Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 8 (2026) 2939

Journal of the Nigerian Society of Physical Sciences

Geographically weighted regression random forest for modeling soil particles

Atiek Iriany[a,*], Wigbertus Ngabu[b], Henny Pramoedyo[a], Amarifai[a]

[a]*Department of Statistics, Faculty of Mathematics and Natural Science, Brawijaya University, Malang, 65144, Indonesia*
[b]*Mathematics Education Study Program, University of Riau, Pekanbaru, 28293, Indonesia*

Abstract

Clay particles play a vital role in determining soil quality, particularly in the fields of agriculture and conservation. However, the complex and non-linear spatial distribution of clay particles is difficult to capture using conventional modeling methods. This study develops a hybrid model, Geographically Weighted Regression Random Forest (GWRRF), which combines the ability of Geographically Weighted Regression (GWR) to capture spatial heterogeneity with the strength of Random Forest (RF) in handling non-linear relationships. The data used in this study were derived from soil texture and local morphologic analysis across 50 observation points in the Kalikonto watershed. The results indicate that the GWRRF model achieved a higher explanatory power (R^2 = 0.735) compared to the conventional GWR model (R^2 = 0.475), demonstrating its better capability in capturing complex spatial variability. However, the RMSE value of the GWRRF model (4.314) was slightly higher than that of the GWR model (3.485), reflecting a trade-off between model flexibility and prediction accuracy. Overall, the integration of GWR and Random Forest in the GWRRF framework provides a more adaptive and context-aware approach for analyzing spatial heterogeneity in clay particle distribution, offering valuable insights for data-driven and sustainable land management practices.

DOI:10.46481/jnsps.2026.2939

Keywords: Clay particles, GWR, GWRRF, Hybrid model, Random forest

Article History :
Received: 17 May 2025
Received in revised form: 27 October 2025
Accepted for publication: 02 November 2025
Available online: 14 March 2026

© 2026 The Author(s). Published by the Nigerian Society of Physical Sciences under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Communicated by: P. Thakur

1. Introduction

Soil is a fundamental component of ecosystems that sustains life, particularly in the fields of agriculture and environmental management. It serves as a growth medium for plants and functions as a reservoir and distributor of essential nutrients [1]. One of the key characteristics that determines land productivity is soil particle composition. Generally, soil consists of sand, silt, clay, and other mineral fractions, each possessing distinct physical and chemical properties [2]. Among these, clay particles have the most significant impact on soil quality due to their extremely small size (<0.002 mm) and high chemical reactivity [3].

Clay particles contribute to several critical soil functions, including water retention, cation exchange capacity, and nutrient retention, all of which are vital for plant growth [4]. Soils with high clay content generally exhibit better moisture retention and nutrient-holding capacity [5]. Thus, understanding the spatial distribution of clay particles is essential for effective land management, particularly in the context of agriculture and soil conservation.

The spatial distribution of clay particles in a given region is influenced by various factors such as land use, climate, topography, and geological processes [6]. Conventional approaches

*Corresponding author Tel. No.: +62-818-535-96.
Email address: atiekiriany@ub.ac.id (Atiek Iriany)

often rely on simple statistical methods that fail to capture the complex spatial variation of soil particles [7]. Therefore, there is a need for more robust methods that can accurately represent local variations and capture intricate non-linear patterns in the data. Geographically Weighted Regression (GWR) is one such statistical method that has been widely applied to various spatial issues, including modeling the distribution of soil particles [8]. However, GWR is limited in its ability to handle non-linear relationships, which are common in soil characteristics.

With advances in spatial data technologies, limitations in addressing non-linear relationships in spatial data can be overcome through machine learning techniques [9]. One promising approach is to develop a hybrid model that combines GWR with machine learning algorithms such as Random Forest (RF). Random Forest is known for its ability to handle non-linear data structures and can yield more accurate predictions by aggregating results from multiple decision trees [10]. Consequently, a hybrid Geographically Weighted Regression Random Forest (GWRRF) model presents a potential solution for accurately modeling the spatial distribution of clay particles.

The GWRRF approach offers a novel contribution by simultaneously addressing spatial heterogeneity and non-linear complexity. Unlike previous studies that utilize either GWR or Random Forest in isolation, the integration of both methods results in a more accurate and context-aware model [11]. By incorporating spatial factors, GWRRF allows for more realistic analysis of clay particle distribution, facilitating the identification of areas with high erosion risk, low water retention capacity, or varying soil fertility potential [10].

Although both GWR and Random forest have been extensively used in geospatial research, their integration for modeling soil particle distribution remains underexplored [12]. GWR is effective at capturing spatial variation but struggles with complex non-linear patterns, whereas Random Forest excels in modeling non-linear relationships but overlooks spatial context [13]. Furthermore, while the GWR Random Forest model conceptually provides a comprehensive approach to addressing non-linearity in soil particle data, it also presents technical challenges, particularly when applied to large and high-quality datasets that demand intensive computational resources [14].

This study seeks to address these gaps by developing the GWRRF hybrid model, combining the strengths of GWR and Random forest. This approach is expected to produce more comprehensive and context-specific predictions of clay particle distribution. The innovation of this research lies not only in the integration of GWR and Random Forest but also in its practical application to support more effective and sustainable land management, through a deeper understanding of spatial variation and non-linear patterns in soil characteristics. The application of GWR Random Forest in modeling clay particle distribution represents not only an innovation in soil research methodology but also contributes to the enrichment of environmental science and spatial geography literature. As the demand for more adaptive approaches to spatial data and environmental data heterogeneity continues to grow, GWRRF offers a novel framework capable of addressing emerging challenges in soil research.

In terms of novelty, the GWR Random Forest model is a relatively recent hybrid approach in geospatial studies, particularly within the context of clay particle distribution modeling. The integration of GWR, which captures spatial aspects, with Random Forest, which excels in handling non-linear data structures, represents a methodological advancement well-aligned with current research needs. The success of this approach is expected to accelerate progress in soil science, especially in understanding the complex dynamics of clay particle distribution and supporting data-driven decision-making in sustainable land management.

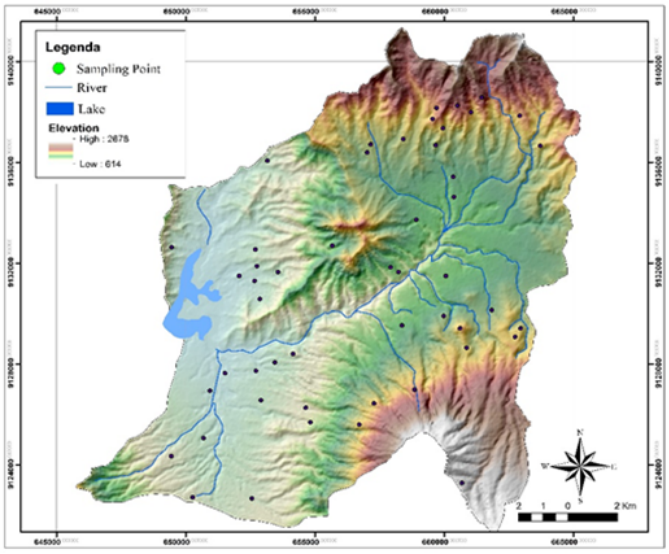


Figure 1: Research location. Source: ArcGIS 10.8 application.

Moreover, this study utilizes high-resolution spatial data, allowing for more detailed analysis of clay particle distribution patterns across various landscape conditions. Most previous studies relied on lower-resolution data or simple interpolation methods, which are less effective in capturing complex spatial variation. By applying the GWR Random Forest approach, the resulting models are not only more accurate but also more applicable in supporting soil conservation strategies and sustainable land management practices. Therefore, this research provides a valuable contribution to the fields of soil science, geospatial analysis, and environmental modeling, and serves as a reference for more precise, data-driven land use planning.

2. Method

2.1. Data

The data used in this study consist of primary data obtained from soil texture measurements and digital terrain modeling (DTM) analysis. The collected data were utilized for model training and validation purposes. A total of 50 observations were analyzed, derived from soil texture analysis conducted in the Kalikonto watershed. The research location within the Kalikonto watershed is presented in Figure 1.

The variables used in this study consist of five Local Morphologic Variables (LMVs) that represent curvature variations of the topographic surface [15]. These LMVs include:

(a) Vertical Curvature (Kv)

$$K_v = \frac{p^2 r + 2pqs + q^2 t}{(p^2 + q^2)\sqrt{(1 + p^2 + q^2)^3}}. \quad (1)$$

(b) Horizontal Curvature (Kh)

$$K_h = \frac{q^2 r - 2pqs + p^2 t}{(p^2 + q^2)\sqrt{(1 + p^2 + q^2)}}. \quad (2)$$

(c) Accumulation Curvature (Ka)

$$K_a = \frac{(q^2 r - 2pqs + p^2 t)(p^2 r + 2pqs + q^2 t)}{[(p^2 + q^2)(1 + p^2 + q^2)]^2}. \quad (3)$$

(d) Ring Curvature (Kr)

$$K_r = \left[\frac{(p^2 - q^2)s - pq(r - t)}{(p^2 + q^2)(1 + p^2 + q^2)}\right]^2. \quad (4)$$

(e) Northness Aspects (An)

$$A_n = \cos\left[-90\,(1 - \sin(q))\,(1 - |\sin(p)|) + 180\,(1 + \sin(p)) - \frac{180}{\pi}\sin(p) \arccos\left(\frac{-q}{\sqrt{p^2 + q^2}}\right)\right].$$

In the equations of *Local Morphologic Variables* (LMVs), the variables p, q, r, s, and t represent the partial derivatives of the topographic surface elevation function $z = f(x, y)$. These variables play an essential role as the basis for determining the curvature and slope direction of a surface derived from *Digital Elevation Model* (DEM) data.

Specifically, p and q denote the first-order derivatives of elevation. The value $p = \frac{\partial z}{\partial x}$ represents the change in elevation or surface slope along the x-axis, while $q = \frac{\partial z}{\partial y}$ describes the slope along the y-axis. Together, these two components indicate the direction and magnitude of the slope at a given point on the surface.

Meanwhile, r, s, and t are second-order derivatives related to the curvature of the surface. The value $r = \frac{\partial^2 z}{\partial x^2}$ indicates the rate of slope change along the x-axis, whereas $t = \frac{\partial^2 z}{\partial y^2}$ explains the slope change along the y-axis. In addition, $s = \frac{\partial^2 z}{\partial x \partial y}$ is a mixed derivative that represents the interaction of elevation changes caused by the combination of the x and y directions.

2.2. Geographically weighted regression

In this study, the Geographically Weighted Regression (GWR) model is employed to capture spatial heterogeneity in soil particle distribution by allowing regression coefficients to vary across locations [16]. Unlike global regression models that assume constant coefficients across the entire study area, GWR accounts for local differences influenced by geographic context [17]. Consequently, GWR provides more accurate and location-specific estimates, which makes it highly relevant for the objectives of this research [18].

y_i	:	Observation value of the response variable at location i
x_{ik}	:	Observation value of the explanatory variable at location i
$\beta_0(u_i, v_i)$	:	Intercept value of the model at location i
$\beta_k(u_i, v_i)$	:	Parameter values for each i-th location
(u_i, v_i)	:	Coordinate point (latitude, longitude) of the i-th location

Geographically Weighted Regression (GWR) is a spatial analysis method based on point data, developed from linear regression by incorporating spatial location [18]. The GWR model is formulated as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + e_i, \quad i = 1, 2, \ldots, n, \quad (5)$$

where the error term e_i represents the random disturbance at location i, which captures the variation in the response variable not explained by the explanatory variables. In the GWR framework, it is generally assumed that e_i is independently and identically distributed with zero mean and constant variance, i.e., $e_i \sim N(0, \sigma^2)$. This assumption ensures that the model residuals are spatially uncorrelated and that the estimated parameters remain unbiased.

2.3. Parameter estimation of the geographically weighted regression (GWR) model regression

Parameter estimation in the Geographically Weighted Regression (GWR) model is performed using the Weighted Least Squares (WLS) method, which assigns different weights to each location. The weight for each location $(u_i, v_i)$is denoted as $w_j(u_i, v_i)$, where j=1,2,...,n, The variation in weights reflects the differing characteristics of each location in the GWR model. The parameter estimation for the GWR model at the observation location (u_i, v_i), based on the application of the weight $w_j(u_i, v_i)$, is expressed as follows [19]:

$$y_i w_j(u_i, v_i) = w_j(u_i, v_i)\left(\beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + e_i\right). \quad (6)$$

By assigning the weight $w_j(u_i, v_i)$ in the GWR model, the goal is to minimize the sum of squared errors, which is expressed by the following equation:

$$\sum_{j=1}^{n} w_j(u_i, v_i) e_i^2 = \sum_{j=1}^{n} w_j(u_i, v_i)\left[y_i - \beta_0(u_i, v_i) - \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik}\right]^2 \quad (7)$$

can be written in matrix form as:

$$\begin{aligned} & e'(u_i, v_i)\mathbf{W}(u_i, v_i)e(u_i, v_i) \\ &= (Y - X\beta(u_i, v_i))' W(u_i, v_i)(Y - X\beta(u_i, v_i)) \\ &= Y' W(u_i, v_i) Y - Y' W(u_i, v_i) X\beta(u_i, v_i) \end{aligned}$$

$$-\beta'(u_i,v_i)X'W(u_i,v_i)Y+\beta'(u_i,v_i)X'W(u_i,v_i)X\beta(u_i,v_i). \quad (8)$$

Expanding the quadratic form in Equation (8) produces three main components, which simplify to the expression shown in Equation (9).

$$\begin{aligned} &e'(u_i,v_i)\mathbf{W}(u_i,v_i)e(u_i,v_i)\\ &=\mathbf{Y}'\mathbf{W}(u_i,v_i)\mathbf{Y}-2\boldsymbol{\beta}'(u_i,v_i)\mathbf{X}'\mathbf{W}(u_i,v_i)\mathbf{Y}\\ &+\boldsymbol{\beta}'(u_i,v_i)\mathbf{X}'\mathbf{W}(u_i,v_i)\mathbf{X}\boldsymbol{\beta}(u_i,v_i). \end{aligned} \quad (9)$$

Equation (9) represents the expanded form of the quadratic expression, where the cross-product terms have been separated into individual components. This step confirms the algebraic consistency of the GWR formulation.

In Equation (9), the term $Y'W(u_i,v_i)Y$ corresponds to the squared component of the response variable, the term $2\beta'(u_i,v_i)X'W(u_i,v_i)Y$ arises from the cross-product between the explanatory variables and the response, and the term $\beta'(u_i,v_i)X'W(u_i,v_i)X\beta(u_i,v_i)$ represents the weighted quadratic form of the explanatory variables. Together, these three components form the complete expansion of the quadratic expression in Equation (8).

$$\begin{aligned} e'W(u_i,v_i)e = Y'W(u_i,v_i)Y-2\beta'(u_i,v_i)X'W(u_i,v_i)Y,\\ \beta'(u_i,v_i)X'W(u_i,v_i)X\beta(u_i,v_i), \end{aligned} \quad (10)$$

where

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

$$W(u_i,v_i)=\operatorname{diag}\left[w_1(u_i,v_i),w_2(u_i,v_i),\ldots,w_n(u_i,v_i)\right]. \quad (11)$$

$$W(u_i,v_i)=\begin{pmatrix} w_1(u_i,v_i) & 0 & 0 & \cdots & 0\\ 0 & w_2(u_i,v_i) & 0 & \cdots & 0\\ 0 & 0 & w_3(u_i,v_i) & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & w_n(u_i,v_i)\end{pmatrix}.$$

$W(u_i,v_i)$ is the spatial weighting matrix for the GWR model with dimensions $n\times n$. The parameter estimation of the GWR model is obtained by deriving with respect to $\beta'(u_i,v_i)$ [8]:

$$\frac{\partial e'W(u_i,v_i)e}{\partial\beta'(u_i,v_i)}=0 \quad (12)$$

$$\frac{\partial\left(Y'W(u_i,v_i)Y-2\beta'(u_i,v_i)X'W(u_i,v_i)Y\right)}{\partial\beta'(u_i,v_i)}+$$

$$\frac{\partial\left(\beta'(u_i,v_i)X'W(u_i,v_i)X\beta(u_i,v_i)\right)}{\partial\beta'(u_i,v_i)}=0, \quad (13)$$

$$0-2X'W(u_i,v_i)Y+X'W(u_i,v_i)X\beta(u_i,v_i)$$

$$+\left(\beta'(u_i,v_i)X'W(u_i,v_i)X\right)'=0, \quad (14)$$

$$-2X'W(u_i,v_i)Y+X'W(u_i,v_i)X\beta(u_i,v_i)$$

$$+X'W(u_i,v_i)X\beta(u_i,v_i)=0, \quad (15)$$

$$2X'W(u_i,v_i)X\beta(u_i,v_i)=2X'W(u_i,v_i)Y, \quad (16)$$

$$\beta(u_i,v_i)=\frac{2X'W(u_i,v_i)Y}{2X'W(u_i,v_i)X}, \quad (17)$$

$$\hat{\beta}(u_i,v_i)=\left(X'W(u_i,v_i)X\right)^{-1}X'W(u_i,v_i)Y. \quad (18)$$

Using Equation (18), the parameter coefficients of the GWR model are obtained, with each location having distinct coefficient values.

2.4. Random forest regression

The term Random Forest was first introduced by Tin Kam Ho in 1995. Random Forest is an ensemble method designed to improve the accuracy of data classification by combining multiple unstable classifiers derived from the same algorithm through a voting process to generate the final classification prediction [20]. It is an extension of the Classification and Regression Tree (CART) method, which incorporates bootstrap aggregating (bagging) and random feature selection [21]. CART is a data exploration technique based on decision trees, where a classification tree is generated for categorical response variables and a regression tree for numerical response variables. The construction of a CART classification tree involves three main steps:

(a) Selection of the splitter (split).

(b) Determination of terminal nodes.

(c) Labeling of class categories.

Bootstrap aggregating (bagging) is a technique used to create bootstrap samples, where each decision tree is built using a bootstrap sample of candidate data attributes, with node splitting based on randomly selected subsets of these attributes [22]. Bagging is a widely applied ensemble method in classification algorithms that aims to enhance the accuracy of classifiers by aggregating multiple weak learners, yielding better results than random sampling. Both bagging and boosting are relatively new ensemble methods that have gained popularity [23].

In Random Forest, randomness is introduced not only in data sampling but also in the selection of predictor variables. As a result, the generated decision trees vary in size and structure [24]. Random Forest consists of multiple decision trees constructed using random vectors [25]. It extends the decision tree method by training multiple decision trees using individual bootstrap samples, with each tree splitting attributes selected from a random subset. Classification is then based on majority voting among the trees [11][26].

The classification process, poorly performing trees may yield weak predictions, but the strongest predictors will still emerge [27]. To achieve more stable importance measures, it is recommended to use a large number of trees, especially when

dealing with many independent variables [28]. Random Forest provides variable importance measures, namely Mean Decrease Accuracy (MCA) and Mean Decrease Gini (MDG).

The Random Forest operator generates a set of random trees, and the final class is determined by the mode (most frequently predicted class) among these trees [29]. A large number of trees are grown in a Random Forest, forming the "forest" to be analyzed. Given a dataset with nnn observations and ppp predictor variables, the Random Forest algorithm is carried out as follows [30]:

(a) A random sample of size n is drawn with replacement from the dataset—this step is known as the bootstrap sampling process.

(b) Using each bootstrap sample, a decision tree is grown to its maximum size (i.e., without pruning). At each node, the best split is chosen from a randomly selected subset of m predictor variables, where m < p. This step is referred to as random feature selection.

(c) Steps 1 and 2 are repeated k times to produce a forest consisting of k decision trees.

To achieve optimal performance, the Random Forest algorithm must define both m, the number of randomly selected predictor variables, and k, the number of trees to be built [31]. a value of k = 50 is sufficient for achieving satisfactory classification results using bagging, suggests that $k \geq 100$ tends to reduce the misclassification rate. The value of m, the number of randomly selected predictor variables, significantly affects the correlation among trees and the strength of each individual tree [32]. To determine m, where p is the total number of independent variables, the following rules of thumb are suggested [30]:

(a) For classification, the value of m is determined using the formula $\left|\sqrt{}\right|$, with the minimum number of nodes set to 1.

(b) For regression, the value of m is determined using the formula $\left|\sqrt{p}\right|$, with the minimum number of nodes set to 5.

3. Results and discussion

3.1. Modeling using GWR

3.1.1. Spatial autocorrelation test (Moran's I test)

The spatial autocorrelation test is conducted to determine whether there is spatial autocorrelation or spatial location effects in the data being analyzed. The results of the Moran's I test are presented in Table 1.

Based on the results of the Moran's I test conducted using R-Studio software, it was found that the variables Kh through M exhibit spatial autocorrelation at a 95% confidence level. Given that the variables in this study contain spatial autocorrelation, the use of a Geographically Weighted Regression model is more appropriate and provides better performance compared to a global regression model.

Table 1: Spatial autocorrelation test.

Variables	P-value	Moran's I Statistics	Decision
Kh	2.343×10^{-7}	0.5837	Reject H_0
Kv	5.722×10^{-5}	0.5936	Reject H_0
H	3.789×10^{-8}	0.5872	Reject H_0
K	6.742×10^{-8}	0.5213	Reject H_0
M	1.297×10^{-6}	0.5773	Reject H_0

Table 2: GWR model parameter estimates at location 1.

Variables	Coefficient	T-test	Information
Intercept	21,506	-	-
Kh	-0,276	-0,503	not significant
Kv	0,109	0,018	not significant
H	-0,297	-0,507	not significant
K	-6,919	-10,214	significant
M	-0,354	-0,591	not significant

Significant if $t\text{-}test > t_{(0.025,95)} = 1.985$
$R^2 = 0.475$
RMSE = 5.813
F-test = 1.714 $< F_{table} = 2.28$

3.1.2. Testing of GWR model parameter estimators

Simultaneous testing of parameter estimation in the GWR model is conducted to assess the impact of weighting on the estimation process for soil type parameters. The results of the simultaneous parameter estimation test are presented in Table 2. This test employs the t-test statistic, based on the following hypothesis:

Based on Table 2, the calculated F-test statistic is lower than the critical F-value, indicating that the clay particle model does not have a statistically significant effect at the 95% confidence level. Therefore, it can be concluded that applying weights in the GWR model does not have a significant effect on the resulting model, and simultaneously, all variables included do not significantly influence the determination of clay particle content.

The next step involves partial parameter estimation testing for clay particle content. This test is conducted to determine the effect of individual independent variables on clay particle concentration. In GWR modeling, partial parameter testing is performed at each observation point, meaning the parameter estimates are local in nature.

The test is carried out by comparing the t-test statistic, where H_0 is rejected if $|t - statistic| > t_{((0.025;50))} = 1.985$. Table 2 presents the results of the partial parameter estimation test at Location 1, using the t-test and comparing it with $t_{((0.025;50))}$. Based on the results shown in Table 2 above, it is found that the K variable has a significant effect on clay particle content, while the other four variables Kh, Kv, H, and M do not significantly affect clay particle content. Thus, it can be concluded that at Location 1, only the K variable significantly influences clay particle concentration. This finding reflects the variability of clay particle content across locations and highlights the influence of multiple factors [32]. The GWR model for clay particle con-

Table 3: Actual data vs GWR model predictions.

Loc.	Clay	Clay	Loc.	Clay	Clay
	Y	$\hat{Y}_{GWR}$		Y	$\hat{Y}_{GWR}$
1	2	4,697,264	26	20	1,598,489
2	2	4,281,628	27	3	1,488,378
3	2	3,153,864	28	21	1,710,606
4	3	1,080,245	29	22	1,704,282
5	2	8,574,699	30	21	156,213
6	1	8,886,942	31	18	1,796,655
7	3	1,156,358	32	23	1,833,851
8	2	6,324,589	33	24	2,163,748
9	2	0.637662	34	27	1,702,784
10	14	1,349,418	35	17	183,751
11	3	7,782,667	36	18	1,953,802
12	3	0.376376	37	18	1,780,913
13	3	7,826,195	38	22	1,712,451
14	2	1,315,539	39	21	1,839,847
15	14	1,624,528	40	36	1,922,481
16	15	1,226,581	41	16	1,919,985
17	16	1,383,928	42	16	1,844,107
18	26	146,788	43	18	1,928,663
19	3	1,213,954	44	24	1,978,391
20	21	1,405,077	45	24	205,225
21	2	1,067,279	46	20	1,908,092
22	24	1,560,551	47	15	1,884,799
23	26	1,818,197	48	24	1,877,566
24	24	1,236,954	49	24	1,971,472
25	16	1,718,686	50	32	2,191,144

tent at Location 1, based on Table 2, can be written as follows: $\hat{Y}_1 = 21,506–0,276Kh + 0,109Kv–0,297H–6,919K–0,354M$

The GWR model illustrates the relationship between each predictor variable and the response variable at Location 1, where a negative sign indicates that the variable decreases the clay particle content, while a positive sign indicates that the variable increases it. Based on the model above, it can be seen that Kv is the only variable that increases clay particle content, whereas Kh, H, K, and M are variables that reduce clay particle content at Location 1.

The coefficient of determination (R^2) obtained from the GWR model for clay particles is 0.475 or 47.5%. This R^2 value indicates that the predictor variables Kh, Kv, H, K, and M explain 47.5% of the variation in clay particle content, while the remaining percentage is attributed to other factors not included in the model.

Table 3 presents a comparison between the actual values of clay content (Y) and the predicted values obtained from the GWR model ($\hat{Y}_{GWR}$). Overall, the predicted values closely follow the observed data, although some differences remain, particularly at locations with relatively high or low clay content. This indicates that the GWR model is able to capture the spatial variation of clay reasonably well, but local discrepancies suggest the presence of additional influencing factors that may not be fully explained by the model.

Figure 2 illustrates the comparison graph between the actual clay particle content and the predicted values obtained from the GWR model. The orange line represents the actual values, while the blue line shows the predicted ones. Based on the figure, it can be observed that the actual values are generally higher than the predicted values. This outcome reflects the GWR model results, where most predictor variables had negative coefficients, leading to lower predicted clay contents.

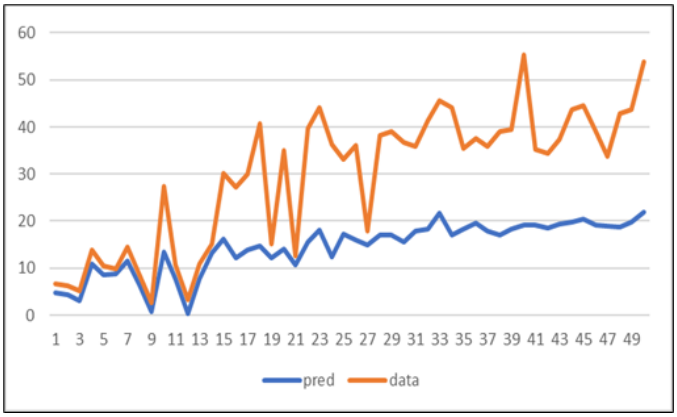


Figure 2: Actual data vs GWR model prediction graph.

Table 4: Random forest modeling results.

Particle	R^2	Concordance	MSE	RMSE
Clay	0.535	0.803	25.231	3.485

The orange line in the graph above represents the actual clay particle content at Location 1, while the blue line indicates the predicted values obtained using the GWR model. Based on the graph, it can be observed that, in general, the actual values are higher than the predicted ones. This graph reflects the previously obtained GWR model, in which nearly all predictor variables have negative coefficients, thus contributing to a decrease in the predicted clay particle content.

3.2. Modeling using random forest

Random Forests are built upon the decision tree method known as Classification and Regression Trees (CART). This technique combines multiple decision trees into an ensemble. The goal of this aggregation is to improve model accuracy by introducing randomness into the modeling process. The ensemble is generated by averaging the predictions of several models, each built from a different bootstrap sample of the original dataset. Additionally, at each tree split, only a random subset of all features is considered to identify the most optimal parameter.

Random Forest can be used to analyze predictors and identify their influence on spatial distribution, including soil characteristics whether on a global, local, or partial scale. Through this approach, various interpretations can be derived regarding soil formation processes and the impact of specific features on the model.

Table 4 summarizes the Random Forest modeling results for clay particle content. In this analysis, five variables were

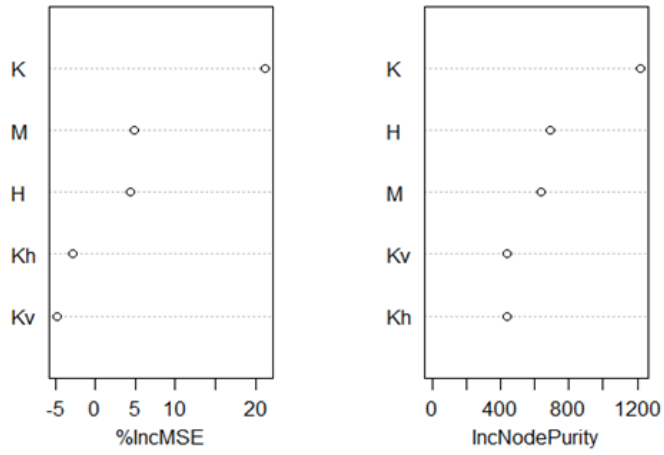


Figure 3: Random forest model predictor ranking.

Table 5: GWR random forest cut off model parameter estimation

Variables	Coefficient	T-test	Decision
Intercept	21.901	35.678	-
H	-0.226	-3.672	Significant
K	-7.301	-9.841	Significant
M	-0.423	-0.676	Not Significant

R^2 = 0.735
RMSE = 5.736
F-test = 2.105*

*) *Significant if t-test* > $t_{(0.025;46)}$ = *2.014. Significant at Level 5%. NS: Not Significant*

randomly selected at each split, and the model was constructed using 1,000 decision trees.

Table 4 shows the performance of the Random Forest model in predicting clay particle content. The model achieved an R^2 value of 0.535, indicating that about 53.5% of the variability in clay content was explained by the predictors. The concordance value of 0.803 suggests a good level of agreement between observed and predicted values. Meanwhile, the error statistics, with MSE = 25.231 and RMSE = 3.485, indicate a moderate prediction error, showing that while the model performs reasonably well, there is still room for improvement in capturing the full variability of clay content.

In Random Forest modeling, predictor ranking can also be determined. This ranking is displayed through a %IncMSE graph, which shows the increase in mean square error when a variable is randomly permuted. The higher the %IncMSE value, the greater the increase in error due to the permutation of that variable, indicating a higher level of importance. The predictor ranking graph is presented in Figure 3. In addition, the analysis results indicate that the percentage of variance explained for clay particle content is 53.5%. This result is relatively unsatisfactory when compared to previous studies that also employed the Random Forest approach, although those studies used different predictor variables. The suboptimal outcome in this study may be attributed to the quality of the input data, which may not have been sufficient to capture the complexity of clay particle distribution accurately.

Table 6: Comparison of GWR and GWRRF cutoff model results.

Particle	GWRRF Cut off		GWR	
	R^2	RMSE	R^2	RMSE
Clay	0.735	4.314	0.475	3.485

Table 7: Predictions of the GWR model and the GWRRF model.

Loc.	Clay Y	Clay $\hat{Y}_{GWRRF}$	Loc.	Clay Y	Clay $\hat{Y}_{GWRRF}$
1	2	5,252,264	26	20	1,673,989
2	2	4,836,628	27	3	1,563,878
3	2	3,708,864	28	21	1,786,106
4	3	1,135,745	29	22	1,779,782
5	2	9,129,699	30	21	163,763
6	1	9,441,942	31	18	1,872,155
7	3	1,211,858	32	23	1,909,351
8	2	6,879,589	33	24	2,239,248
9	2	1,192,662	34	27	1,778,284
10	14	1,404,918	35	17	191,301
11	3	8,337,667	36	18	2,029,302
12	3	0.931376	37	18	1,856,413
13	3	8,381,195	38	22	1,787,951
14	2	1,371,039	39	21	1,915,347
15	14	1,680,028	40	36	1,997,981
16	15	1,282,081	41	16	1,995,485
17	16	1,439,428	42	16	1,919,607
18	26	152,338	43	18	2,004,163
19	3	1,269,454	44	24	2,053,891
20	21	1,460,577	45	24	212,775
21	2	1,122,779	46	20	1,983,592
22	24	1,616,051	47	15	1,960,299
23	26	1,873,697	48	24	1,953,066
24	24	1,292,454	49	24	2,046,972
25	16	1,774,186	50	32	2,266,644

3.3. Evaluation of PSF modeling results using GWRRF

Particle Size Fraction refers to the proportion of soil particles categorized into clay, silt, and sand fractions, which are used as response variables in this study. Model evaluation was conducted to determine the best model for predicting clay particle content. In this evaluation process, the type of predictor variables had to remain consistent between the GWR and Random Forest models. Therefore, predictor selection was based on a cutoff determined by the predictor ranking percentages from the RF modeling results. Predictors were selected if their %IncMSE values were greater than zero. Based on this criterion, three predictors met the requirement: K, M, and H.

Based on Table 5, the simultaneous test of the three predictor variables H, K, and M, shows that they have a significant

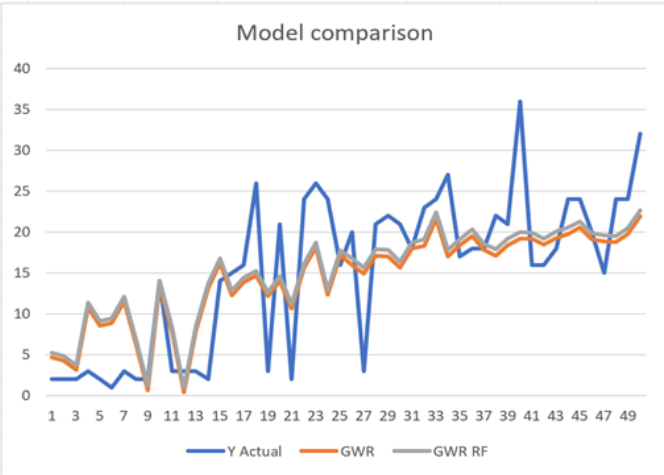


Figure 4: Comparison of actual data, GWR prediction, and GWRRF prediction.

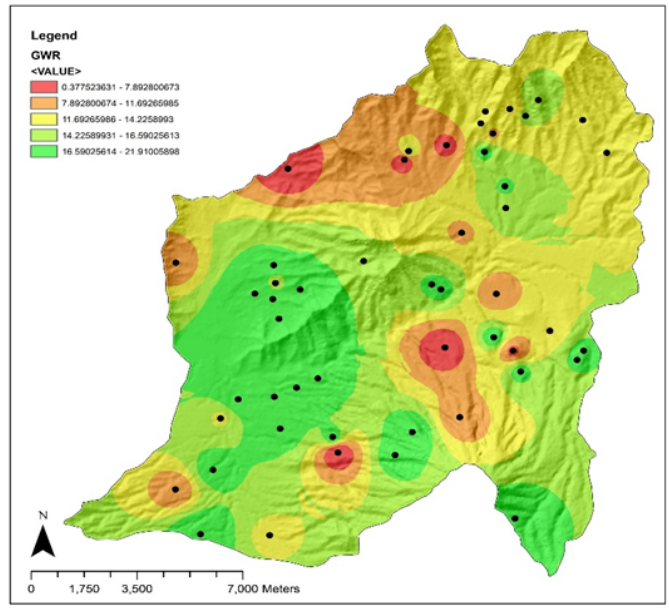
effect on clay particle content. However, in the partial test, only H and K are significant, while M is not. In addition, the values of R-squared and RMSE in the GWR model with Random Forest-based cutoff are lower compared to the original GWR model without cutoff. This indicates that the cutoff-based GWR model performs worse than the full GWR model, and therefore does not provide a better fit.

Table 6 presents a comparison of the GWR and Random Forest modeling results before and after applying the cut-off. The results indicate a significant difference between the GWR model and the GWRRF cut-off model. This can be observed in the cut-off GWRRF model, where an increase in accuracy is reflected by a higher R^2 value for clay particle content.

Table 6 shows that the GWRRF Cutoff model achieved a higher (0.735) compared to GWR (0.475), indicating that it explains more variance in clay content. However, the RMSE of GWRRF Cutoff (4.314) is slightly higher than that of GWR (3.485), which suggests that despite capturing more overall variability, the model tends to produce larger errors at certain locations. This reflects a trade-off between variance explained and error magnitude, and the results remain valid as both metrics highlight different aspects of model performance.

The variation in RMSE values across the models reflects the differences in modeling frameworks and parameter selections. While the GWRRF model with variable cutoff achieved a higher explanatory power (R^2 = 0.735), it also resulted in a higher RMSE (4.314), indicating a trade-off between variance explained and prediction error. This suggests that although the hybrid GWRRF model better captures spatial heterogeneity, it may produce larger residuals at specific locations with extreme values.

The prediction of Particle Size Fraction (clay particle content) appears to yield better results when modeled using the GWRRF method compared to GWR alone. Therefore, it can be concluded that the GWRRF model performs better than the GWR model in predicting clay particles. The prediction values for all 50 locations are shown in Table 7.



source: ArcGIS 10.8 application

Figure 5: Visualization of GWR predictions of clay particle content.

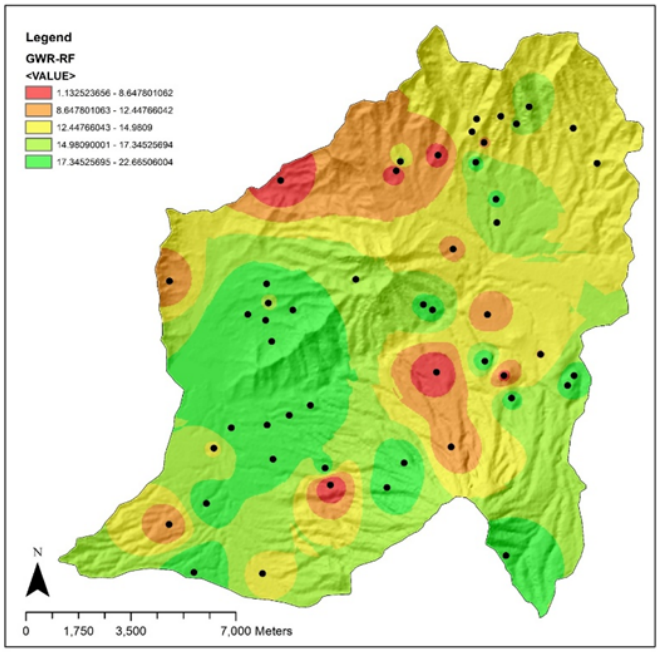
Table 7 presents the prediction results of the GWRRF model compared with the observed values of clay content. The predicted values generally follow the actual observations, indicating that the GWRRF model is able to capture spatial variability in the data, although some discrepancies appear at locations with very high or very low clay content.

Figure 4. The following graph provides a clearer illustration of the comparison between the actual data, the predicted values using the GWR model, and the predicted values using the GWRRF hybrid model.

Based on Figure 4, it can be observed that the predictions from the GWR and Random forest models do not differ significantly from each other, but both show noticeable differences when compared to the actual data. The graph pattern indicates that the predicted values are smoother than the actual values. The Random Forest predictions, represented by the grey line, appear slightly higher than those generated by the GWR model. These prediction results support the overall modeling findings, indicating that the GWR-Random Forest model is more suitable for analyzing clay particle content than the standard GWR model.

The prediction results from the GWR and GWRRF models are visualized separately in Figures 5 and 6. Figure 5 shows the distribution of clay content predicted using the GWR model, while Figure 6 presents the corresponding predictions using the GWRRF model.

Based on the visualization results in Figure 5, the GWR model displays a smoother distribution of clay particle content with less distinct spatial variation, whereas the GWRRF model captures more contrasted and detailed spatial variations. The red areas indicating high clay content and the green areas indicating low content are more clearly visible in the GWRRF



source: ArcGIS 10.8 application

Figure 6: Visualization of GWRRF predictions of clay particle content.

model, highlighting its superior ability to capture the complexity of non-linear relationships and spatial heterogeneity. This comparison reinforces earlier findings that the GWRRF hybrid model is more adaptive and accurate in predicting the distribution of clay particles than the standalone GWR model.

4. Conclusion

This study demonstrated the effectiveness of the GWRRF model in predicting the spatial distribution of clay particles. Compared with the conventional GWR model ($R^2 = 0.475$), the GWRRF approach produced a higher coefficient of determination ($R^2 = 0.735$) and was able to capture more detailed spatial variability, although some discrepancies remained at extreme values. The prediction results showed that the GWR-RF model generally followed the observed data more closely than GWR alone, and the visualization further confirmed its superior ability to represent spatial heterogeneity and non-linear patterns in soil characteristics. Therefore, the integration of GWR and Random forest offers a more adaptive and accurate modeling framework for analyzing soil particle distribution, supporting data-driven strategies in land management and soil conservation.

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgment

The author also extends heartfelt thanks to the Directorate of Research and Community Service (DRPM), Universitas Brawijaya, for the financial support provided through the Flagship Fundamental Research Grant. This funding played a vital role in enabling and successfully carrying out this study.

References

[1] K. Yin, A. Fauchille, E. Di Filippo, P. Kotronis & G. Sciarra, "A review of sand-clay mixture and soil-structure interface direct shear test", Geotechnics **1** (2021) 260. https://doi.org/10.3390/geotechnics1020014.

[2] Y. Li, E. Padoan & F. Ajmone-Marsan, "Soil particle size fraction and potentially toxic elements bioaccessibility: A review", Ecotoxicology and Environmental Safety **209** (2021) 111806. https://doi.org/10.1016/j.ecoenv.2020.111806.

[3] F. J. Matus, "Fine silt and clay content is the main factor defining maximal C and N accumulations in soils: a meta-analysis", Scientific Reports **11** (2021) 6438. https://doi.org/10.1038/s41598-021-84821-6

[4] D. Kim, B. H. Nam & H. Youn, "Effect of clay content on the shear strength of clay-sand mixture", International Journal of Geo-Engineering **9** (2018) 19. https://doi.org/10.1186/s40703-018-0087-x

[5] J. Zhang, J. E. Amonette & M. Flury, "Effect of biochar and biochar particle size on plant-available water of sand, silt loam, and clay soil", Soil and Tillage Research **212** (2021) 104992. https://doi.org/10.1016/j.still.2021.104992.

[6] H. Pramoedyo, W. Ngabu, S. Riza & A. Iriany, "Spatial analysis using geographically weighted ordinary logistic regression (GWOLR) method for prediction of particle-size fraction in soil surface", in *IOP Conference Series: Earth and Environmental Science*, vol. 1299, no. 1, IOP Publishing, 2024, pp. 012005. https://doi.org/10.1088/1755-1315/1299/1/012005.

[7] H. Yu, A. S. Fotheringham, Z. Li, T. Oshan, W. Kang & L. J. Wolf, "Inference in multiscale geographically weighted regression", Geographical Analysis **52** (2020) 87. https://doi.org/10.1111/gean.12189

[8] L. Chao, K. Zhang, Z. Li, Y. Zhu, J. Wang & Z. Yu, "Geographically weighted regression based methods for merging satellite and gauge precipitation", Journal of Hydrology **558** (2018) 275. https://doi.org/10.1016/j.jhydrol.2018.01.038.

[9] Y. Gao, J. Zhao & L. Han, "Exploring the spatial heterogeneity of urban heat island effect and its relationship to block morphology with the geographically weighted regression model", Sustainable Cities and Society **76** (2022) 103431. https://doi.org/10.1016/j.scs.2022.103431.

[10] S. Georganos & S. Kalogirou, "A forest of forests: a spatially weighted and computationally efficient formulation of geographical random forests", ISPRS International Journal of Geo-Information **11** (2022) 471. https://doi.org/10.3390/ijgi11090471.

[11] J. L. Speiser, M. E. Miller, J. Tooze & E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling", Expert Systems with Applications **134** (2019) 93. https://doi.org/10.1016/j.eswa.2019.05.048.

[12] A. Iriany, W. Ngabu, D. Ariyanto & H. Pramoedyo, "Kriging prediction and simulation model: analysis of surface soil particle size distribution", Mathematical Modelling of Engineering Problems **12** (2025) 408. https://doi.org/10.18280/mmep.120408.

[13] W. Ngabu, R. Fitriani, H. Pramoedyo & A. B. Astuti, "Cluster fast double bootstrap approach with random effect spatial modeling", BAREKENG: Jurnal Ilmu Matematika dan Terapan **17** (2023) 0945. https://doi.org/10.24123/barekeng.v17i2.900.

[14] F. Deng, W. Liu, M. Sun, Y. Xu, B. Wang, W. Liu *et al.*, "Fine estimation of water quality in the yangtze river basin based on a geographically weighted random forest regression model", Remote Sensing **17** (2025) 731. https://doi.org/10.3390/rs17040731.

[15] P. A. Shary, "Land surface in gravity points classification by a complete system of curvatures", Mathematical Geology **27** (1995) 373. https://doi.org/10.1007/BF02101691.

[16] M. Charlton, S. Fotheringham & C. Brunsdon, *Geographically weighted regression*, White paper. National Centre for Geocomputation. National University of Ireland Maynooth, 2009, vol. 2. Available online: https://www.ncge.ie/wp-content/uploads/2019/01/GWR_whitepaper.pdf.

[17] C. Brunsdon, S. Fotheringham & M. Charlton, "Geographically weighted regression", Journal of the Royal Statistical Society: Series D (The Statistician) **47** (1998) 431. https://doi.org/10.1111/1467-9884.00142.

[18] D. C. Wheeler, "Geographically weighted regression", in *Handbook of Regional Science*, Springer, 2021, pp. 1895–1921. https://doi.org/10.1007/978-3-642-23416-3_143.

[19] A. Iriany, W. Ngabu & D. Ariyanto, "Rainfall modeling using the geographically weighted poisson regression method", BAREKENG: Jurnal Ilmu Matematika dan Terapan **18** (2024) 0627. https://doi.org/10.24123/barekeng.v18i1.914.

[20] M. Van Wezel & R. Potharst, "Improved customer choice predictions using ensemble methods", European Journal of Operational Research **181** (2007) 436. https://doi.org/10.1016/j.ejor.2006.07.043.

[21] L. Breiman, "Random forests", Mach Learn **45** (2001) 5. Available online: https://link.springer.com/article/10.1023/A:1010933404324.

[22] A. Sekulić, M. Kilibarda, G. Heuvelink, M. Nikolić & B. Bajat, "Random forest spatial interpolation", Remote Sensing **12** (2020) 1687. https://doi.org/10.3390/rs12101687.

[23] N. H. A. Malek, W. F. W. Yaacob, Y. B. Wah, S. A. Md Nasir, N. S. Shaadan & S. W. Indratno, "Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data", Indones. J. Elec. Eng. Comput. Sci **29** (2023) 598. https://doi.org/10.11591/ijeecs.v29.i3.598-608.

[24] A. Liaw & M. Wiener, "Classification and regression by randomForest", R News **2** (2002) 18. Available online: https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf.

[25] A. B. Shaik & S. Srinivasan, "A brief survey on random forest ensembles in classification model", in *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2*, Springer, 2019, pp. 253–260. https://doi.org/10.1007/978-981-15-0232-3_23.

[26] T. Hengl, M. Nussbaum, M. N. Wright, G. M. Heuvelink & B. Gräler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables", PeerJ **6** (2018) e5518. https://doi.org/10.7717/peerj.5518.

[27] F. Santos, V. Graw & S. Bonilla, "A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon", PloS one **14** (2019) e0226224. https://doi.org/10.1371/journal.pone.0226224.

[28] B. P. O. Lovatti, M. H. C. Nascimento, Á. C. Neto, E. R. Castro & P. R. Filgueiras, "Use of Random forest in the identification of important variables", Microchemical Journal **145** (2019) 1129. https://doi.org/10.1016/j.microc.2018.10.035.

[29] D. Denisko & M. M. Hoffman, "Classification and interaction in random forests", Proceedings of the National Academy of Sciences **115** (2018) 1690. https://doi.org/10.1073/pnas.1722310115.

[30] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang & X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees", Expert Systems with Applications **237** (2024) 121549. https://doi.org/10.1016/j.eswa.2023.121549.

[31] J. Hu & S. Szymczak, "A review on longitudinal data analysis with random forest", Briefings in bioinformatics **24** (2023) bbad002. https://doi.org/10.1093/bib/bbad002.

[32] Z. Chen, S. Zhang, W. Geng, Y. Ding & X. Jiang, "Use of geographically weighted regression (GWR) to reveal spatially varying relationships between Cd Accumulation and soil properties at field scale", Land **11** (2022) 635. https://doi.org/10.3390/land11050635.