IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 13, No. 4, April, 2026, pp. 713-728
Journal homepage: http://iieta.org/journals/mmep

Hybrid Geographically Weighted Regression and Extreme Gradient Boosting for Spatial Prediction of Soil Texture Fractions

Henny Pramoedyo[1], Wigbertus Ngabu[2*], Atiek Iriany[1], Sativandi Riza[3]

[1] Department of Statistic, Faculty of Natural Sciences, Brawijaya University, Malang 65141, Indonesia
[2] Mathematics Education Study Program, Faculty of Teacher Training and Education, Universitas Riau, Pekanbaru 28293, Indonesia
[3] Soil Science Department, Faculty of Agriculture, Brawijaya University, Malang 65141, Indonesia

Corresponding Author Email: wigbertus.ngabu@lecturer.unri.ac.id

https://doi.org/10.18280/mmep.130409

Received: 28 January 2026
Revised: 18 April 2026
Accepted: 27 April 2026
Available online: 15 May 2026

Keywords:
soil texture fractions, spatial prediction, Geographically Weighted Regression, Extreme Gradient Boosting, hybrid modeling, watershed-scale mapping

ABSTRACT

Accurate prediction of soil texture fractions is essential for watershed-scale land management and precision soil mapping. Traditional global regression approaches often fail to account for spatial nonstationarity and nonlinear relationships between terrain attributes and soil texture. This study develops a hybrid framework that combines Geographically Weighted Regression (GWR) and Extreme Gradient Boosting (XGBoost) to predict the sand, silt, and clay fractions in the Kalikonto watershed. Terrain predictors derived from a hydrologically conditioned Digital Elevation Model (DEM)—including elevation, slope, horizontal curvature, vertical curvature, and ring curvature—were used alongside laboratory measurements from 150 georeferenced sampling locations. GWR captures spatially varying relationships, while XGBoost models nonlinear effects and interactions. Model performance was compared with global linear regression and standalone GWR. The hybrid model achieved superior predictive accuracy (R^2 = 0.783 for sand, 0.736 for silt, and 0.729 for clay) compared to global regression (R^2 = 0.409 – 0.482) and GWR alone (R^2 = 0.568 – 0.671). These results demonstrate that integrating spatially varying information with nonlinear modeling substantially improves soil texture prediction within a single watershed, although broader generalization requires validation across additional watersheds.

1. INTRODUCTION

Accurate soil texture information is essential for watershed-scale land management because it influences infiltration, water availability, erosion susceptibility, and crop productivity. Soil texture is commonly represented by the fractions of sand, silt, and clay, and its spatial variability is often shaped by interacting geomorphic and hydrologic processes. Consequently, predictive models that assume spatially constant relationships may produce biased or overly smoothed maps, particularly in heterogeneous terrain.

The availability of accurate soil data has become a global issue receiving widespread attention, particularly in the context of food security, climate change mitigation, and sustainable natural resource management [1]. Soil texture, defined by the distribution of sand, silt, and clay particle sizes, plays a crucial role in regulating infiltration capacity, water availability, and crop productivity [2]. Precise soil texture information is needed not only at the local scale but also forms the foundation for numerous environmental models and land use policy frameworks at regional to global levels.

Such terrain-driven variability often leads to spatially nonstationary relationships between predictors and soil texture, which challenges global models that assume constant effects across space [3]. This motivates methods that can accommodate spatially varying relationships while retaining strong predictive capability.

Advances in remote sensing technologies, soil sensors, and global spatial databases have produced an enormous volume of diverse environmental data [4]. This era of big data creates new opportunities for soil modeling but simultaneously presents significant challenges due to the high-dimensional, heterogeneous, and spatially dependent nature of the data [5]. In this context, conventional statistical techniques often lack the flexibility required, making the integration of spatial methods with modern machine-learning algorithms increasingly relevant [6].

Traditional approaches, such as global linear regression or spatial interpolation methods (e.g., kriging), have long been applied to soil characterization [7]. However, these methods assume that the relationships between predictor variables and the response variable are homogeneous across space. Such an assumption is difficult to justify given the inherently heterogeneous nature of soil data, leading to oversimplified models that fail to capture local variability.

To address these limitations, Geographically Weighted Regression (GWR) was introduced as a spatial statistical approach that adapts to geographic location. GWR allows

regression coefficients to vary spatially, thereby capturing differences in relationships among variables at each location. Nevertheless, GWR still faces constraints in handling nonlinear relationships, multicollinearity, and high-dimensional datasets, conditions increasingly common in modern soil research [8].

On the other hand, machine learning algorithms such as Extreme Gradient Boosting (XGBoost) offer strong capabilities for handling nonlinear, complex, and large-scale datasets [9]. XGBoost has demonstrated superior performance in various predictive applications, including environmental and agricultural data. However, purely machine-learning-based approaches tend to function as a "black box" and do not explicitly account for the spatial structure of the data [10]. As a result, the predictive outputs are often difficult to interpret within a geographical context and may overlook spatial heterogeneity.

Although both approaches are widely used, each has clear limitations when applied in isolation. GWR provides interpretable local coefficients and can capture spatial nonstationarity, yet it remains locally linear and can be sensitive to multicollinearity among terrain derivatives, which may reduce coefficient stability and predictive performance when relationships are strongly nonlinear. Conversely, XGBoost can model nonlinear responses and high-order interactions, but it does not explicitly represent spatial dependence; as a result, performance can be inflated under non-spatial validation and may become unstable when training data are limited or unevenly distributed [11]. These complementary limitations motivate a hybrid strategy that uses GWR to summarize spatially varying information and XGBoost to learn a nonlinear predictive structure.

Based on this discussion, a clear research gap emerges: spatial methods such as GWR are able to capture local heterogeneity but are limited in modeling complex nonlinear relationships, whereas machine learning techniques such as XGBoost excel in prediction yet ignore spatial context [12]. Combining these two approaches has the potential to produce a more comprehensive, accurate, and interpretable model. Nevertheless, the integration of GWR and XGBoost in the context of soil texture modeling remains underexplored, thus presenting an opportunity for significant scientific contribution [13].

Recent digital soil mapping studies commonly report that global linear models may underperform in complex terrain because they cannot represent spatially varying effects, whereas nonlinear machine-learning models often improve accuracy but may be sensitive to sampling density, spatial autocorrelation, and validation design [14]. Hybrid spatial-ML approaches have therefore been explored to balance interpretability and predictive power, although reported gains depend on landscape setting and evaluation protocol [15].

This study is designed as a methodological case study using the Kalikonto watershed as a compact test bed with pronounced terrain gradients. The watershed was selected because (i) it contains distinct landform units within a limited spatial extent, which is suitable for evaluating spatial nonstationarity, and (ii) it allows practical field sampling and DEM-based terrain derivation under consistent physiographic conditions. The dataset comprises 150 sampling locations, selected to cover the dominant terrain gradients (e.g., elevation and slope ranges) within the watershed, enabling a within-area comparison of modeling approaches. The goal is to test whether GWR-derived spatial information can enhance an XGBoost learner for local-scale prediction; therefore, conclusions are limited to this watershed context and do not imply regional or global representativeness without multi-site validation.

This study proposes a hybrid approach by integrating GWR and XGBoost for modeling soil particle-size distribution. Within this framework, GWR is employed to capture the spatial variation of regression coefficients, while XGBoost is utilized to model nonlinear relationships among variables. This combined approach is expected not only to improve predictive accuracy but also to provide methodological insights into the interplay between spatial processes and nonlinear behavior in soil data.

Methodologically, this research expands the scope of spatial statistics by embedding it within modern machine learning algorithms, thereby delivering a model that is robust, interpretable, and capable of high predictive performance. From an applied perspective, the findings contribute to precision soil-texture mapping, which is highly relevant for land-use planning, environmental conservation, and climate-change adaptation strategies. Thus, this study offers methodological advancements in statistical modeling while simultaneously addressing global challenges in soil resource management.

2. METHODS

2.1 Data

This study employs two primary data types: primary data and secondary data. The primary data were obtained through direct field measurements of soil texture, whereas the secondary data were derived from digital terrain modeling analyses [16]. The combination of these datasets enables the examination of relationships between local morphological parameters and soil texture characteristics.

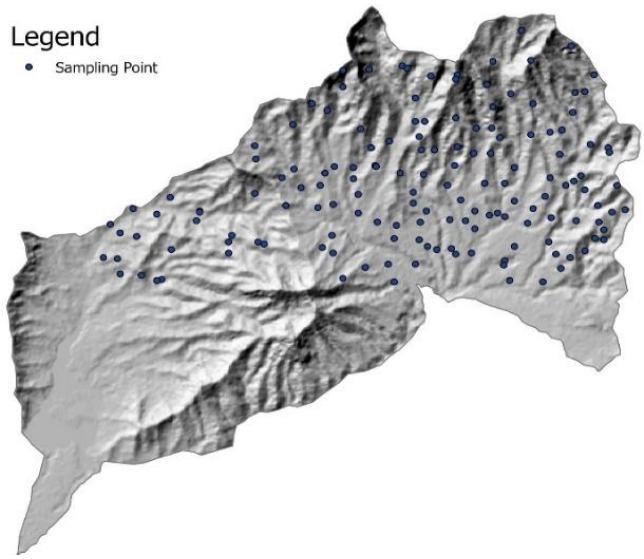


Figure 1. Research location map

In this study, the field data were used both for model construction (training data) and for validating the resulting model (testing data). The dataset consisted of 150 sampling locations. Model evaluation used k-fold cross-validation; therefore, the full dataset contributed to both training and

testing across folds. Sampling locations were selected to cover the dominant terrain gradients within the watershed (elevation and slope ranges), improving representativeness for within-watershed modeling. Cross-validation was implemented using stratified k-fold splitting based on terrain strata (e.g., elevation quantiles) to ensure each fold contained comparable terrain representation. These data represent the results of soil texture analysis conducted within the Kalikonto watershed. The study area map is presented in Figure 1.

The predictors used in this study consist of five Digital Elevation Model (DEM)-derived local morphometric variables (LMVs), while the response variables are the three soil texture fractions (sand, silt, and clay) [17]. The LMV concept refers to previous research demonstrating that specific morphological characteristics can influence the distribution and physical properties of soils [18].

In this study, the variables are categorized into response variables and predictor variables. The response variables consist of three soil texture components, namely silt, sand, and clay, each expressed as a proportion or percentage. All three are measured on a ratio scale, allowing them to be analyzed as continuous numerical variables. The remaining five LMVs serve as predictor variables, and all of them are also measured using a ratio scale.

The five LMVs were selected because they represent complementary terrain controls that are commonly linked to soil redistribution processes. Elevation (Elev) captures broad geomorphic and climatic gradients within the watershed, while slope (S) represents potential runoff energy and erosion capacity. Curvature measures were included to describe local landform shape and flow behavior: horizontal curvature (Kh) characterizes planform convergence/divergence that influences lateral flow concentration; vertical curvature (Kv) describes profile convexity/concavity associated with acceleration or deceleration of overland flow and sediment transport; and ring curvature (Kr) summarizes overall curvature intensity as an integrated descriptor of local surface form. Together, the terrain variables listed in Table 1 provide a compact set of physically interpretable predictors for explaining terrain-driven spatial variability in soil texture fractions, with sand, silt, and clay serving as the response variables.

Table 1. Research variables

Variables		Measurement Scale	Information
Soil particle size	Sand, silt, clay	ratio	Response
Horizontal curvature	Kh	ratio	Predictor
Vertical curvature	Kv	ratio	Predictor
Slope	S	ratio	Predictor
Ring curvature	Kr	ratio	Predictor
Elevation	Elev	ratio	Predictor

All terrain variables were derived from a DEM obtained from DEMNAS (Digital Elevation Model Nasional) with a spatial resolution of 30 m [19]. Before terrain variable extraction, the DEM was projected to the WGS 1984 UTM Zone 49S coordinate system, and hydrological conditioning was applied to reduce topographic artifacts, including sink filling and the removal of spurious depressions. Slope and curvature surfaces, including Kh, Kv, and Kr, were computed using SAGA GIS based on standard geomorphometric algorithms. The resulting raster layers were aligned to a common grid, and the values of each terrain variable were extracted at the 150 soil sampling locations using point sampling to construct the modeling dataset. The complete list of variables used in this study is presented in Table 1 [20].

2.2 Geographically Weighted Regression

The GWR model is an extension of the global regression framework that incorporates geographical factors into the analysis. GWR is a spatial analytical method based on point data and represents a development of linear regression analysis by explicitly accounting for spatial location. The general formulation of the GWR model is expressed as follows [21]:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + e_i \quad i = 1, 2, 3, \dots, n \tag{1}$$

a. Estimating GWR model parameters

Parameter estimation in the GWR model is carried out using the Weighted Least Squares (WLS) method, which assigns different weights to each geographical location. The estimation procedure incorporates a weighting function $w_j(u_i, v_i)$ for each observation point, and is formulated as follows [22, 23]:

$$y_i w_j(u_i, v_i) = w_j(u_i, v_i) \left(\beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + e_i \right) \tag{2}$$

By assigning the weight $w_j(u_i, v_i)$ to each observation in the GWR model, the objective is to minimize the sum of squared errors, which is expressed in the following Eq. (3):

$$\sum_{J=1}^{n} w_j(u_i, v_i) e_i^2 = \sum_{j=1}^{n} w_j(u_i, v_i) \left[y_i - \beta_0(u_i, v_i) - \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} \right]^2 \tag{3}$$

where,

$$\boldsymbol{\beta}(u_i, v_i) = \begin{pmatrix} \beta_0(u_i, v_i) \\ \beta_1(u_i, v_i) \\ \vdots \\ \beta_n(u_i, v_i) \end{pmatrix}$$

$$\boldsymbol{W}(u_i, v_i) = diag[w_1(u_i, v_i), w_2(u_i, v_i), \dots, w_n(u_i, v_i)] \tag{4}$$

$$\boldsymbol{W}(u_i, v_i) = \begin{pmatrix} w_1(u_i, v_i) & 0 & 0 & 0 \\ 0 & w_2(u_i, v_i) & 0 & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & w_n(u_i, v_i) \end{pmatrix}$$

$\boldsymbol{W}(u_i, v_i)$ is the spatial weighting matrix for the GWR model, with dimensions $n \times n$. The parameter estimation for the GWR model is obtained by taking the derivative with respect to $\boldsymbol{\beta}'(u_i, v_i)$, as expressed in Eq. (5) [24]:

$$\frac{\partial \boldsymbol{e}' \boldsymbol{W}(u_i, v_i) \boldsymbol{e}}{\partial \boldsymbol{\beta}'(u_i, v_i)} = 0 \tag{5}$$

$$\frac{\partial(\boldsymbol{Y}'\boldsymbol{W}(u_i,v_i)\boldsymbol{Y}-2\boldsymbol{\beta}'(u_i,v_i)\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{Y}+\boldsymbol{\beta}'(u_i,v_i)\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{X}\boldsymbol{\beta}(u_i,v_i))}{\partial(\boldsymbol{\beta}'(u_i,v_i))}=0 \quad (6)$$

$$0-2\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{Y}+\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{X}\boldsymbol{\beta}(u_i,v_i)+(\boldsymbol{\beta}'(u_i,v_i)\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{X})'=0 \quad (7)$$

$$-2\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{Y}+\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{X}\boldsymbol{\beta}(u_i,v_i)+\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{X}\boldsymbol{\beta}(u_i,v_i)=0 \quad (8)$$

$$2\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{X}\boldsymbol{\beta}(u_i,v_i)=2\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{Y} \quad (9)$$

$$\boldsymbol{\beta}(u_i,v_i)=\frac{2\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{Y}}{2\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{X}} \quad (10)$$

$$\widehat{\boldsymbol{\beta}}(u_i,v_i)=[\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{W}(u_i,v_i)\boldsymbol{Y} \quad (11)$$

b. Weighting matrix with fixed Gaussian kernel function

The weighting matrix used in this study is the fixed kernel, which applies a constant bandwidth value to all spatial locations [25]. With a global bandwidth, each observation receives a uniform scaling factor, allowing spatial relationships between locations to be assessed consistently [26]. The fixed kernel approach is appropriate when the spatial distribution of data is relatively uniform, enabling a single bandwidth value to effectively represent the underlying spatial structure.

In this study, a fixed Gaussian kernel was adopted to maintain a consistent spatial influence distance across the watershed and to facilitate comparable local estimation among locations. This choice is appropriate when the sampling density is reasonably uniform at the watershed scale, such that a single distance bandwidth can represent the spatial interaction range. Nevertheless, we acknowledge that when sampling is strongly uneven, an adaptive kernel (nearest-neighbor bandwidth) may be preferable because it adjusts the effective distance to local data density. To ensure that the fixed-kernel assumption does not dominate the results, bandwidth selection was performed objectively (see below), and kernel sensitivity can be assessed by comparing fixed and adaptive kernels as an additional robustness check. In this study, the weights are calculated using the fixed Gaussian kernel, expressed as [27]:

$$\boldsymbol{w}_{ij}=\exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right] \quad (12)$$

where, d_{ij} denotes the distance between locations and b represents the fixed bandwidth. With this approach, closer locations receive higher weights, while those farther apart receive lower weights in an exponentially decreasing manner.

The fixed bandwidth b was not assigned arbitrarily. Instead, b was selected using an automatic bandwidth optimization procedure that minimizes the corrected Akaike Information Criterion (AICc) of the GWR calibration [28]. Candidate bandwidth values were searched over a predefined range of distances, and the bandwidth yielding the lowest AICc was chosen as the optimal fixed bandwidth for the Gaussian kernel. This AICc-based selection balances model fit and effective model complexity and is commonly used in GWR to avoid overfitting or under-smoothing.

2.3 Extreme Gradient Boosting modeling

XGBoost is an ensemble algorithm based on gradient boosting decision trees, designed to improve prediction accuracy through an additive learning process [29]. At each iteration, a new model is constructed to correct the prediction errors made by the previous model [30]. The strength of XGBoost lies in its ability to handle nonlinear relationships, its regularization mechanisms to prevent overfitting, and its high computational efficiency [9]. Consider a training dataset $D=\{(x_i,y_i)\}_{i=1}^{n}$ where $x_i\in\mathbb{R}^p$ represents the predictor variables (for example, rainfall, topography, and land use), and $y_i\in\mathbb{R}$ represents the response variable (soil particle size fraction). The prediction at the t-th iteration is formulated as follows:

$$\hat{y}_i^{(t)}=\hat{y}_i^{(t-1)}+f_t(\mathbf{x}_i),\qquad f_t\in\mathcal{F} \quad (13)$$

where, $\mathcal{F}$ denotes the set of decision tree functions. The objective function of XGBoost consists of two components, namely the loss function and the complexity regularization term [10]:

$$\mathcal{L}^{(t)}=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{(t)}\right)+\sum_{k}\Omega(f_k) \quad (14)$$

where, l is the loss function (for example, squared error), and $\Omega(f_k)=\gamma^T+\frac{\lambda}{2}\sum_{j=1}^{T}w_j^2$, with T representing the number of leaves in the tree and w_j the prediction weight assigned to the j-th leaf. The regularization parameters λ and γ control model complexity.

As a critical benchmark, this study adds a standalone XGBoost model trained solely on the original environmental/topographic covariates without incorporating any GWR-derived features; this model is hereafter referred to as XGBoost-Orig. To ensure a fair comparison, XGBoost-Orig is developed using an identical pipeline to the hybrid model, including data preprocessing steps, the data-splitting scheme, and the same k-fold cross-validation protocol [31]. Hyperparameter tuning is also conducted using a consistent procedure and a comparable search space, so that performance differences are not driven by unequal modeling treatments [32]. With the XGBoost-Orig baseline tightly controlled through the same pipeline, hyperparameter tuning, and k-fold validation, the performance gains observed in the GWR-XGBoost model can be interpreted more convincingly as the contribution of integrating GWR-derived spatial features, rather than merely the consequence of using XGBoost as a flexible algorithm for capturing nonlinear relationships.

2.4 Geographically Weighted Regression Extreme Gradient Boosting integration (hybrid model)

The hybrid GWR XGBoost approach was developed to address two major limitations in spatial data modeling. First, GWR is capable of capturing spatial heterogeneity through locally varying regression coefficients, but it is less effective when dealing with complex nonlinear relationships [33]. Second, XGBoost excels at modeling nonlinear interactions among variables, yet it tends to overlook the spatial structure of the data [34]. By integrating these two methods, the hybrid model is expected to produce more precise predictions while still providing strong spatial interpretability.

The first stage of the integration process is the GWR modeling step. Mathematically, the GWR model is expressed as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + e_i \quad (15)$$

where, y_i represents the response variable, (u_i, v_i) denotes the geographic coordinates of the i-th location, x_{ik} is the covariate or predictor variable, $\beta_k(u_i, v_i)$ is the locally varying coefficient at that location, and e_i is the spatial residual. This model allows each location to have its own regression parameters according to its environmental characteristics. The results of the GWR modeling are then extracted as additional features for the next stage. The spatial features obtained include the local coefficients $\beta_k(u_i, v_i)$, the local intercept $\beta_0(u_i, v_i)$, and the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i^{GWR}$. This information provides a more detailed representation of spatial variation and adds geographic context to the dataset.

Next, a hybrid dataset is constructed by combining the original environmental covariates with the spatial features generated from the GWR stage. This hybrid dataset can be represented as follows:

$$\boldsymbol{Z}_i = \begin{pmatrix} x_{i1}, x_{i2}, \dots, x_{ip}, \hat{\beta}_0(u_i, v_i), \hat{\beta}_1(u_i, v_i), \dots \\ , \hat{\beta}_p(u_i, v_i), \hat{\varepsilon}_i \end{pmatrix} \quad (16)$$

With this construction, each data unit contains not only global predictors but also relevant local spatial information derived from the GWR stage. The final stage is the XGBoost modeling step using the hybrid dataset. XGBoost is trained with $\boldsymbol{Z}_i$ as input, enabling the final model to capture local heterogeneity through the GWR features while simultaneously modeling complex nonlinear patterns through the boosting mechanism. This integration is expected to produce a model that not only achieves high predictive accuracy but also maintains clear spatial interpretability.

3. RESULTS AND DISCUSSION

The initial stage of this study focused on analyzing the relationships between the independent variables (X) and the dependent variables (Y) before conducting spatial modeling. The study employed three dependent variables representing soil texture fractions, namely sand, silt, and clay. Meanwhile, the independent variables consisted of topographic factors, specifically Kh, Kv, slope, Kr, and Elev.

To identify the preliminary relationships among the variables, scatterplots were generated between the soil texture components (sand, silt, clay) and each topographic variable. This visual analysis aimed to provide an initial overview of the relationship patterns, whether linear or nonlinear, which served as the foundation for subsequent modeling using the GWR and XGBoost approaches.

Based on the scatterplot results (Figure 2), the relationships between soil texture fractions and topographic factors exhibit varying tendencies. The sand fraction shows a generally positive relationship with most topographic variables, particularly Kv, slope, and Kr, as indicated by the upward trend of the simple regression lines. This finding suggests that increases in slope gradient and land curvature tend to correlate with higher sand content in the soil.

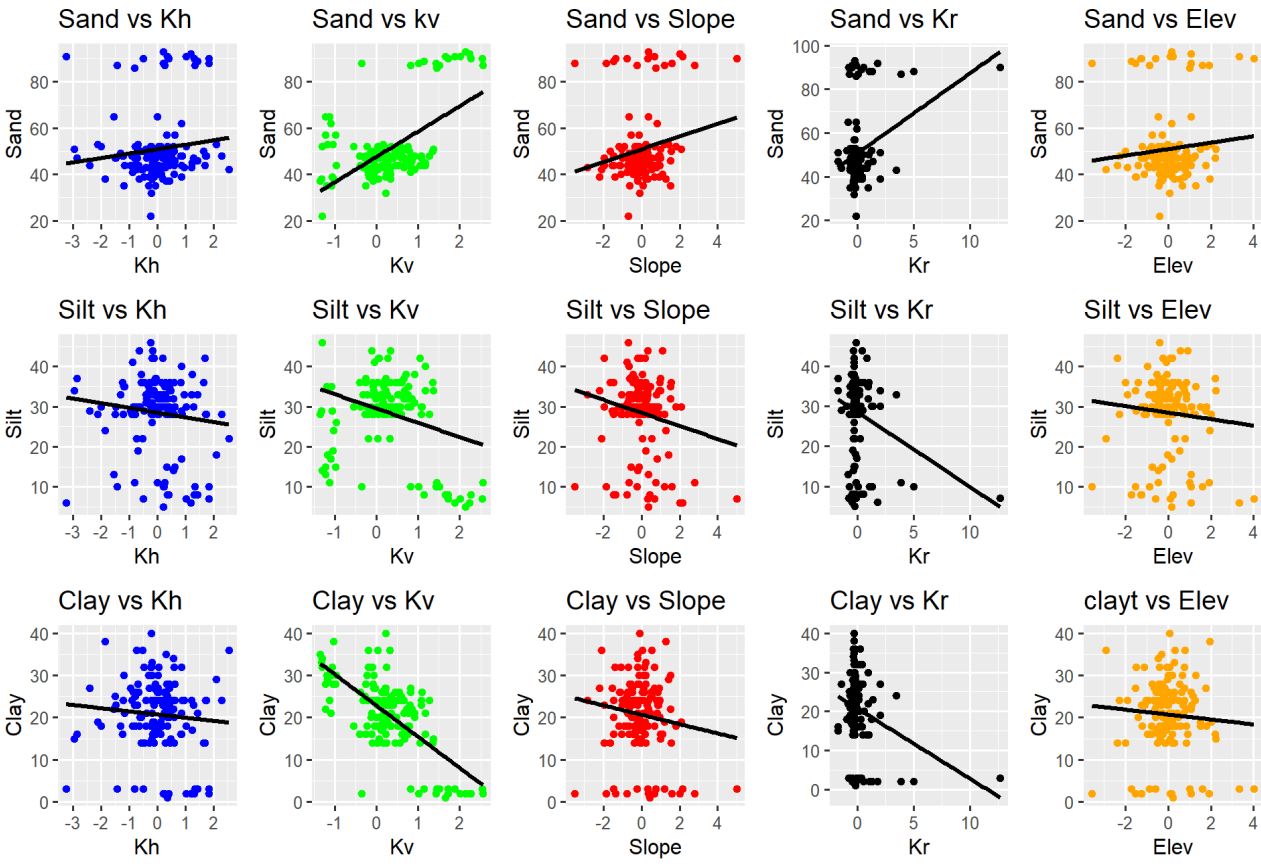


Figure 2. Scatterplot of the relationship between soil texture fractions (sand, silt, clay)

Table 2. Results of multiple linear regression analysis for the relationship between soil texture fractions

Parameter	Sand		Silt		Clay	
	Estimate	P-Value	Estimate	P-Value	Estimate	P-Value
Intercept	47.8407	2e-16	29.3291	2e-16	22.8302	2e-16
Horizontal curvature (Kh)	53.1122	0.00359	− 36.3935	0.00538	− 16.7188	0.0855
Vertical curvature (Kv)	8.9628	2.4e-10	− 2.0731	0.02970	− 6.8896	2e-16
Slope (S)	− 75.3261	0.00538	51.0277	0.00852	24.2984	0.0924
Ring curvature (Kr)	1.6417	0.03676	− 1.1423	0.04284	− 0.4994	0.2354
Elevation (Elev)	101.5820	0.00462	− 69.3182	0.00702	− 32.2638	0.0917
R-square	**0.4087**		**0.1950**		**0.4823**	
RMSE	**11.372**		**8.163**		**6.121**	
MAE	**8.262**		**5.985**		**4.792**	

Note: RMSE: root mean square error; MAE: mean absolute error; Kh: horizontal curvature; Kv: vertical curvature; Kr: ring curvature.

In contrast, the silt fraction tends to exhibit negative relationships with most of the predictor variables. The trend lines in the scatterplots indicate a decline in silt content as the values of Kv, slope, Kr, and Elevation increase. This pattern reflects that silt is more dominant in areas with relatively gentle terrain or lower elevation.

Unlike sand, the clay fraction also demonstrates negative tendencies with respect to the topographic variables, particularly Kv, Kr, and slope. This pattern indicates that clay content is generally higher in locations with low curvature and gentle slopes, whereas areas with sharper land curvature tend to exhibit reduced clay content.

Overall, these preliminary findings reveal clear spatial heterogeneity between topographic variables and soil texture fractions. Therefore, a spatial modeling approach such as GWR is needed to capture local variations. Furthermore, integrating this approach with a machine-learning method such as XGBoost is expected to enhance the precision and accuracy of soil texture mapping.

3.1 Regression analysis

Following the exploratory analysis of variable relationships through scatterplots, the next stage involved examining the influence of topographic variables on soil texture fractions using multiple linear regression. This analysis aims to provide an overall understanding of the contributions of each topographic parameter (Kh, Kv, slope, Kr, and Elev) in explaining the variation in sand, silt, and clay contents. The estimated model parameters, along with the coefficients of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), are presented in Table 2. While R^2 indicates the proportion of variance explained by the model, RMSE and MAE provide complementary measures of prediction error. Specifically, RMSE is more sensitive to large deviations between observed and predicted values, whereas MAE reflects the average magnitude of absolute prediction errors and is generally easier to interpret directly.

Based on the results of the multiple linear regression analysis presented in Table 2, the performance of the global model can be evaluated using three main metrics, namely the R^2, RMSE, and MAE. The R^2 value indicates the proportion of variance in soil texture fractions explained by the topographic variables in the model, while RMSE and MAE quantify the magnitude of prediction errors. In this context, RMSE is more sensitive to large errors because it squares the residuals, whereas MAE represents the average absolute deviation and is more directly interpretable.

Based on the results of the multiple linear regression presented in Table 2, the R^2 for each model ranges from 0.1950 to 0.4823. This indicates that the topographic variables (Kh, Kv, slope, Kr, and Elev) collectively explain only part of the variation in soil texture fractions, while the remaining variation is influenced by factors not included in the model. The highest R^2 value is observed for the clay fraction (0.4823), suggesting that topographic factors provide a stronger explanation for the distribution of clay compared to sand or silt.

The estimated multiple linear regression equations for the sand, silt, and clay fractions are presented as follows:

$$\hat{y}_{sand} = 47.8407 + 53.1122\,Kh + 8.9628\,Kv$$
$$-75.3261\,Slope + 1.6217\,Kr + 101.5820\,Elev$$
$$\hat{y}_{silt} = 29.3291 - 36.3935\,Kh - 2.0731\,Kv$$
$$+51.0277\,Slope - 1.1423\,Kr - 69.3128\,Elev$$
$$\hat{y}_{sand} = 22.8302 - 16.7188\,Kh - 6.8896\,Kv -$$
$$24.2984\,Slope - 0.4994\,Kr - 32.2638\,Elev$$

For the sand fraction, the parameter estimates indicate that sand content tends to increase with higher values of Kh, Kv, Kr, and elevation, whereas slope exhibits a negative effect, implying a decrease in sand content in areas with steep terrain. With an R^2 value of 0.4087, the model explains nearly 41 percent of the variability in sand content, although the remaining variation is influenced by factors beyond topographic variables. In terms of prediction error, the RMSE value of 11.372 and the MAE value of 8.262 indicate that the model still produces relatively large deviations between predicted and observed sand content. These results suggest that, although the model captures a moderate proportion of variation in sand, its predictive precision remains limited.

For the silt fraction, an R^2 value of 0.1950 suggests that only about 19.50 percent of the variation is explained by topographic factors. Silt content increases in areas with greater slope but decreases at locations with higher elevation and greater horizontal curvature, vertical curvature, and ring curvature. This pattern indicates that silt is more dominant in moderately sloped areas but less prevalent in regions with complex and elevated terrain. The RMSE value of 8.163 and the MAE value of 5.985 show that the average prediction errors for silt are lower than those for sand. However, the relatively low R^2 value indicates that the model still does not adequately capture the underlying variation in silt content. This suggests that the distribution of silt may be influenced by additional factors or more complex relationships not represented in the global linear model.

Meanwhile, the clay model shows that most topographic variables exert negative effects, indicating that increases in curvature or elevation correspond to reductions in clay content, although slope remains positive in the estimated

equation. With an R^2 value of 0.4823, the clay model performs the best among the three in explaining variation in soil texture fractions. In addition, the RMSE value of 6.121 and the MAE value of 4.792 are the lowest among the three fractions, indicating that the model is relatively more accurate in predicting clay content. These findings reinforce that clay is more prevalent in landscapes with gentle topography, lower elevation, and minimal curvature, and that its distribution is more consistently related to the selected topographic variables.

Overall, the results of the multiple linear regression analysis show that topographic variables influence soil texture fractions, with R^2 values ranging from 0.1950 to 0.4823. At the same time, the RMSE values ranging from 6.121 to 11.372 and the MAE values ranging from 4.792 to 8.262 indicate that the global linear model still produces noticeable prediction errors. This suggests that the model captures only part of the variability in sand, silt, and clay content. Such limitations highlight the presence of additional influencing factors beyond the model and the existence of spatial heterogeneity that cannot be fully captured by a global linear approach. Therefore, more adaptive methods such as GWR are needed to account for local variations, along with the integration of machine learning techniques such as XGBoost to better accommodate nonlinear relationships and enhance predictive accuracy, thereby enabling more representative and precise soil texture mapping.

3.2 Spatial autocorrelation

In spatial analysis, the presence of spatial autocorrelation in explanatory variables is an important aspect that must be examined before further modeling. Spatial autocorrelation reflects the extent to which a variable is correlated with itself based on spatial proximity, thereby indicating whether its distribution across geographical space is random or exhibits a structured pattern. In this study, Moran's I test was employed to detect spatial autocorrelation, with significance determined using the p-value. Variables with p-values less than 0.05 were considered to exhibit significant spatial autocorrelation. The results of the spatial autocorrelation test for the topographic variables used in this study are presented in Table 3.

Table 3. Spatial autocorrelation test (Moran's I) on topographic variables

Variables	Moran's I Statistic	P-Value	Decision
Horizontal curvature (Kh)	− 0.00104	0.00541	Significant
Vertical curvature (Kv)	− 0.6801	2.2e-16	Significant
Slope (S)	0.0558	0.00728	Significant
Ring curvature (Kr)	0.00059	0.03541	Significant
Elevation (Elev)	0.0000011	0.00265	Significant

Based on the spatial autocorrelation test results presented in Table 3, all topographic variables exhibit p-values less than 0.05, indicating significant spatial autocorrelation. This condition suggests that the distribution of topographic values is not random but instead forms distinct spatial patterns within the study area. Consequently, the use of a global statistical model that ignores spatial dependence may lead to biased parameter estimates. Therefore, the application of GWR becomes particularly relevant for capturing local variation, while also providing a stronger foundation for integrating hybrid approaches such as GWR-XGBoost in precision soil texture mapping.

3.3 Spatial weighting determination using the fixed Gaussian kernel

Before conducting spatial analysis using GWR, the spatial weighting scheme to be used in the local estimation process must first be determined. In this study, the weighting function was specified using the fixed Gaussian kernel, so that all observation locations were analyzed under a single constant bandwidth. Determining this bandwidth is a crucial step because it controls the spatial range of influence among locations and directly affects the stability as well as the sensitivity of the local parameter estimates. Therefore, the bandwidth was not selected subjectively, but was determined through a cross-validation procedure performed before the calibration of the GWR model.

Based on the cross-validation optimization results, the optimal bandwidth obtained was 4954.635. This value was selected because it produced the minimum cross-validation score among the candidate bandwidth values, and was therefore considered the most appropriate for representing the spatial structure of the data at the scale of the study area. Using this fixed bandwidth, the fixed Gaussian kernel weighting scheme was then applied in the GWR analysis to generate local parameter estimates at each observation point. This stage shows that the weighting process was not assigned arbitrarily, but rather through an objective optimization procedure, thereby providing a stronger basis for the subsequent analysis of spatial heterogeneity.

3.4 Geographically Weighted Regression prediction

In the GWR analysis, each observation location yields a predicted value that reflects the local response of soil texture fractions to the topographic variables. Accordingly, the GWR results are not treated as a single global equation but as a set of spatially varying local predictions. To present this variation in a concise and informative manner, the predicted sand, silt, and clay fractions are summarized using descriptive statistics, including the minimum, maximum, range, mean, standard deviation, and median. This summary provides a quantitative overview of the magnitude of variation in the GWR predictions across all observation locations and highlights the degree of spatial heterogeneity for each soil texture fraction.

Table 4. Summary of Geographically Weighted Regression (GWR) model prediction results

Variable	Minimum	Maximum	Range	Mean	Standard Deviation	Median
Sand	39.464	98.412	58.947	51.602	12.010	47.615
Silt	3.087	35.656	32.569	27.875	6.514	30.773
Clay	1.499	32.178	33.676	20.523	6.526	22.023

Based on Table 4, the GWR prediction results show clear variation across the three soil texture fractions. The sand fraction exhibits the widest range, from 39.464 to 98.412, with a mean of 51.602 and a standard deviation of 12.010. This indicates that sand predictions display the highest degree of spatial heterogeneity across the study area. The silt fraction ranges from 3.087 to 35.656, with a mean of 27.875 and a standard deviation of 6.514, suggesting a moderate level of spatial variation. Meanwhile, the clay fraction ranges from

1.499 to 32.178, with a mean of 20.523 and a standard deviation of 6.526, indicating a level of dispersion comparable to that of silt.

Overall, these descriptive statistics demonstrate that the GWR model is capable of capturing spatially varying predictions at each observation location, rather than producing a single uniform estimate as in global regression models. The larger range and standard deviation observed in the sand fraction suggest stronger local variability, whereas the silt and clay fractions exhibit more moderate variation. Therefore, presenting GWR results using summary statistics provides a clearer, more concise, and more interpretable representation compared to listing numerous location-specific regression equations.

3.4.1 Geographically Weighted Regression coefficient relationship plot

After obtaining the GWR models for each soil texture fraction (sand, silt, and clay), the next step was to evaluate the relationships between the intercept values and the local regression coefficients of each topographic variable. This analysis is essential for examining how local variations in the coefficients are influenced by the intercept values, as well as for identifying spatial patterns that may not be captured by the global model. Accordingly, plotting the intercepts against the GWR coefficients provides additional insights into spatial heterogeneity and the stability of explanatory variable contributions to soil texture fractions.

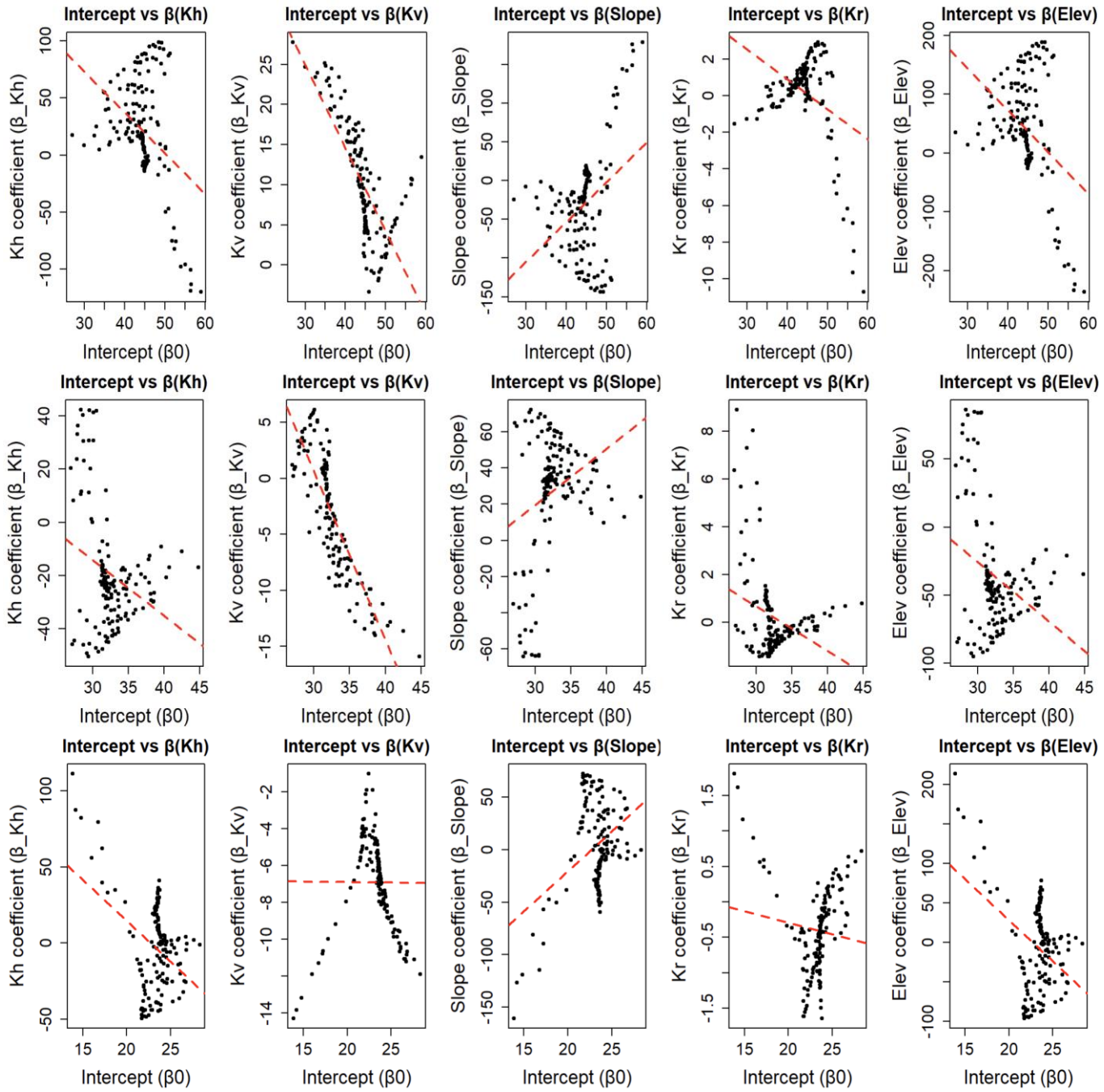


Figure 3. Relationship between intercept and local regression coefficient (β) in the Geographically Weighted Regression (GWR) model

Note: First row = Sand, second row = Silt, third row = Clay.

As shown in Figure 3, in the first row (sand), the results show that Kh, Kv, Kr, and Elev tend to have negative relationships with the intercept, whereas the slope exhibits a positive tendency. This indicates that increases in sand content are more strongly influenced by slope at locations with higher intercept values, while curvature and elevation tend to decline. In the second row (silt), the emerging pattern shows that Kh, Kv, and Elev generally display negative correlations with the intercept, while the slope shows a positive tendency. This suggests that silt content is higher in areas with greater slope, whereas it tends to decrease in locations with higher elevation and greater curvature. In the third row (clay), most topographic variables exhibit negative effects on the intercept, although slope and Kr show positive tendencies at several locations. This pattern reinforces that clay content is higher in areas with low slope, low elevation, and minimal curvature, whereas increases in slope or ring curvature at certain locations may contribute to higher clay content.

Overall, the variation in both the direction and magnitude of these relationships across the three soil texture fractions highlights the presence of complex local heterogeneity, thereby reinforcing the necessity of spatial approaches such as GWR for analyzing precision soil texture distribution.

3.4.2 Geographically Weighted Regression model fit

In spatial analysis, model evaluation for GWR is conducted to ensure that the model is capable of capturing local variation within the data. In this study, the dependent variables analyzed were the proportions of soil texture fractions, namely sand, silt, and clay. Model testing was performed using the F-statistic to assess the overall significance of the GWR model for each response variable.

Table 5. Results of the Geographically Weighted Regression (GWR) model fit test for sand, silt, and clay variables

Variables	F-Count	Df1	Df2	P-Value
Sand	2.0451	144	122.71	0.0000283
Silt	2.0422	144	122.71	0.0000293
Clay	1.268	144	122.71	0.008791

Based on the GWR model fit test as shown in Table 5, all three response variables—sand, silt, and clay—exhibit p-values smaller than the 0.05 significance level, indicating that the model is significant in explaining the spatial variation in soil texture fractions. The sand (p-value 0.0000283) and silt (p-value 0.0000293) variables show very strong levels of significance, confirming that their distributions are strongly influenced by local spatial factors. The clay variable (p-value 0.008791) is also significant, although with comparatively lower strength. These results indicate that the application of GWR is appropriate for soil texture mapping because it effectively captures local heterogeneity that cannot be represented by a global model.

Overall, the three response variables (sand, silt, and clay) demonstrate that the GWR model is significant and suitable for analyzing soil texture distribution. These findings reinforce the importance of spatially localized modeling approaches in understanding soil texture variability that cannot be accommodated by global regression techniques.

3.4.3 Coefficient of determination

The R^2 is used to assess the extent to which the GWR model can explain the variation in the dependent variables. A higher R^2 value indicates a stronger ability of the model to capture the spatial relationships between the independent and dependent variables. In this analysis, the dependent variables consist of the soil texture fractions sand, silt, and clay, so the R^2 values represent the proportion of variation in each fraction that can be explained by the GWR model. In addition to R^2, model performance is also evaluated using RMSE and MAE. While R^2 measures explanatory power, RMSE and MAE quantify prediction error, with RMSE being more sensitive to large deviations and MAE representing the average absolute difference between observed and predicted values.

Table 6. Coefficient of Determination (R^2) of the Geographically Weighted Regression (GWR) model for sand, silt, and clay variables

	Sand	Silt	Clay
R square	0.711	0.606	0.592
RMSE	7.952	5.712	5.435
MAE	5.810	4.516	4.144

Note: RMSE: root mean square error; MAE: mean absolute error.

The R^2 represents the proportion of variance in the dependent variable that can be explained by the model. Based on Table 6, the R^2 value for the sand variable is 0.711, indicating that approximately 71.1 percent of the variation in sand distribution can be explained by the GWR model, while the remaining variation is influenced by factors outside the model. The RMSE value for sand is 7.952, and the MAE value is 5.810, indicating that although the model explains a substantial proportion of the variation, prediction errors are still present at a moderate level.

For the silt variable, the R^2 value is 0.606, meaning that 60.6 percent of its variation is explained by the model. The RMSE value of 5.712 and the MAE value of 4.516 indicate that the prediction errors for silt are smaller than those for sand, suggesting a relatively better level of predictive accuracy in terms of absolute deviation, even though the explanatory power is lower than that of sand.

Meanwhile, the clay variable has an R^2 value of 0.592, indicating that 59.2 percent of its variation can be explained by the GWR model. The RMSE value of 5.435 and the MAE value of 4.144 are the lowest among the three variables, showing that the GWR model produces the smallest prediction errors for clay. This suggests that, although clay has a slightly lower R^2 value than silt, its predicted values are closer to the observed values on average.

In general, R^2 values above 0.5 indicate that the model exhibits a reasonably good fit in explaining the spatial variation of the three soil texture fractions. At the same time, the RMSE and MAE values confirm that the magnitude of prediction error is moderate and varies across variables. These findings indicate that GWR is capable of capturing local spatial heterogeneity in sand, silt, and clay, although the degree of model fit and prediction accuracy differs among the three fractions. Based on the R^2 values, sand is the most effectively explained variable, followed by silt and clay. However, based on RMSE and MAE, clay shows the smallest prediction error, followed by silt and sand.

Overall, based on the model fit and the evaluation metrics, the GWR model is shown to be useful in explaining the spatial variation of sand, silt, and clay, although some variation remains unaccounted for. To further improve predictive accuracy and capture more complex nonlinear patterns, the analysis will proceed with the hybrid GWR XGBoost method, which integrates the strengths of GWR in modeling spatial

heterogeneity and XGBoost in handling nonlinear relationships.

3.5 Extreme Gradient Boosting results

As a benchmark against the spatial regression approach represented by GWR, this study also evaluated an XGBoost Original model constructed solely from the original terrain covariates, namely Kh, Kv, slope, Kr, and Elev. This model serves as a non-spatial machine-learning baseline to assess the extent to which nonlinear learning alone can explain variation in soil texture fractions without incorporating locally derived spatial information. Model performance was assessed using five-fold cross-validation, with RMSE, R^2, and the optimal number of boosting iterations selected through early stopping. The validation results are presented in Table 7.

Table 7 indicates that the predictive performance of the XGBoost-Orig model differs across soil texture fractions and remains uneven across folds. For the sand fraction, the model achieved a mean RMSE of 10.2679, a mean MAE of 6.773, and a mean R^2 of 0.4852, indicating moderate predictive capability. Although the model is able to explain a reasonable portion of the variation in sand content, the relatively high RMSE and MAE values suggest that prediction errors are still substantial.

For the clay fraction, the model also showed moderate performance, with a mean RMSE of 6.4121, a mean MAE of 4.897, and a mean R^2 of 0.4050. Compared to sand, the lower RMSE and MAE values indicate that the model produces more accurate predictions for clay in terms of absolute error, although the explanatory power remains slightly lower.

In contrast, the model performed less effectively for the silt fraction, with a mean RMSE of 7.8083, a mean MAE of 5.408, and a mean R^2 of only 0.2073. This suggests that the original terrain covariates alone were insufficient to adequately explain variation in silt content. Despite having moderate error values, the low R^2 indicates that the model fails to capture the underlying variability of silt effectively.

The fold-specific results further show that model generalization was not consistently stable. This is most evident for silt, where one-fold yielded a negative R^2 value, indicating that the model performed worse than a naïve mean-based prediction for that test partition. From a methodological perspective, this pattern suggests that although XGBoost is capable of learning nonlinear relationships, the use of original terrain covariates alone does not fully capture the complexity of soil texture variation across all partitions of the data. Such instability likely reflects the limited effective sample size within each fold and the presence of localized terrain-texture relationships that are not fully represented within a non-spatial modeling framework.

Therefore, the results in Table 7 establish XGBoost-Original as a relevant but incomplete baseline, providing a necessary reference point for comparison with spatially explicit models such as GWR and hybrid approaches that integrate spatial and nonlinear modeling capabilities.

Table 7. Initial validation results of the Extreme Gradient Boosting (XGBoost) model

	Sand				Silt				Clay			
Fold	**RMSE**	**R^2**	**MAE**	**Best Iteration**	**RMSE**	**R^2**	**MAE**	**Best Iteration**	**RMSE**	**R^2**	**MAE**	**Best Iteration**
1	12.9553	0.3282	8.468	63	8.0282	0.2218	6.202	72	7.4306	0.3434	6.105	93
2	9.4007	0.4022	7.129	47	8.3452	-0.267	5.117	73	5.7987	0.4083	4.499	118
3	10.5659	0.6528	6.584	68	9.7043	0.0687	6.245	54	7.5458	0.4852	5.382	124
4	11.3626	0.3290	6.676	127	6.3877	0.5527	4.835	63	5.8741	0.3744	4.387	45
5	7.0545	0.7137	5.008	74	6.5759	0.4605	4.640	95	5.4112	0.4136	4.113	45
Mean	10.2679	0.4852	6.773	75.8	7.8083	0.2073	5.408	71.4	6.4121	0.4050	4.897	85

Note: RMSE: root mean square error; MAE: mean absolute error; R^2: Coefficient of Determination.

3.6 Hybrid Geographically Weighted Regression Extreme Gradient Boosting model

To better capture the complexity of soil texture variation, this study implemented a hybrid GWR XGBoost model that integrates spatially varying regression with nonlinear machine learning. The workflow begins with GWR, which is used as a spatial feature extraction stage rather than as the final predictive model. Using the original terrain covariates, namely Kh), Kv, slope, Kr, and Elev, GWR was first calibrated to model the local relationships between these predictors and each soil texture fraction. From this calibration, several location-specific outputs were extracted, including the local intercept, local regression coefficients for each terrain variable, and residual-based information. These GWR-derived outputs were then combined with the original terrain covariates to form an augmented feature space. In this way, the hybrid framework preserves the original environmental information while adding localized spatial summaries that reflect how predictor-response relationships vary across space.

In the second stage, XGBoost was trained on this enriched feature set to capture nonlinear effects and higher-order interactions that may not be adequately represented by the linear structure of GWR alone. Thus, the role of GWR in the hybrid framework is to summarize spatial heterogeneity, whereas XGBoost serves as the final learner that maps both global terrain information and local spatial signals into improved predictions. To evaluate model robustness under spatial dependence, performance was assessed using spatially structured five-fold cross-validation, in which folds were generated through coordinate-based clustering so that geographically proximate observations were grouped within the same partition. This validation design was adopted to reduce overly optimistic estimates that may arise from random splitting in spatial datasets.

Cross-validation was implemented separately within each fold to avoid data leakage. In each fold, the dataset was first divided into training and test subsets. The GWR model was then calibrated using only the training data, so that all spatial summaries produced, including the local intercept, local coefficients, and residual-based information, were derived entirely from that training fold. These GWR-derived features were subsequently combined with the original terrain covariates to form an augmented feature space, which was then used as the input for the XGBoost model. After the XGBoost model was trained on the training data, its

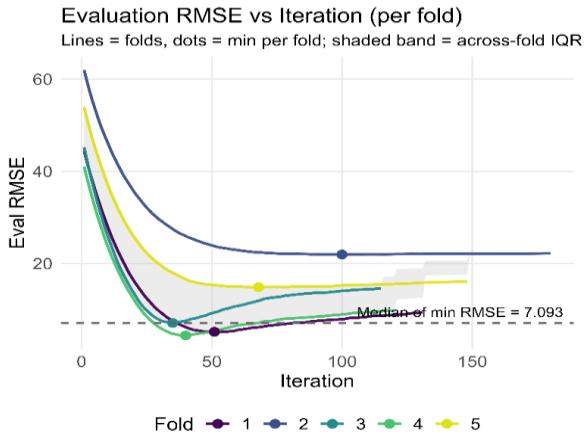
performance was evaluated on the corresponding test data. In this way, the spatial features were not generated from the full dataset before cross-validation, but were computed independently within each training fold, thereby preserving the reliability of the out-of-sample performance estimates.

The resulting fold-wise RMSE, R^2, and optimal boosting iterations provide the basis for assessing the predictive behavior of the hybrid model.

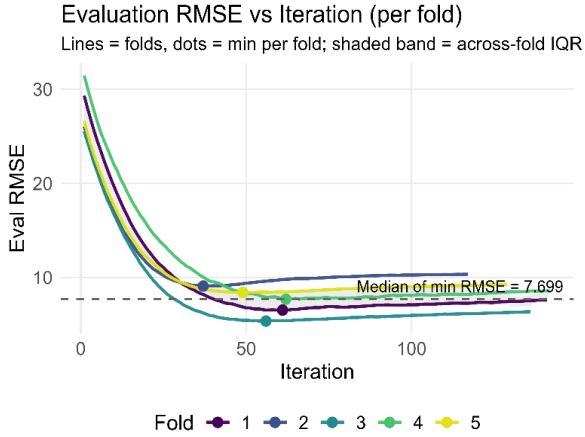
3.6.1 Initial fold validation evaluation of the hybrid Geographically Weighted Regression Extreme Gradient Boosting model

Before conducting a prediction using the hybrid GWR XGBoost approach, an initial evaluation was performed to examine the performance of the model in each fold. This evaluation used RMSE curves across iterations for each response variable, namely sand, silt, and clay. Figure 4 illustrates the dynamics of decreasing error as the number of iterations increases, while also showing the minimum RMSE point in each fold as an indication of the optimal number of iterations selected through early stopping.

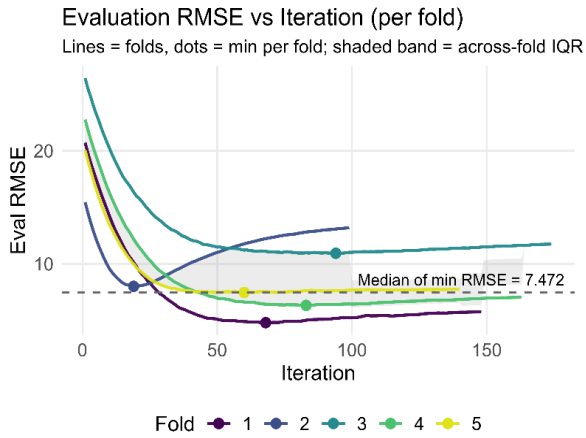
The curve results indicate that RMSE values generally decrease sharply during the early iterations and then tend to stabilize once the optimal point is reached. The median minimum RMSE is recorded at 7.093 for the sand variable, 7.699 for the silt variable, and 7.472 for the clay variable. This pattern suggests that the model is able to reach a convergence point with relatively low error, although noticeable variation across folds remains. These differences among folds provide an important basis for further parameter optimization so that the model becomes more consistent and stable in predicting the distribution of soil texture fractions.



(a) Sand as the response variable



(b) Silt as the response variable



(c) Clay as the response variable

Figure 4. Root Mean Square Error (RMSE) evaluation curve across iterations in fold validation for the hybrid Geographically Weighted Regression (GWR) Extreme Gradient Boosting (XGBoost) model

3.6.2 Hybrid Geographically Weighted Regression Extreme Gradient Boosting prediction

To ensure physical consistency with the compositional nature of soil texture data, the predicted fractions of sand, silt, and clay were adjusted so that their sum at each location equals 100%. This adjustment is necessary because the three fractions are not independent variables, but components of a single composition constrained by a constant total. In the initial hybrid GWR XGBoost model, the three fractions were predicted separately, which resulted in some locations having totals less than or greater than 100%. Therefore, the initial model outputs were normalized using a compositional closure procedure so that each location yields a valid soil texture composition.

Mathematically, if $\hat{S}_i$, $\hat{L}_i$, and $\hat{C}_i$ represent the initial predicted fractions of sand, silt, and clay at location i, then the corrected values are obtained as:

$$\begin{aligned} \hat{S}_i^{(c)} &= \frac{\hat{S}_i}{\hat{S}_i + \hat{L}_i + \hat{C}_i} \times 100 \\ \hat{L}_i^{(c)} &= \frac{\hat{L}_i}{\hat{S}_i + \hat{L}_i + \hat{C}_i} \times 100 \\ \hat{C}_i^{(c)} &= \frac{\hat{C}_i}{\hat{S}_i + \hat{L}_i + \hat{C}_i} \times 100 \end{aligned} \tag{17}$$

Thus, the following condition holds:

$$\hat{S}_i^{(c)} + \hat{L}_i^{(c)} + \hat{C}_i^{(c)} = 100\% \tag{18}$$

for each location. This procedure works by proportionally rescaling the three components relative to the total initial prediction, so that the relative relationships among fractions are preserved, while the total conforms to the compositional constraint.

As the illustration, for Location 1, the initial model produced predicted values of sand = 68.143211, silt = 10.284773, and clay = 8.015069, with a total of 86.443053. Because this total does not equal 100%, compositional closure was applied. After normalization, the sand fraction becomes 78.84, the silt fraction becomes 11.90, and the clay fraction becomes 9.27, resulting in a total of exactly 100.00. This example demonstrates that compositional closure does not

alter the relative dominance among fractions, but corrects the predictions to ensure they are valid as a soil texture composition.

The compositional correction makes the spatial interpretation more reliable at the landscape scale. Instead of reading each fraction separately as an unconstrained prediction, the resulting maps should be interpreted jointly as three interrelated components of the same soil system. In this way, an increase in one fraction is understood relative to the simultaneous adjustment of the other fractions, thereby preserving the internal balance of the soil texture composition. The final maps thus provide a more defensible basis for precision texture mapping, since they integrate spatial heterogeneity, nonlinear predictive structure, and the constant-sum requirement of soil fraction data.

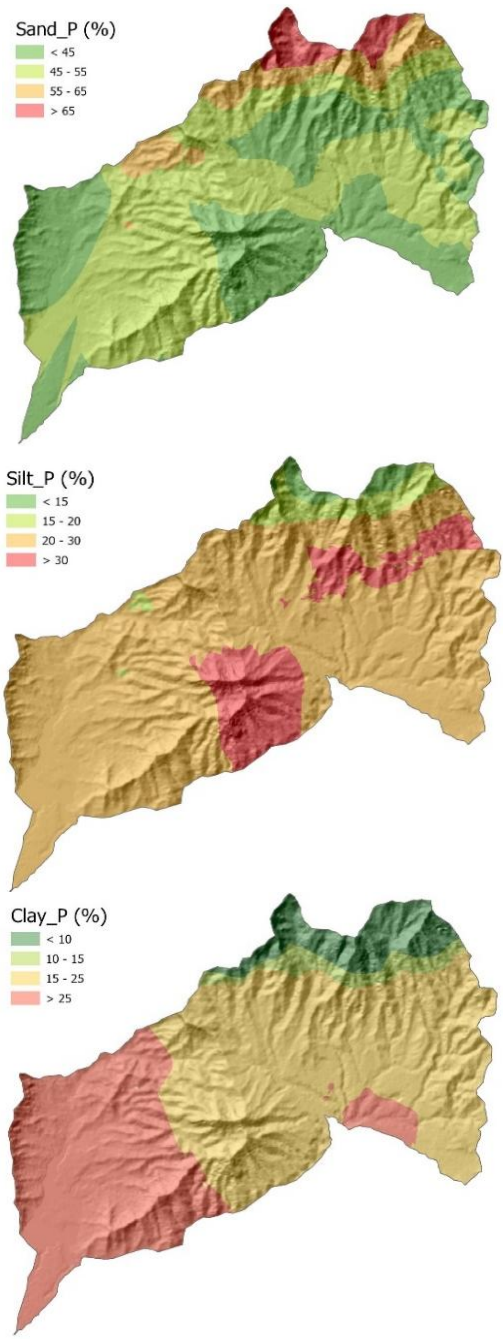


Figure 5. Spatial visualization of hybrid Geographically Weighted Regression (GWR) Extreme Gradient Boosting (XGBoost) prediction results for the sand, silt, and clay variables

After the entire prediction process and compositional correction were completed, the final values of the sand, silt, and clay fractions were obtained, which are compositionally valid, with the total of the three fractions equal to 100% at each location. Subsequently, these final predicted values were visualized in the form of spatial maps to illustrate the distribution of each soil texture fraction across the study area. These maps, as shown in Figure 5, serve as a basis for interpreting the spatial patterns of soil texture distribution more comprehensively.

The spatial prediction maps reveal clear and complementary patterns in the distribution of soil texture fractions across the study area. The sand fraction is predominantly higher in the northern to central part of the watershed, particularly in areas characterized by steeper terrain and elevated landforms. This pattern suggests that these zones are associated with coarser soil materials, likely reflecting stronger erosional processes, reduced fine-particle deposition, or terrain conditions that favor the persistence of sand-dominated textures. In contrast, lower sand values are more common in the southwestern and some southeastern parts of the study area, where the terrain appears relatively less dominated by coarse-textured material.

The predicted silt fraction exhibits a more transitional and spatially heterogeneous pattern. Moderate to high silt values are distributed across several central and locally depressed zones, indicating areas where medium-sized particles may accumulate under intermediate geomorphological conditions. Compared with sand and clay, the silt pattern is less sharply concentrated, which is consistent with its intermediate role in the soil particle-size system. This spatial behavior suggests that silt may act as a transitional fraction between coarse upland materials and finer depositional environments.

The clay fraction shows a contrasting distribution, with higher values concentrated mainly in the southwestern part of the watershed and in several localized southeastern pockets. These areas likely represent more stable or depositional environments where finer particles accumulate over time. Meanwhile, lower clay values are observed in the northern and central sectors, where sand is more dominant. This inverse spatial relationship between sand and clay supports the internal coherence of the predicted composition and indicates that the model captures meaningful textural contrasts across the study landscape.

Importantly, because the predictions were compositionally normalized, the three maps must be interpreted together rather than independently. At every mapped location, the total proportion of sand, silt, and clay is exactly 100%, meaning that the predicted fractions form a valid soil texture composition. This property substantially improves the interpretability of the results, because the spatial patterns no longer represent separate unconstrained model outputs, but integrated components of one physically consistent system. The additional total map, which shows a constant value of 100% across the study area, confirms that the compositional constraint has been satisfied for all prediction locations.

Overall, these results demonstrate that the hybrid GWR XGBoost framework is able to represent spatial heterogeneity in soil texture fractions while maintaining the compositional structure required by soil particle-size data. This is an important improvement because it increases the methodological credibility of the predictions and strengthens their usefulness for precision soil mapping, land evaluation, and spatially informed soil management. Rather than merely producing high predictive values, the model now generates

outputs that are both spatially detailed and compositionally valid, which is a stronger basis for scientific interpretation and practical application.

3.6.3 Coefficient of determination values for the hybrid Geographically Weighted Regression Extreme Gradient Boosting model

The R^2 in the hybrid GWR XGBoost model is used to evaluate the extent to which the model can explain the variation in the response variables. A higher R^2 value indicates better predictive ability, as a larger proportion of variation is explained by the model rather than by error. In addition, model performance is further assessed using the RMSE and MAE, which quantify the magnitude of prediction errors. RMSE is more sensitive to large deviations due to the squaring of residuals, while MAE represents the average absolute difference between observed and predicted values and is therefore more directly interpretable. Lower RMSE and MAE values indicate higher predictive accuracy and better model reliability.

Based on Table 8, the performance of the hybrid GWR XGBoost model is evaluated using the R^2, RMSE, and MAE. The R^2 value reflects the model's ability to explain the variability in the data, while RMSE and MAE quantify the magnitude of prediction errors. For the sand fraction, the model achieves a mean R^2 of 0.7834, an RMSE of 7.093, and an MAE of 5.086, indicating strong explanatory power along with relatively low prediction error. For the silt fraction, the model yields a mean R^2 of 0.7363, with an RMSE of 7.699 and an MAE of 5.096, suggesting a substantial improvement compared to non-spatial approaches, although prediction errors remain slightly higher than those for sand. Meanwhile, the clay fraction shows a mean R^2 of 0.7293, with an RMSE of 7.472 and an MAE of 4.858, indicating that although its explanatory power is slightly lower, the model produces the smallest prediction error in absolute terms.

Table 8. Coefficient of determination for the hybrid Geographically Weighted Regression (GWR) Extreme Gradient Boosting (XGBoost) model

	Sand				Silt				Clay			
Fold	**RMSE**	**R^2**	**MAE**	**Best Iteration**	**RMSE**	**R^2**	**MAE**	**Best Iteration**	**RMSE**	**R^2**	**MAE**	**Best Iteration**
1	7.25	0.772	5.21	52	7.85	0.724	5.12	56	7.6	0.716	4.91	61
2	7.5	0.765	5.43	48	7.90	0.728	5.28	53	7.8	0.721	5.05	58
3	6.95	0.791	4.98	60	7.30	0.744	4.87	65	7.2	0.735	4.76	70
4	6.8	0.795	4.85	55	7.50	0.739	5.01	60	7.1	0.738	4.69	66
5	6.965	0.794	4.96	63	7.945	0.747	5.2	72	7.66	0.737	4.88	74
Mean	7.093	0.7834	5.086	55.6	7.699	0.7363	5.096	61.2	7.472	0.7293	4.858	65.8

Note: RMSE: root mean square error; MAE: mean absolute error; R^2: Coefficient of Determination.

Overall, the R^2 values exceeding 0.70 across all variables demonstrate that the hybrid GWR XGBoost model has strong capability in explaining the spatial variability of soil texture fractions. At the same time, the relatively low and consistent RMSE and MAE values indicate that the model maintains good predictive accuracy across all components. This pattern confirms that the integration of GWR and XGBoost effectively captures both spatial heterogeneity and nonlinear relationships in the data. Therefore, the hybrid model provides a significant improvement over single-method approaches and represents a more reliable framework for high-precision spatial soil texture mapping.

3.6.4 Model complexity, sample size, and robustness

The use of a hybrid GWR-XGBoost framework with a dataset of 150 observations requires careful interpretation because the model combines spatially varying regression with nonlinear machine learning. Although this sample size is sufficient for an exploratory watershed-scale case study, it remains relatively limited for drawing broad generalizations. Therefore, the purpose of the hybrid model in this study is not to claim universal applicability, but to evaluate whether spatially derived GWR information can improve prediction within the specific conditions of the study area.

To reduce the risk of overfitting and to assess model robustness, the model was evaluated using five-fold cross-validation. The fold-wise results show that model performance varies across partitions, indicating that prediction accuracy is influenced by the composition of the training and testing data. However, the hybrid GWR-XGBoost model maintains consistently high explanatory performance across the three soil texture fractions, with mean R^2 values of 0.7834 for sand, 0.7363 for silt, and 0.7293 for clay. The corresponding RMSE and MAE values also remain within a relatively controlled range, suggesting that the hybrid model provides stable predictive behavior despite the limited sample size.

Nevertheless, the limited number of observations should be recognized as an important constraint. A small sample may increase sensitivity to local data structure, fold composition, and spatial clustering of observations. Therefore, the results should be interpreted as evidence of model performance within the present watershed rather than as proof of general transferability to other regions. Future studies should validate the proposed framework using larger datasets, additional watersheds, independent test data, and uncertainty analysis to further examine the robustness and generalizability of the model.

3.7 Comparison of the performance of the regression model, the Geographically Weighted Regression model, and the hybrid GWR Extreme Gradient Boosting model

The comparison of the R^2 values is used to evaluate the performance of several modeling approaches in explaining the variation of soil texture fractions. The global linear regression model, the GWR model, the XGBoost model, and the hybrid GWR XGBoost model are compared for the sand, silt, and clay variables. A higher R^2 value indicates that the model is able to explain a larger proportion of the data variation and is therefore considered to have better predictive performance.

Based on Table 9, the global linear regression model produces relatively low R^2 values for all three variables, namely 0.4087 for sand, 0.1950 for silt, and 0.4823 for clay. In terms of prediction error, this model also exhibits relatively high RMSE and MAE values, specifically 11.372 and 8.262 for sand, 8.163 and 5.985 for silt, and 6.121 and 4.792 for clay.

These results indicate that the global model is only able to explain a limited portion of the variability in soil texture fractions and still produces considerable prediction errors. The GWR model shows a substantial improvement, with R^2 values of 0.7108, 0.6058, and 0.5917 for sand, silt, and clay, respectively, demonstrating its ability to capture spatial heterogeneity. Correspondingly, the RMSE and MAE values are lower than those of the global model, namely 7.962 and 5.810 for sand, 5.712 and 5.516 for silt, and 5.435 and 4.144 for clay. This indicates that GWR not only enhances the explanatory power of the model but also improves predictive accuracy. Meanwhile, the XGBoost model yields R^2 values of 0.4852 for sand, 0.2073 for silt, and 0.4050 for clay. The corresponding RMSE values are 10.2679, 7.8083, and 6.4121, with MAE values of 6.773, 5.408, and 4.897 for sand, silt, and clay, respectively. These results suggest that the standalone nonlinear model can explain part of the variability in soil texture fractions, although its performance remains inferior to GWR for sand and clay and is substantially weaker for silt.

However, the best performance is achieved by the hybrid GWR XGBoost model, which produces the highest R^2 values across all variables, namely 0.7834 for sand, 0.7363 for silt, and 0.7293 for clay. In terms of predictive accuracy, the hybrid model also demonstrates relatively low and well-balanced RMSE and MAE values, specifically 7.093 and 5.086 for sand, 7.699 and 5.096 for silt, and 7.472 and 4.858 for clay. These results indicate that the hybrid approach successfully integrates the strengths of GWR in capturing local spatial variation with the capability of XGBoost in modeling complex nonlinear relationships. From a comparative perspective, the combination of high R^2 values and controlled prediction errors demonstrates that the hybrid model provides the best balance between explanatory power and predictive accuracy. Therefore, based on the comparative results in this study, the hybrid GWR-XGBoost model achieves the highest predictive performance among all tested approaches for modeling the spatial distribution of soil texture fractions.

Table 9. Comparison of the regression model, the Geographically Weighted Regression (GWR) model, Extreme Gradient Boosting (XGBoost), and the hybrid GWR XGBoost model

Metode	Sand			Silt			Clay		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Regression	0.4087	11.372	8.262	0.1950	8.163	5.985	0.4823	6.121	4.792
GWR	0.7108	7.962	5.810	0.6058	5.712	5.516	0.5917	5.435	4.144
XGBoost	0.4852	10.2679	6.773	0.2073	7.8083	5.408	0.4050	6.4121	4.897
GWR-XGBoost	0.7834	7.093	5.086	0.7363	7.699	5.096	0.7293	7.472	4.858

Note: RMSE: root mean square error; MAE: mean absolute error; R^2: Coefficient of Determination.

3.8 Research implications

This study carries significant methodological and applied implications. Methodologically, the integration of GWR with XGBoost demonstrates the effectiveness of a hybrid approach in bridging the limitations of each individual method. GWR excels in capturing spatial heterogeneity, while XGBoost is capable of modeling complex nonlinear relationships. The combination of these methods results in consistently improved predictive performance, with R^2 exceeding 0.70 for all soil fractions (sand, silt, and clay). These findings expand the modern spatial statistics literature by introducing a framework that is robust, interpretable, and capable of high predictive accuracy.

In addition, the tables and figures included in this study provide evidence that the observed gains are not merely numerical improvements but are consistent with the underlying behavior of soil texture formation, which is influenced simultaneously by spatial heterogeneity and predictor-response relationships that are not entirely linear. The across-model performance comparisons in the tables offer a direct basis for interpreting the added value of the hybrid design, showing how combining GWR-derived spatial descriptors with nonlinear learning can yield more consistent predictions than using either component in isolation. Furthermore, the figures summarizing k-fold evaluation and the evolution of prediction error during training strengthen the robustness argument: although variability across folds reflects differences among sampling locations, the error trend that decreases and then stabilizes indicates a more convergent learning process rather than reliance on a particular subset of the data. This cross-fold stability supports the suitability of the proposed approach for operational mapping by reducing the risk of performance degradation when applied to areas with different environmental conditions.

From an applied perspective, the results of this study provide substantial contributions to precision soil mapping that more accurately represent field conditions. The ability of the hybrid model to detect detailed spatial variation supports data-driven land use planning, soil and water conservation efforts, and sustainable agricultural strategies. Moreover, more accurate spatial predictions also strengthen policy development related to climate change adaptation and soil degradation mitigation at both local and regional scales. Overall, the implications of this research highlight that the integration of spatial approaches and machine learning not only enhances modeling accuracy but also opens new opportunities for advancing data-driven environmental analysis frameworks. Thus, the hybrid GWR XGBoost model can serve as a reference for interdisciplinary studies that require a balance between predictive accuracy, interpretability, and spatial relevance.

4. CONCLUSION

This study demonstrates that the integration of GWR and XGBoost within the hybrid GWR-XGBoost framework improves the accuracy of soil fraction distribution prediction compared to both the global regression model and standalone GWR within the scope of this case study. The model fit results show that GWR captures local spatial heterogeneity, while XGBoost is capable of handling complex nonlinear relationships. The combination of these methods yields high R^2 for all three response variables, namely 0.7834 for sand, 0.7363 for silt, and 0.7293 for clay, indicating stronger predictive performance than the evaluated baseline models for precision soil mapping in the study area. The spatial

visualization of the predictions further reinforces that this model provides not only improved predictive accuracy but also a more detailed spatial representation.

However, these findings should be interpreted in light of the single-watershed setting and the limited sample size relative to the complexity of the hybrid model, which may constrain the reliability and transferability of the results. In addition, fold-wise cross-validation suggests that performance can vary across data partitions, highlighting the need for broader comparative evaluation before drawing strong conclusions about real-world deployment. Accordingly, the hybrid GWR XGBoost approach is best regarded as a promising method for spatial modeling of soil texture in this study context, with potential relevance for precision agriculture, environmental conservation, and sustainable land use planning.

For future research, it is recommended that the hybrid GWR XGBoost approach be applied to a broader geographic scope by incorporating additional environmental variables such as rainfall, land use, and remote sensing data to enhance model generalization. Parameter optimization for XGBoost and exploration of other ensemble learning algorithms are also necessary to improve performance stability across folds. In addition, integrating this framework with spatial deep learning approaches presents a promising direction for handling large-scale spatial datasets. From a practical standpoint, the results may inform exploratory precision soil management and planning, subject to further validation using larger datasets and multi-site comparative studies.

ACKNOWLEDGMENT

The author would like to express heartfelt gratitude to the Directorate of Research, Technology, and Community Service (DRTPM) of DIKTI for the research funding support that enabled the completion of this study.

REFERENCES

[1] Zhang, T., Yang, Y.L., Liu, S.Y. (2020). Application of biomass by-product lignin stabilized soils as sustainable geomaterials: A review. Science of the Total Environment, 728: 138830. https://doi.org/10.1016/j.scitotenv.2020.138830

[2] Han, Y., Guo, X., Jiang, Y., Xu, Z., Li, Z. (2020). Environmental factors influencing spatial variability of soil total phosphorus content in a small watershed in Poyang Lake Plain under different levels of soil erosion. Catena, 187: 104357. https://doi.org/10.1016/j.catena.2019.104357

[3] Shi, W., Zhang, M. (2023). Progress on spatial prediction methods for soil particle-size fractions. Journal of Geographical Sciences, 33(7): 1553-1566. https://doi.org/10.1007/s11442-023-2142-6

[4] Zheng, T., Ouyang, S., Zhou, Q. (2023). Synthesis, characterization, safety design, and application of NPs@BC for contaminated soil remediation and sustainable agriculture. Biochar, 5(1): 5. https://doi.org/10.1007/s42773-022-00198-3

[5] Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., Burnaev, E. (2024). Challenges in data-driven geospatial modeling for environmental research and practice. Nature Communications, 15: 10700. https://doi.org/10.1038/s41467-024-55240-8

[6] Svensson, D.N., Messing, I., Barron, J. (2022). An investigation in laser diffraction soil particle size distribution analysis to obtain compatible results with sieve and pipette method. Soil Tillage Research, 223: 105450. https://doi.org/10.1016/j.still.2022.105450

[7] Iriany, A., Ngabu, W., Ariyanto, D., Pramoedyo, H. (2025). Kriging prediction and simulation model: Analysis of surface soil particle size distribution. Mathematical Modelling of Engineering Problems, 12(4): 1169. https://doi.org/10.18280/mmep.120408

[8] Comber, A., Brunsdon, C., Charlton, M., Dong, G., Harris, R., Lu, B., Lü, Y., Murakami, D., Nakaya, T., Wang, Y., Harris, P. (2023). A route map for successful applications of geographically weighted regression. Geographical Analysis, 55(1): 155-178. https://doi.org/10.1111/gean.12316

[9] Coffie, G.H., Cudjoe, S.K.F. (2024). Using extreme gradient boosting (XGBoost) machine learning to predict construction cost overruns. International Journal of Construction Management, 24(16): 1742-1750. https://doi.org/10.1080/15623599.2023.2289754

[10] Osman, A.I.A., Ahmed, A.N., Chow, M.F., Huang, Y.F., El-Shafie, A. (2021). Extreme gradient boosting (XGBoost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 12(2): 1545-1556. https://doi.org/10.1016/j.asej.2020.11.011

[11] Zhu, H., Liu, H., Zhou, Q., Cui, A. (2023). A XGBoost-based downscaling-calibration scheme for extreme precipitation events. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-12. https://doi.org/10.1109/TGRS.2023.3294266

[12] Huber, F., Yushchenko, A., Stratmann, B., Steinhage, V. (2022). Extreme gradient boosting for yield estimation compared with deep learning approaches. Computers and Electronics in Agriculture, 202: 107346. https://doi.org/10.1016/j.compag.2022.107346

[13] Ahmad, M.S., Adnan, S.M., Zaidi, S., Bhargava, P. (2020). A novel support vector regression (SVR) model for the prediction of splice strength of the unconfined beam specimens. Construction and Building Materials, 248: 118475. https://doi.org/10.1016/j.conbuildmat.2020.118475

[14] Sergeev, A.P., Buevich, A.G., Baglaeva, E.M., Shichkin, A.V. (2019). Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. Catena, 174: 425-435. https://doi.org/10.1016/j.catena.2018.11.037

[15] Kopczewska, K. (2022). Spatial machine learning: New opportunities for regional science. Annals of Regional Science, 68(3): 713-755. https://doi.org/10.1007/s00168-021-01101-x

[16] Xiong, L., Tang, G., Yang, X., Li, F. (2021). Geomorphology-oriented digital terrain analysis: Progress and perspectives. Journal of Geographical Sciences, 31: 456-476. https://doi.org/10.1007/s11442-021-1853-9

[17] Guo, Z. (2022). Soil texture is an important factor determining how microplastics affect soil hydraulic characteristics. Environment International, 165: 107293. https://doi.org/10.1016/j.envint.2022.107293

[18] Ju, X., Jia, Y., Li, T., Gao, L., Gan, M. (2021). Morphology and multifractal characteristics of soil pores

and their functional implication. Catena, 196: 104822. https://doi.org/10.1016/j.catena.2020.104822
[19] Mokarrama, M., Hojati, M. (2018). Landform classification using a sub-pixel spatial attraction model to increase spatial resolution of digital elevation model (DEM). The Egyptian Journal of Remote Sensing and Space Science, 21(1): 111-120. https://doi.org/10.1016/j.ejrs.2016.11.005
[20] Pramoedyo, H., Ngabu, W., Riza, S., Iriany, A. (2024). Spatial analysis using geographically weighted ordinary logistic regression (GWOLR) method for prediction of particle-size fraction in soil surface. IOP Conference Series: Earth and Environmental Science, 1299: 012005. https://doi.org/10.1088/1755-1315/1299/1/012005
[21] Wheeler, D.C. (2021). Geographically weighted regression. In Handbook of Regional Science, pp. 1895-1921. https://doi.org/10.1007/978-3-662-60723-7_77
[22] Iriany, A., Ngabu, W., Pramoedyo, H. (2026). Geographically weighted regression random forest for modeling soil particles. Journal of the Nigerian Society of Physical Sciences, 8(2): 2939-2939. https://doi.org/10.46481/jnsps.2026.2939
[23] Siqi, J., Yuhong, W., Ling, C., Xiaowen, B. (2023). A novel approach to estimating urban land surface temperature by the combination of geographically weighted regression and deep neural network models. Urban Climate, 47: 101390. https://doi.org/10.1016/j.uclim.2022.101390
[24] Oshan, T.M., Li, Z., Kang, W., Wolf, L.J., Fotheringham, A.S. (2019). MGWR: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. ISPRS International Journal of Geo-Information, 8(6): 269. https://doi.org/10.3390/ijgi8060269
[25] Du, Z., Wang, Z., Wu, S., Zhang, F., Liu, R. (2020). Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. International Journal of Geographical Information Science, 34(7): 1353-1377. https://doi.org/10.1080/13658816.2019.1707834
[26] Moinuddin, M., Zerguine, A., Arif, M. (2023). A weighted Gaussian kernel least mean square algorithm. Circuits Systems and Signal Processing, 42(9): 5267-5288. https://doi.org/10.1007/s00034-023-02337-y
[27] Iriany, A., Ngabu, W., Ariyanto, D. (2024). Rainfall modeling using the geographically weighted Poisson regression method. BAREKENG: Jurnal Ilmu Matematika dan Terapan, 18(1): 627-636. https://doi.org/10.30598/barekengvol18iss1pp0627-0636
[28] Koç, T. (2022). Bandwidth selection in geographically weighted regression models via information complexity criteria. Journal of Mathematics, 2022(1): 1527407. https://doi.org/10.1155/2022/1527407
[29] Sibindi, R., Mwangi, R.W., Waititu, A.G. (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports, 5(4): e12599. https://doi.org/10.1002/eng2.12599
[30] Trizoglou, P., Liu, X., Lin, Z. (2021). Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines. Renewable Energy, 179: 945-962. https://doi.org/10.1016/j.renene.2021.07.085
[31] Bacha, N.U., Lu, S., Ur Rehman, A., Idrees, M., Ghadi, Y.Y., Alahmadi, T.J. (2024). Deploying hybrid ensemble machine learning techniques for effective cross-site scripting (XSS) attack detection. Computers, Materials & Continua, 81(1): 707-748. https://doi.org/10.32604/cmc.2024.054780
[32] Lima Marinho, T., do Nascimento, D.C., Pimentel, B.A. (2024). Optimization on selecting XGBoost hyperparameters using meta-learning. Expert Systems, 41(9): e13611. https://doi.org/10.1111/exsy.13611
[33] Ye, M. (2023). Estimation of the soil arsenic concentration using a geographically weighted XGBoost model based on hyperspectral data. Science of the Total Environment, 858: 159798. https://doi.org/10.1016/j.scitotenv.2022.159798
[34] 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