BAREKENG: Journal of Mathematics and Its Applications
December 2024 Volume 18 Issue 4 Page 2577–2588
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss4pp2577-2588

GEOGRAPHICALLY WEIGHTED MACHINE LEARNING MODEL FOR ADDRESSING SPATIAL HETEROGENEITY OF PUBLIC HEALTH DEVELOPMENT INDEX IN JAVA ISLAND

Muhammad Azis Suprayogi[1*], Bagus Sartono[2], Khairil Anwar Notodiputro[3]

[1,2,3]*Department of Statistics, Faculty of Mathematics and Natural Science, IPB University*
Jalan Meranti, Kampus IPB Dramaga, Bogor, West Java, 16680, Indonesia

*Corresponding author's e-mail: * **azissuprayogi@apps.ipb.ac.id***

Article History:

Received: 16th, May 2024
Revised: 4th, July 2024
Accepted: 9th,August 2024
Published:14th, October 2024

Keywords:

Geographically Weighted;
GW-RF;
Random Forest;
Spatial.

ABSTRACT

Random Forest (RF) machine learning models have emerged as a prominent algorithm, addressing problems arising from the sole use of decision trees, such as overfitting and instability. However, conventional RF has global coverage that may need to capture spatial variations better. Based on the analysis of the level of public health development, the relationship between the level of health development and risk factors can vary spatially. We use a modified RF algorithm called Geographically Weighted Random Forest (GW-RF) to address this challenge. GW-RF, as a tree-based non-parametric machine learning model, can help explore and visualize relationships between the Public Health Development Index (PHDI) as response variables and factors that are indicators at the district level. GW-RF output is compared with global output, which is RF in 2018 using the percentage of the population with access to clean/decent water (X1), consumption of eggs and milk per capita per week (X2), number of healthcare facilities per 1000 people (X3), number of doctors per 1000 people (X4), pure participation rate ratio female/male (X5), percentage of households that have hand washing facilities with soap and water (X6) as independent variables. Our results show that the non-parametric GW-RF model shows high potential for explaining spatial heterogeneity and predicting PHDI versus a global model when including six major risk factors. However, some of these predictions mean little. Findings of spatial heterogeneity using GW-RF show the need to consider local factors in approaches to increasing PHDI values. Spatial analysis of PHDI provides valuable information for determining geographic targets for areas whose PHDI values need to be improved.

How to cite this article:

M. A. Suprayogi, B. Sartono, and K. A. Notodiputro., "GEOGRAPHICALLY WEIGHTED MACHINE LEARNING MODEL FOR ADDRESSING SPATIAL HETEROGENEITY OF PUBLIC HEALTH DEVELOPMENT INDEX IN JAVA ISLAND," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2577-2588, December, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Machine Learning (ML) models have high predictive capabilities from data mining and are often flexible and non-linear. However, it is often less than optimal in capturing geographic relationships, making it less sensitive to spatial context. This challenge is significant given that spatial data usually exhibit heterogeneity, leading to variations in the relationship between dependent and independent variables across regions. Conventional ML models need help dealing with such complexity because they produce a single output for the entire study area without considering the spatial variations that may exist. Research on handling spatial heterogeneity in population modeling based on geographic data is still limited [1]. Previous research on introducing a framework for modeling spatial data using the Random Forest (RF) algorithm with distance maps of spatial covariates as additional input showed improved performance compared to models that do not pay attention to spatial context [2]. However, their approach focuses more on spatial-temporal interpolation rather than providing insight into potential spatial heterogeneity.

One commonly used method is geographically weighted regression (GWR), an extension of conventional statistical regression methods that considers the spatial influence of various factors. Approaches considering spatial variation involve developing several local regression models simultaneously by incorporating spatial distance weights [3]. However, the main drawback is the susceptibility of the linear model to data that deviates from the general pattern and relies on strong assumptions about the linear relationship between the explanatory variables and the dependent variable, as well as the existence of multicollinearity between the explanatory variables. With the spread of machine learning technologies, approaches have emerged that combine spatially weighted structures with machine learning models [4][5].

Geographically weighted random forest (GW-RF) is a tree-based non-parametric ensemble model developed recently to overcome the limitations of GWR models and improve the predictive performance of non-geographically weighted random forest (RF) models. The basic concept of GW-RF is similar to the GWR model in that the model is calibrated locally rather than globally [4]. GW-RF draws inspiration from spatially varying coefficient models, wherein a global process is decomposed into multiple local sub-models, serving as both predictive and explanatory tools [6]. GW-RF does not need to consider multicollinearity; it can analyze all independent variables without filtering, provide better predictive power, and evaluate the relationship between spatial independent and dependent variables better than GWR [4].

Previous research using GW-RF has been done in a few cases. Research on addressing spatial heterogeneity in remote sensing and population modeling results showed that GW-RF can be more predictive when an appropriate spatial scale is selected to model the data so that it will reduce residual autocorrelation and lower Root Mean Squared Error and Mean Absolute Error values [7]. Research on analyzing the spatial variability of type 2 diabetes mellitus (T2D) prevalence in the USA by comparing GWR Ordinary Least Square and GW-RF results showed that the GW-RF model outperformed the GW-OLS model also the GW-RF model may be suitable for spatial analyses where multicollinearity across different geographical locations is a significant issue [4]. Another research on modeling spatial heterogeneity in traffic crash frequency and its determinants in the US compares the GW-RF, GWR, and global RF models. The results showed that the GW-RF model demonstrates better predictive accuracy compared to the global RF as GW-RF has lower MSE value than that of global RF model, also GW-RF model has improved in the overall performances compared to the GWR model with higher average R^2 value [8].

Health development entails a deliberate and sustainable endeavor to enhance a community's overall health status through diverse strategies, policies, and interventions. Its primary objective is to attain optimal health levels for the entire populace, encompassing enhanced access to healthcare services, disease prevention, health advocacy, and overall quality of life [9]. To realize the objective of sustainable health development, quantifiable health metrics are essential. Monitoring health through these metrics enables nations to evaluate the attainment of internationally set health objectives [10]. Consequently, Indonesia's Ministry of Health, through its Health Research and Development Agency (BALITBANGKES), has compiled the Public Health Development Index (PHDI).

The PHDI comprises a set of health indicators that can be easily and directly measured to describe health problems. The underlying principles guiding the selection of indicators for the PHDI emphasize simplicity, ease of measurement, utility, reliability, and timeliness. In the 2018 edition of the PHDI, 30 indicators are categorized into 7 groups, covering aspects such as child health, reproductive health, healthcare services, health-related behaviors, non-communicable diseases and associated risks, infectious diseases, and environmental health. The PHDI serves as a foundational tool for planning health development initiatives at

the district/city level and serves as a basis for determining the allocation of health funding from central to provincial levels, further down to district/city levels, and from provinces to regencies/cities [11].

On a global scale, Indonesia's health index, according to The Legatum Prosperity Index 2017, ranks 101st out of 149 countries. According to the Global Health Security Index (GHSI) report, Indonesia's global health security ranks 13th among G20 countries. In this ranking, the United States holds the top position with 75.9 points, while Indonesia scores 50.4 points out of 100 [12]. Given these statistics, the significance of the PHDI as a tool for gauging health development achievements in Indonesia becomes apparent. Enhancing the PHDI involves modeling it with several variables closely linked to public health. In this study, PHDI modeling will be conducted based on four independent variables: access to clean water, consumption of eggs and milk, the ratio of community health centers to population, and the ratio of doctors to population. This modeling aims to identify significant variables influencing PHDI improvement, enabling policymakers to target interventions effectively. Access to clean water is selected as a crucial independent variable due to its role in preventing infectious diseases, meeting nutritional requirements, and ensuring personal and environmental hygiene [13]. Consuming eggs and milk, which are rich in nutrients, is another independent variable under consideration. Research indicates that egg consumption can boost good cholesterol levels without adversely affecting lousy cholesterol levels in most individuals [14]. In contrast, dairy consumption may contribute to bone health and reduce the risk of osteoporosis [15]. Lastly, the ratios of health centers and doctors per population are chosen as variables reflecting the availability and distribution of healthcare facilities and professionals. Adequate and well-distributed healthcare infrastructure can enhance accessibility to essential healthcare services, ensuring prompt diagnosis and treatment [16].

Previous research has been conducted by using the Geographically Weighted Logistic Regression (GWLR) Model applied to PHDI data in East Java Province in 2018 [17], which concluded that the GWLR model with Adaptive Gaussian Kernel produced the smallest AIC than the Fixed Gaussian Kernel and the Adaptive Bisquare Kernel. The research by M. Fathurahman [18] using Geographically Weighted Multivariate Logistic Regression (GWMLR) model and Multivariate Logistic Regression (MLR) model applied to PHDI and Human Development Index in Kalimantan Island concluded that GWMLR has the lower AIC, AICC, and BIC values compared with the MLR. The other research [19] analyzed PHDI in Sumatra Island in 2018 with the Geographically Weighted Regression and Linear Regression, in which the Breusch-Pagan statistical test indicated spatial heterogeneity.

Based on this background, we are interested in studying the GW-RF model to address spatial heterogeneity in the PHDI case on Java Island. We aim to apply the GW-RF model to a dataset on PHDI of Java Island, train the model, and compare its performance with global RF implementations. Additionally, we explore the influence of geographic scale and unique GW-RF outcomes, such as spatial feature importance of independent variables, to show the impact of the level of importance of local variables.

2. RESEARCH METHODS

In this research, two models will be used, namely Random Forest (RF) and Geographically Weighted Random Forest (GW-RF). Data will be processed using R software.

2.1 Data

The research data used is 2018 PHDI data sourced from the Central Statistics Agency of Indonesia and the Provincial Level Health Office. Data was collected from 119 regencies/cities on the island of Java. According to the Ministry of Health data in 2018, the average PHDI was 0.6087. PHDI denoted as Y, is the dependent variable, while the other variables, denoted as X1, X2, X3, X4, X5, and X6, are independent variables, and variable xy is the coordinate. The details of the research variables are explained in **Table 1** below:

Table 1. Dependent Variables, Independent Variables Used in The Analysis.

Variables	Variables Name	Type
Y	Public Health Development Index (PHDI)	Ratio
X1	Percentage of households that have access to improved water	Ratio
X2	Expenditure on egg and milk consumption per capita per week (in thousands)	Ratio
X3	Percentage of public health centers per thousand population	Ratio
X4	Percentage of doctors per thousand population	Ratio
X5	Pure Participation Rate Ratio (APM) at elementary school/equivalent level	Ratio
X6	Percentage of households having hand washing facilities with soap and water	Ratio
xy	Coordinates of cities/regencies	Spatial Point

2.2 Exploratory Data Analysist

To investigate the global relationship between PHDI and the factors that influence it, we used histograms, boxplots, and distribution of factors on maps and calculated the Pearson correlation coefficient between PHDI and the six factors.

2.3 Random Forest (RF)

The "randomForest" package in the R Statistical Computing Environment was used for this purpose. RF is a collection of multiple Classification and Regression Trees (CARTs). These were devised to address significant drawbacks of employing a single CART, such as the risk of overfitting [20]. In the conventional RF approach, each decision tree is randomly generated by sampling approximately two-thirds of the training data with replacement. At the same time, the remaining one-third is held out of training (bagging of training data) [21][22]. Random Forest (RF) performs effectively with high-dimensional variables even when the sample size is small. The RF algorithm follows this process[23]:

1. The datasets $D_1, D_2, \ldots, D_n$ are created by repeatedly applying the bootstrap method to randomly sample from the entire dataset *D*, resulting in the generation of the corresponding n decision trees $H_1, H_2, \ldots, H_n$.
2. At each node of the decision tree, a random selection of *m* variables where $m < k$ is made from the total *k* variables available. The node is then split using the selected *m* variables, applying the optimal segmentation method based on a specific criterion.
3. The value of *m* remains constant as the forest expands. Each tree grows to its maximum size without pruning until further splitting is no longer possible.

In the first step of constructing the RF, whether with or without replacement, some data samples are not used to grow the tree, which is called the out-of-bag (OOB) for the tree. The accuracy of the RF model can be estimated from the OOB data as equation:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{\hat{y}}_i\right)^2 \quad (1)$$

where *N* is the number of samples from the OOB data, y_i is the actual value of the *i*th sample, and $\bar{\hat{y}}_i$ is the average prediction for the *i*th sample from all trees. Random forest can also be used to produce feature importance, which shows how much a predictor contributes to predicting the response [24].

2.4 Geographically Weighted Random Forest (GW-RF)

The geographic random forest RF concept remains a generalized and aspatial global model, which may need to be revised to resolve spatial heterogeneity. Therefore, we propose an extension of RF by decomposing the model into several local sub-models. This concept is partly similar to GWR [25], where the focus shifts from global to local computing. A local RF is calculated for each location but only using n nearest

observations. This results in an RF calculation of each training data point, including its respective performance, predictive ability, and feature importance. Thus, we increase RF flexibility with local rather than global calibration. To explain the difference between the two approaches, we can use a simple regression equation:

$$Y_i = ax_i + e, i = 1, \ldots, n \quad (2)$$

where Y_i is the value of the dependent variable for the *i*th observation, ax_i is the prediction of non-linear of RF based on the independent variables of x, and e is error term. The equation is formed with all data without considering their spatial distribution. In the GRF approach, the above equation is expanded to:

$$Y_i = a(u_i, v_i)x_i + e, i = 1, \ldots, n \quad (3)$$

where $a(u_i, v_i)x$ is the prediction of the RF model calibrated at location i, and (u_i, v_i) are its coordinates. Sub-models are built for each data location, considering only nearby observations. The area in which the sub-model operates is called the neighborhood (or kernel), and the maximum distance between a data point and its kernel is called the bandwidth **[7][26]**. In this study, we use an adaptive kernel. Adaptive kernels are beneficial when sampling densities vary across space, as the size of census units can vary greatly. We combine global and local estimates using the weight parameter (a). Combining predictions allows us to extract local heterogeneous signals (low bias) from local sub-models and combine them into a global model that uses more data. The weight parameters are user-definable, and for the scope of this research, we experimented with five settings i) a = 0, which means the weight for the local model is zero and 100% using the global model, ii) a = 1, which means the weight for the local model is 100%, iii) a = 0.5, which means the same weight for the local model and the global model, iv) a = 0.25, which means a smaller weight for the local model, namely 25%, while the weight for the global model is 75%, and v) a = 0.75, which means a greater weight for the local model, namely 75%, while the weight for the global model is 25%. Implementation of GW-RF and RF analysis was carried out using the newly developed R package 'SpatialML' **[21][27][28]**.

2.5 Predictive Performance

We first evaluated the predictive performance of GW-RF using K-fold cross-validation. Cross-validation statistics usually better indicate how a model will perform on unseen data. In K-fold cross-validation, the data set was randomly divided into a test and training set k different times, and model evolution was repeated k times. Each time, one of the k subsets was used as the test set, and the other k-1 subsets are put together to form a training set. Then the average error across all k trials was computed. We employ two established and robust error measurement metrics to evaluate the accuracy of the models, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), as in **Equation (4)** and **Equation (5) [29]**:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (4)$$

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n} \quad (5)$$

where y_i is the observed variable, $\hat{y}_i$ is the predicted value, and n is the sample size.

3. RESULTS AND DISCUSSION

3.1 Exploratory Data Analysist

As for our target dataset, the total number of data instances was 119 observations from regencies/cities on the island of Java in 2018. This dataset size is small, so we may not have enough cases for accurate prediction, but the random forest has been shown to handle challenges arising from small sample sizes. **Figure 1** (a) shows a class comparison of the response variable, namely the public health development index (PHDI). In contrast, **Figure 1**(b) shows the descriptive statistics of the explanatory variables considered in this study.

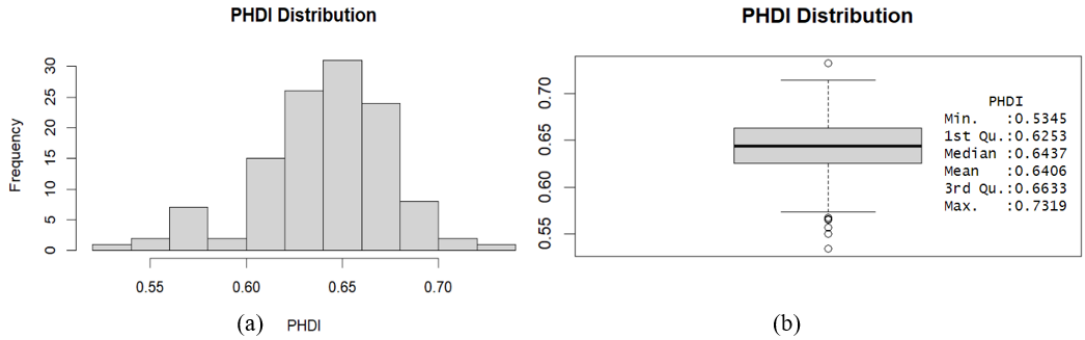


Figure 1. (a) Histogram of PHDI distribution, (b) Boxplot of PHDI distribution.

The PHDI distribution in Figure 1(a) looks bell-shaped or normally distributed. The lower the PHDI value, the smaller the frequency, and vice versa. The higher the PHDI value, the smaller the frequency, which means that only a small portion of areas have high PHDI values and low PHDI values. Most of the data has an average PHDI value of around 0.65. This is supported by Figure 1(b) where the median value of 0.6437 is in the middle of the box plot, and the whiskers on both sides of the box are almost the same, indicating that the data is normally distributed. There are several outliers, but they do not affect the data distribution much, as noted in the mean value, which is not too different from the median value.

The values of predictor variables on response variables can be seen as in Figure 2.

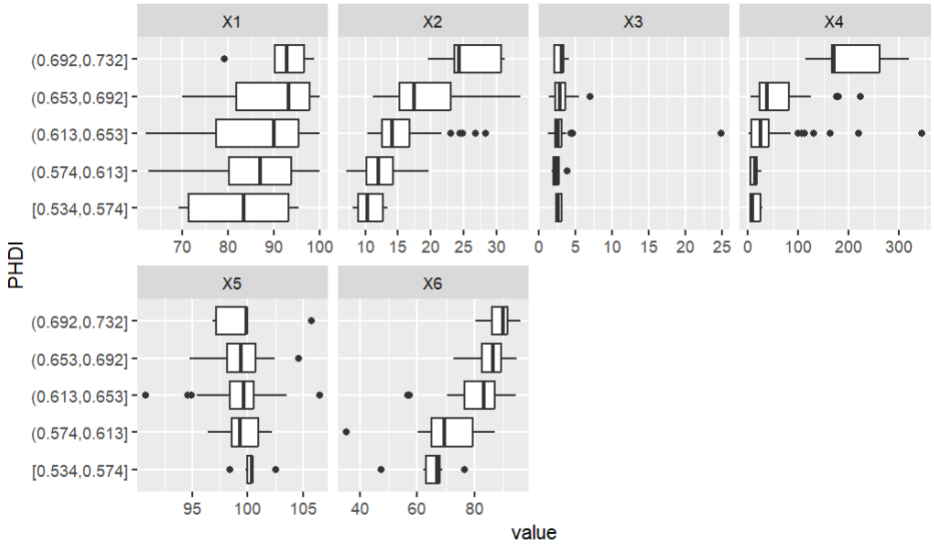


Figure 2. Values of variables based on PHDI

The Percentage of households with access to adequate water (X1) has a higher median at a high level of PHDI. This means that the more households have access to improved water, the higher the PHDI level. This also happens in the variable Expenditure on egg and milk consumption per capita per week (in thousands) (X2); the median boxplot of X2, which is at a high PHDI level, is greater than the median boxplot, which is at a low PHDI level. This means that the greater the egg and milk consumption, the higher the PHDI level. As for the variable number of community health centers per thousand population (X3), the box plot is not much different for both those with high and low PHDI, both of which still have low values with only a few outlier values. The variable number of doctors per thousand population (X4) has a low value both at the high PHDI level and at the low PHDI level; it's just that the number of doctors per population is still more significant at the high PHDI level, while at the low PHDI level, the number of doctors is very small, which means that the more doctors per population, the higher the PHDI level. The median box plot of Pure Participation Rate Ratio (APM) at elementary school/equivalent level (X5) is not much different for both high and low PHDI; even the box plot shows a negative correlation between the variable Pure Participation Rate Ratio (APM) at elementary school/equivalent level and PHDI. The Percentage of households having hand washing facilities with soap and water (X6) has a higher median at a high level of PHDI, which means that the more households have hand washing facilities, the higher the PHDI level. Outliers exist in all variables, with the most outliers being in X4.

Figure 3. Global correlation coefficients

This research uses global correlation coefficients to find out how significant the correlation between PHDI and the six main factors in general in numbers. **Figure 3** shows that the percentage of households having hand washing facilities with soap and water (X6), Expenditure on egg and milk consumption per capita per week in thousands (X2), and number of doctors per thousand population (X4) are the three variables that have the most significant correlation values among others, followed by the percentage of households having access to adequate water (X1), expenditure on egg and milk consumption per capita per week in thousands (X2), and pure Participation Rate Ratio (APM) at elementary school/equivalent level (X5).

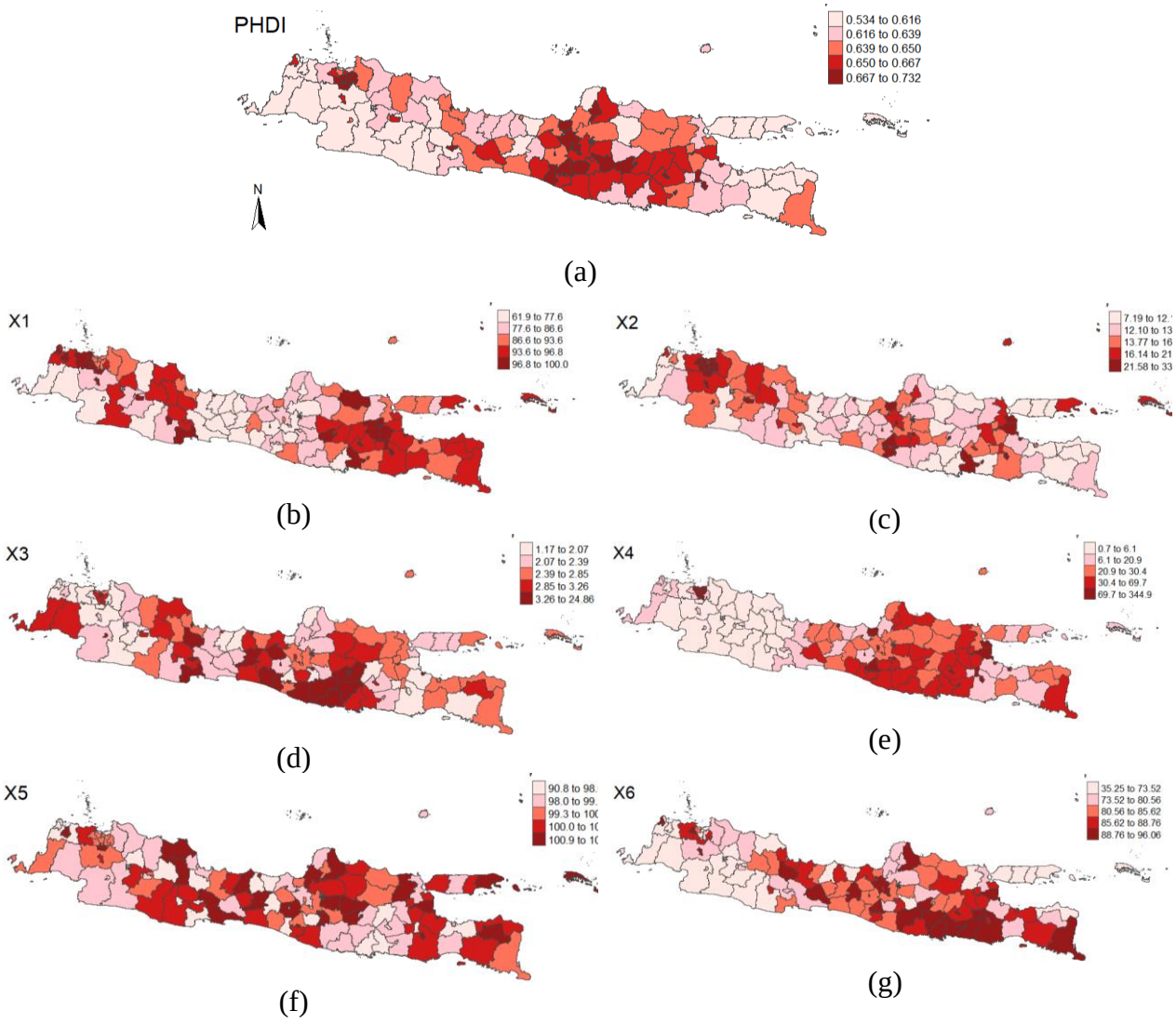


Figure 4. Distribution map of PHDI and six independent variables: (a) PHDI, (b) X1, (c) X2, (d) X3, (e) X4, (f) X5, (g) X6

Figure 4 shows the PHDI distribution map and the factors that influence it. The small PHDI index is mainly in the western part of the island of Java, namely the provinces of Banten and West Java, some areas in Central Java Province, and East Java Province. This may be due to the lowest value percentage of households having hand washing facilities with soap and water (X6) and the number of doctors per thousand population (X4) spread across most of the district or cities on western Java Island, as the X4 and X6 have the most significant correlation with the PHDI. The smallest percentage of households with access to improved water is similar to the largest PHDI in Banten Province and West Java Province. The smallest expenditure on egg and milk consumption per capita per week is found in Central Java and East Java Provinces. The percentage of public health centers per thousand population and the Pure Participation Rate Ratio (APM) at elementary school/equivalent level are spread throughout all Provinces. In comparison, the Percentage of doctors per thousand population and the Percentage of households having hand washing facilities with soap and water are low in the Province West Java.

3.2 Predictive Performances

The methodological workflow is showed in flowchart of **Figure 5**.

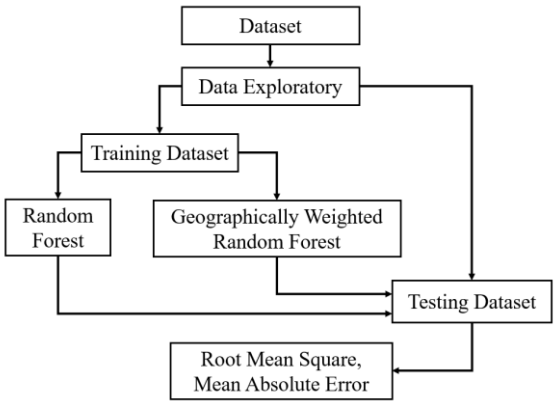


Figure 5. Flowchart of methodological framework

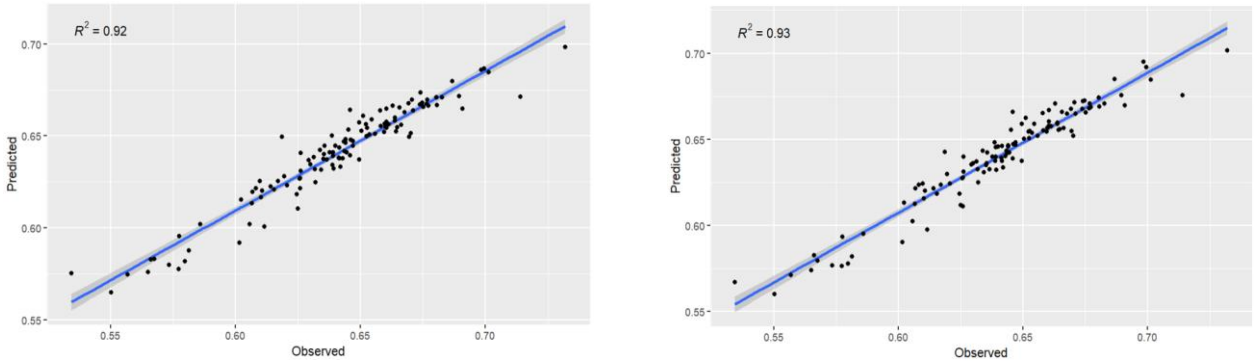
The dataset used is 119 examples of observation results from cities and districts. The first thing we do is conduct exploratory data by making a box plot to determine the relationship between the PHDI variable and the six factors and identify which influences PHDI most. We also created a distribution of the PHDI response variable and six factors to determine which regions on the island of Java have PHDI values from the highest to the lowest values and identify possible causes. All data will be used as training data, as each belongs to a city or regency. Next, we used random forest and geographically weighted random forest methods to model the training data. The final step determines the best performance by calculating each model's RMSE and MAE values. The comparison between RF and GW-RF is carried out based on the values of RMSE and MAE.

Table 2. Comparison of Models Based on RMSE and MAE

Models	Weight of local model (a)	Bandwidth	RMSE	MAE
RF	-	-	0.0201	0.0162
GW-RF	0	27	0.0173	0.0137
GW-RF	1	27	0.0168	0.0131
GW-RF	0.5	27	0.0171	0.0134
GW-RF	0.25	27	0.0172	0.0135
GW-RF	0.75	27	0.0170	0.0132

The comparison of the RF and GW-RF models is shown in **Table 2**. The model with a smaller RMSE and MAE value indicates a better model. **Table 2** shows that giving significant weight to local predictions (a = 1) in all five GW-RF modeling designs can produce better predictions based on the lowest RMSE and MAE. In particular, using an optimal bandwidth of 27 systematically shows the lowest RMSE and MAE

values across all five approaches. The weight value of a = 1, which means 100% weight for the local model, is the most optimal choice for minimizing RMSE and MAE and consistently predicts better than global R.



(a) **(b)**

Figure 6. Plots of observed PHDI versus predicted PHDI (a) RF, (b) GW-RF

Figure 6 shows a 1:1 plot comparing the observed PHDI with the predicted PHDI using the RF and GW-RF models. The plot shows consistent improvement in accounting for PHDI (R^2) variability when moving from global RF to GW-RF models. The GW-RF model makes a small contribution to more variability in the PHDI (R^2 = 0.93) than in the RF model (R^2 = 0.92).

3.3 Feature Importance

The global feature importance from RF model is shown in **Figure 7**.

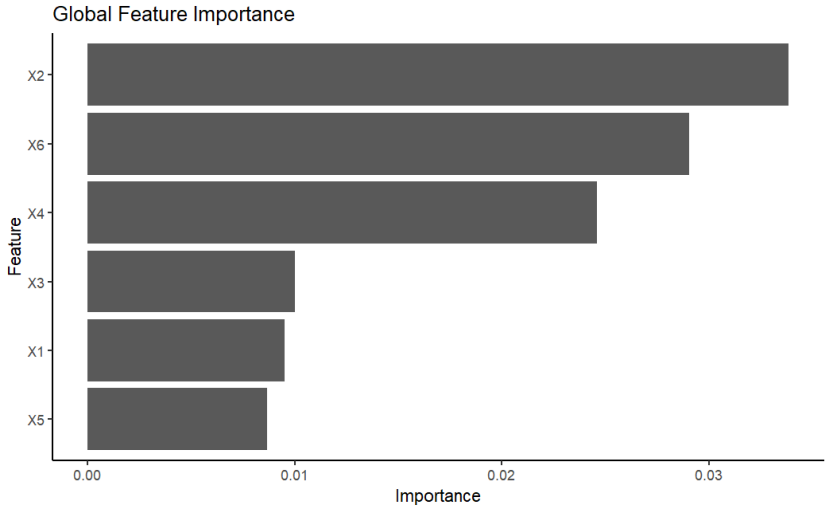
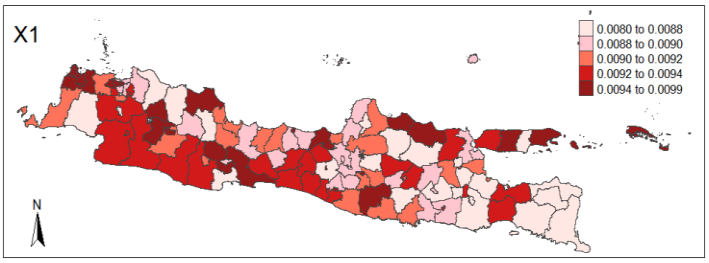


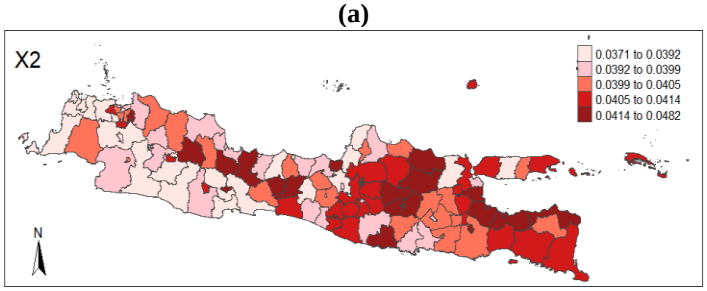
Figure 7. Global feature importance

The global feature importance from the RF model places the expenditure on egg and milk consumption per capita per week as the most important variable, followed by the percentage of households having hand washing facilities with soap and water, percentage of doctors per thousand population, percentage of public health centers per thousand population, percentage of households that have access to improved water, and pure participation rate ratio at elementary school/equivalent level.

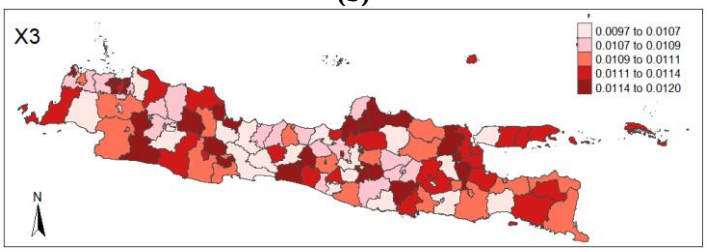
GW-RF can be utilized solely as an exploratory tool. Since GW-RF is a local decomposition of RF, its results can be mapped. Using the whole dataset without training/testing splits for visualization, the spatial variation of feature importance of each independent variable can be shown in **Figure 8**.



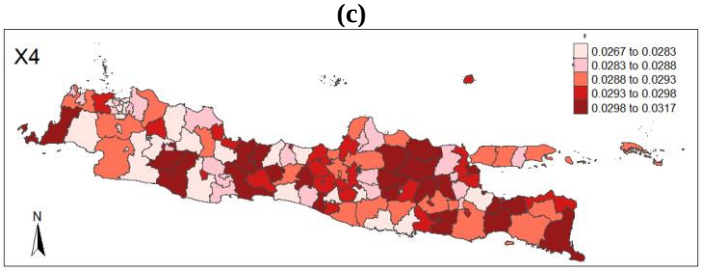
(a)



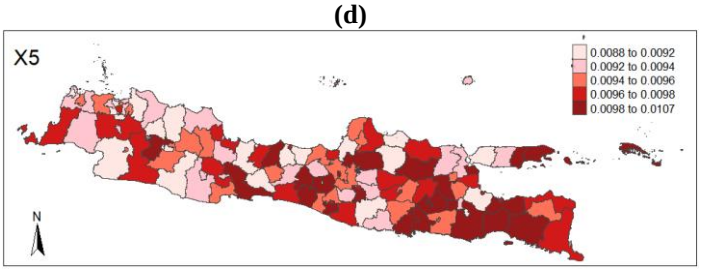
(b)



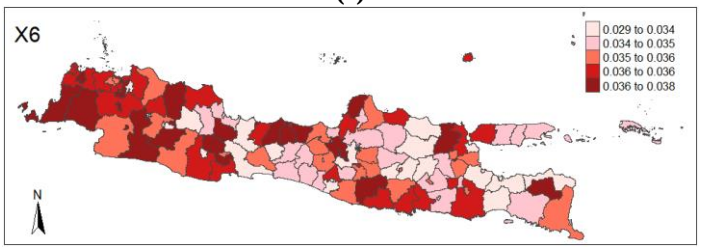
(c)



(d)



(e)



(f)

Figure 8. Plots of local feature importance of independent variables. (a) Percentage of the population with access to clean/decent water, (b) Consumption of eggs and milk per capita per week, (c) Number of healthcare facilities per 1000 people, (d) Number of doctors per 1000 people, (e) Pure participation rate ratio female/male, (f) Percentage of households that have hand washing facilities with soap and water.

The influence of the level of importance of local variables on PHDI varies quite widely in each location. However, some variables appear to be concentrated in certain areas. The high consumption of eggs and milk per capita per week is focused on the eastern part of Java Island. The high value of the percentage

of households that have hand washing facilities with soap and water is seen concentrated in the western part of Java Island. The high variable of the number of doctors per 1000 people tends to congregate in the eastern part of Java Island. This shows the existence of spatial heterogeneity, where the geographic location of a region influences the variables that play a role in PHDI. As for the remaining variables, namely the percentage of the population with access to clean/decent water, consumption of eggs and milk per capita per week, and number of healthcare facilities per 1000 people, they appear to vary widely in various regions so that they do not reflect spatial heterogeneity.

3.4 Discussion

Empirical results from the application of GW-RF show good performance as a prediction and exploration tool. GW-RF outperforms RF with more accurate predictions based on lower RMSE and MAE values. Then, using geographic coordinates as features seems to be a good practice when using machine learning algorithms with spatial data. This confirms previous research **[2]**. Since RF is a decision tree (DT) algorithm, the explicit use of spatial features such as coordinates can support a degree of spatial interaction in tree development and, at least in part, overcome spatial non-stationarity. Although RF is a very flexible and non-linear algorithm, it is a model that does not account for spatial heterogeneity. Additionally, identifying and addressing spatial heterogeneity in machine learning models can be challenging because they are not based on strict parametric distributions like generalized linear models. Instead, GW-RF can illustrate these spatial effects very practically along with other information, such as the local performance of independent variables as feature importance. However, GW-RF has important limitations, such as increasing the complexity on the computational side.

Based on the application of GW-RF in the PHDI case, the optimal performance of GW-RF was obtained at a ratio of local model weights to global model weights of 1:0, which means using local model components without global model components. However, the results obtained depend on the available data set. The results may be different for different data sets. For example, in other cases with varying degrees of spatial heterogeneity, adjusting the weights of global model components to local model components in a given comparison may provide more robust estimates.

4. CONCLUSIONS

This research is the first to apply the GW-RF regression model compared to the global RF model to explore the spatial heterogeneity of PHDI on the Indonesian island of Java concerning various influencing factors. It has been proven that the performance of GW-RF, which includes coordinates as the geographic scale, is better than that of the RF model. Thus, the geographic scale has the effect of improving the prediction performance of the model. The optimal performance of GW-RF depends on selecting the appropriate weight parameters when combining local and global estimates as GW-RF output, and the result can vary depending on the dataset. GW-RF models with higher local model weights perform better than those with low local model weights. The unique GW-RF outcome is the influence of the level of importance of the local variables (features) on PHDI, which varies quite widely in each location of the city/regency. GW-RF may be applicable in spatial models where multicollinearity across geographic locations is a significant concern. Understanding the spatial heterogeneity of the relationship between PHDI and its influencing factors allows further research and development of fundamental and spatially varying PHDI improvement policies on Java Island.

REFERENCES

[1] C. F. Cockx K, *Incorporating spatial non-stationarity to improve dasymetric mapping of popu lation*. Appl Geogr, 2015.
[2] G. B. Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, *Random forest as a generic frame work for predictive modeling of spatial and spatio-temporal variables*. PeerJ, 2018.
[3] B. Liu, J., Khattak, A.J., Wali, "Do safety performance functions used for predicting crash frequency vary across space? Applying geographically weighted regressions to account for spatial heterogeneity," *Accid. Anal*, vol. Prev. 109, pp. 132–142, 2017, doi: http://dx.doi.org/10.1016/j.aap.2017.10.012.
[4] S. Quiñones, A. Goyal, and Z. U. Ahmed, "Geographically weighted machine learning model for untangling spatial

heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021, doi: 10.1038/s41598-021-85381-5.
[5] S. Santos, F., Graw, V., Bonilla, "A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon," *PLoS One*, vol. 14, no. 12, p. e0226224, 2019.
[6] K. W. Fotheringham AS, Yang W, *Multiscale geographically weighted regression (MGWR)*. Ann Am Assoc Geogr., 2017.
[7] S. Georganos and S. Kalogirou, "A Forest of Forests: A Spatially Weighted and Computationally Efficient Formulation of Geographical Random Forests," *ISPRS Int. J. Geo-Information*, vol. 11, no. 9, 2022, doi: 10.3390/ijgi11090471.
[8] S. Wang, K. Gao, L. Zhang, B. Yu, and S. M. Easa, "Geographically weighted machine learning for modeling spatial heterogeneity in traffic crash frequency and determinants in US," *Accid. Anal. Prev.*, vol. 199, no. March, 2024, doi: 10.1016/j.aap.2024.107528.
[9] R. Labonte and G. Laverack, *Capacity Building in Health Promotion, Part I: For Whom? And for What Purpose?* vol. 11: Crit. Public Health, 2001.
[10] W. H. O. (WHO), *World Health Statistics 2021: Monitoring Health for SDGs*. 2021.
[11] D. H. Tjandrarini and dkk, *Indeks Pembangunan Kesehatan Masyarakat 2018*. Jakarta: Lembaga Penerbit Badan Penelitian dan Pengembangan Kesehatan, 2019.
[12] A. M. H. Putri, "Perhatian! Indeks Ketahanan Kesehatan RI Masih Jauh di Bawah," *CNBC Indonesia, 2023*, 2023. https://www.cnbcindonesia.com/research/20230315085601-128-421755/perhatian-indeks-ketahanan-kesehatan-ri-masih-jauh-di-bawah (accessed Nov. 28, 2023).
[13] World Health Organization (WHO), *Water, Sanitation, Hygiene, and Health*. 2022.
[14] J. Y. Shin, P. Xun, Y. Nakamura, and K. He, "Egg consumption in relation to risk of cardiovascular disease and diabetes: a systematic review and meta-analysis.," *Am. J. Clin. Nutr.*, vol. 98, no. 1, pp. 146–159, 2013, doi: 10.3945/ajcn.112.051318.
[15] D. Feskanich, W. C. Willett, and G. A. Colditz, "Calcium, vitamin D, milk consumption, and hip fractures: a prospective study among postmenopausal women," *Am. J. Clin. Nutr.*, vol. 77, no. 2, pp. 504–511, 2003, doi: 10.1093/ajcn/77.2.504.
[16] World Health Organization (WHO), "Everybody's business: Strengthening health systems to improve health outcomes: WHO's framework for action.," *World Health Organization (WHO)*, 2016. https://www.who.int/healthsystems/strategy/everybodys_business.pdf (accessed Nov. 28, 2023).
[17] Q. S. Wardhani, S. S. Handajani, and I. Susanto, "Masyarakat Jawa Timur dengan metode," *J. Apl. Stat. dan Komputasi*, vol. 14, no. 2, pp. 1–12, 2022, [Online]. Available: https://doi.org/10.34123/jurnalasks.v14i2.333
[18] M. Fathurahman, Purhadi, Sutikno, and V. Ratnasari, "Geographically Weighted Multivariate Logistic Regression Model and Its Application," *Abstr. Appl. Anal.*, vol. 2020, 2020, doi: 10.1155/2020/8353481.
[19] U. K. Krismayanto and E. Pasaribu, "Analisis Regresi Spasial Indeks Pembangunan Kesehatan Masyarakat dan Paradoks Simpson Kabupaten/Kota di Pulau Sumatera Tahun 2018," *Semin. Nas. Off. Stat.*, vol. 2022, no. 1, pp. 1037–1052, 2022, doi: 10.34123/semnasoffstat.v2022i1.1330.
[20] Breiman L., *Random forests*. Machine Learn, 2001.
[21] S. Georganos *et al.*, "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling," *Geocarto Int.*, vol. 36, no. 2, pp. 121–136, 2021, doi: 10.1080/10106049.2019.1595177.
[22] S. N. Khan, D. Li, and M. Maimaitijiang, "A Geographically Weighted Random Forest Approach to Predict Corn Yield in the US Corn Belt," *Remote Sens.*, vol. 14, no. 12, pp. 1–21, 2022, doi: 10.3390/rs14122843.
[23] Y. Luo, J. Yan, and S. McClure, "Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis," *Environ. Sci. Pollut. Res.*, vol. 28, no. 6, pp. 6587–6599, 2021, doi: 10.1007/s11356-020-10962-2.
[24] H. Ilma, K. A. Notodiputro, and B. Sartono, "Association Rules in Random Forest for the Most Interpretable Model," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 1, pp. 0185–0196, 2023, doi: 10.30598/barekengvol17iss1pp0185-0196.
[25] C. M. Fotheringham AS, Brunsdon C, *Geographically weighted regression: the analysis of spa tially varying relationships*. Hoboken, NJ: John Wiley & Sons., 2003.
[26] C. M. Brunsdon C, Fotheringham S, *Geographically weighted regression*. J Royal Stat Soc D, 1998.
[27] K. S. Georganos S, Abdi AM, Tenenbaum DE, "Examining the NDVI-rainfall relationship in the semi-arid Sahel using geographically weighted regression," *J Arid Env.*, no. 146, pp. 64–74, 2017.
[28] G. S. Kalogirou S, *SpatialML*. R Foundation for Statistical Computing., 2018.
[29] D. Wu, Y. Zhang, and Q. Xiang, "Geographically weighted random forests for macro-level crash frequency prediction," *Accid. Anal. Prev.*, vol. 194, no. November 2023, p. 107370, 2024, doi: 10.1016/j.aap.2023.107370.