XGBoost, stems from the observation that the relationship between real estate features and their appraisal values is inconsistent across the entire dataset (Guliker et al., 2022). For example, a property's appraisal may decrease with an older construction date, while in other cases, it may increase due to heritage value (Soltani et al., 2024).

To achieve the research aim of improving appraisal performance by adapting the hybrid GWR-XGBoost model, three research questions are posed: What features influence real estate appraisal within Al Bireh city? How can the hybrid GWR-XGBoost model be designed to improve performance? How can the model's performance be evaluated? The contribution of this research lies not only in adapting the hybrid GWR-XGBoost model for the first time in the real estate industry but also in the design of the hybrid model itself. Improving real estate appraisal performance has been

shown to directly impact the ability to make well-informed decisions across a wide spectrum of business processes at both the individual and policymaker levels. (Trindade Neves et al., 2024).

The structure of this paper is outlined as follows. Section 1 presents the introduction, detailing the research focus, significance, research questions, and implications. Section 2 is dedicated to the literature review, highlighting key remarks and observations that identified the gap in implementing the hybrid GWR-XGBoost model within the real estate domain, along with recent studies that have successfully employed GWR and XGBoost models individually. Section 3 describes the methodology employed in conducting the research. Section 4 provides an overview of the dataset. The analysis is conducted in Section 5, with results presented in Section 6 and the discussion in Section 7. Finally, the summary and conclusions are presented in Section 8.

2. Literature review

Given the critical importance of real estate appraisal values in various economic activities and financial decisions—including property sales, purchases, mortgages, and insurance improving its performance has been a consistent and prominent demand for decades. This demand, along with addressing other significant challenges such as interpretability, stability, reliability, scalability, flexibility, simplicity, adaptability, applicability, generalizability, comprehensibility, data availability, and evaluation metrics, has received significant attention from both academia and stakeholders (Mathotaarachchi et al., 2024; Song and Ma, 2024; Hurley and Sweeney, 2024; Elnaeem Balila and Shabri, 2024; Hoxha, 2024; Jin et al., 2024; Chen et al., 2024b; Stang et al., 2023; Hoang and Wiegratz, 2023; Lahmiri et al., 2023; Çilgin and Gökçen, 2023; Zhan et al., 2023; Geerts and De Weerd, 2023; Oust et al., 2023; Das et al., 2021). Given the absence of a universal model that consistently outperforms others across all settings and contexts of real estate appraisals (Elnaeem Balila and Shabri, 2024; Chen et al., 2024a, Hoxha, 2024; Stang et al., 2023; Mora-Garcia et al., 2022), numerous studies have been conducted to address this challenge with the goal of enhancing performance. This research represents a new attempt within these ongoing efforts through the adaptation of a hybrid model that combines GWR and XGBoost in the field of real estate appraisal. The notion of adapting the hybrid GWR-XGBoost model emerged from multiple observations and remarks in the literature. First, it is inspired by a study conducted by Wang et al. (2022), where the proven superiority of the ANN model led to combining it with GWR to improve appraisal performance. Their findings indicated that the performance of the GWR-ANN surpassed that of GWR alone, though information on the performance of ANN alone was lacking. Secondly, Hoxha (2024) highlighted that the effectiveness of machine learning models can be enhanced when combined with GWR, allowing both the nonstationarity and nonlinearity aspects of real estate to be addressed. Additionally, Mathotaarachchi et al. (2024) noted that the field of real estate appraisal still requires hybrid models that blend advanced approaches with conventional ones. Furthermore, while XGBoost, a machine learning model that addresses nonlinearity has demonstrated superiority in real estate appraisal (Çilgin and Gökçen, 2024; Jafary et al., 2024; Mathotaarachchi et al., 2024; Sharma et al., 2024; Trindade Neves et al., 2024; Chen et al., 2023; Çilgin and Gökçen, 2023; Oust et al., 2023; Zhang et al., 2023; Stang et al., 2023; Guliker et al., 2022; Hjort et al., 2022; Iban, 2022; Jin et al., 2022; Zaki et al., 2022; Vargas-Calderón and Camargo, 2022; Li et al., 2021; Taecharungroj, 2021; Xu and Li, 2021), it has not yet been combined with GWR, the most commonly used nonstationarity model (He et al., 2024; Lee, 2024; Yang et al., 2023; Kopczewska, 2022; Li and Niu, 2022). This is despite the fact that both nonstationarity (Renigier-Bilozor et al., 2022, Guliker et al., 2022) and nonlinearity (Guliker et al., 2022) are acknowledged as inherent aspects of the real estate industry. This gap in studies combining GWR and XGBoost presents an opportunity for this research to enhance real estate appraisal performance by developing a hybrid GWR-XGBoost model.

While the primary focus is on real estate literature, the identified gap concerning adapting a hybrid GWR-XGBoost model warrants further exploration to assess whether this model has been successfully applied in other fields. Evidence suggests that similar hybrid models, though designed

differently, have shown success in areas like soil contamination (Ye et al., 2023) and air pollution (Fan et al., 2020). Specifically, Ye et al. (2023) combine GWR and XGBoost by using the intercept and coefficients from GWR as inputs for XGBoost, while Fan et al. (2020) employ XGBoost based on the bandwidth concept derived from GWR. Despite the positive outcomes observed in these studies, Deppner and Cajias (2024) caution that a model's success in one field does not necessarily ensure similar results in others due to unique characteristics, conditions, interactions, and dynamics inherent to each domain. This caution is particularly pertinent in real estate, widely recognized as one of the least transparent, most complex, heterogeneous, sentiment-driven, with high variability, and volatile sectors globally (Trindade Neves et al., 2024; Krämer et al., 2023a; Xu and Zhang, 2023; Jin et al., 2022, Wahid et al., 2022). Consequently, this affirms that the identified gap remains valid.

In essence, the extant literature supports the proposition that this research provides novel insights to the field. The novelty of the research lies in its adaptation of the hybrid GWR-XGBoost model, which integrates GWR and XGBoost in real estate appraisal through the application of Cook's D to enhance performance. The hybrid model design aligns with the definition of a hybrid system as outlined by Özögür Akyüz et al. (2023), which states that the outputs of one model serve as inputs for another. In this study, GWR functions as the initial model, XGBoost serves as the second model, and Cook's D is an output feature derived from the GWR model that is utilized as an input for the XGBoost model.

Although the gap is clearly defined, examining how GWR and XGBoost are applied independently in real estate appraisal is beneficial, as demonstrated in two recent studies. Valdez Gómez de la Torre and Chen (2024) conducted an analysis using GWR and LR with a dataset of 11,446 houses and 17 features, targeting house prices in Quito, Ecuador. Their findings reveal that GWR significantly outperforms LR, suggesting that the local model, GWR, is more effective than global models like LR. Trindade Neves et al. (2024) utilized XGBoost, RF, AdaBoost, and ANN in Lisbon, Portugal, working with a dataset of 22,470 properties and 25 features, focusing on the transacted price as the target variable. XGBoost demonstrated superior performance compared to all other models when tested on an unseen dataset representing 20% of the original data size.

3. Methodology

The research methodology for adapting the hybrid GWR-XGBoost model to improve real estate appraisal performance is guided by research questions that include identifying features influencing real estate appraisal within Al Bireh city, designing the hybrid GWR-XGBoost model, and evaluating its performance, as elaborated in the following sections.

3.1. Features Identification

Features identified as influencing real estate appraisal have a direct impact on its performance (Aydinoglu and Sisman, 2024; Hoxha, 2024; Oust et al., 2023; Glumac and Des Rosiers, 2021). Four main approaches are commonly used to achieve this. The first approach involves the recognition of pertinent features from existing studies in the literature, termed the literature-based approach (Metzner and Kindt, 2018). The second method, an extension of the first, entails refining the features identified from the literature through the outcomes of a questionnaire survey, referred to as the literature and questionnaire-based approach (Numan and Yusoff, 2024b; Mohamed et al., 2023; Yap et al., 2019; Kamaruzzaman et al., 2018). The third method initiates with extracting features from the literature but refines them through data analysis; this is labeled as the literature-data analysis approach (Lawal Dano, 2023; Hong et al., 2020). The fourth method directly extracts features from an appropriate and sufficient dataset without the prerequisite of initially identifying them from the literature, known as the data-driven approach (Yilmazer and Kocaman, 2020).

In this study, the third approach is utilized, beginning with identifying 30 features drawn from a literature review, local laws and regulations, and the author's expertise. The emphasis is on features commonly referenced in the literature, comply with local laws and regulations, and are pertinent to the Al Bireh city context. These features include area, construction date, block, number

of apartments in the building, number of apartments on each floor, number of floors in the building, floor level, floor-to-parcel ratio, number of bedrooms, number of bathrooms, number of toilets, number of balconies, number of facades, central heating availability, elevator availability, parking availability, storage availability, adjacent street type, adjacent street width, adjacent street aesthetics, number of adjacent streets, wall construction material, view, proximity to hospitals, proximity to schools, proximity to the city center, proximity to main roads, proximity to area C, proximity to colonies, and the appraisal value as a target variable. This set of potential features guides the data collection process to focus on aspects relevant to real estate appraisal. Subsequently, the set undergoes further refinement through data analysis using Shapley Additive Explanations (SHAP) within the XGBoost model, with the goal of selecting the top features from those potentially identified.

3.2. Hybrid Model Design

Since the hybrid GWR-XGBoost model combines GWR with the XGBoost model using Cook's D, the technical structure of these two component models is thoroughly explained in the following sections, laying the groundwork for the design of the hybrid model. Additionally, the LR model is discussed in detail, as it serves as a reference model.

3.2.1. LR

The LR constructs the real estate appraisal equation based on the relationship between the real estate features, which represent the independent variables, and their appraisal values, which represent the target variable, contained in a dataset of observations. The intercept and coefficients of the features in the LR equation are determined using arithmetic techniques such as Ordinary Least Squares (OLS). The shortcoming of LR is that it does not address the nonlinearity and nonstationarity issues of real estate data (Ryu et al., 2024). However, LR serves as a baseline and reference for comparing other models. The LR model is mathematically represented as:

$$y = \beta_0 + \sum_{k=1}^m \beta_k x_k + \varepsilon \quad (1)$$

where y is the predicted value (representing the real estate appraisal or predicted price), β_0 is the intercept that stands for independent features absent in the model (corresponding to the predicted value when all other features equal zero), β_k is the coefficient of the k^{th} independent feature, x_k is the value of the observation for the k^{th} independent feature, ε is the error, and m is the total number of independent features.

3.2.2. GWR

The GWR, first developed by Brunsdon et al. (1996), constructs a linear appraisal equation at a specific location, referred to as the prediction point, by capturing the relationship between the weighted target variable, which represents the appraisal value, and the real estate features within the data subset that fall within the bandwidth. The weights for the target variable observations are determined by their distance from the prediction point. In this sense, GWR adheres to Tobler's first law of geography, which asserts that closer observations have a greater impact than those farther away (Yacim and Boshoff, 2019). This implies that GWR accounts for nonstationarity while remaining linear (Guliker et al., 2022; Fan et al., 2020). Mathematically, the GWR model equation is written as follows:

$$y = \beta_0^i + \sum_{k=1}^m \beta_k^i x_k^i + \varepsilon_i \quad i: 1, 2, \dots, n \quad (2)$$

where y is the prediction at the i^{th} location, β_0^i is the intercept at the i^{th} location representing the prediction in case all other independent features are zero, β_k^i is the coefficient of the k^{th} independent feature at the i^{th} location, x_k^i is the value of the k^{th} independent feature at the i^{th} location, ε_i is the error of the prediction at the i^{th} location, k is the independent feature, m is the total number of independent features, and n is the total number of observations within the bandwidth of the i^{th} prediction point.

In addition to the basic results such as the intercept and coefficients offered by GWR, it is essential to analyze supplementary diagnostic indicators like local R^2 , residuals, and Cook's D, which evaluates the influence of each data point on the evaluation and is computed using the following equation:

$$D_i = \frac{\sum_{i=1}^{n-1} (\hat{Y}_i - \hat{Y}_i)^2}{m * MSE} \quad (3)$$

where D_i is the Cook's D of the i^{th} observation, \hat{Y}_i is the predicted value of the i^{th} observation derived from using the model constructed with all observations, \hat{Y}_i is the predicted value of the i^{th} observation using the model constructed from all the observations except the i^{th} observation, m is the number of features used in the models, and MSE is the mean squared error.

3.2.3. XGBoost

The XGBoost, developed by [Chen and Guestrin \(2016\)](#), is a tree-based machine learning model that captures the relationship between the features and their appraisal by training sequential trees on a dataset, where each successive tree aims to minimize the prediction error of the previous tree. The first tree is trained on the actual values of the target variable, while the successor trees are trained on the prediction error (residual) produced by the preceding tree. The mathematical representation of the XGBoost model is as follows:

$$\hat{y}_i = \sum_{p=1}^P f_p(x_i) \quad (4)$$

where \hat{y}_i is the prediction of the i^{th} observation, $f_p(x_i)$ is the prediction of the i^{th} observation in the p^{th} tree, and P is the total number of trees. The objective function to be optimized is given by

$$Obj(\theta) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{p=1}^P \Omega(f_p) \quad (5)$$

where $Obj(\theta)$ is the objection function, L is the loss function, and Ω is the regularizing item that prevents overfitting and complexity of the models. $\Omega(f_p)$ is defined as:

$$\Omega(f_p) = \gamma T + \frac{1}{2} \lambda W \quad (6)$$

where T is the number of leaves, γ is a parameter that controls T , w is the weight of each leaf, and λ is a parameter that controls w .

There are multiple hyperparameters that define how XGBoost works. Some common candidate values for the basic hyperparameters of XGBoost are outlined in Table 1. These hyperparameters can be fine-tuned by choosing an optimization strategy such as grid search, a validation method such as k-fold cross-validation, and a statistical metric like R^2 ([Alzubi, 2024](#)).

Table 1 A list of typical XGBoost hyperparameters, their definitions, and candidate values

Hyperparameter	Definition	Candidate values
n_estimators	Number of iterations	50, 100, 150
subsample	Number of observations (as a percent)	0.8, 0.9, 1.0
colsample_bytree	Number of columns (as percent)	0.8, 0.9, 1.0
Learning_rate	Learning rate	0.01, 0.1, 0.2
max_depth	Number of levels of a tree	3, 4, 5
lambda	Regularization item	0.1, 0.17, 0.2
alpha	Regularization item	9.0, 9.5, 10.0

To improve model performance and interpretability, it is common practice to select the top influential features from those potentially identified. The SHAP technique, developed by [Lundberg and Lee \(2017\)](#), is an agnostic approach used to assess the individual contributions of features to a model's prediction. The SHAP value for a specific feature is derived by aggregating the weighted

differences in predictions across all conceivable subsets of features, encompassing both the inclusion and exclusion of that particular feature. Mathematically, the SHAP formula is expressed as follows:

$$\varphi_k = \sum_{q=1}^Q \frac{|s_q|! (m! - |s_q| - 1)!}{m!} [f(X_{s_q \cup k}) - f(X_{s_q \setminus k})] \quad (7)$$

where φ_k is the SHAP value of the k^{th} feature, s_q is the q^{th} subset of feature, $|s_q|$ is the number of features in the q^{th} subset, "!" denotes the mathematical factorial, $X_{s_q \cup k}$ is the data containing the q^{th} subset of features including the k feature, $X_{s_q \setminus k}$ is the data containing the q^{th} subset of features excluding the k feature, f is the model, q is a subset of features, Q is the total number of all possible feature subsets for the k^{th} feature, and m is the total number of features.

3.2.4. The Hybrid GWR-XGBoost Model

There is no single, unique form for designing a hybrid GWR-XGBoost model. Two design variations of this hybrid model have already been implemented in different sectors: soil contamination (Ye et al., 2023) and air pollution (Fan et al., 2020). Uniquely, the hybrid model in this research is designed by integrating GWR and XGBoost through Cook's D.

Once potential features affecting real estate appraisal are identified, the corresponding data is gathered, preprocessed, and integrated with spatial data into a single dataset. XGBoost is then trained on the entire dataset, considering all 30 features, with the appraisal value as the target variable, along with a set of hyperparameters and k -fold cross-validation. The next section elaborates on the predicted appraisal value is evaluated based on five metrics. Typically, to reduce computational time, optimize hardware resources, and potentially improve performance (Theng and Bhojar, 2024; Chanasit et al., 2021), the number of features is reduced to select only the most influential ones. To accomplish this, the SHAP technique within XGBoost is utilized. With these top features selected, XGBoost is trained solely on them, with the expectation that performance remains approximately identical while minimizing computational time and hardware resource usage.

Subsequently, GWR is trained on the selected top influential features, with Cook's D calculated as a diagnostic measure. Following this, XGBoost is re-trained on these top features, incorporating Cook's D as an additional feature. The inclusion of Cook's D aims to account for the nonstationarity captured by the GWR model, as well as to provide an indication of outlier observations. Examining the inputs of Cook's D formula presented in Equation (3), it is clear that it considers the difference between the prediction of each observation under two scenarios: one where the GWR model includes all observations, and another where the GWR model excludes the specific observation for which Cook's D is being calculated. This indicates that the higher the Cook's D value, the greater the likelihood of the observation being an outlier. By training XGBoost on both the top influential features and Cook's D, the model effectively addresses both nonstationarity and nonlinearity.

Regarding software, LR and GWR are run in ESRI ArcGIS Pro 3.0.2, which includes off-the-shelf tools designed to handle these two models. Conversely, XGBoost is executed using open-source Python programming language libraries such as pandas, sklearn, numpy, xgboost, and matplotlib, which are imported into the JupyterLab environment provided by Anaconda.

3.3. Model Evaluation

The empirical study conducted by Steurer et al. (2021), in which five models are evaluated by 48 metrics clearly indicates that judging model performance depends on which metrics are utilized. They found out that each model is superior according to at least one metric. This research assesses the hybrid GWR-XGBoost model's performance based on five metrics detailed in Table 2.

Table 2 The list of most common statistical evaluation metrics to measure the accuracy of prediction models. e_i is the error in the prediction of the i^{th} observation which is the difference between the actual value and predicted value, Y_i is the actual value of the i^{th} observation of the target variable, \hat{Y}_i is the predicted value of the i^{th} observation of the dependent variable, \bar{Y} is the average of the actual values of the target variable for all observations, and n is the number of observations. [Sharma et al. \(2024\)](#), [Krämer et al. \(2023b\)](#), [Steurer et al. \(2021\)](#)

Metric	Definition	Equation
Mean Squared Error (MSE)	This metric involves averaging the square of the errors, magnifying larger errors while minimizing smaller ones. This means that its value is strongly affected by outliers.	$\frac{\sum_{i=1}^n e_i^2}{n}$ $e = Y_i - \hat{Y}_i$
Root Mean Squared Error (RMSE)	This measure involves taking the square root of the MSE to bring it back to the scale of the target variable.	$\sqrt{\frac{\sum_{i=1}^n e_i^2}{n}}$
Mean Absolute Error (MAE)	This metric involves averaging the absolute error which, treating all errors equally. This means that its value is not affected by outliers	$\frac{\sum_{i=1}^n e_i }{n}$
Mean Absolute Percentage Error (MAPE)	This measure involves averaging the absolute error relative to the actual value, leading to different treatment for errors of similar magnitude. Overestimated values lead to a higher MAPE, while underestimated values lead to a lower value.	$\frac{\sum_{i=1}^n \left(\frac{ e_i }{Y_i} \right)}{n}$
Coefficient of Determination (R^2)	This metric measures the percentage of the squared sum of errors relative to the squared sum of errors in the mean, subtracted from 1.	$1 - \frac{\sum_1^n e_i^2}{\sum_1^n (Y_i - \bar{Y})^2}$

The first four metrics—MSE, RMSE, MAE, and MAPE—are error-based, indicating that lower values reflect better model predictions, whereas the fifth metric, R^2 , represents the goodness-of-fit and model explanatory power, with higher values denoting improved model predictions ([Hoxha, 2024](#); [Soltani et al., 2024](#); [Soltani and Lee, 2024](#); [Gunes, 2023](#); [Torres-Pruñonosa et al., 2021](#)). To demonstrate performance improvement, the metrics of the hybrid model are compared with those of its component models: GWR and XGBoost, as well as with LR, which serves as a reference model. Additionally, a comparative analysis is conducted with previous studies that employed a similar hybrid model in different domains, such as those by [Ye et al. \(2023\)](#) and [Fan et al. \(2020\)](#), applied in soil contamination and air pollution, respectively. In general, statistical evaluation metrics can be computed for three subsets of the original dataset: training, testing, and validation. In this research, only the k-fold cross-validation technique applied to the entire dataset is used to evaluate performance, due to the relatively small data size available for analysis ([Horvath et al., 2021](#)).

3.4. Conceptual Framework

To comprehensively detail the study, Figure 1 presents a conceptual framework outlining the process. The research identifies potential features influencing real estate appraisal to guide data collection. Non-spatial data is collected and stored in an Excel file in tabular format, followed by typical preprocessing steps to remove missing values, duplicates, and invalid entries ([Arvianti, 2021](#)). The cleaned data is then combined with spatial data using Geographic Information System (GIS) tools, resulting in a single dataset saved in one Excel file. After this, XGBoost is trained on the full dataset with assumed values for basic hyperparameters. The SHAP technique within XGBoost

is applied to select the top influential features from those initially identified. To address multicollinearity among the top-selected features, Spearman's rank correlation coefficient is checked.

All symbols and abbreviations in Figure 1 adhere to previously established definitions. The i is the observation number, y_i is the value of the target variable of the i^{th} observation, x_i^k is the value of the k^{th} feature for the i^{th} observation, N is the number of all observations, and m is the number of all features. Symbols indicated with an apostrophe (') represent the raw data before undergoing the cleaning process, $\hat{y}_i(\text{GWR})$ is the predicted value of the i^{th} observation using the GWR model, e is the residual, and $\hat{y}_i(\text{XGBoost})$ is the predicted value using the XGBoost

Afterward, GWR is trained on the dataset containing only the top influential features, with Cook's D calculated as a diagnostic measure. The choice of Cook's D over other GWR outputs—such as intercept, coefficients, predictions, residuals, and local R^2 —as a feature to transfer nonstationarity awareness into XGBoost, a nonlinearity-aware model, is based on trial and error to determine which GWR-derived outputs can enhance XGBoost performance. Subsequently, the dataset containing the top influential features, along with Cook's D , is fed into XGBoost to generate predictions. The predictive performance of the GWR-XGBoost model is assessed in three ways: using statistical metrics, comparing it with its component models, and evaluating it against similar hybrid models implemented in other sectors.

4. Data

4.1. Study Area

The study area takes place in Al Bireh, a prominent city situated in the center of the West Bank within Palestine. Al Bireh is located in close proximity to Ramallah, which currently functions as the provisional political capital of Palestine. Additionally, Al Bireh is located approximately 15 kilometers north of Jerusalem, the occupied capital of Palestine. The geographic coordinates of Al Bireh are $31^\circ 54' \text{ N}$ and $35^\circ 12' \text{ E}$. The city encompasses an approximate total area of 22 square kilometers and is administered by the Palestinian National Authority, in accordance with the Oslo Accords agreement. Its urban master plan and municipal boundary cover an area of around 10 square kilometers. The Palestinian Central Bureau of Statistics (PCBS) reports that the population size of Al Bireh in 2017 was approximately 45,000 people. The city has a number of schools, hospitals, and governmental departments, and the presence of important central governmental departments has contributed to the growth of the real estate market. Al Bireh is a dynamic and diverse city that hosts cultural and artistic events and has a thriving commercial sector. Birzeit University, a well-respected institution, is located 10 km from Al Bireh. The city of Al Bireh is divided into 29 blocks. The study area is a mix of residential and commercial zones, and it contains a variety of building types such as multi-story buildings, detached houses, and amenities such as schools and hospitals. It also has main roads and important governmental institutions. The boundary of the study area, along with the locations of the collected dataset of apartment buildings, area C, and colonies, are depicted in Figure 2.

4.2. Data Collection

The corresponding data for the 30 features initially identified as potential factors influencing real estate appraisal, specifically for residential apartments in Al Bireh city, are found not to be consolidated in a single agency but are available across three official agencies: Al Bireh Municipality (BM), Palestine Land Authority (PLA), and the Ministry of Local Government (MOLG). During the initial data collection phase, information for 23 features is obtained from BM.

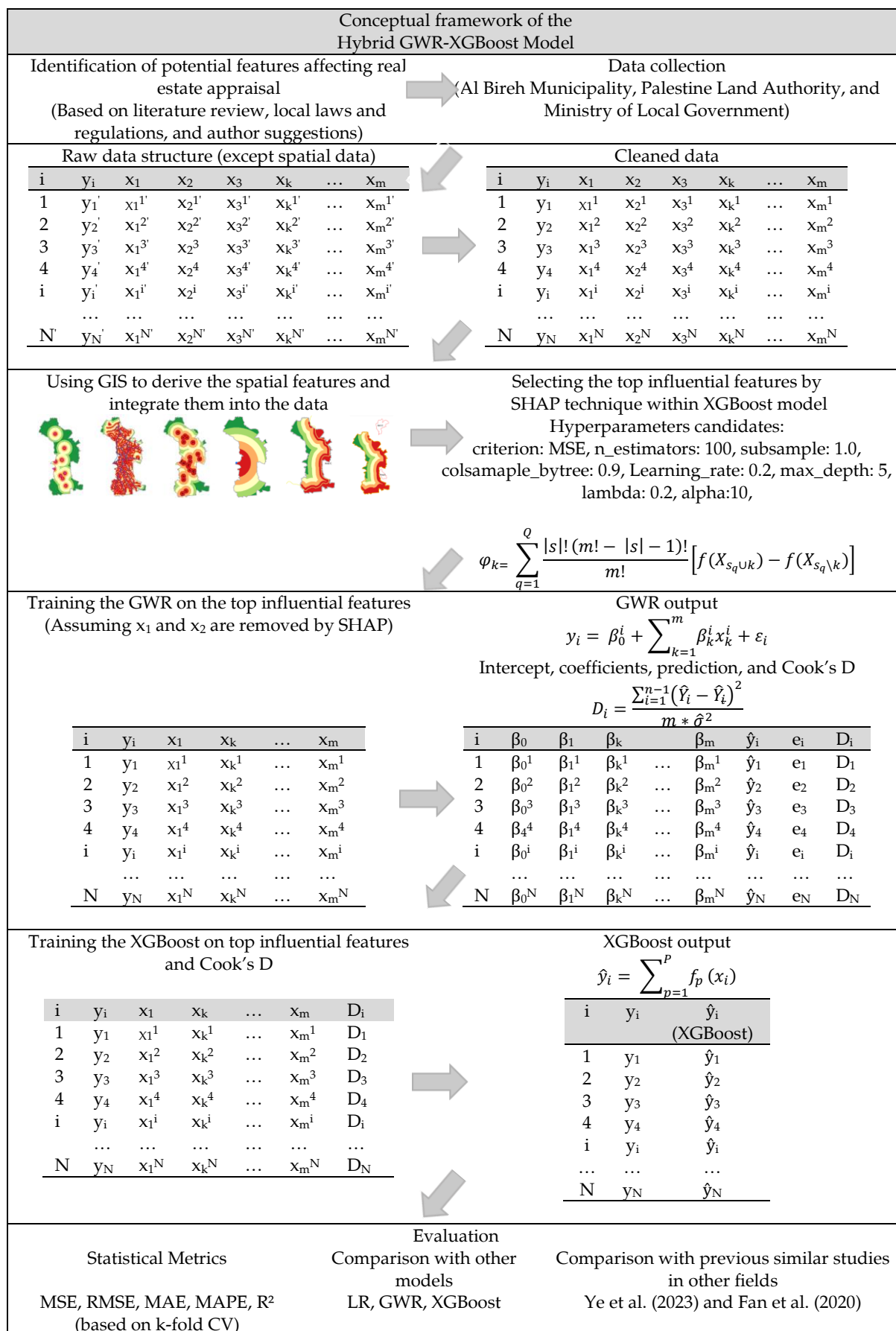


Figure 1 Conceptual framework illustrating how the hybrid GWR-XGBoost model works through the Cook's D

As this data is not presented in a tabular format, it is extracted from the municipality engineering plans database through manual data entry. The sampling strategy involves randomly selecting records from engineering plans for residential buildings within the study area and designated as appraised. Identifier information, such as parcel and quarter numbers, is included to facilitate linkage with other data sources. Furthermore, x and y coordinates are recorded for mapping purposes, resulting in a dataset comprising 5,586 entries stored in an Excel spreadsheet table. Moving to the second phase, the focus is on collecting two features: appraisal value and appraisal date, for each residential apartment observation. This information is sourced from PLA, where details are available in separate Microsoft Word files. These attributes are added to the Excel data file and populated with corresponding values based on common parcel, quarter, and block numbers. Only 2,354 observations have appraisal values. With the conclusion of this phase, data collection is completed for 24 features, with the appraisal value as the target variable. In the third phase, the aim is to acquire the remaining six features related to proximity to amenities, such as schools, hospitals, the city center, main roads, area C, and colonies. This data is downloaded from the Geomolg website (geomolg.ps), administered by MOLG. These spatial layers are integrated into the cleaned Excel data table using GIS techniques. By the end of this phase, a dataset containing all 30 features, with the appraisal value as the target variable and encompassing 2,354 observations, is complete.

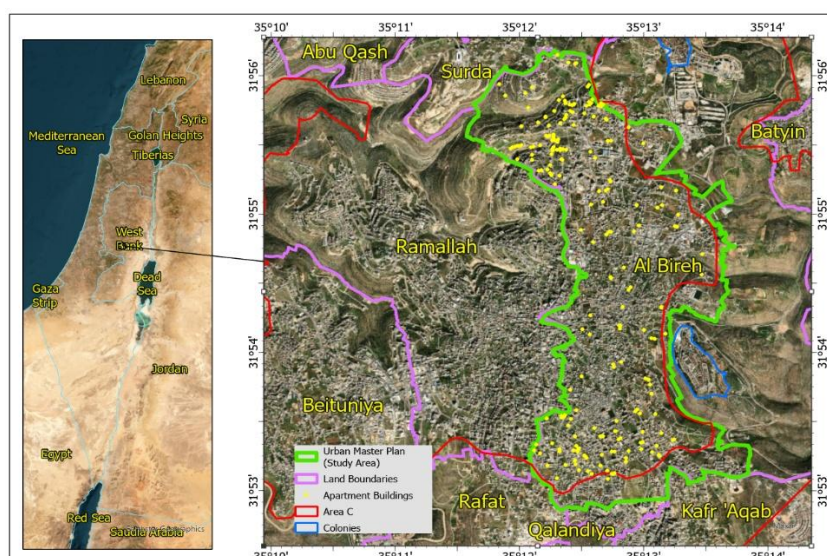


Figure 2 The study area in Al Bireh city. Source: Layers captured from Geomolg Portal and Imagery inserted from ArcGIS Pro

5. Analysis

Prior to commencing the analysis, it is standard practice to perform data cleaning as a preprocessing step. Employing descriptive statistics and scatter plots is an effective approach to understanding and evaluating the quality of the data. This procedure aids in detecting data errors such as duplications, missing values, outliers, invalid entries, and typo errors, enabling their subsequent correction. The data is rectified accordingly, with 3,232 observations being removed due to lacking appraisal values. Consequently, the apartments dataset, consisting of 2,354 observations with 24 features, with the appraisal value as the target variable, located in 211 residential buildings, is stored in an Excel file and ready for analysis. The analysis is detailed in the following steps:

5.1. Deriving the values of the six locational features using GIS

In this step, locational features—represented by proximities to amenities are derived and added to the Excel file. More specifically, the proximity of apartments to schools, hospitals, the city center,

main roads, area C, and colonies are derived using the GIS and integrated into the data table. The "Euclidean Distance" tool, available in ESRI ArcGIS Pro 3.0.2, is employed. The output is then processed through the "Reclassify" tool, which categorizes the distance from these amenities into five classes according to the natural breaks method, ranging from 1 to 5, where 1 signifies a short distance to the amenity, and 5 denotes a longer distance. The "Spatial Join" tool is applied to assign the class to each apartment projected as a point according to its x and y coordinates. Consequently, apartments closer to the amenities are designated as class 1, while those farther away are assigned class 5, as shown in Figure 3. By the end of this stage, six attributes are added to the Excel data table, resulting in a final table of 30 features, with the appraisal value as the target variable.

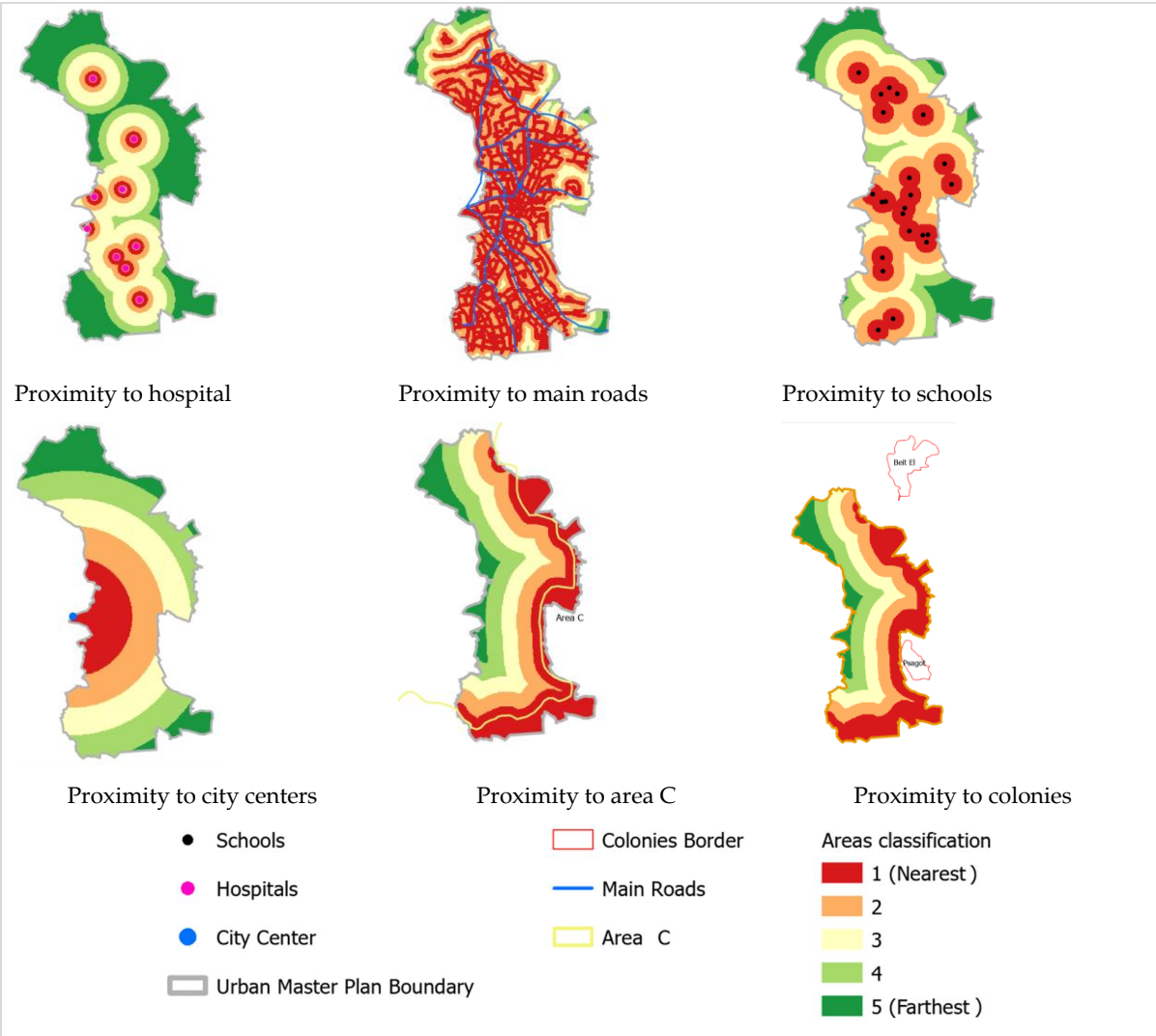


Figure 3 Maps showing proximity classes for locational features, including schools, hospitals, the city center, main roads, area C, and colonies

5.2. Descriptive statistics

With all the information pertaining to apartment characteristics, appraisals, and proximity to amenities consolidated into a single table, the descriptive statistics, comprising maximum, minimum, mean, and standard deviation, have been computed and are displayed in Table 3.

Table 3 Descriptive statistics of potential features affecting residential apartment appraisal

No	Feature	Minimum	Maximum	Mean	Standard Deviation
0.	Appraisal value	30,961	79,992	63,986	10,401
1.	Appraisal date	2008	2023	2016	3
2.	Area	91	207	146	20
3.	Construction date	1995	2022	2011	5.33
4.	Block	7	28	18	8
5.	Number of apartments in building	2	38	17.05	6.22
6.	Number of apartments in floor	1	5	2.62	0.80
7.	Number of floors in building	2	11	6.69	1.47
8.	Floor level	-6	8	1.68	2.13
9.	Floor to parcel ratio	0.069	0.968	0.45	0.09
10.	Number of bedrooms	1	5	2.96	0.23
11.	Number of bathrooms	1	3	1.76	0.44
12.	Number of toilets	0	2	0.74	0.45
13.	Number of balconies	0	4	1.39	0.69
14.	Number of facades	1	4	2.58	0.54
15.	Central heating availability	0	1	0.02	0.14
16.	Elevator availability	0	1	0.95	0.22
17.	Parking availability	0	1	0.91	0.29
18.	Storage availability	0	1	0.13	0.33
19.	Adjacent street type	1	4	2.27	0.64
20.	Adjacent street width	3	30	10.30	3.76
21.	Adjacent street aesthetics	1	5	3.20	1.06
22.	Number of adjacent streets	1	3	1.45	0.60
23.	Construction material	1	2	1.00	0.07
24.	View	0	5	2.67	0.96
25.	Proximity to hospitals	1	5	2.37	1.02
26.	Proximity to schools	1	4	1.70	0.70
27.	Proximity to city center	1	5	3.53	0.90
28.	Proximity to main roads	1	4	1.51	0.80
29.	Proximity to area C	1	4	2.12	1.03
30.	Proximity to colonies	1	5	3.25	0.91

6. Features selection using SHAP within XGBoost trained on 30 features

To streamline model analysis, it is common to opt for selecting the most influential features from those potentially identified. Consequently, XGBoost is trained on the complete dataset comprising 30 features with the appraisal value as the target variable. From the candidate hyperparameter values listed in Table 1, the input hyperparameters are selected using the random search method and further refined through trial and error, resulting in the following: $n_estimators = 100$, $subsample = 1.0$, $colsample_bytree = 0.9$, $learning_rate = 0.2$, $max_depth = 5$, $lambda = 0.2$, and $alpha = 10$, with MSE as the criterion. A k-fold cross-validation approach is applied to evaluate performance, with k set to 5. Under these settings, the resulting metrics for MSE, RMSE, MAE, MAPE, and R^2 are 21,515,010; 4,638; 2,215; 3.7%; and 0.80, respectively, as summarized in Table 4.

Table 4 Performance evaluation metrics for the XGBoost model with 30 features, using appraisal value as the target variable and a dataset size of 2,354 observations of residential apartments

MSE	RMSE	MAE	MAPE	R^2
21,515,010	4,638	2,215	3.7	0.80

The SHAP technique is applied within this XGBoost model, resulting in the generation of bar and beeswarm plots as depicted in Figure 4. In the bar plot (located on the left side of Figure 4), the features are sorted in descending order, with the most influential feature on top and the least influential at the bottom. This means that the higher the value of SHAP for a feature, the greater its impact on the prediction. In return, the beeswarm plot (shown on the right side of Figure 4), also sorts the features using the same technique as in the bar plot. However, with one additional piece of information, it shows the direction of influence of each feature on the prediction. The red color represents high values of the feature, while blue indicates low values. The plot illustrates how the SHAP value changes with high and low values of each feature. By analyzing the results from both plots, the top five features with the highest SHAP values are selected for further analysis. This selection of the five top features is based on the observation that choosing fewer features reduces performance, while adding more features does not improve it. The top five features selected include area, appraisal date, number of apartments in the building, block, and construction date.

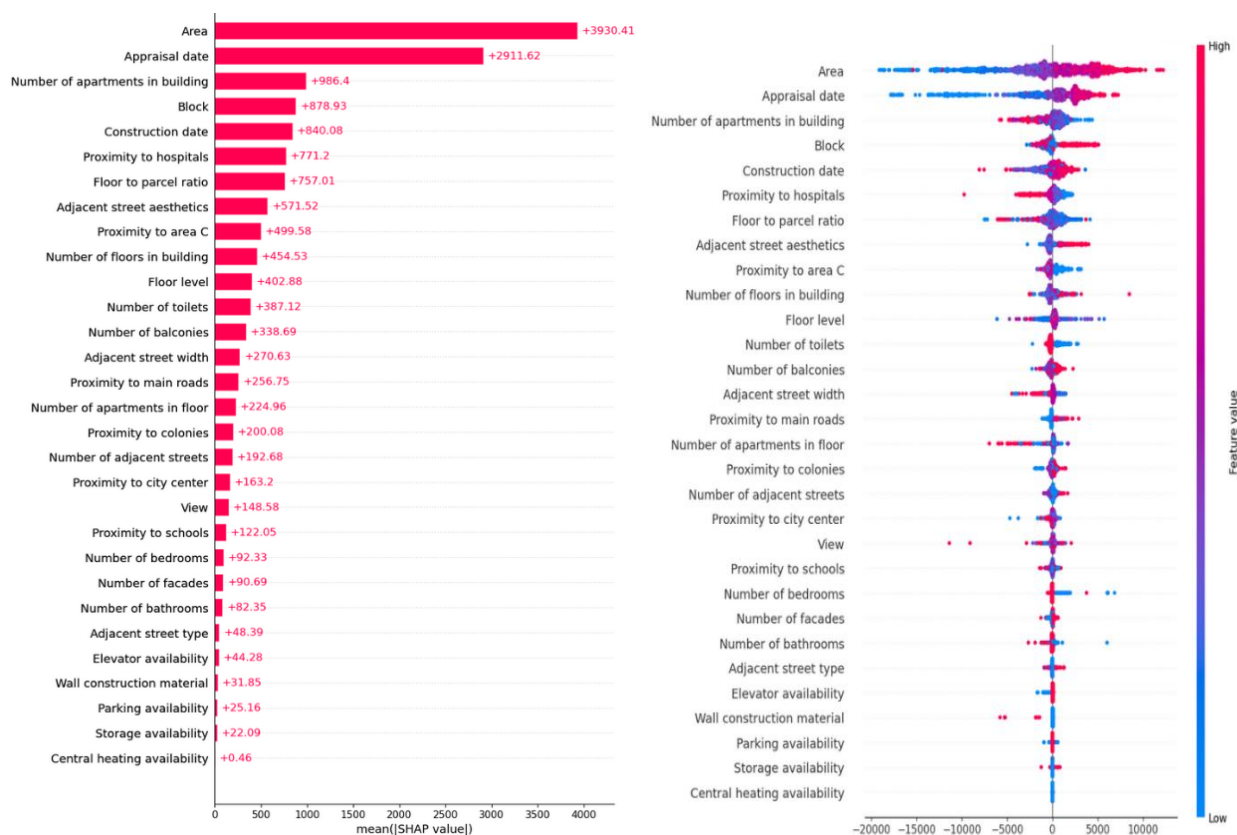


Figure 4 The XGBoost-based SHAP summary plot (left) and beeswarm plot (right) for the features affecting residential apartment appraisal

7. Multicollinearity check among the five selected features

At this stage of the analysis, examining multicollinearity among the top five selected features is crucial, even though XGBoost incorporates its own regularization techniques to address this issue. A scatter plot can effectively illustrate both the relationships among the features and their relationship with the appraisal value as the target variable, as shown in Figure 5. This figure provides visual context, clearly indicating the absence of multicollinearity and the presence of nonlinearity.

With the scatter plot revealing nonlinear relationships between each pair of features, Spearman's rank correlation coefficient is utilized, with the results presented in Figure 6. The Spearman's rank values are low, reinforcing the lack of multicollinearity.

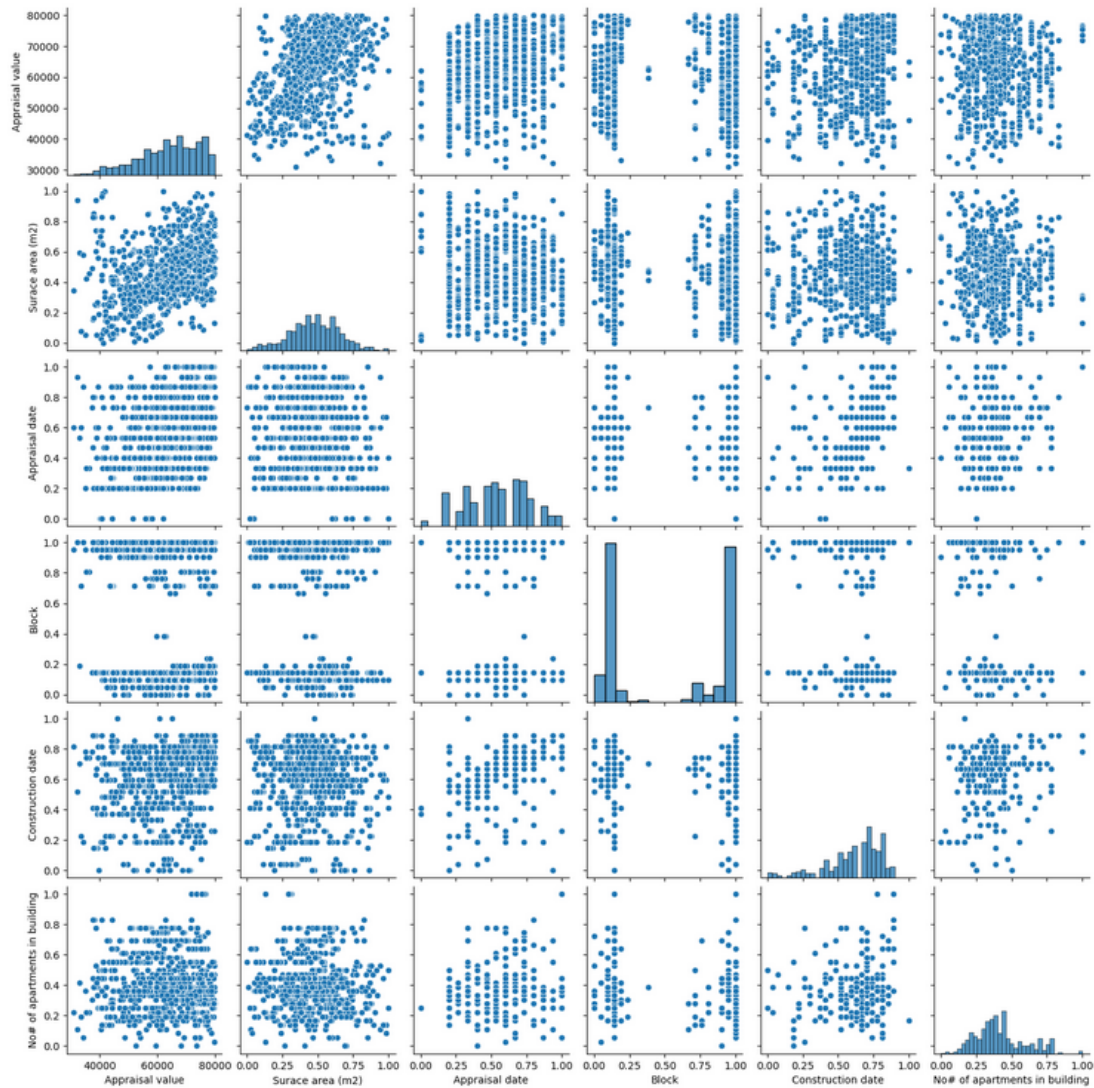


Figure 5 Scatter plot showing the relationships among the top five selected features and their relationship with the appraisal value as the target variable

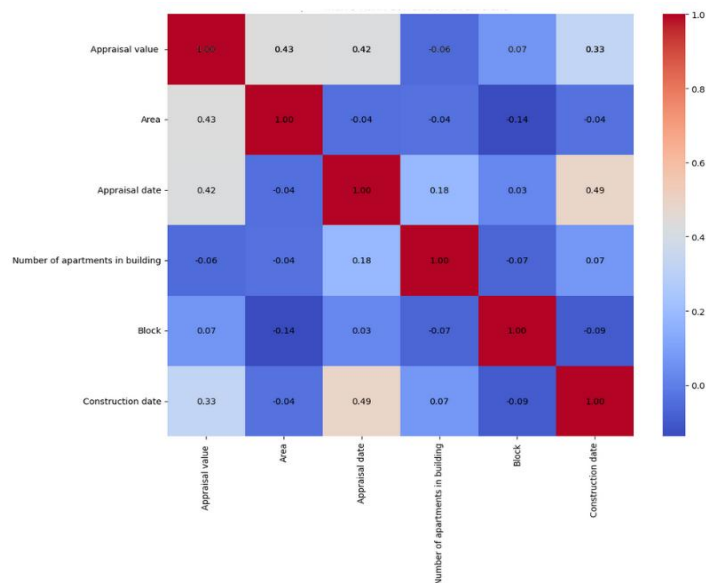


Figure 6 The Spearman's rank correlation coefficient among the top five selected features and their relationship with the appraisal value as the target variable

8. Training the XGBoost model using the top five selected features

The XGBoost model is retrained on the top five features, with the appraisal value as the target variable. The results for MSE, RMSE, MAE, MAPE, and R^2 are 20,268,307, 4,489, 2,158, 3.7, and 0.81, respectively, as depicted in Table 5. This indicates that the model's performance with the top five features is comparable to that obtained when the XGBoost model is trained on all 30 features. The reduction in dimensionality from 30 to 5 features enhances the model's efficiency and facilitates interpretation, all while preserving appraisal performance. This streamlined model offers improved performance alongside heightened explanatory power preserved.

Table 5 Performance evaluation metrics for the XGBoost model with the five top features, using appraisal value as the target variable and a dataset size of 2,354 observations of residential apartments

MSE	RMSE	MAE	MAPE	R^2
20,268,307	4,489	2,158	3.7	0.81

9. Training the GWR model using the top five selected features

The GWR is trained on five features with the appraisal value as the target variable. The bandwidth is set to be adaptive with 500 as the number of neighboring observations. The choice of the adaptive bandwidth is more suitable due to the sparsity of observations within the dataset. The decision to select 500 as the number of observations, which is relatively high, is due to multiple apartments within the same building, sharing identical locations. Each building typically includes multiple floors, and each floor may contain multiple apartments. The kernel function is set as Gaussian. The results of the model performance measured by the five metrics of MSE, RMSE, MAE, MAPE, and R^2 are 52,540,381, 7,248, 5,266, 8.9, and 0.53, respectively, as depicted in

Table 6. Along with the five evaluation metrics, other outputs from the GWR model, such as intercept, coefficients, local R^2 , and Cook's D, are obtained.

Table 6 Performance evaluation metrics for the GWR model with the top five features, using appraisal value as the target variable and a dataset size of 2,354 observations of residential apartments

MSE	RMSE	MAE	MAPE	R^2
52,540,381	7,248	5,266	8.9	0.53

10. Training the XGBoost on the six features

In this step, XGBoost is trained on six features: the top five most influential features and Cook's D obtained from the GWR, with the appraisal value as the target variable. The decision to include Cook's D in addition to the top five features for training XGBoost is based on trial and error, testing which output from the GWR can be added to XGBoost as a feature to achieve the highest performance, reflecting the nonstationarity aspect. Among all outputs of the GWR, incorporating Cook's D as a feature into XGBoost, along with the top five features, resulted in the highest performance. This model addresses both nonstationarity and nonlinearity simultaneously. More precisely, the hybrid GWR-XGBoost model is designed by transferring the effect of nonstationarity captured by GWR into XGBoost, a model aware of nonlinearity. The evaluation performance metrics of XGBoost are shown in Table 7.

Table 7 Performance evaluation metrics for the XGBoost model with 6 features, using appraisal value as the target variable and a dataset size of 2,354 observations of residential apartments

MSE	RMSE	MAE	MAPE	R^2
9,213,573	3,032	1,335	2.3	0.92

11. Results

The analysis addresses the research questions by identifying the features influencing real estate appraisal within the study area, selecting the most influential ones, designing the hybrid model, and evaluating its performance. Specifically, the top five selected features include area, appraisal date, number of apartments in the building, block, and construction date. The hybrid model is designed by combining GWR and XGBoost through Cook's D, and its performance is evaluated based on five metrics: MSE, RMSE, MAE, MAPE, and R^2 , as demonstrated in Figure 7. The figure also includes the performance results of GWR and XGBoost as component models and LR as a reference model for ease of comparison.

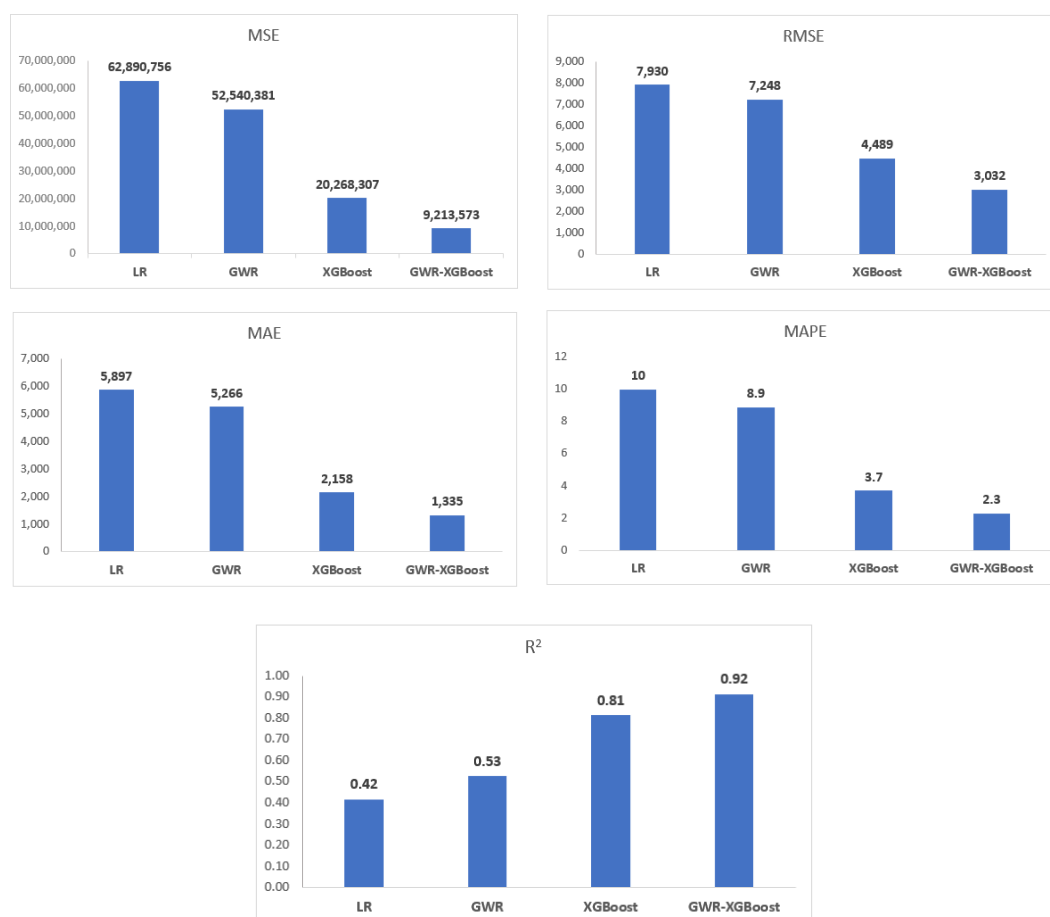


Figure 7 Bar chart presentation of the performance results for LR, GWR, XGBoost, and GWR-XGBoost using MSE, RMSE, MAE, MAPE, and R^2 metrics

12. Discussion

It is noteworthy that the selected top five features align closely with findings from other research papers. Area, appraisal date, block, and construction date are recognized as influential features in several studies, including those by [Elnaeem Balila and Shabri \(2024\)](#), [Hjort et al. \(2024\)](#), [Trindade Neves et al. \(2024\)](#), [Metek and Yomralioglu \(2023\)](#), [Mora-Garcia et al. \(2022\)](#), [Rampini and Re Cecconi \(2022\)](#), and [Koochpayma and Argany \(2021\)](#). While few studies include the number of apartments in a building as an influential feature, [Taecharungroj \(2021\)](#) ranked it as third.

The design of the hybrid model, which combines GWR with XGBoost through Cook's D, achieved superior results. This improvement is attributed to Cook's D, which not only reflects the nonstationarity aspect of real estate but also provides information about outliers and influential observations, helping XGBoost better understand patterns and mitigate the influence of outliers on predictions. Employing the most commonly used evaluation metrics in the literature ([Numan and](#)

[Yusoff, 2024a](#)) demonstrates that the hybrid GWR-XGBoost model outperforms all other models across all five evaluation metrics, as the results are presented in bar charts in Figure 7.

The hybrid GWR-XGBoost model presents notable distinctions from LR and GWR, yet demonstrates a relatively modest deviation from XGBoost. The degree of performance enhancement of the GWR-XGBoost hybrid over XGBoost regarding MSE, RMSE, MAE, MAPE, and R^2 is 11,054,734, 1,457, 823, 1.4%, and 11%, respectively. The reason LR yields the least favorable outcome is its disregard for nonstationarity and nonlinearity ([Hoxha, 2024](#)). The slight improvement of GWR over LR appears to stem from its ability to tackle nonstationarity issues. XGBoost attains superior outcomes, indicating that addressing nonlinearity is more crucial than nonstationarity. The hybrid GWR-XGBoost model, which incorporates both aspects, has been identified as the most efficient approach.

Conversely, when comparing the results of the hybrid GWR-XGBoost model in this study with those from the applications by [Ye et al. \(2023\)](#) and [Fan et al. \(2020\)](#), as depicted in Table , it becomes evident that the hybrid GWR-XGBoost model consistently surpasses the performance of both the GWR and XGBoost models in all three studies, as indicated by the employed statistical metrics.

Table 8 The result of performance evaluation metrics of the hybrid GWR-XGBoost in the studies of [Ye et al. \(2023\)](#) and [Fan et al. \(2020\)](#)

Author	Field	Model	RMSE	MAE	MAPE	R^2
Ye et al. (2023)	Soil contamination	GWR	5.58	2.82	---	0.74
		XGBoost	7.98	4.7	---	0.48
		GWR-XGBoost	3.43	1.72	---	0.90
Fan et al. (2020)	Air pollution	GWR	33.67	---	21.79	0.71
		XGBoost	27.01	---	19.52	0.85
		GWR-XGBoost	24.08	---	16.90	0.88

In the three investigations carried out by the author, [Ye et al. \(2023\)](#), and [Fan et al. \(2020\)](#), the R^2 serves as a common metric for performance assessment. Utilizing a bar chart to present the values provides a more effective representation of the results for the three models, as illustrated in Figure 8. The results indicate that the R^2 value for GWR-XGBoost in all three studies exceeds its corresponding values in both the GWR and XGBoost models. Notably, the study conducted by [Ye et al. \(2023\)](#) showcases a substantial 42% enhancement in the R^2 for GWR-XGBoost over XGBoost, while the improvements in the studies by [Fan et al. \(2020\)](#) and the author are 3% and 11%, respectively. Interestingly, in the study implemented by [Ye et al. \(2023\)](#), the performance of GWR surpasses XGBoost. However, the enhancements in R^2 for GWR-XGBoost in the three studies conducted by the author, [Ye et al. \(2023\)](#), and [Fan et al. \(2020\)](#) are 11%, 3%, and 16%, respectively.

One may question whether the observed differences in performance between these three studies are influenced by factors beyond the methodology, such as dataset size and the timeframe of data collection. [Ye et al. \(2023\)](#) study was executed with 222 observations gathered from 2018 to 2020, whereas [Fan et al. \(2020\)](#) research utilized a significantly larger dataset with hourly readings spanning from March 2016 to February 2017 across 35 sites. The dataset sizes range from relatively small to extensive, and the time of data collection is comparatively longer in our research, spanning from 2008 to 2023. The quality of the data itself strongly influences the results. Additionally, there is a fundamental distinction in the sectors studied, each possessing unique characteristics, dynamics, and interactions. Furthermore, despite the three authors using the same model, the detailed mechanisms in each model differ. In all instances, it is crucial to acknowledge that comparing our research results with those of other similar studies is helpful for a more comprehensive evaluation of the findings.

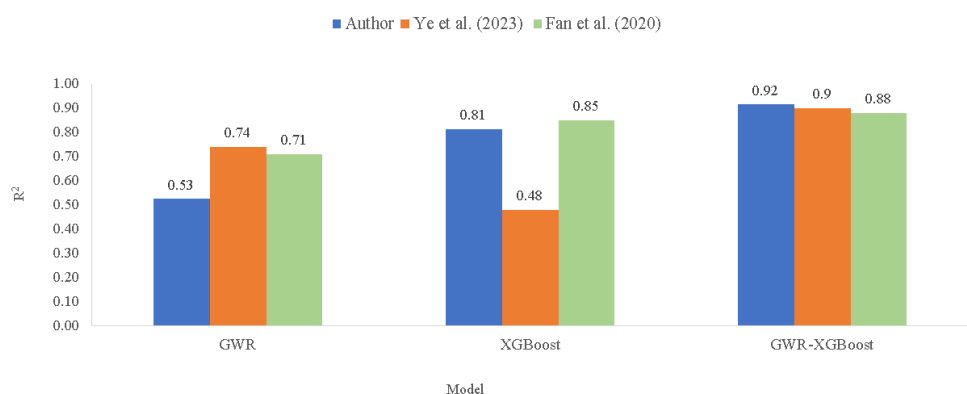


Figure 8 The comparison of the performance results of the GWR, XGBoost, and hybrid GWR-XGBoost model in the three studies: author, [Ye et al. \(2023\)](#) and [Fan et al. \(2020\)](#) studies

13. Summary and Conclusion

Given the challenges in model-based real estate appraisal methods, including issues related to performance, interpretability, stability, reliability, scalability, flexibility, simplicity, adaptability, applicability, generalizability, comprehensibility, data availability, and evaluation metrics, this study focuses specifically on addressing the performance challenge. The novelty of this research lies not only in adapting the hybrid GWR-XGBoost model for the first time in real estate appraisal but also in the unique way the component models, GWR and XGBoost, are combined through Cook's D, setting this approach apart from previous studies with the aim of enhancing performance. To achieve this aim, three research questions were formulated: (1) which features impact residential apartment appraisal within the context of Al Bireh city, (2) how the hybrid GWR-XGBoost model can be designed, and (3) how model performance can be evaluated. The rationale for this hybrid model lies in its capacity to concurrently address two fundamental aspects of the real estate sector: nonstationarity and nonlinearity, which are managed by Geographically Weighted Regression (GWR) and eXtreme Gradient Boosting (XGBoost), respectively.

Initially, a set of 30 features was identified as potential factors influencing real estate appraisal values in Al Bireh city, Palestine. Corresponding data was gathered from three agencies: MB, PLA, and MOLG. Data from the first two agencies was consolidated into a single Excel table and cleaned. Spatial data from MOLG was then integrated into this cleaned data table. The SHAP technique within XGBoost was employed to select the most significant features, resulting in the selection of the top five features: area, appraisal date, number of apartments in the building, block, and construction date. This selection process addresses the first research question.

Using a dataset comprising 2,354 observations with five features and the appraisal value as the target variable, a consolidated Excel file was prepared as input for the hybrid GWR model. Training the GWR on this data enabled the calculation of Cook's D for each observation. Cook's D was then added to the data table, resulting in a dataset with six features and the appraisal value as the target variable. This combined dataset was used as input for the subsequent execution of the XGBoost model, generating predictions as a result. This addresses the second research question.

Five commonly used statistical metrics were selected to evaluate the performance of the hybrid GWR-XGBoost model: MSE, RMSE, MAE, MAPE, and R^2 . Calculations for the hybrid model yielded values of 9,213,573, 3,032, 1,335, 2.3, and 0.92, respectively. These results were then compared with those from LR, GWR, and XGBoost, where LR serves as the reference model and GWR and XGBoost serve as the component models. The analysis reveals that the hybrid GWR-XGBoost model outperforms LR, GWR, and XGBoost across all five metrics. Incorporating Cook's D, derived from GWR, into XGBoost enables the model to grasp nonstationarity and detect outliers and influential observations, effectively mitigating their impact and thus improving performance. This improvement is particularly pronounced when compared to LR and GWR, with a noticeable, albeit smaller, advantage over XGBoost. LR shows the least favorable results due to its limited ability to

account for nonstationarity and nonlinearity. The slight improvement of GWR over LR can be attributed to its capability to address nonstationarity issues, while XGBoost demonstrates superior performance, indicating that addressing nonlinearity is more critical than nonstationarity. The hybrid GWR-XGBoost model, which addresses both aspects, is the most effective approach. In reciprocation, the results from prior research on a similar model concerning soil contamination and air pollution, conducted by [Ye et al. \(2023\)](#) and [Fan et al. \(2020\)](#), respectively, align with the findings of our current study. This confirms the superior performance of the hybrid GWR-XGBoost model compared to both GWR and XGBoost. Regarding the R^2 metric, the hybrid GWR-XGBoost model achieved an improvement of 11% over XGBoost, compared to improvements of 2% and 4% over the GWR-XGBoost models by [Ye et al. \(2023\)](#) and [Fan et al. \(2020\)](#), respectively. This concludes the answer to the third research question.

Upon examining the constraints, this research reveals numerous limitations. The study area, representing Al Bireh city, is adjacent to Ramallah city, making it more appropriate to analyze their real estate data to consider associations, dynamics, and interactions. However, data unavailability has constrained the research to focus solely on Al Bireh city. In a developing country and a state under military occupation, such as Palestine, accessing real estate data for public use is challenging. Obtaining this data involves sending official letters to relevant agencies to obtain permission. Due to the same constraint, the set of explanatory features does not include econometric factors like gross domestic product, interest rates, and exchange rates. On the other hand, the dataset used in the analysis is relatively small, particularly in the context of machine learning techniques that require massive data for training to better capture knowledge, patterns, and relationships. Therefore, there is a need to investigate the model's performance with a larger volume of data. Additionally, the data itself was not tabulated; rather, data entry was required for its collection, which could lead to human error and impact the performance.

The implications are threefold: first, the key features affecting real estate appraisal are identified, providing a useful checklist and reference for stakeholders such as real estate sellers, buyers, agents, developers, investors, creditors, auditors, brokers, consumers, economists, and planners. Second, the improved appraisal accuracy supports more informed and mature economic, financial, and business decisions, which can lead to beneficial and impactful outcomes in practice. Finally, this research within the context of Palestine paves the way for examining advanced technologies, particularly machine learning, as alternatives to traditional manual or field-based real estate approaches appraisals. Specifically, a significant takeaway from the research is the potential to develop an online real-time application for real estate appraisal in Palestine, comparable to Zillow in the US and Zoopla in the UK. This development could save time, reduce costs, enhance efficiency, and serve as a valuable decision-support tool.

Future research endeavors may explore specific aspects for consideration. One avenue involves shifting the study's focus to the transacted prices of residential apartments as the target variable, rather than appraised prices. Additionally, expanding the analysis to include neighboring cities to Al Bireh, such as Ramallah and Beitunia, could provide a more comprehensive and nuanced understanding. Similar studies employing the same approach could be conducted in other cities within Palestine or globally to validate its generalizability.

Author Contributions

Jamal A.A. Numan: Conceptualization, methodology, data collection, formal analysis, software, validation, visualization, writing—original draft.

Izham Mohamad Yusoff: Review and editing, project administration, guidance on methodology and structure, and overall research oversight.

Conflict of Interest

The authors declare no conflict of interest.

References

- Alzain, E, Alshebami, AS, Aldhyani, TH & Alsubari, SN 2022, 'Application of artificial intelligence for predicting real estate prices: the case of Saudi Arabia', *Electronics*, vol. 11, p. 3448, <https://doi.org/10.3390/electronics11213448>
- Alzubi, Y 2024, 'Comparison of various machine learning models for estimating construction projects sales valuation using economic variables and indices', *Journal of Soft Computing in Civil Engineering*, vol. 8, pp. 1-32
- Arvianti, T 2021, 'An analysis of implicit value of property characteristics in residential property prices using a hedonic value approach', *International Journal of Technology*, vol. 12, <https://doi.org/10.14716/ijtech.v12i6.5216>
- Aydinoglu, AC & Sisman, S 2024, 'Comparing modelling performance and evaluating differences of feature importance on defined geographical appraisal zones for mass real estate appraisal', *Spatial Economic Analysis*, vol. 19, pp. 225-249, <https://doi.org/10.1080/17421772.2023.2242897>
- Bogin, AN & Shui, J 2020, 'Appraisal accuracy and automated valuation models in rural areas', *The Journal of Real Estate Finance and Economics*, vol. 60, pp. 40-52, <https://doi.org/10.1007/s11146-019-09712-0>
- Brunsdon, C, Fotheringham, AS & Charlton, ME 1996, 'Geographically weighted regression: a method for exploring spatial nonstationarity', *Geographical analysis*, vol. 28, pp. 281-298, <https://doi.org/10.1111/j.1538-4632.1996.tb00936.x>
- Chanasit, K, Chuangsuwanich, E, Suchato, A & Punyabukkana, P 2021, 'A real estate valuation model using boosted feature selection', *IEEE Access*, vol. 9, pp. 86938-86953
- Chen, C, Ma, X & Zhang, X 2024a, 'Empirical study on real estate mass appraisal based on dynamic neural networks', *Buildings*, vol. 14, p. 2199, <https://doi.org/10.3390/buildings14072199>
- Chen, T & Guestrin, C 2016, 'Xgboost: A scalable tree boosting system', *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794, <https://doi.org/10.1145/2939672.2939785>
- Chen, Y, Jiao, J & Farahi, A 2023, 'Disparities in affecting factors of housing price: A machine learning approach to the effects of housing status, public transit, and density factors on single-family housing price', *Cities*, vol. 140, p. 104432, <https://doi.org/10.1016/j.cities.2023.104432>
- Chen, Y, Yang, Q, Geng, L & Yin, W 2024b, 'Analysis of factors influencing housing prices in mountain cities based on multiscale geographically weighted regression—demonstrated in the central urban area of chongqing', *Land*, vol. 13, p. 602, <https://doi.org/10.3390/land13050602>
- Cheng, C-H & Tsai, M-C 2022, 'An intelligent homogeneous model based on an enhanced weighted kernel self-organizing map for forecasting house prices', *Land*, vol. 11, article 1138, <https://doi.org/10.3390/land11081138>
- Çilgin, C & Gökçen, H 2023, 'Machine learning methods for prediction real estate sales prices in Turkey', *Revista de la construcción*, vol. 22, pp. 163-177, <http://dx.doi.org/10.7764/rdlc.22.1.163>
- Çilgin, C & Gökçen, H 2024, 'A hybrid machine learning model architecture with clustering analysis and stacking ensemble for real estate price prediction', *Computational Economics*, pp. 1-52, <https://doi.org/10.1007/s10614-024-10703-4>
- Das, SSS, Ali, ME, Li, Y-F., Kang, Y-B & Sellis, T 2021, 'Boosting house price predictions using geo-spatial network embedding', *Data Mining and Knowledge Discovery*, vol. 35, pp. 2221-2250, <https://doi.org/10.1007/s10618-021-00789-x>
- Deppner, J & Cajias, M 2024, 'Accounting for spatial autocorrelation in algorithm-driven hedonic models: A spatial cross-validation approach', *The Journal of Real Estate Finance and Economics*, vol. 68, pp. 235-273, <https://doi.org/10.1007/s11146-022-09915-y>
- Droj, G, Kwartnik-Pruc, A & Droj, L 2024, 'A comprehensive overview regarding the impact of gis on property valuation', *ISPRS International Journal of Geo-Information*, vol. 13, article 175, <https://doi.org/10.3390/ijgi13060175>
- Elnaeem Balila, A & Shabri, AB 2024, 'Comparative analysis of machine learning algorithms for predicting Dubai property prices', *Frontiers in Applied Mathematics and Statistics*, vol. 10, article 1327376, <https://doi.org/10.3389/fams.2024.1327376>
- Fan, Z, Zhan, Q, Yang, C, Liu, H & Bilal, M 2020, 'Estimating PM2. 5 concentrations using spatially local xgboost based on full-covered SARA AOD at the urban scale', *Remote Sensing*, vol. 12, article 3368, <https://doi.org/10.3390/rs12203368>

- Gabrielli, L & French, N 2021, 'Pricing to market: property valuation methods—a practical review', *Journal of property investment & finance*, vol. 39, pp. 464-480, <https://doi.org/10.1108/JPIF-09-2020-0101>
- Geerts, M & De Weerd, J 2023, 'A survey of methods and input data types for house price prediction', *ISPRS International Journal of Geo-Information*, vol. 12, p. 200, <https://doi.org/10.3390/ijgi12050200>
- Glumac, B & Des Rosiers, F 2021, 'Practice briefing—Automated valuation models (AVMs): their role, their advantages and their limitations', *Journal of Property Investment & Finance*, vol. 39, pp. 481-491
- Guliker, E, Folmer, E & Van Sinderen, M 2022, 'Spatial determinants of real estate appraisals in the Netherlands: A machine learning approach', *ISPRS international journal of geo-information*, vol. 11, article 125
- Gunes, T 2023, 'Model agnostic interpretable machine learning for residential property valuation', *Survey Review*, pp. 1-16, <https://doi.org/10.1080/00396265.2023.2293366>
- He, W, Zhao, R & Gao, S 2024, 'Exploring the impact of multimodal access on property and land economies in shanghai's inner ring districts: Leveraging advanced spatial analysis techniques', *Land*, vol. 13, article 311, <https://doi.org/10.3390/land13030311>
- Hjort, A, Pensar, J, Scheel, I & Sommervoll, DE 2022, 'House price prediction with gradient boosted trees under different loss functions', *Journal of Property Research*, vol. 39, pp. 338-364, <https://doi.org/10.1080/09599916.2022.2070525>
- Hjort, A, Scheel, I, Sommervoll, DE & Pensar, J 2024, 'Locally interpretable tree boosting: An application to house price prediction', *Decision Support Systems*, vol. 178, article 114106, <https://doi.org/10.1016/j.dss.2023.114106>
- Hoang, D & Wiegatz, K 2023, 'Machine learning methods in finance: Recent applications and prospects', *European Financial Management*, vol. 29, pp. 1657-1701, <https://doi.org/10.1111/eufm.12408>
- Hong, J, Choi, H & 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, pp. 140-152, <https://doi.org/10.3846/ijspm.2020.11544>
- Horvath, S, Soot, M, Zaddach, S, Neuner, H & Weitkamp, A 2021, 'Deriving adequate sample sizes for ANN-based modelling of real estate valuation tasks by complexity analysis', *Land Use Policy*, vol. 107, p. 105475, <https://doi.org/10.1016/j.landusepol.2021.105475>
- Hoxha, V 2024, 'Comparative analysis of machine learning models in predicting housing prices: a case study of Prishtina's real estate market', *International Journal of Housing Markets and Analysis*, vol. 18, no. 3, pp. 694-711, <https://doi.org/10.1108/IJHMA-09-2023-0120>
- Hurley, AK & Sweeney, J 2024, 'Irish property price estimation using a flexible geo-spatial smoothing approach: What is the Impact of an address?', *The Journal of Real Estate Finance and Economics*, vol. 68, pp. 355-393, <https://doi.org/10.1007/s11146-022-09888-y>
- Iban, MC 2022, 'An explainable model for the mass appraisal of residences: The application of tree-based Machine Learning algorithms and interpretation of value determinants', *Habitat International*, vol. 128, p. 102660, <https://doi.org/10.1016/j.habitatint.2022.102660>
- Jafary, P, Shojaei, D, Rajabifard, A & 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, p. 105115, <https://doi.org/10.1016/j.cities.2024.105115>
- Jin, S, Zheng, H, Marantz, N & Roy, A 2024, 'Understanding the effects of socioeconomic factors on housing price appreciation using explainable AI', *Applied Geography*, vol. 169, p. 103339, <https://doi.org/10.1016/j.apgeog.2024.103339>
- Jin, T, Cheng, L, Liu, Z, Cao, J, Huang, H & Witlox, F 2022, 'Nonlinear public transit accessibility effects on housing prices: Heterogeneity across price segments', *Transport Policy*, vol. 117, pp. 48-59, <https://doi.org/10.1016/j.tranpol.2022.01.004>
- Kamaruzzaman, SN, Razali, A, Zawawi, EMA, Basir, SA & Riley, M 2018, 'Residents' Satisfaction Towards the Indoor Environmental Quality of Re-engineered Affordable Housing Scheme in Malaysia', *International Journal of Technology*, vol. 9, pp. 501-512, <https://doi.org/10.14716/ijtech.v9i3.826>
- Koohpayma, J & Argany, M 2021, 'Estimating the price of apartments in Tehran using extracted compound variables', *International Journal of Housing Markets and Analysis*, vol. 14, pp. 569-595, <https://doi.org/10.1108/IJHMA-05-2020-0050>
- Kopczewska, K 2022, 'Spatial machine learning: new opportunities for regional science', *The Annals of Regional Science*, vol. 68, pp. 713-755, <https://doi.org/10.1007/s00168-021-01101-x>
- Krämer, B, Nagl, C, Stang, M & Schäfers, W 2023a, 'Explainable AI in a real estate context—exploring the determinants of residential real estate values', *Journal of Housing Research*, vol. 32, pp. 204-245

Krämer, B, Nagl, C, Stang, M & Schäfers, W 2023b, 'Explainable AI in a real estate context—exploring the determinants of residential real estate values', *Journal of Housing Research*, pp. 1-42, <https://doi.org/10.1080/10527001.2023.2170769>

Lahmiri, S, Bekiros, S & Avdoulas, C 2023, 'A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization', *Decision Analytics Journal*, vol. 6, article 100166, <https://doi.org/10.1016/j.dajour.2023.100166>

Lawal Dano, U 2023, 'Analyzing the spatial determinants of housing prices in Dammam, Saudi Arabia: an AHP approach', *International Journal of Housing Markets and Analysis*, vol. 18, no. 2, pp. 354-377, <https://doi.org/10.1108/IJHMA-07-2023-0101>

Lee, C 2024, 'Exploring property orientation preferences and local variations', *Property Management*, vol. 42, pp. 353-362, <https://doi.org/10.1108/PM-06-2023-0048>

Li, S, Jiang, Y, Ke, S, Nie, K & Wu, C 2021, 'Understanding the effects of influential factors on housing prices by combining extreme gradient boosting and a hedonic price model (XGBoost-HPM)', *Land*, vol. 10, p. 533, <https://doi.org/10.3390/land10050533>

Li, Z & Niu, X 2022, 'Exploring spatial nonstationarity in determinants of intercity commuting flows: a case study of Suzhou–Shanghai, China', *ISPRS International Journal of Geo-Information*, vol. 11, article 335, <https://doi.org/10.3390/ijgi11060335>

Lundberg, SM & Lee, S-I 2017, 'A unified approach to interpreting model predictions', *Advances in neural information processing systems*, vol. 30, <https://doi.org/10.48550/arXiv.1705.07874>

Mankad, MD 2022, 'Comparing OLS based hedonic model and ANN in house price estimation using relative location', *Spatial Information Research*, vol. 30, pp. 107-116, <https://doi.org/10.1007/s41324-021-00416-3>

Mathotaarachchi, KV, Hasan, R & Mahmood, S 2024, 'Advanced machine learning techniques for predictive modeling of property prices', *Information*, vol. 15, p. 295, <https://doi.org/10.3390/info15060295>

Mete, MO & Yomralioglu, T 2023, 'A hybrid approach for mass valuation of residential properties through geographic information systems and machine learning integration', *Geographical Analysis*, vol. 55, pp. 535-559, <https://doi.org/10.1111/gean.12350>

Metzner, S & Kindt, A 2018, 'Determination of the parameters of automated valuation models for the hedonic property valuation of residential properties: A literature-based approach', *International Journal of Housing Markets and Analysis*, vol. 11, pp. 73-100, <https://doi.org/10.1108/IJHMA-02-2017-0018>

Mohamed, HH, Ibrahim, AH & Hagrass, OA 2023, 'Forecasting the real estate housing prices using a novel deep learning machine model', *Civil Engineering Journal*, vol. 9, pp. 46-64

Mora-Garcia, R-T, Cespedes-Lopez, M-F & Perez-Sanchez, VR 2022, 'Housing price prediction using machine learning algorithms in COVID-19 times', *Land*, vol. 11, article 2100, <https://doi.org/10.3390/land11112100>

Numan, JA & Yusoff, IM 2024a, 'Identifying the Current Status of Real Estate Appraisal Methods', *Real Estate Management and Valuation*, pp. 12-27, <https://doi.org/10.2478/remav-2024-0032>

Numan, JA & Yusoff, IM 2024b, 'Identifying the key variables affecting condominium real estate appraisal in the context of Al Bireh City, Palestine: a literature and questionnaire-based approach', *Journal of Facilities Management*, <https://doi.org/10.1108/JFM-01-2024-0006>

Oust, A, Hansen, SN & Pettrem, TR 2020, 'Combining property price predictions from repeat sales and spatially enhanced hedonic regressions', *The Journal of Real Estate Finance and Economics*, vol. 61, pp. 183-207, <https://doi.org/10.1007/s11146-019-09723-x>

Oust, A, Westgaard, S, Waage, JE & Yemane, NK 2023, 'Assessing the explanatory power of dwelling condition in automated valuation models', *Journal of Real Estate Research*, pp. 1-27, <https://doi.org/10.1080/08965803.2023.2280280>

Özögür Akyüz, S, Eygi Erdogan, B, Yıldız, Ö & Karadayı Ataş, P 2023, 'A novel hybrid house price prediction model', *Computational economics*, vol. 62, pp. 1215-1232, <https://doi.org/10.1007/s10614-022-10298-8>

Potrawa, T & 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, pp. 50-65, <https://doi.org/10.1016/j.jbusres.2022.01.027>

Rampini, L & Re Cecconi, F 2022, 'Artificial intelligence algorithms to predict Italian real estate market prices', *Journal of Property Investment & Finance*, vol. 40, pp. 588-611, <https://doi.org/10.1108/JPIF-08-2021-0073>

Renigier-Bilozor, M, Žróbek, S & Walacik, M 2022, 'Modern technologies in the real estate market—opponents vs. proponents of their use: Does new category of value solve the problem?', *Sustainability*, vol. 14, p. 13403, <https://doi.org/10.3390/su142013403>

Ryu, D, Hong, J & Jo, H 2024, 'Capturing locational effects: application of the K-means clustering algorithm', *The Annals of Regional Science*, pp. 1-25, <https://doi.org/10.1007/s00168-024-01263-4>

Sharma, H, Harsora, H & Ogunleye, B 2024, 'An optimal house price prediction algorithm: XGBoost', *Analytics*, vol. 3, pp. 30-45, <https://doi.org/10.3390/analytics3010003>

Sisman, S & Aydinoglu, AC 2022, 'Improving performance of mass real estate valuation through application of the dataset optimization and spatially constrained multivariate clustering analysis', *Land Use Policy*, vol. 119, p. 106167, <https://doi.org/10.1016/j.landusepol.2022.106167>

Soltani, A & Lee, CL 2024, 'The non-linear dynamics of South Australian regional housing markets: A machine learning approach', *Applied Geography*, vol. 166, article 103248, <https://doi.org/10.1016/j.apgeog.2024.103248>

Soltani, A, Zali, N, Aghajani, H, Hashemzadeh, F, Rahimi, A & Heydari, M 2024, 'The nexus between transportation infrastructure and housing prices in metropolitan regions', *Journal of Housing and the Built Environment*, vol. 39, pp. 787-812, <https://doi.org/10.1007/s10901-023-10085-3>

Song, Y & Ma, X 2024, 'Exploration of intelligent housing price forecasting based on the anchoring effect', *Neural Computing and Applications*, vol. 36, pp. 2201-2214, <https://doi.org/10.1007/s00521-023-08823-3>

Stang, M, Krämer, B, Nagl, C & Schäfers, W 2023, 'From human business to machine learning—methods for automating real estate appraisals and their practical implications', *Zeitschrift Für Immobilienökonomie*, vol. 9, pp. 81-108, <https://doi.org/10.1365/s41056-022-00063-1>

Steurer, M, Hill, RJ & Pfeifer, N 2021, 'Metrics for evaluating the performance of machine learning based automated valuation models', *Journal of Property Research*, vol. 38, pp. 99-129, <https://doi.org/10.1080/09599916.2020.1858937>

Taecharungroj, V 2021, 'Google Maps amenities and condominium prices: Investigating the effects and relationships using machine learning', *Habitat International*, vol. 118, article 102463, <https://doi.org/10.1016/j.habitatint.2021.102463>

Theng, D & Bhoyar, KK 2024, 'Feature selection techniques for machine learning: a survey of more than two decades of research', *Knowledge and Information Systems*, vol. 66, pp. 1575-1637, <https://doi.org/10.1007/s10115-023-02010-5>

Torres-Pruñonosa, J, Garcia-Estevez, P & Prado-Roman, C 2021, 'Artificial neural network, quantile and semi-log regression modelling of mass appraisal in housing', *Mathematics*, vol. 9, article 783, <https://doi.org/10.3390/math9070783>

Trindade Neves, F, Aparicio, M & De Castro Neto, M 2024, 'The Impacts of Open Data and eXplainable AI on Real Estate Price Predictions in Smart Cities', *Applied Sciences*, vol. 14, article 2209, <https://doi.org/10.3390/app14052209>

Valdez Gómez De La Torre, FM & Chen, X 2024, 'Housing price determinants in Ecuador: A spatial hedonic analysis', *International Journal of Housing Markets and Analysis*, vol. 17, no. 6, pp. 1461-1487, <https://doi.org/10.1108/IJHMA-09-2023-0121>

Vargas-Calderón, V & Camargo, JE 2022, 'Towards robust and speculation-reduction real estate pricing models based on a data-driven strategy', *Journal of the Operational Research Society*, vol. 73, pp. 2794-2807, <https://doi.org/10.1080/01605682.2021.2023672>

Wahid, A, Kowalewski, O & Mantell, EH 2022, 'Determinants of the prices of residential properties in Pakistan', *Journal of Property Investment & Finance*, vol. 41, pp. 35-49, <https://doi.org/10.1108/JPIF-06-2021-0051>

Wang, Z, Wang, Y, Wu, S & Du, Z 2022, 'House price valuation model based on geographically neural network weighted regression: The case study of Shenzhen, China', *ISPRS International Journal of Geo-Information*, vol. 11, p. 450, <https://doi.org/10.3390/ijgi11080450>

Xu, L & Li, Z 2021, 'A new appraisal model of second-hand housing prices in China's first-tier cities based on machine learning algorithms', *Computational Economics*, vol. 57, pp. 617-637, <https://doi.org/10.1007/s10614-020-09973-5>

Xu, X & Zhang, Y 2021, 'House price forecasting with neural networks', *Intelligent Systems with Applications*, vol. 12, p. 200052, <https://doi.org/10.1016/j.iswa.2021.200052>

Xu, X & Zhang, Y 2023, 'Composite property price index forecasting with neural networks', *Property Management*, vol. 42, no. 3, pp. 388-411, <https://doi.org/10.1108/PM-11-2022-0086>

Yacim, JA & Boshoff, DGB 2019, 'A comparison of bandwidth and kernel function selection in geographically weighted regression for house valuation', *International Journal of Technology*, vol. 10, pp. 58-68, <https://doi.org/10.14716/ijtech.v10i1.975>

Yang, X, Yang, Y, Xu, S, Han, J, Chai, Z & Yang, G 2023, 'A new algorithm for large-scale geographically weighted regression with k-nearest neighbors', *ISPRS International Journal of Geo-Information*, vol. 12, p. 295, <https://doi.org/10.3390/ijgi12070295>

Yap, JBH, Yong, PS & Skitmore, M 2019, 'Analysing the desired quality of housing in the Klang Valley region, Malaysia', *Pacific Rim Property Research Journal*, vol. 25, pp. 125-140, <https://doi.org/10.1080/14445921.2019.1626542>

Ye, M, Zhu, L, Li, X, Ke, Y, Huang, Y, Chen, B, Yu, H, Li, H & Feng, H 2023, 'Estimation of the soil arsenic concentration using a geographically weighted XGBoost model based on hyperspectral data', *Science of The Total Environment*, vol. 858, p. 159798, <https://doi.org/10.1016/j.scitotenv.2022.159798>

Yilmazer, S & Kocaman, S 2020, 'A mass appraisal assessment study using machine learning based on multiple regression and random forest', *Land use policy*, vol. 99, article 104889, <https://doi.org/10.1016/j.landusepol.2020.104889>

Zaki, J, Nayyar, A, Dalal, S & Ali, ZH 2022, 'House price prediction using hedonic pricing model and machine learning techniques', *Concurrency and computation: practice and experience*, vol. 34, e7342, <https://doi.org/10.1002/cpe.7342>

Zhan, C, Liu, Y, Wu, Z, Zhao, M & Chow, TW 2023, 'A hybrid machine learning framework for forecasting house price', *Expert Systems with Applications*, vol. 233, article 120981, <https://doi.org/10.1016/j.eswa.2023.120981>

Zhang, H, Li, Y & Branco, P 2023, 'Describe the house and I will tell you the price: House price prediction with textual description data', *Natural Language Engineering*, pp. 1-35, <https://doi.org/10.1017/S1351324923000360>