

Evaluation of G-Xgboost Model Performance on Data with Different Levels of Heterogeneity Spatial and Linearity Patterns

Zamrah Mutmainah¹, Budi Susetyo^{1*}, Septian Rahardianto¹

¹Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics,
IPB University, Indonesia

budisu@apps.ipb.ac.id

ABSTRACT

Article History:

Received : 09-06-2026

Revised : 24-06-2026

Accepted : 24-06-2026

Online : 01-10-2026

Keywords:

G-XGBoost;

GW-XGBoost;

Simulation;

Spatial Heterogeneity.



Spatial data analysis often faces challenges arising from spatial heterogeneity and nonlinear relationships among variables. Conventional methods, such as Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR), are limited in capturing complex nonlinear patterns, while machine learning approaches, such as XGBoost, do not explicitly account for spatial variation. To address these limitations, Geographical-XGBoost (G-XGBoost) was developed by directly incorporating spatial information into the XGBoost learning process. This study evaluates the performance of G-XGBoost against OLS, GWR, XGBoost, and Geographically Weighted XGBoost (GW-XGBoost) through a simulation study using data generated at 119 locations on Java Island. Each observation included two predictor variables drawn from a standard normal distribution, with response variables generated through linear combinations or hyperbolic tangent (tanh) transformations of spatially varying coefficients. Four scenarios were constructed by combining two levels of relationship linearity (linear and nonlinear) and spatial heterogeneity (low and high), yielding coefficient patterns ranging from smooth gradients to highly dynamic local fluctuations. Model evaluation was based on 100 replications using independently generated training and testing datasets, assessed via R^2 , RMSE, MAE, and the training-testing performance gap as a measure of generalization stability. Under linear scenarios, GWR achieved the highest R^2 values on the test data (up to 0.94 and 0.92, respectively) with the lowest error metrics. In nonlinear scenarios, tree-based methods outperformed regression-based approaches, with GW-XGBoost achieving the highest test R^2 values of up to 0.89 and 0.85. Although G-XGBoost did not always achieve the highest predictive accuracy, it consistently produced training-testing R^2 gaps centered near zero across all scenarios, with an interquartile range of approximately -0.05 to 0.05, the narrowest among all competing methods, indicating more stable generalization and greater resistance to overfitting across varying levels of linearity and spatial heterogeneity.



<https://doi.org/10.31764/jtam.v10i4.40121>



This is an open access article under the [CC-BY-SA](#) license

A. INTRODUCTION

Spatial data analysis is a statistical approach that incorporates geographic location into regression modeling, allowing the relationship between variables to vary across regions. Unlike classical linear regression, which assumes parameter homogeneity across all observations, spatial data frequently exhibit variation in the effect of explanatory variables on the response

variable across locations, a phenomenon known as spatial heterogeneity (Anselin, 2022). Ignoring spatial heterogeneity can lead to inaccurate parameter estimation.

Comber et al. (2020) demonstrated that neglecting spatial structure can weaken a model's ability to capture genuine variation, particularly in regions with diverse characteristics. To address this limitation, Geographically Weighted Regression (GWR) was developed to allow parameters to vary across locations, thereby enabling more contextually representative estimates of local effects (Yang et al., 2026). Nevertheless, GWR has fundamental limitations: it can only handle linear relationships, is susceptible to multicollinearity, and lacks flexibility for large-scale and complex data, which can lead to unstable coefficients and reduced predictive accuracy.

The rapid advancement of machine learning over the past decade has produced a wide range of algorithms capable of modeling nonlinear relationships and complex interactions among variables, including Random Forest, Artificial Neural Networks, and XGBoost (Chen & Guestrin, 2016; Goodfellow et al., 2016; Salman et al., 2024). Among these methods, XGBoost has demonstrated strong predictive performance across various applications by capturing nonlinear relationships, modeling complex interactions, and mitigating overfitting through regularization (Rimal et al., 2024). However, in its global form, XGBoost treats observations independently and does not explicitly account for spatial dependence or spatial heterogeneity, thereby overlooking geographically varying relationships. In contrast, Wiedemann et al. (2023) showed that local models, such as Geographically Weighted Regression (GWR), tend to perform better on datasets with high spatial heterogeneity and limited sample sizes, whereas machine learning models generally achieve superior predictive performance on larger and more complex datasets. These findings suggest that the choice of modeling approach should consider the underlying spatial characteristics of the data and highlight the potential benefits of integrating the strengths of local spatial models and machine learning algorithms.

Recognizing the complementary strengths of GWR and machine learning algorithms, several studies have developed hybrid approaches that incorporate spatial weighting into machine learning models. Examples include the Geographically Weighted Artificial Neural Network (GWANN) proposed by Hagenauer & Helbich (2022), Geographically Weighted Random Forest (GWRF) developed by Suprayogi et al. (2024), GWBoost introduced by H. Wang et al. (2022), and Geographically Weighted XGBoost (GW-XGBoost) frameworks proposed by Ye et al. (2023) and Schimohr et al. (2023). These studies consistently reported improved predictive performance compared with their corresponding global models, highlighting the importance of incorporating spatial heterogeneity into machine learning frameworks.

Among these approaches, GW-XGBoost proposed by Ye et al. (2023) is one of the most widely cited. This method operates sequentially: first estimating local parameters with GWR, then using these coefficients to generate spatially weighted features that serve as inputs to the XGBoost model. In estimating soil arsenic concentrations, GW-XGBoost achieved an R^2 of 0.90, outperforming both GWR ($R^2 = 0.74$) and standard XGBoost ($R^2 = 0.48$). These findings demonstrate the potential benefits of integrating spatial information into machine learning models.

Despite its strong predictive performance, GW-XGBoost has several structural limitations. Khosravi and Ouarda (2026) noted that spatial information is incorporated through a separate preprocessing stage rather than being directly embedded within the XGBoost learning algorithm. Consequently, geographic information does not directly influence the tree-building process, and spatial heterogeneity is only indirectly represented in the final model. This design may also increase the risk of overfitting, as the feature transformation is performed prior to model training rather than integrated into the optimization process itself.

To address these limitations, Grekousis (2025) introduced the Geographical-XGBoost (G-XGBoost) model. Unlike GW-XGBoost, which incorporates spatial information before model training, G-XGBoost integrates geographic information directly into the XGBoost objective function and tree construction process. As a result, the formation of decision-tree splits explicitly considers geographic context, enabling the model to better represent spatially varying relationships. In addition, G-XGBoost generates local models and local feature importance measures, providing greater interpretability of how predictor effects vary across locations. These characteristics suggest that G-XGBoost may offer advantages in capturing spatial heterogeneity while retaining XGBoost's predictive capabilities. However, whether these theoretical advantages translate into consistently superior predictive performance over GW-XGBoost across varying spatial conditions has not yet been empirically verified.

Although both GW-XGBoost and G-XGBoost were developed to address spatial heterogeneity, they employ fundamentally different mechanisms for incorporating geographic information, which may have important implications for their practical use. G-XGBoost is theoretically more integrated, embedding spatial context directly into the learning process, whereas GW-XGBoost relies on a preprocessing stage. However, it remains unclear whether this theoretical difference leads to meaningful differences in predictive performance across varying spatial conditions. Without such evidence, researchers and practitioners lack a principled basis for choosing between the two methods. Moreover, existing studies have evaluated the two approaches separately on different empirical datasets, making it impossible to attribute performance differences to the modeling framework rather than to dataset-specific characteristics. A systematic and controlled comparison is therefore necessary to determine the conditions under which each method performs best and to guide model selection in applied spatial analysis.

In addition, most previous studies evaluating GW-XGBoost and G-XGBoost have relied on empirical datasets. For example, although these studies demonstrated the practical applicability of geographically weighted machine learning methods, they were conducted on datasets in which the true degree of spatial heterogeneity and the underlying functional form of the relationships were unknown (Chen et al., 2026; Han et al., 2026; Schimohr et al., 2023; Ye et al., 2023; Zhou et al., 2025). Consequently, the effects of spatial heterogeneity and nonlinearity are often intertwined with other dataset-specific factors, making it difficult to isolate the mechanisms responsible for differences in model performance. Because the ground truth is unobservable in empirical settings, it is impossible to determine whether a model performs well by genuinely capturing spatial heterogeneity or simply by overfitting the noise structure of a particular dataset. A simulation framework provides a controlled environment in which these characteristics can be manipulated independently, enabling a more rigorous

assessment of how GW-XGBoost and G-XGBoost respond to varying levels of spatial heterogeneity and nonlinear relationships.

Therefore, this study aims to evaluate the predictive performance of GW-XGBoost and G-XGBoost across varying levels of spatial heterogeneity and relationship linearity using a simulation study. The models are further compared with OLS, GWR, and standard XGBoost to provide a broader evaluation of spatial and non-spatial modeling approaches. In addition, this study investigates model stability and potential overfitting by comparing performance on training and testing datasets.

B. METHODS

This study employs a simulation approach designed to evaluate the predictive performance of the G-XGBoost model against other competing methods. The simulation was conducted using a real spatial structure that included the coordinates of 119 regencies/cities across Java Island. The selection of this location provides a spatial framework to ensure that the distribution of locations in the simulation reflects realistic geographic proximity patterns, rather than a purely random distribution of points. The use of real locations aims to preserve the characteristics of geographic distribution and realistic spatial neighborhood, so that the simulation scenarios can reflect conditions commonly encountered in spatial data.

1. Generate Data Processing (GDP)

The spatial framework was constructed using coordinates for 119 regencies/cities across Java Island. Each point represents the location of a regency/city in Java, with two independent variables X_1 and X_2 , each drawn from a standard normal distribution $N(0,1)$. Each point was also assigned an error component ε derived from a $N(0,1)$ distribution. Spatial heterogeneity was controlled through coefficients β that vary across locations using small-scale and large-scale cosine functions. Based on the combination of relationship form (linear/nonlinear) and level of heterogeneity (low/high), four simulation scenarios were constructed (Hagenauer & Helbich, 2022).

The nonlinear simulation design was adapted from Hagenauer & Helbich, (2022), who generated nonlinear relationships using the function $4\tanh(z/3)$. In the present study, the divisor inside the tanh function was removed, resulting in the use of $4\tanh(z)$. This modification increases the steepness of the S-shaped curve and accelerates the response's saturation. Consequently, the generated relationships exhibit stronger nonlinear characteristics and depart more substantially from linearity, providing a more demanding setting for evaluating the ability of GW-XGBoost and G-XGBoost to model nonlinear spatial processes.

$$\beta_0 = 1 \quad (1)$$

$$\beta_1(u_i, v_i) = 1 + \frac{u_i + v_i}{12} \quad (2)$$

$$\beta_2^{(1)}(u_i, v_i) = 1 + 2 \left(\cos\left(\frac{\pi u_i}{24}\right) \cos\left(\frac{\pi v_i}{24}\right) \right) \quad (3)$$

$$\beta_2^{(2)}(u_i, v_i) = 1 + 2 \left(\cos\left(\frac{\pi u_i}{12}\right) \cos\left(\frac{\pi v_i}{12}\right) \right) \quad (4)$$

The scenarios used are as follows:

- a. The first scenario contains a linear relationship with low spatial heterogeneity, where the coefficient $\beta_1(u_i, v_i)$ varies linearly with respect to location.

$$y_i = \beta_0 + \beta_1(u_i, v_i)x_{1i} + \beta_2^{(1)}(u_i, v_i)x_{2i} + \varepsilon_i \quad (5)$$

- b. The second scenario contains a linear relationship with high spatial heterogeneity, where the coefficient $\beta_1(u_i, v_i)$ varies linearly with respect to location

$$y_i = \beta_0 + \beta_1(u_i, v_i)x_{1i} + \beta_2^{(2)}(u_i, v_i)x_{2i} + \varepsilon_i \quad (6)$$

- c. The third scenario covers a nonlinear relationship with low spatial variation, using a hyperbolic tangent activation function to introduce non-linearity.

$$y_i = \beta_0 + 4 \tanh(\beta_1(u_i, v_i)x_{1i}) + 4 \tanh(\beta_2^{(1)}(u_i, v_i)x_{2i}) + \varepsilon_i \quad (7)$$

- d. The fourth scenario represents a combination of nonlinear relationships and high spatial heterogeneity.

$$y_i = \beta_0 + 4 \tanh(\beta_1(u_i, v_i)x_{1i}) + 4 \tanh(\beta_2^{(1)}(u_i, v_i)x_{2i}) + \varepsilon_i \quad (8)$$

All XGBoost models, whether used as a global model or as a local component within G-XGBoost, were run with default hyperparameters. This approach was chosen because the use of default values simplifies the simulation process, allowing it to remain focused on the primary objective of the study, namely, to evaluate model performance in a more objective manner, as the results obtained are not influenced by parameter configurations that may favor one model over another.

All simulations and model implementations were conducted in Google Colab using *Python* and the most recent stable versions of the required packages at the time of the study. The GWR model was implemented with the '*mgwr*' package, and the global XGBoost model with the *xgboost* package. The GW-XGBoost framework was implemented sequentially: first, estimating local coefficients and bandwidths with '*mgwr*', then using the resulting spatially weighted features as inputs to the standard XGBoost algorithm. The G-XGBoost model was implemented using the '*geoxgboost*' package. To ensure a fair comparison among competing methods, all XGBoost-based models were estimated using the default hyperparameter settings provided by the corresponding package implementations, without additional tuning or optimization. This approach was adopted because the study's primary objective was to evaluate the intrinsic performance of each modeling framework rather than the effects of hyperparameter optimization. Consequently, differences in predictive performance can be more directly attributed to the models' methodological characteristics rather than to customized parameter configurations (Rimal et al., 2024).

For the spatial models, an adaptive bi-square kernel was employed following the default settings of the respective packages. In the GWR model, the optimal bandwidth was selected using the automatic bandwidth selection procedure implemented in the mgwr package. The same bandwidth obtained from GWR was subsequently used in GW-XGBoost to ensure consistency in the spatial weighting process. For G-XGBoost, bandwidth optimization was performed via a grid search with candidate bandwidth values from 10 to 100 in increments of 3. To ensure reproducibility of the simulation results, all random processes were initialized using a fixed random seed of 42.

2. Extreme Gradient Boosting (XGBoost)

XGBoost is a supervised learning algorithm proposed by Chen & Guestrin (2016) that extends the GBDT framework. The model sequentially adds decision trees to improve prediction accuracy while controlling model complexity through regularization. Its objective function consists of a loss function and a complexity penalty:

$$\mathcal{L}(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k), \quad (9)$$

The first component, $l(\hat{y}_i, y_i)$ is the loss function that measures the discrepancy between the predicted and actual values. To determine whether a split is worthwhile, XGBoost uses the following gain score:

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (10)$$

where G_L, H_L and G_R, H_R represent the sums of the gradients and Hessians for the left and right branches resulting from the split, respectively. If the gain is greater than zero, the split is performed because it improves the objective function value Chen & Guestrin (2016).

3. Geographically Weighted Regression (GWR)

Geographically Weighted Regression (GWR) is a regression method that incorporates spatial location effects into the relationship between independent and dependent variables. The basic GWR model can be expressed as:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^p \beta_k(u_i, v_i) x_{ik} + \varepsilon_i, \quad (11)$$

In GWR, bandwidth is a key parameter that determines the spatial extent of influence among observations through the weighting function (Z. Wang et al., 2022). Spatial weights are estimated using a kernel function, where observations closer to a location receive higher weights than those farther away. A small bandwidth uses fewer neighboring observations, resulting in higher variance and less stable parameter estimates.

4. Geographically Weighted- XGBoost (GW-XGBoost)

GW-XGBoost was developed to address the limitations of traditional regression and conventional machine learning models in handling spatially non-stationary data. The method integrates GWR with XGBoost by first estimating local regression coefficients β_k , which are then used as spatial weights for each input feature. This process produces a spatially weighted input matrix (x^w) that incorporates both feature values and their geographic influence, expressed as:

$$x^w = \begin{bmatrix} w_1^1 V_1^1 & w_1^2 V_1^2 & \dots & w_1^m V_1^m \\ \vdots & \vdots & \ddots & \vdots \\ w_i^1 V_1^i & w_i^2 V_2^i & \dots & w_i^m V_m^i \end{bmatrix} \quad (12)$$

In this matrix, V_i^m represents the m -th feature of the i -th observation, while w_i^m denotes the spatial weight derived from the corresponding GWR coefficient. The resulting spatially weighted matrix, open paren x to the w, close paren, (x^w) is then used as input to the XGBoost algorithm for ensemble-based learning (Ye et al., 2023).

5. Geographical - XGBoost (G-XGBoost)

G-XGBoost extends XGBoost by incorporating spatial proximity into the training process. For each location, the model is calibrated using a subset of neighboring observations selected through a bandwidth b and spatial weights w_s derived from a kernel function. This local learning strategy enables the model to capture spatial heterogeneity more effectively. Unlike the global XGBoost model, G-XGBoost constructs n local models, one for each spatial unit. For location i , the prediction $\hat{y}_{i-loc}$ is generated using decision trees f_k trained on locally weighted neighboring data x_{ib} , and can be expressed as follows:

$$\hat{y}_{i-loc} = \sum_{k=1}^K (u_i, v_i) f_k(w_s, x_{ib}), \quad f_k \in \mathcal{F} \quad (13)$$

where (u_i, v_i) denotes the coordinates of the i -th spatial unit serving as the local reference point. Spatial weights w_s are computed using a distance-based bi-square kernel with bandwidth b . G-XGBoost evaluates node splits using a spatially weighted gain function that measures the improvement in model performance for each feature, calculated as follows:

$$\text{Gain} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} w_{s_i} g_i)^2}{\sum_{i \in I_L} w_{s_i} h_i + \lambda} + \frac{(\sum_{i \in I_R} w_{s_i} g_i)^2}{\sum_{i \in I_R} w_{s_i} h_i + \lambda} - \frac{(\sum_{i \in I} w_{s_i} g_i)^2}{\sum_{i \in I} w_{s_i} h_i + \lambda} \right] - \gamma \quad (14)$$

The parameters λ and γ are regularization hyperparameters that control tree complexity. In G-XGBoost, the gain function incorporates spatial weights w_{s_i} , which represent the geographic proximity of observation i to the local focal point. As a result, nearby observations receive greater influence, leading to more spatially relevant splits (Grekousis, 2025).

In addition to local modeling, G-XGBoost combines local and global predictions through an ensemble approach to enhance prediction accuracy and interpretability. The final prediction is determined by the ensemble weight α_i , which controls the relative contribution of the local and global models:

$$\hat{y}_{i-ens} = \alpha_i \hat{y}_{i-loc} + (1 - \alpha_i) \hat{y}_{i-gl} \quad (15)$$

where $\hat{y}_{i-loc}$ and $\hat{y}_{i-gl}$ are the local and global predictions, respectively, and $\alpha_i \in [0,1]$ is the ensemble weight. This study considers two weighting schemes: fixed weights, which use the same α for all locations, and varying weights, where α_i adapts according to the relative performance of the local and global models at each spatial location. The final prediction in G-XGBoost is obtained by combining local and global model outputs through an ensemble weighting parameter (α), as illustrated in Figure 1.

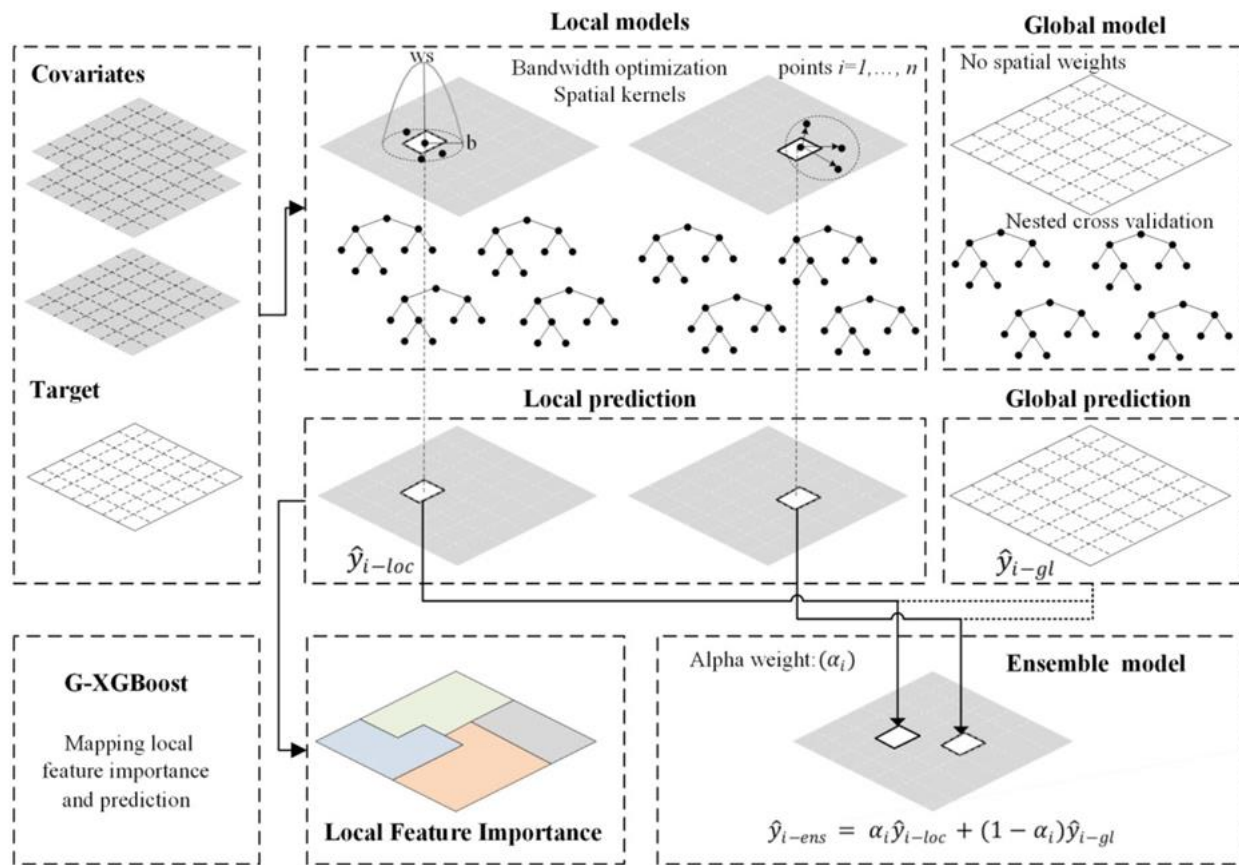


Figure 1. G-XGBoost Framework

C. RESULT AND DISCUSSION

1. Generate Processing Data

This study employs four simulation scenarios constructed by combining different levels of linearity and spatial heterogeneity. Based on the spatial variables u and v , the coefficient parameters β were generated to create simulation scenarios representing both linear and nonlinear relationships at varying levels of spatial heterogeneity. The spatial distributions of the coefficient parameters β used in the simulation scenarios are illustrated in Figure X. These

visualizations depict how the coefficient values vary across geographic locations throughout Java Island, as shown in Figure 2.

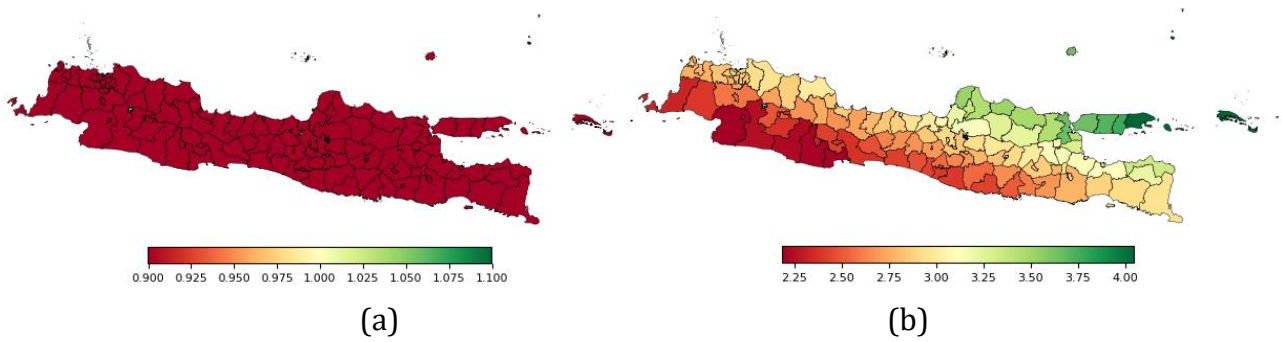


Figure 2. Spatial Surfaces of Simulated Coefficients: (a) β_0 , (b) β_1

Figure 2a shows the intercept coefficient (β_0), which is held constant across all locations. As a result, there is no spatial variation, and the map exhibits a nearly uniform color pattern without a noticeable gradient. This specification was intended to create a controlled simulation setting in which the intercept remains spatially invariant, allowing the effects of other spatially varying coefficients to be examined independently. In contrast, Figure 2b presents the coefficient (β_1), which was generated using a structured linear spatial pattern. This produces a smooth west-to-east color gradient, reflecting a systematic change in the variable's influence across geographic locations, as shown in Figure 3.

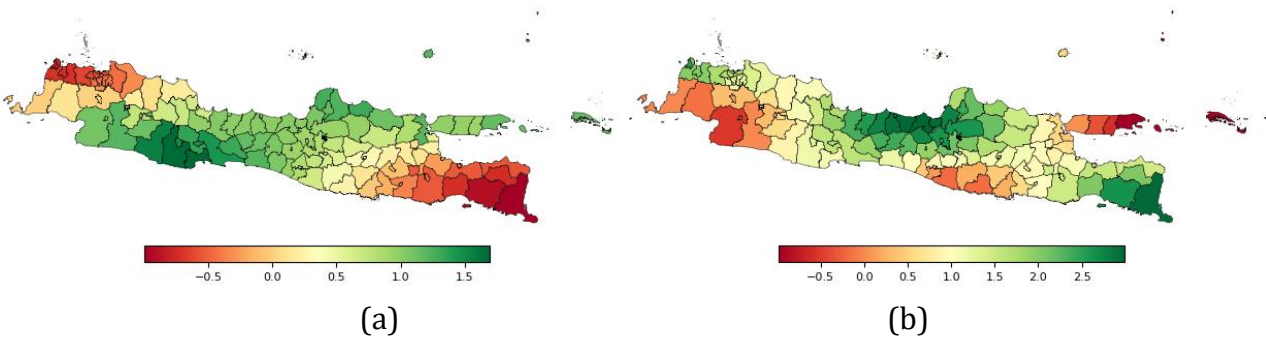


Figure 3. Spatial Surfaces of β_2 Coefficients: (a) Low Spatial Heterogeneity ($\beta_2^{(1)}$), (b) High Spatial Heterogeneity ($\beta_2^{(2)}$)

For the low spatial heterogeneity scenario, the coefficient β_2 is generated using a low-frequency cosine function, producing a smooth wave-like pattern in which coefficient values change gradually across regions. This pattern exhibits broad spatial clustering and introduces turning points (peaks and troughs) that are not present in the linear coefficient pattern. In contrast, the high spatial heterogeneity scenario employs a higher-frequency cosine function, resulting in rapid and highly dynamic fluctuations in coefficient values, even between neighboring regions. Together, these coefficient configurations are designed to evaluate the robustness of the models under varying spatial conditions, ranging from smooth global trends to highly complex local variations (Comber et al., 2024).

2. Parameter settings

The evaluation of predictive performance began by generating training and testing data independently, from the same distribution and at the same spatial locations in each replication (Kattenborn et al., 2022). Although the coordinates remained fixed, the predictor variables and error terms were regenerated independently, ensuring that the training and testing sets did not originate from the same observations or share direct spatial dependence. This approach enables the assessment of model generalization to new data realizations while preserving the underlying spatial structure. Furthermore, it provides a consistent evaluation framework for comparing five methods with different characteristics, namely OLS and GWR as parametric regression-based models, and XGBoost, GW-XGBoost, and G-XGBoost as nonparametric tree-based models. Differences in the use of spatial information can also be evaluated more objectively, as OLS and XGBoost do not incorporate location information, whereas GWR, GW-XGBoost, and G-XGBoost explicitly utilize geographic coordinates through spatial weighting mechanisms.

To ensure stable performance estimates and reduce the influence of random variation, each simulation scenario was repeated 100 times. This replication strategy provides a more robust evaluation of model performance based on R^2 , RMSE, and MAE (Chicco et al., 2021). Candidate bandwidth values ranged from 10 to 100 in increments of 3, yielding a discrete set of bandwidth candidates evaluated across all simulation replications. The search interval was selected based on preliminary analyses indicating that optimal bandwidth values generally fell within this range, while also maintaining a balance between predictive accuracy and computational efficiency. Unlike the continuous optimization procedure used in GWR, a grid search was adopted for G-XGBoost due to the substantial computational burden of 100 simulation replications. Although this approach may not always identify the exact optimal bandwidth, applying the same procedure consistently across all scenarios ensures a fair comparison among methods (J. Wang et al., 2021; Z. Wang et al., 2022).

In addition to bandwidth optimization, G-XGBoost requires the ensemble parameter α , which controls the contribution of the global and local models. Five candidate values were evaluated: $\alpha = 0.00, 0.25, 0.50, 0.75$, and 1.00 . The results showed a highly consistent pattern across all scenarios, with $\alpha = 1.00$ achieving the best predictive performance, characterized by the highest R^2 values and the smallest variability across replications.

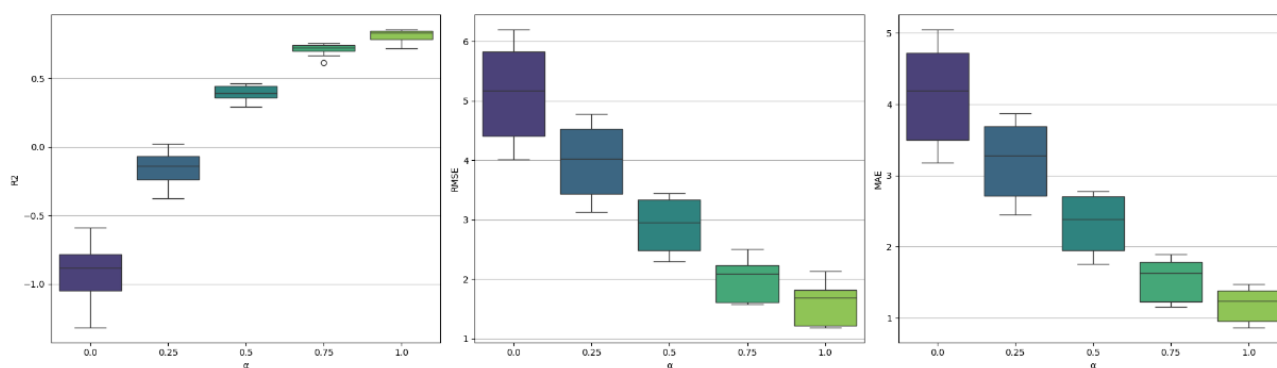


Figure 4. Boxplots of G-XGBoost performance metrics across different α values.

Figure 4 shows a substantial improvement in predictive performance when $\alpha = 1.0$, indicating that G-XGBoost's predictive capability relies entirely on the local modeling component. This finding suggests spatial heterogeneity in the simulated data, with relationships varying across locations. Consequently, local models are better suited than global models for capturing location-specific characteristics and producing more accurate predictions (Lessani & Li, 2024).

3. Comparison of Model Predictive Performance

This section compares the predictive performance of OLS, GWR, XGBoost, GW-XGBoost, and G-XGBoost under different combinations of relationship linearity and spatial heterogeneity. Because the characteristics of the simulated data vary across scenarios, differences in model performance can provide insights into the conditions under which each method is most effective and the extent to which spatial information contributes to predictive accuracy, as shown in Table 1 and Figure 5.

Table 1. Descriptive Statistics of testing data performance (Mean $\pm$ SD)

Scenario	Metric	Mean $\pm$ SD				
		OLS	GWR	XGBoost	GW-XGBoost	G-XGBoost
Linear-Low	R^2	0.90 $\pm$ 0.02	0.94 $\pm$ 0.01	0.84 $\pm$ 0.03	0.89 $\pm$ 0.02	0.84 $\pm$ 0.03
	RMSE	0.94 $\pm$ 0.07	0.73 $\pm$ 0.06	1.21 $\pm$ 0.10	0.98 $\pm$ 0.10	1.21 $\pm$ 0.12
	MAE	0.72 $\pm$ 0.05	0.57 $\pm$ 0.04	0.92 $\pm$ 0.07	0.75 $\pm$ 0.07	0.92 $\pm$ 0.08
Linear-High	R^2	0.89 $\pm$ 0.02	0.92 $\pm$ 0.02	0.81 $\pm$ 0.04	0.85 $\pm$ 0.03	0.80 $\pm$ 0.04
	RMSE	1.09 $\pm$ 0.10	0.94 $\pm$ 0.08	1.45 $\pm$ 0.16	1.28 $\pm$ 0.13	1.49 $\pm$ 0.15
	MAE	0.80 $\pm$ 0.07	0.69 $\pm$ 0.05	1.07 $\pm$ 0.10	0.95 $\pm$ 0.09	1.11 $\pm$ 0.10
Nonlinear-Low	R^2	0.62 $\pm$ 0.05	0.72 $\pm$ 0.04	0.66 $\pm$ 0.05	0.84 $\pm$ 0.03	0.78 $\pm$ 0.04
	RMSE	2.48 $\pm$ 0.16	2.15 $\pm$ 0.16	2.35 $\pm$ 0.15	1.62 $\pm$ 0.16	1.89 $\pm$ 0.15
	MAE	1.98 $\pm$ 0.13	1.73 $\pm$ 0.13	1.76 $\pm$ 0.12	1.23 $\pm$ 0.12	1.42 $\pm$ 0.11
Nonlinear-High	R^2	0.73 $\pm$ 0.05	0.73 $\pm$ 0.05	0.84 $\pm$ 0.04	0.84 $\pm$ 0.04	0.84 $\pm$ 0.03
	RMSE	2.25 $\pm$ 0.16	2.26 $\pm$ 0.17	1.75 $\pm$ 0.16	1.74 $\pm$ 0.17	1.75 $\pm$ 0.14
	MAE	1.79 $\pm$ 0.12	1.79 $\pm$ 0.12	1.28 $\pm$ 0.12	1.27 $\pm$ 0.12	1.30 $\pm$ 0.11

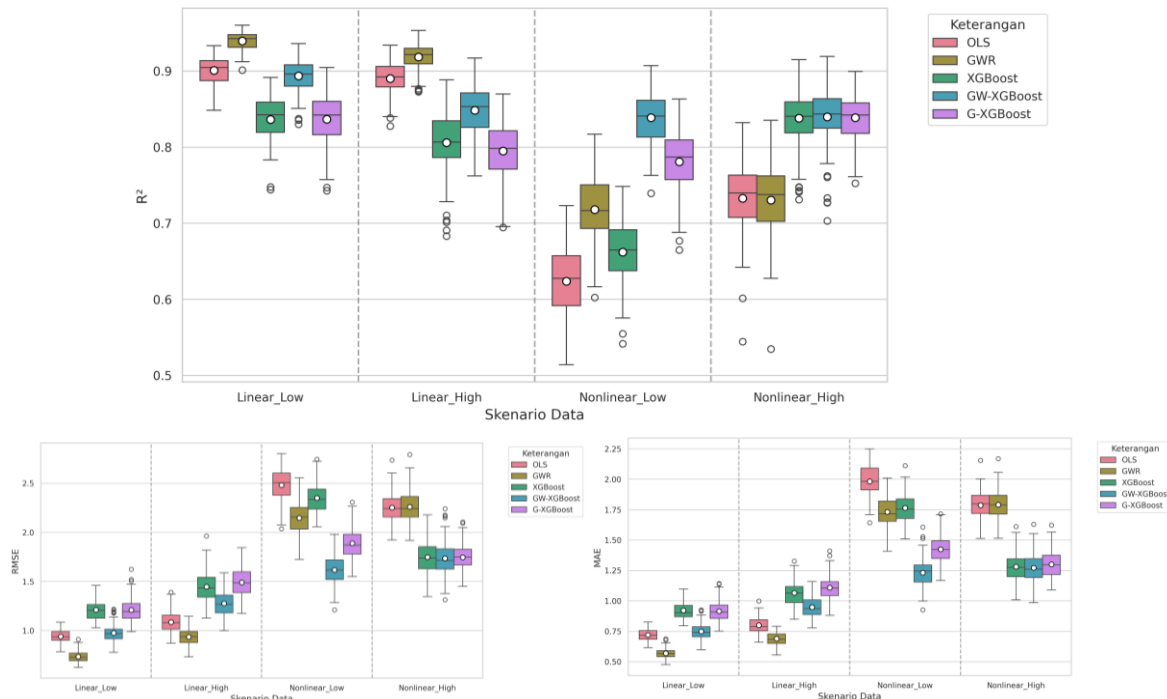


Figure 5. Boxplots of model performance evaluation on the testing data

Table 1 and Figure 5 present the descriptive statistics and boxplots of the testing R^2 , RMSE, and MAE values across 100 replications for each method and simulation scenario. Both summaries reveal consistent patterns across all evaluation metrics. Under the linear scenarios, OLS and GWR achieved the best predictive performance, indicating that regression-based approaches remain highly effective when the underlying relationships are linear. In contrast, their performance deteriorated considerably under the nonlinear scenarios, where the tree-based methods demonstrated superior predictive capability. Among these methods, GW-XGBoost generally produced the best results across most nonlinear scenarios, suggesting that incorporating spatially weighted information into the boosting framework improves its ability to capture complex nonlinear and spatially heterogeneous relationships.

Although GW-XGBoost consistently achieved the highest average performance in most nonlinear settings, its advantage over G-XGBoost was relatively small across the evaluation metrics. The substantial overlap observed in the boxplots also suggests that the apparent differences may not necessarily reflect statistically significant performance improvements. Therefore, a formal statistical comparison is required to determine whether the observed differences in ranking are meaningful. To address this issue, the Friedman test was conducted to evaluate whether significant performance differences exist among the competing methods across the simulation replications. The results of the Friedman test are presented in Table 2.

Table 2. Friedman test results for performance testing data

Metric	Chi-Square (p-value)			
	Linear-Low	Linear-High	Nonlinear-Low	Nonlinear-High
R^2	361.640 (p < 0.001)	367.088 (p < 0.001)	356.032 (p < 0.001)	301.744 (p < 0.001)
RMSE	361.640 (p < 0.001)	367.088 (p < 0.001)	356.032 (p < 0.001)	301.744 (p < 0.001)
MAE	362.312 (p < 0.001)	370.696 (p < 0.001)	361.800 (p < 0.001)	303.560 (p < 0.001)

Table 2 presents the results of the Friedman test comparing the performance of the evaluated methods across all simulation scenarios. The results indicate that the p-values for all scenarios are below the 5% significance level, leading to the rejection of the null hypothesis that there are no performance differences among the competing methods. Therefore, statistically significant differences in performance are observed among methods across all simulation scenarios. Overall, these findings reinforce previous results, showing that OLS and GWR tend to perform best in linear settings, whereas tree-based methods, particularly GW-XGBoost and G-XGBoost, are more effective at capturing nonlinear relationships and spatial heterogeneity. However, the descriptive statistics reveal that the performance differences among the tree-based methods, especially in the *nonlinear-high* scenario, are relatively small, making it difficult to identify a clearly superior method based solely on the average evaluation metrics. Therefore, a further pairwise comparison using the Wilcoxon signed-rank test is required to identify which pairs of methods differ significantly in performance, particularly in nonlinear scenarios where competition among the tree-based methods is most pronounced (Table 3).

Table 3. Wilcoxon post-hoc comparison of performance testing data

Scenario	Method 1	Method 2	Wilcoxon Statistic	p-Holm
Linear-Low	XGBoost	G-XGBoost	2421.0	0.721
Nonlinear-High	OLS	GWR	2080.0	0.504
Nonlinear-High	XGBoost	GW-XGBoost	2377.0	1.000
Nonlinear-High	XGBoost	G-XGBoost	2435.0	1.000
Nonlinear-High	GW-XGBoost	G-XGBoost	2300.0	1.000

Table 3 presents the results of the Wilcoxon post-hoc test, showing only method pairs that do not exhibit statistically significant performance differences ($p\text{-Holm} > 0.05$). The results are reported using RMSE, as R^2 and MAE yielded the same conclusions regarding the significance of pairwise comparisons. Therefore, method pairs not listed in the table can be considered to differ significantly in performance.

Based on Table 3, XGBoost and G-XGBoost do not differ significantly in performance under the *Linear-Low* scenario. Meanwhile, under the *Nonlinear-High* scenario, OLS and GWR also do not differ significantly. Furthermore, the three tree-based methods, namely XGBoost, GW-XGBoost, and G-XGBoost, do not differ statistically from one another. These findings indicate that although differences in average evaluation metrics are observed in Table 1, several methods still achieve statistically comparable performance under certain conditions, particularly in scenarios characterized by high levels of nonlinearity and spatial heterogeneity.

Overall, the Friedman and Wilcoxon test results suggest that no single method consistently outperforms all others with statistically significant superiority across all simulation scenarios. Therefore, the subsequent evaluation focuses not only on predictive accuracy on the test data but also on the model's performance stability. This stability is assessed by analyzing the performance gap between the training and testing datasets, which helps evaluate model generalization and identify potential overfitting. Such an analysis is particularly important, given that Khosravi and Ouarda (2026) reported that GW-XGBoost may exhibit a tendency toward overfitting, necessitating an investigation of whether a similar phenomenon emerges in

the simulation data used in this study (Burt et al., 2025; Guan & Burton, 2022). The following are the descriptive statistics for gap performance (mean $\pm$ SD), as shown in Table 4 and Figure 6.

Table 4. Descriptive Statistics of gap performance (Mean $\pm$ SD)

Scenario	Metric	Mean $\pm$ SD				
		OLS	GWR	XGBoost	GW-XGBoost	G-XGBoost
Linear-Low	R^2	0.01 ± 0.02	0.02 ± 0.01	0.16 ± 0.03	0.11 ± 0.02	0.02 ± 0.04
	RMSE	-0.03 ± 0.08	-0.10 ± 0.08	-1.21 ± 0.10	-0.97 ± 0.10	-0.06 ± 0.14
	MAE	-0.02 ± 0.06	-0.07 ± 0.06	-0.92 ± 0.07	-0.75 ± 0.07	-0.04 ± 0.10
Linear-High	R^2	0.01 ± 0.03	0.02 ± 0.02	0.19 ± 0.04	0.15 ± 0.03	0.02 ± 0.05
	RMSE	-0.03 ± 0.13	-0.13 ± 0.11	-1.44 ± 0.16	-1.27 ± 0.13	-0.07 ± 0.20
	MAE	-0.02 ± 0.08	-0.09 ± 0.07	-1.06 ± 0.10	-0.95 ± 0.09	-0.05 ± 0.14
Nonlinear-Low	R^2	0.01 ± 0.07	0.07 ± 0.05	0.34 ± 0.05	0.16 ± 0.03	0.00 ± 0.05
	RMSE	-0.09 ± 0.21	-0.30 ± 0.19	-2.34 ± 0.15	-1.61 ± 0.16	-0.04 ± 0.20
	MAE	-0.07 ± 0.17	-0.23 ± 0.16	-1.76 ± 0.12	-1.23 ± 0.12	-0.04 ± 0.15
Nonlinear-High	R^2	0.01 ± 0.06	0.04 ± 0.06	0.16 ± 0.04	0.16 ± 0.04	0.01 ± 0.04
	RMSE	-0.06 ± 0.24	-0.19 ± 0.25	-1.74 ± 0.16	-1.73 ± 0.17	-0.07 ± 0.18
	MAE	-0.05 ± 0.18	-0.14 ± 0.19	-1.28 ± 0.12	-1.27 ± 0.12	-0.06 ± 0.14

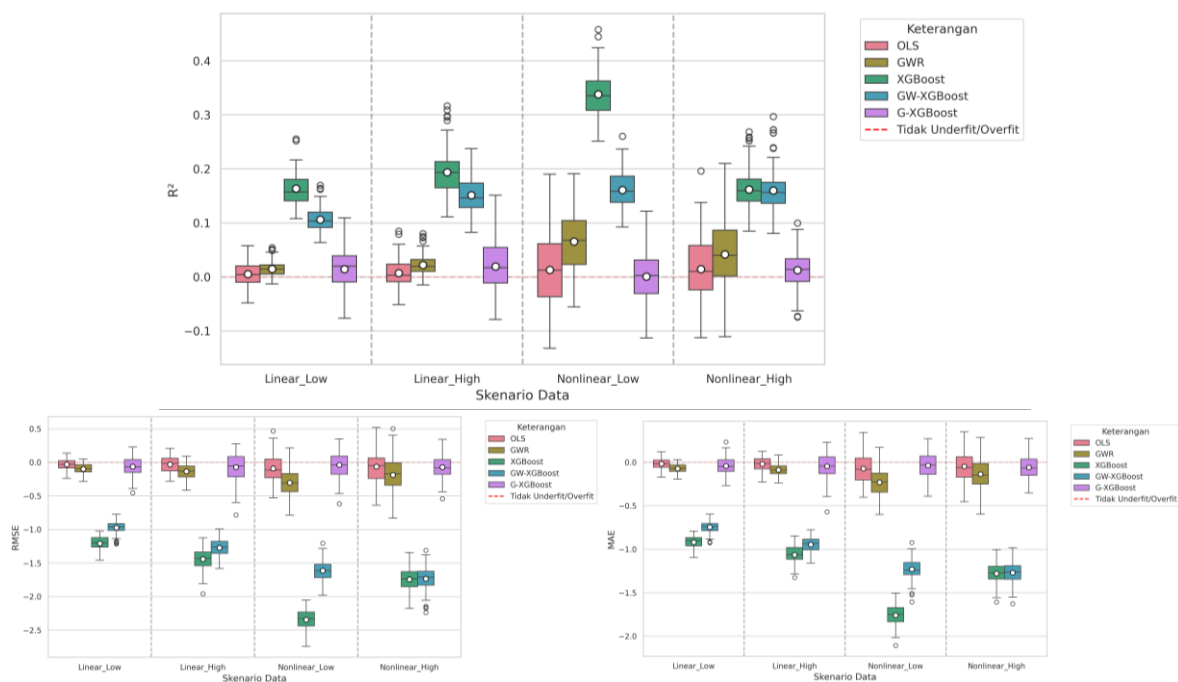


Figure 6. Boxplots of the performance gap training and testing data

Table 4 and Figure 6 present the descriptive statistics and boxplots of the training–testing performance gaps across all simulation scenarios. The performance gap reflects model stability and generalization ability, where values closer to zero indicate more consistent performance between the training and testing datasets, while larger gaps suggest a greater tendency toward overfitting. The observed patterns are highly consistent across all evaluation metrics (R^2 , RMSE, and MAE).

From a model stability perspective, the five methods exhibit distinct patterns. G-XGBoost and OLS generally produce relatively small performance gaps centered around zero, indicating strong consistency between training and testing performance. In contrast, XGBoost and, in particular, GW-XGBoost exhibit substantially larger performance gaps, suggesting that the performance improvements achieved during training are not fully maintained when the models are applied to unseen data. This finding is noteworthy because Khosravi and Ouarda (2026) also reported relatively large performance gaps for GW-XGBoost, which they attributed to the dataset's specific characteristics. However, the results of the present study reveal a similar pattern consistently across simulation scenarios with varying levels of nonlinearity and spatial heterogeneity. Therefore, the large performance gaps observed for GW-XGBoost may not be solely attributable to dataset-specific characteristics, but may also reflect an inherent tendency of the method to overfit relative to the competing approaches. However, the observed patterns are based solely on descriptive evidence. Therefore, a Friedman test was subsequently performed to formally assess whether the differences in performance gaps among the competing methods were statistically significant (Table 5).

Table 5. Friedman Test Results for Performance Gaps

Metric	Chi-Square (p-value)			
	Linear-Low	Linear-High	Nonlinear-Low	Nonlinear-High
R^2	326.552 (<0.001)	326.744 (<0.001)	344.360 (<0.001)	320.312 (<0.001)
RMSE	338.056 (<0.001)	337.304 (<0.001)	355.592 (<0.001)	328.120 (<0.001)
MAE	335.192 (<0.001)	340.096 (<0.001)	350.696 (<0.001)	320.816 (<0.001)

As shown in Table 5, the Friedman test yielded significant results across all simulation scenarios and evaluation metrics (all $p < 0.001$), indicating that at least one method differed significantly from the others in terms of the training–testing performance gap. Nevertheless, the descriptive statistics presented in Table 4 and Figure 6 suggest that the stability differences between G-XGBoost and OLS are relatively small, with both methods producing performance gaps centered close to zero. Therefore, a paired Wilcoxon signed-rank test was conducted as a follow-up analysis to determine whether the observed stability differences between OLS and G-XGBoost were statistically significant (Table 6).

Table 6. Wilcoxon post-hoc comparison of OLS and G-XGBoost performance gaps

Metric	OLS vs G-XGBoost			
	Linear-Low	Linear-High	Nonlinear-Low	Nonlinear-High
R^2	0.007	0.018	0.114	1.000
RMSE	0.010	0.040	0.073	1.000
MAE	0.014	0.042	0.045	0.652

Table 6 presents the results of the paired Wilcoxon test comparing the training–testing performance gaps between OLS and G-XGBoost. The results indicate that the stability differences between the two methods are generally small. Under the Nonlinear-High scenario, all evaluation metrics yielded p-values greater than 0.05, indicating that OLS and G-XGBoost do not differ significantly in model stability. This finding is consistent with the descriptive

statistics in Table 4 and Figure 6, which show that the performance-gap distributions for both methods are centered near zero.

Nevertheless, when considered alongside the testing performance results in Table 1, G-XGBoost demonstrated a favorable balance between predictive performance and model stability. G-XGBoost maintained relatively consistent predictive performance across all simulation scenarios, with R^2 values generally remaining around 0.80. In contrast, although OLS achieved excellent performance under linear scenarios, with R^2 values approaching 0.90, its predictive accuracy declined substantially under nonlinear scenarios, where R^2 dropped to approximately 0.70. Therefore, while OLS and G-XGBoost exhibited comparable stability, G-XGBoost maintained more consistent predictive performance across varying levels of nonlinearity and spatial heterogeneity.

Furthermore, the Wilcoxon post-hoc results presented in Table 3 indicate that the predictive performance of G-XGBoost did not differ significantly from that of XGBoost and GW-XGBoost across several scenarios. However, the performance-gap analysis revealed that G-XGBoost generally produced smaller training-testing performance gaps than both methods. This finding may be related to the way spatial information is incorporated into the model structure. In GW-XGBoost, spatial information is incorporated via a preprocessing step that transforms the predictor variables into spatially weighted features prior to model training. While this approach may enhance the model's ability to capture highly localized spatial patterns and achieve slightly higher predictive accuracy, it may also increase sensitivity to variations in the training data, resulting in larger training-testing performance gaps (Khosravi & Ouarda, 2026).

In contrast, G-XGBoost incorporates spatial information directly into the tree-construction process through a spatially weighted gain function. Consequently, spatial context influences the selection of decision-tree splits rather than modifying the input features themselves, leading to a more integrated learning process (Grekousis, 2025). This mechanism may reduce the tendency to fit highly localized patterns that are specific to the training data, resulting in more stable predictive performance across different data realizations. However, the same mechanism may also make G-XGBoost less aggressive in exploiting local variations, which can limit the maximum predictive accuracy achieved in some scenarios. These findings suggest a trade-off between predictive accuracy and generalization: GW-XGBoost tends to prioritize exploiting local spatial patterns, whereas G-XGBoost tends to provide greater stability and robustness across varying levels of nonlinearity and spatial heterogeneity.

D. CONCLUSION AND SUGGESTIONS

The simulation study demonstrates that no single method consistently achieved the best predictive performance across all data conditions. Under the linear scenarios, the parametric regression models, namely OLS and GWR, produced the highest predictive accuracy on the testing data. In contrast, under the nonlinear scenarios, the tree-based methods outperformed the regression-based approaches, with GW-XGBoost generally achieving higher predictive performance than G-XGBoost and standard XGBoost. However, when model stability was evaluated by the performance gap between the training and testing datasets, GW-XGBoost

showed stronger signs of overfitting, whereas G-XGBoost consistently produced smaller performance gaps across most simulation scenarios.

These findings were supported by descriptive statistics, boxplot visualizations, and both Friedman and Wilcoxon post-hoc tests conducted on testing performance and training–testing performance gaps. Overall, G-XGBoost emerged as the most favorable method because it achieved competitive predictive accuracy without sacrificing model generalization. As a result, G-XGBoost provides a better balance between predictive performance and robustness than the other tree-based approaches. Therefore, G-XGBoost can be recommended as an effective method, particularly for applications that require both the flexibility of machine learning and the ability to capture spatially varying relationships.

This study is subject to several limitations. The simulation design was intentionally restricted to variations in spatial heterogeneity and relationship linearity, without explicitly incorporating spatial autocorrelation. Furthermore, nonlinear relationships were generated exclusively using the hyperbolic tangent (*tanh*) function. Future research should therefore extend the simulation framework by incorporating varying levels of spatial autocorrelation and exploring alternative nonlinear functions to assess the robustness of the competing methods across a wider range of spatial data conditions.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Department of Statistics and Data Science, IPB University, for providing academic support and research facilities that contributed to the completion of this study. The authors also acknowledge the use of ChatGPT (OpenAI), Claude (Anthropic), and Grammarly solely for language editing, grammar checking, and improving the manuscript's clarity and readability. These tools were not used for data analysis, scientific interpretation, conceptual development, or drawing research conclusions. The authors take full responsibility for the accuracy, originality, and integrity of the content presented in this work.

REFERENCES

- Anselin, L. (2022). Spatial econometrics. *Handbook of Spatial Analysis in the Social Sciences*, 101–122. <https://doi.org/10.4337/9781789903942.00014>
- Burt, D. R., Shen, Y., & Broderick, T. (2025). Consistent Validation for Predictive Methods in Spatial Settings. *Proceedings of Machine Learning Research*, 258, 820–828. <https://doi.org/10.48550/arXiv.2402.03527>
- Chen, M., Liu, Y., Grekousis, G., & Dong, G. (2026). Nonlinear and spatially heterogenous effects of air pollution and greenspace exposure on chronic respiratory disease mortality in Guangzhou, China: A Geographical-XGBoost approach. *Health & Place*, 100, 103689. <https://doi.org/https://doi.org/10.1016/j.healthplace.2026.103689>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Aug, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, 1–24. <https://doi.org/10.7717/PEERJ-CS.623>
- Comber, A., Brunsdon, C., Charlton, M., Dong, G., Harris, R., Lu, B., Lü, Y., Murakami, D., Nakaya, T., Wang, Y., & Harris, P. (2020). *The GWR route map: a guide to the informed application of Geographically Weighted Regression*. 1–34. <https://doi.org/10.48550/arXiv.2004.06070>

- Comber, A., Harris, P., & Brunson, C. (2024). Multiscale spatially varying coefficient modelling using a Geographical Gaussian Process GAM. *International Journal of Geographical Information Science*, 38(1), 27–47. <https://doi.org/10.1080/13658816.2023.2270285>
- Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, Number 2). MIT press Cambridge.
- Grekousis, G. (2025). Geographical-XGBoost: a new ensemble model for spatially local regression based on gradient-boosted trees. *Journal of Geographical Systems*, 27(2), 169–195. <https://doi.org/10.1007/s10109-025-00465-4>
- Guan, X., & Burton, H. (2022). Bias-variance tradeoff in machine learning: Theoretical formulation and implications to structural engineering applications. *Structures*, 46, 17–30. <https://doi.org/https://doi.org/10.1016/j.istruc.2022.10.004>
- Hagenauer, J., & Helbich, M. (2022). A geographically weighted artificial neural network. *International Journal of Geographical Information Science*, 36(2), 215–235. <https://doi.org/10.1080/13658816.2021.1871618>
- Han, Y., Yin, J., Zhu, D., Tan, J., Zhao, Y., Liu, W., & Mu, Y. (2026). Geographical-XGBoost and SHAP reveal ecosystem service supply-demand responses to landscape patterns in karst regions. *Journal of Environmental Management*, 404, 129550. <https://doi.org/https://doi.org/10.1016/j.jenvman.2026.129550>
- Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., & Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5(June), 100018. <https://doi.org/10.1016/j.ophoto.2022.100018>
- Khosravi, Y., & Ouara, T. B. M. J. (2026). A geographically weighted XGBoost framework for Pixel-Level modeling of vegetation responses using Multi-Source Earth Observation data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 235(March), 105–132. <https://doi.org/10.1016/j.isprsjprs.2026.03.006>
- Lessani, M. N., & Li, Z. (2024). SGWR: similarity and geographically weighted regression. *International Journal of Geographical Information Science*, 38(7), 1232–1255. <https://doi.org/10.1080/13658816.2024.2342319>
- Rimal, Y., Sharma, N., & Alsadoon, A. (2024). The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. *Multimedia Tools and Applications*, 83(30), 74349–74364. <https://doi.org/10.1007/s11042-024-18426-2>
- Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 2024, 69–79. <https://doi.org/10.58496/BJML/2024/007>
- Schimohr, K., Doeblner, P., & Scheiner, J. (2023). Prediction of Bike-sharing Trip Counts: Comparing Parametric Spatial Regression Models to a Geographically Weighted XGBoost Algorithm. *Geographical Analysis*, 55(4), 651–684. <https://doi.org/10.1111/gean.12354>
- Suprayogi, M. A., Sartono, B., & Notodiputro, K. A. (2024). Geographically Weighted Machine Learning Model for Addressing Spatial Heterogeneity of Public Health Development Index in Java Island. *Barekeng*, 18(4), 2577–2588. <https://doi.org/10.30598/barekengvol18iss4pp2577-2588>
- Wang, H., Huang, Z., Yin, G., Bao, Y., Zhou, X., & Gao, Y. (2022). Gwrboost: a geographically weighted gradient boosting method for explainable quantification of spatially-varying relationships. <https://doi.org/10.48550/arXiv.2212.05814>
- Wang, J., Du, H., Li, X., Mao, F., Zhang, M., Liu, E., Ji, J., & Kang, F. (2021). Remote sensing estimation of bamboo forest aboveground biomass based on geographically weighted regression. *Remote Sensing*, 13(15), 1–18. <https://doi.org/10.3390/rs13152962>
- Wang, Z., Zhao, Y., & Zhang, F. (2022). Simulating the Spatial Heterogeneity of Housing Prices in Wuhan, China, by Regionally Geographically Weighted Regression. *ISPRS International Journal of Geo-Information*, 11(2). <https://doi.org/10.3390/ijgi11020129>
- Wiedemann, N., Martin, H., & Westerholt, R. (2023). Benchmarking Regression Models Under Spatial Heterogeneity. *Leibniz International Proceedings in Informatics, LIPIcs*, 277, 0–15. <https://doi.org/10.4230/LIPIcs.GIScience.2023.11>

- Yang, R., Shen, T., & Yin, W. (2026). *Geographically Weighted Regression: A Scoping Review of Methods, Development, and Applications*. Land, 15(6), 1–21. <https://doi.org/10.3390/land15060915>
- 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*, 858(April 2022), 159798. <https://doi.org/10.1016/j.scitotenv.2022.159798>
- Zhou, S., Ji, Q., Zhang, L., Wu, J., Li, P., & Zhang, Y. (2025). Exploration of Differences in Housing Price Determinants Based on Street View Imagery and the Geographical-XGBoost Model: Improving Quality of Life for Residents and Through-Travelers. *ISPRS International Journal of Geo-Information*, 14(10). <https://doi.org/10.3390/ijgi14100391>