

Geographically Weighted Machine Learning Model for Untangling Spatial Heterogeneity of Dengue Incidence in West Java

Auralia Putri Astutiningsih¹, Gangga Anuraga^{2*}, Muhammad Athoillah³, Hani Brilianti Rochmanto⁴

^{1,2,3, 4}*Department of Statistics, Faculty of Engineering and Science, Universitas PGRI Adi Buana Surabaya, Indonesia*

*corr-author: g.anuraga@unipasby.ac.id

Abstract - Dengue hemorrhagic fever continues to pose a significant public health challenge, particularly in West Java Province, Indonesia, which consistently reports the highest incidence rates in the country. This study examined the factors influencing dengue fever incidence using Random Forest Regression (RFR) and Geographically Weighted Random Forest (GWRF) methodologies. Utilizing secondary data from 2022 to 2024 across 27 districts/cities, the data from 2022 to 2023 served as training data, while the 2024 data were used for testing. The findings revealed that the optimal RFR model, with $n_{tree} = 1000$ and $m_{try} = 1$, achieved an RMSE of 1796.409, a MAPE of 0.482, and an R^2 of 0.685. Conversely, the GWRF model, which employed an adaptive kernel and an optimal bandwidth of 45 nearest neighbors, exhibited superior performance, with an RMSE of 1,756.713, MAPE of 0.466, and R^2 of 0.700. This enhancement in the model performance suggests that spatial weighting improves the model's capacity to capture spatial heterogeneity. In addition, variations in local feature importance indicate spatial non-stationarity across regions. These results imply that the GWRF is more effective in modeling dengue fever outbreaks and can inform the development of region-specific public health interventions.

Keywords: Dengue; random forest regression; geographically weighted random forest; spatial analysis; West Java.

I. INTRODUCTION

Dengue virus infection is one of the most common arboviral diseases in tropical and subtropical regions and remains an ongoing global public health issue. This disease is caused by the dengue virus (DENV), which is transmitted through the bite of an infected *Aedes aegypti* or *Aedes albopictus* mosquito [1-3]. Transmission via these mosquito bites is the primary cause of Dengue Hemorrhagic Fever, with an incubation period of approximately 4-10 days. In more severe cases, the infection can lead to plasma leakage and organ damage [4, 5]. The dengue virus is a single-stranded RNA virus

of the genus *Flavivirus*, comprising four serotypes (DEN-1, DEN-2, DEN-3, and DEN-4), which can cause a wide spectrum of diseases. Infection by a single serotype generally confers lifelong immunity against that serotype but does not protect against other serotypes, leaving individuals susceptible to reinfection, which can worsen clinical outcomes [6]. According to the World Health Organization (WHO), the global number of dengue cases continues to rise annually, with an estimated 50 million cases and approximately 2.5 billion people worldwide living in endemic areas [7].

Southeast Asia accounts for more than half of the global dengue burden. As a tropical country, Indonesia is an endemic region for dengue, with an upward trend in cases over the past few decades, particularly in 2024, when 91.9 cases per 100,000 population were reported [8, 9]. The disease is widespread across nearly all provinces, driven by warm and humid climatic conditions that are highly conducive to the breeding of *Aedes* mosquito vectors [10]. Data indicate that all four dengue virus serotypes have been circulating simultaneously in various regions of Indonesia, leading to hyperendemic conditions and increasing the risk of recurrent infections with greater severity [11]. This is reflected in the high number of dengue cases and deaths in recent years. The disease has spread to nearly all provinces in Indonesia, particularly West Java Province [12]. Dengue Fever cases in Indonesia reached 203,921, with 1,210 deaths across 36 provinces. West Java Province has consistently recorded the highest number of cases, with 36,608 in 2022, 19,328 in 2023, and 61,423 in 2024. This pattern indicates that dengue remains a serious public health issue with a persistently high and fluctuating risk.

Several spatial modeling studies have highlighted the importance of accounting for inter-regional heterogeneity in health case analyses, including Dengue Hemorrhagic Fever (DHF). Previous research using Geographically Weighted Negative Binomial Regression

(GWNBR) has shown that environmental factors and health services significantly influence the incidence of DHF [13]. However, conventional spatial regression models, such as Geographically Weighted Regression, have limitations in capturing non-linear relationships and are prone to multicollinearity issues. To address these limitations, machine learning approaches have been increasingly developed. One such method is the Geographically Weighted Random Forest (GWRF). This non-parametric ensemble model combines GWR and Random Forests and has been shown to perform better than global models, as evidenced by lower error and higher R^2 values [14,15]. The GWRF method can accommodate spatial heterogeneity while capturing non-linear relationships without requiring prior variable selection or tests for multicollinearity [16, 17].

However, research on the application of GWRF to explore the non-stationarity of the relationship between risk factors and Dengue Hemorrhagic Fever (DHF) cases, particularly in West Java Province, remains limited. Therefore, this study aims to analyze the local relationship between risk factors and DHF cases and to evaluate the performance of GWRF compared to Random Forest. This study contributes by applying the GWRF model to capture spatial non-stationarity in dengue cases, representing one of the earliest applications in this context; by comparing its performance with global Random Forest models; and by

providing spatial insights to support targeted public health interventions.

II. METHOD

A. Data Source

The present study utilizes secondary data on Dengue Hemorrhagic Fever (DHF) in West Java Province, sourced from BPS-Statistics Indonesia Jawa Barat Province (Badan Pusat Statistik Jawa Barat Indonesia—BPS Jawa Barat) and the West Java Provincial Health Office (Dinas Kesehatan Provinsi Jawa Barat—Dinkes Jawa Barat). The dataset encompasses 18 districts and nine cities within the West Java Province for 2024. The dependent variable in this study is the number of Dengue Hemorrhagic Fever (DHF) cases in West Java Province, while the independent variables comprise seven factors. These variables were selected based on data availability from official sources and their relevance to dengue risk factors as identified in previous studies. Specifically, the selected variables represent key aspects of health facilities, environmental conditions, and socio-economic factors that are theoretically and empirically associated with dengue incidence, as presented in Table I.

B. Analysis Procedure

The analysis procedure in this study comprises several stages, from data preparation to model evaluation. The overall workflow of this study is illustrated in Fig. 1.

TABLE I
RESEARCH VARIABLES, MEASUREMENT SCALES, AND DATA SOURCES

Variable Indicator	Variable	Data Scale	Data Source
Number of Dengue Hemorrhagic Fever Cases	Y	Ratio	The West Java Provincial Health Office
Number of Community Health Centers (<i>Puskesmas</i>)	X_1	Ratio	The West Java Provincial Health Office
Number of General Hospitals	X_2	Ratio	BPS-Statistics Indonesia Jawa Barat Province
Clean and Healthy Living Behavior	X_3	Ratio	The West Java Provincial Health Office
Percentage of Households with an Adequate Sanitation	X_4	Ratio	BPS-Statistics Indonesia Jawa Barat Province
Average Years of Schooling	X_5	Ratio	BPS-Statistics Indonesia Jawa Barat Province
Expected Years of Schooling	X_6	Ratio	BPS-Statistics Indonesia Jawa Barat Province
Population Density	X_7	Ratio	BPS-Statistics Indonesia Jawa Barat Province

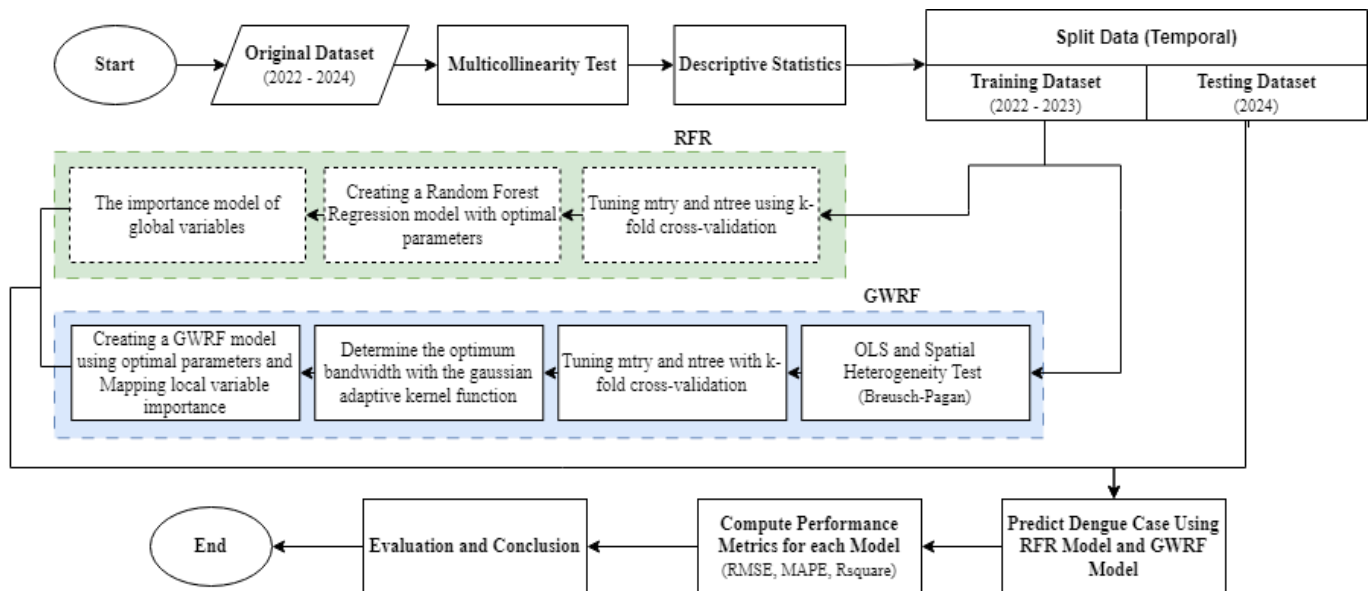


Fig. 1 Illustrates the research process for predicting dengue cases using Random Forest Regression and Geographically Weighted Random Forest

The research procedure in this study was conducted in several stages, as illustrated in Fig. 1. The analysis began with data preparation and preliminary statistical analysis, including descriptive statistics, multicollinearity testing, and heteroscedasticity testing. Next, the main parameters of Random Forest Regression (RFR) [18, 19], namely the number of variables selected randomly at each split (*mtry*) and the number of trees (*ntree*), were tuned to identify the optimal parameter combination [20]. After obtaining the optimal values, bootstrap sampling was performed by randomly selecting samples D_i from the original dataset D with replacement. These bootstrap samples were used to construct decision trees, and the process was repeated the specified number of times, resulting in an RFR ensemble model capable of capturing complex nonlinear relationships between dependent and independent variables.

The optimized RFR configuration was then used as the basis for the Geographically Weighted Random Forest (GWRF) analysis [21]. At this stage, the spatial weights for each location were determined using a kernel function, and the optimal bandwidth controlling the size of the local environment was selected. For each region i , the neighboring observations within the bandwidth were selected. Using this local dataset and the previously obtained RFR parameters, a local RFR model was constructed for each region. Region-specific variable importance values were calculated for each local model. This process was repeated for all regions to obtain spatially varying variable importance, enabling the

analysis of how each independent variable influences different locations.

Predictions were generated using both the RFR and GWRF models, and model performance was evaluated using RMSE, MAPE, and R^2 . The results were then compared to identify the model with the best predictive performance and to support the further interpretation of the findings.

C. Multicollinearity Test

Multicollinearity occurs when independent variables are linearly related, which can increase variances and complicate parameter estimation [22]. It can be detected using the Variance Inflation Factor (VIF), with values below 10 indicating no multicollinearity. The VIF is calculated as shown in (1).

$$VIF = \frac{1}{1-R^2} \quad (1)$$

Where R^2 is the coefficient of determination between the dependent and independent variables.

D. Spatial Heterogeneity

Spatial heterogeneity refers to variations in data characteristics across locations, often indicated by non-constant error variance (heteroscedasticity) in the global model. It can be tested using the Breusch-Pagan statistic, which evaluates the constancy of residual variance [23]. The test statistic is calculated as shown in (2).

$$BP = \frac{1}{2} f^T Z (Z^T Z)^{-1} Z^T f \quad (2)$$

where the elements of the vector f are defined as $f_i = \left(\frac{e_i^2}{\sigma^2} - 1\right)$. Furthermore, BP denotes the Breusch-Pagan test statistic, and e_i represents the residual for the i -th observation. Z represents the standardized matrix of X with dimension $n \times (p + 1)$, while σ^2 is the variance of the residuals.

E. Random Forest Regression

Random Forest Regression (RFR) was applied in this study as a global model to estimate the relationship between the dependent variable Y and the set of independent variables ($X_1, X_2, \dots, X_n$). The model was constructed using a bootstrap aggregating (bagging) approach, in which multiple decision trees were trained on different bootstrap samples of the training data. In this process, approximately 36.8% of the observations were excluded from each bootstrap sample and treated as Out-Of-Bag (OOB) data [21]. These OOB samples were used to evaluate model performance internally using the mean squared error (MSE), as shown in (3).

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (3)$$

where N is the number of observations in the data, y_i is the actual value of the dependent variable for the i -th county, and $\hat{y}_i$ is the prediction value generated by all decision trees for the i -th observation in the Random Forest Regression model. As a nonparametric method, Random Forest Regression does not require assumptions about the functional form of the relationship between variables and can capture complex nonlinear patterns [24]. These characteristics make it robust to noise, more stable than a single decision tree, and less prone to overfitting.

F. Geographically Weighted Random Forest

The Geographically Weighted Random Forest (GWRF) model was applied in this study as a local extension of the global Random Forest Regression (RFR) model by incorporating spatial information [21]. For each observation location i , a local RFR model was constructed using neighboring observations within a specified bandwidth. The optimal bandwidth was determined based on the highest OOB performance. An adaptive kernel approach was used, where the number of nearest neighbors varies across locations. Observations closer to the target location were assigned higher weights, while more distant observations received lower weights. The difference between the global and local models is expressed in (4) and (5).

$$Y_i = \alpha x_i + e, \quad i = 1, 2, \dots, n \quad (4)$$

$$Y_i = \alpha_i(u_i, v_i)x_i + e, \quad i = 1, 2, \dots, n \quad (5)$$

where Y_i is the dependent variable value for observation i and αx_i is the non-linear prediction of RFR based on a set of independent variables, with e as the error, and $\alpha_i(u_i, v_i)$ is the prediction of the RFR model calibrated at location i , and (u_i, v_i) are the coordinates of that location. The distance between locations i and j was calculated using the Euclidean distance (6).

$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \quad (6)$$

where d_{ij} is defined as the Euclidean distance between locations i and j . Spatial weights were then assigned using an adaptive Gaussian kernel, which gives higher weights to closer observations and lower weights to more distant ones. The weighting function is defined as (7).

$$w_{ij}(u_i, v_i) = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{b_i}\right)^2\right) \quad (7)$$

where w_{ij} denotes the spatial weight between locations i and j , and b_i represents the adaptive bandwidth at location i .

G. Variable Importance

Variable importance was evaluated using a permutation-based approach based on the increase in Mean Squared Error (MSE) [21]. In this method, the values of each predictor variable were randomly permuted, and the resulting increase in prediction error was calculated. First, the MSE value was computed using the Out-Of-Bag (OOB) data for each decision tree, as shown in (8) [21].

$$MSE_t = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_i - \hat{y}_{i,t})^2 \quad (8)$$

In equation (8), N_t represents the number of observations included in the OOB data in tree t , while $\hat{y}_{i,t}$ is the predicted value of the response variable for observation i in tree t . In the second step, the target variable l is randomly permuted, and the MSE value for tree t , which has undergone randomization of the variable, is recalculated as shown in (9) [21].

$$MSE_t(l) = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_i - \hat{y}_{i,t}(l))^2 \quad (9)$$

In this equation, $\hat{y}_{i,t}(l)$ is the predicted value of the dependent variable for observation i in the new tree t when predictor variable l has been randomized. The difference between MSE_t and $MSE_t(l)$ represents the importance of variable l . The magnitude of this MSE difference in tree t is used as a measure of the importance

of variable l in the decision tree. The final importance value was obtained by averaging the MSE differences across all trees, as shown in (10).

$$MSE(l) = \frac{1}{n} \sum_{t=1}^n (MSE_t - MSE_t(l)) \quad (10)$$

H. Evaluation Performance

In assessing the accuracy of the model, three widely used performance evaluation metrics were employed, namely Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-Square (R^2) [25]. The RMSE is the square value of the Mean Squared Error (MSE). The RMSE formula is given by (11).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (11)$$

where n is the number of observations, y_i is the observed value, and $\hat{y}_i$ indicates the predicted value. The MAPE value indicates the average percentage error between the predicted and observed values.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (12)$$

R-Square or the coefficient of determination is a value that shows how much influence the independent variable has on the dependent variable.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (13)$$

where $\bar{y}$ indicates the average value of the observation results.

III. RESULT AND DISCUSSION

A. Descriptive Statistics

Descriptive statistics in this study are presented through spatial mapping of the response variable (Y) and the mean values of the predictor variables (X) for each observation area. This approach aims to provide an initial understanding of spatial distribution, identify clustering patterns, and reveal regional heterogeneity. The spatial distribution of DHF cases for the 2022 – 2024 period is presented in Fig. 2. This figure illustrates the variation in dengue incidence across regions and over time.

The spatial distribution of Dengue Hemorrhagic Fever (DHF) cases in West Java Province during the 2022 – 2024 period, as shown in Fig. 2, exhibits a consistent pattern of temporal heterogeneity. Urban areas with high levels of urbanization, such as Bandung, Bekasi, and Depok, consistently fall into the high-case category. In contrast, areas such as Banjar and Pangandaran tend to fall into the low-case category and remain relatively stable. This pattern indicates strong spatial clustering, with high-case areas not randomly distributed but concentrated in specific regions. To better understand the factors associated with dengue incidence, the spatial distributions of the predictor variables are presented in Fig. 3.

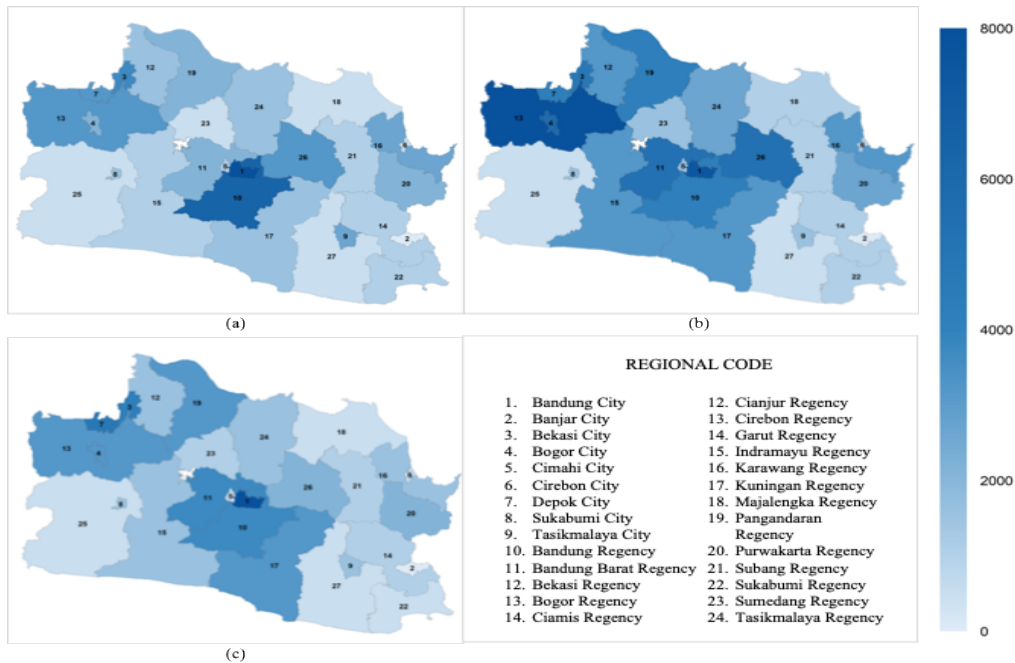


Fig. 2 The spatial distribution of dengue hemorrhagic fever (DHF) cases across cities and regencies in West Java Province from 2022 to 2024. (a) DHF cases in 2022, (b) DHF cases in 2023, and (c) DHF cases in 2024

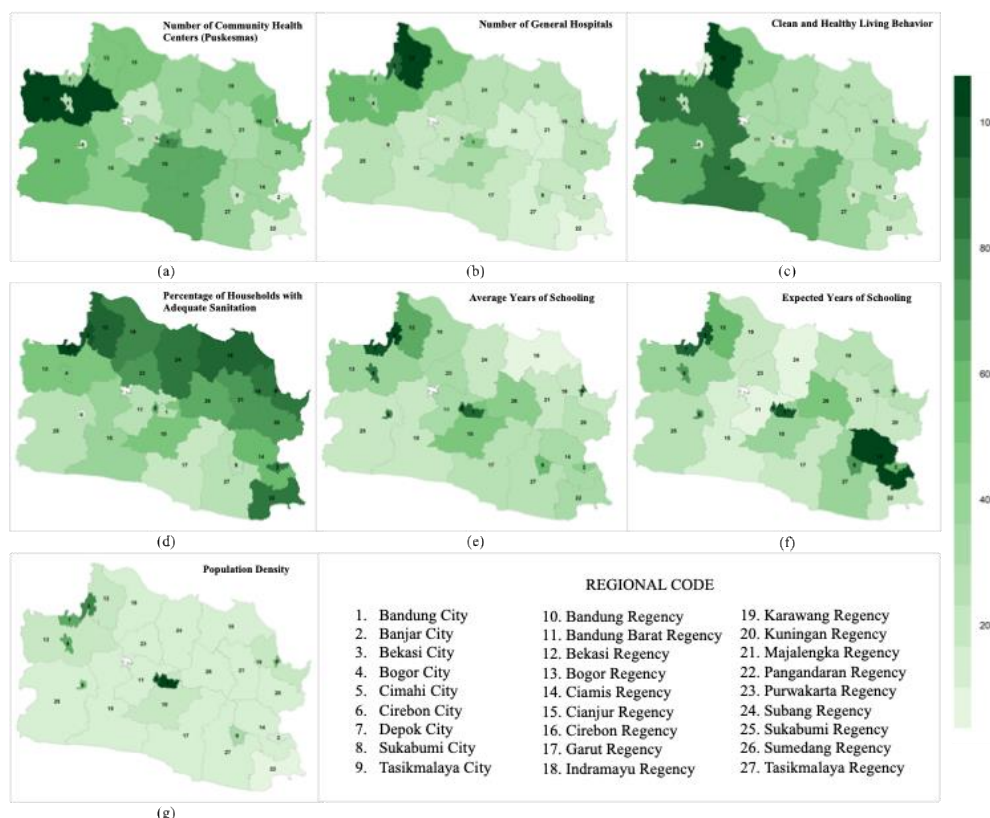


Fig. 3 Spatial distribution of average research variables in West Java Province: (a) number of community health centers (Puskemas); (b) Number of general hospitals; (c) clean and healthy living behavior; (d) percentage of households with adequate sanitation; (e) average years of schooling; (f) expected years of schooling; and (g) population density

Fig. 3 shows that each predictor variable has a different spatial distribution pattern and is not uniform across regions. Variables such as population density and health facilities tend to be concentrated in urban areas, while social indicators such as sanitation and education exhibit more complex variations. This heterogeneity indicates that the relationship between the predictor variables and dengue fever cases is non-stationary, so a global model is insufficient to capture these variations. This further supports the need for spatially adaptive models such as GWRF.

B. Multicollinearity Test

The multicollinearity test based on VIF values was conducted using (1), and the results are presented in Table II.

The multicollinearity test based on VIF values in Table II indicates that most variables do not exhibit multicollinearity, as their VIF values are < 10 . However,

the variable Average Years of Schooling has a VIF value of 12.844, indicating high multicollinearity. Nevertheless, the use of machine learning-based methods such as Random Forest and GWRF remains relevant, as they are not sensitive to multicollinearity.

C. Breusch-Pagan Test

The Breusch–Pagan test for heteroscedasticity was conducted based on the residuals of the global OLS model using (2). The test results were obtained by computing the BP statistic and corresponding p-value, which are presented in Table III.

The test results show that the p-value is $0.020 < 0.05$. This indicates that residual variation differs across regions. This finding reinforces the indication of spatial heterogeneity, as evidenced by the case distribution pattern in Fig. 2. Thus, the use of local models such as GWRF provides strong methodological justification.

TABLE II
THE VIF VALUE

Variable	X_1	X_2	X_3	X_4	X_5	X_6	X_7
VIF	2.788	3.130	2.866	1.449	12.844	3.524	8.876

TABLE III
THE BREUSCH-PAGAN VALUE

Statistics	df	p-value
16.569	7	0.020

D. Random Forest and Geographically Weighted Random Forest

Simulations were conducted to determine the optimal values of *mtry* and *ntree* using 10-fold cross-validation based on the smallest RMSE calculated using (11). The parameters tested consisted of seven *mtry* values (1, 2, 3, 4, 5, 6, 7) and five *ntree* values (50, 200, 300, 500, 1000). The results of the parameter tuning process are presented in Table IV. From Table IV, the parameter values *ntree* = 1000 and *mtry* = 1 were obtained, which produced the smallest RMSE value of 727.788. Therefore, this combination was selected as the final Random Forest Regression (RFR) model. The global RFR model was then used to evaluate the variable importance using a permutation approach, in which the importance of each predictor was measured by the increase in the mean squared error (MSE) after the variable was randomized, as formulated in (10). In addition, the marginal influence of each predictor on the response variable was analyzed using partial dependence plot. The results of the global model and the development of the Geographically Weighted Random Forest (GWRF) model are presented in Fig. 4.

The results of the global Random Forest and GWRF models are presented in Fig. 4. At the global level, permutation-based feature importance (Fig. 4a) identified the most influential variables. The partial dependence plots (Fig. 4b – e) indicate that the relationships between the predictors and the response variable are nonlinear. The optimal bandwidth selection based on the OOB values is shown in Fig. 4f, where the highest performance is achieved at 45 nearest neighbors, suggesting that incorporating information from nearby regions improves the model performance and effectively captures the spatial dependencies. At the local level, the spatial variation in feature importance (Fig. 4g – m) demonstrates that the influence of predictors differs across regions, confirming the presence of spatial non-stationarity. This indicates that a global model may not adequately capture local variations, underscoring the advantage of the GWRF model for modeling region-specific relationships.

The GWRF model was then used to capture spatial heterogeneity, with the local modeling results shown in the distribution of local R^2 in Fig. 5a and the spatial distribution map of feature importance in Fig. 5b. Areas with higher feature importance values tend to have higher local R^2 values in Fig. 5a, indicating that the

model explains variation in dengue fever cases well in those areas. Conversely, areas with lower importance values show smaller local R^2 values. As shown in Fig. 5b, the clustering of regions is based on a combination of variables that are most influential on dengue fever incidence. Each color represents a group of regions with similar feature importance patterns. Regions in West Java Province were divided into several distinct groups, indicating that the dominant factors causing dengue fever were not the same across all regions. Some adjacent regions belonged to the same group, suggesting similarities in risk factor characteristics. However, within different groups, some regions are geographically close, suggesting that the factors influencing dengue fever incidence are local and do not always follow a pattern of geographical proximity.

E. Method Comparison

Model performance was evaluated using RMSE, MAPE, and R^2 , as described (11), (12), and (13). Lower RMSE and MAPE values indicate better prediction accuracy, while higher R^2 values indicate better explanatory power. The results of this evaluation are presented in Table V.

Based on Table V, the GWRF method performs better than the RFR model, as evidenced by lower RMSE and MAPE values and a higher R^2 . Specifically, the GWRF model yields an RMSE of 1756.713 and a MAPE of 0.466, which are lower than those of the RF model, while the R^2 value of 0.700 is higher than that of the RF model. This indicates that the GWRF model provides more accurate predictions and better explains the variation in dengue hemorrhagic fever (DHF) cases.

TABLE IV
OPTIMAL PARAMETER TUNING BASED ON RMSE

<i>mtry</i>	Number of <i>ntree</i>				
	50	200	300	500	1000
1	771.208	803.616	742.657	789.201	727.788
2	800.371	819.538	754.582	795.197	741.555
3	827.843	840.782	802.204	809.251	755.969
4	846.589	862.635	806.523	838.849	764.143
5	844.975	896.730	821.295	849.054	775.540
6	888.044	893.809	820.089	856.716	786.559
7	882.397	910.859	860.209	858.270	788.274

TABLE V
GOODNESS OF FIT OF THE METHODS

Methods	RMSE	MAPE	R^2
Random Forest Regression (RFR)	1796.409	0.482	0.685
Geographically Weighted Random Forest (GWRF)	1756.713	0.466	0.700

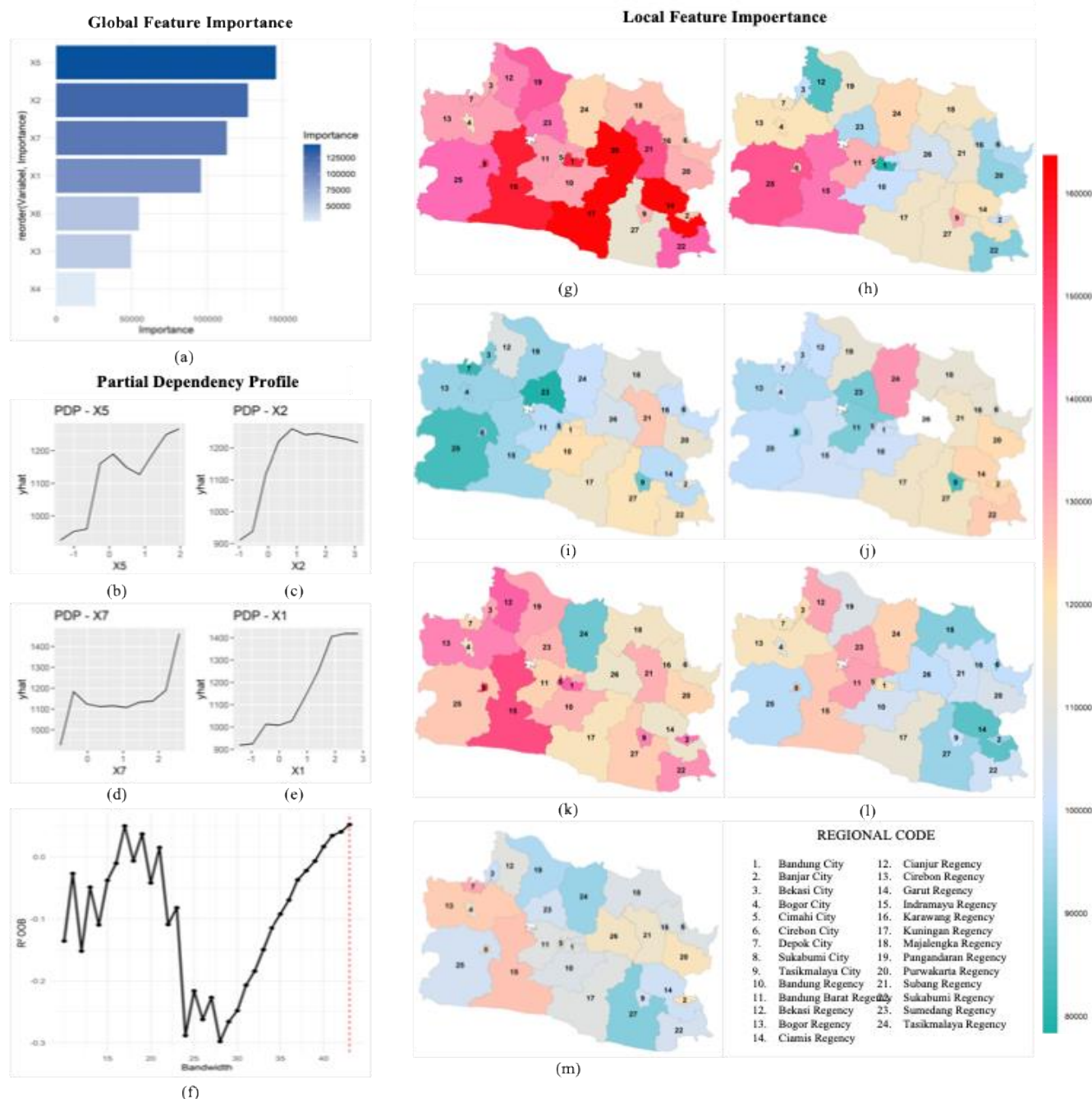


Fig. 4 (a) Permutation-based feature importance from global random forest, **(b-e)** partial dependency profiles of the first four important variables of the global random forest model, **(f)** R^2 OOB value on optimum bandwidth search, **(g-m)** spatial variation of local feature importance (incMSE) in geographically weighted random forest regression models

This performance improvement, particularly the increase in R^2 , indicates that spatial weighting enhances the model's ability to capture spatial heterogeneity. In other words, the increase in R^2 from the global to spatial model provides strong empirical evidence that spatial

factors play a significant role in modeling dengue fever cases. This is because the global RF model assumes that all observations are spatially homogeneous, whereas the GWRF model accommodates local variations through weighting based on geographical proximity.

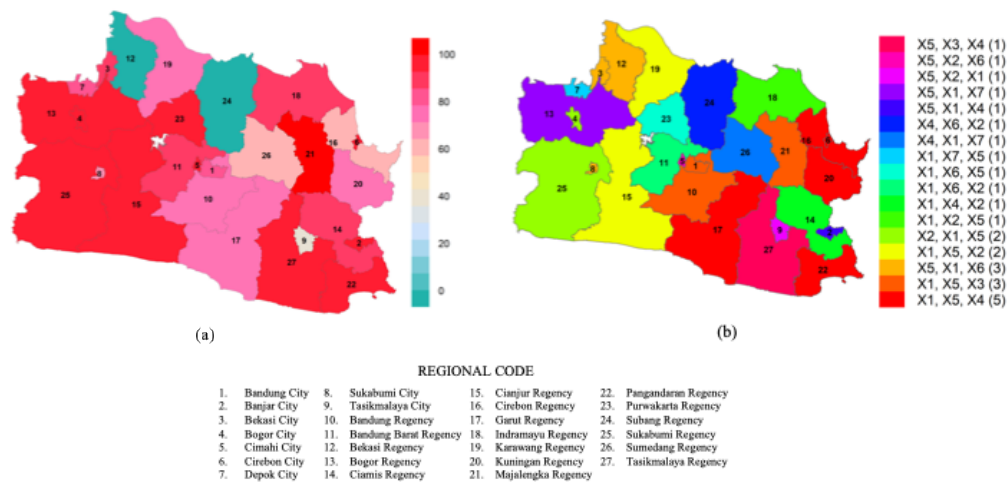


Fig. 5 Spatial distribution of local model performance and feature importance groupings in West Java Province. (a) Spatial distribution of local R^2 values obtained from the Geographically Weighted Random Forest Regression (GWRf) model. (b) Spatial distribution of feature importance groupings derived from the GWRf model.

These results are consistent with those of previous studies showing that the GWRf is more effective at capturing spatial heterogeneity than global models. A study on stunting in East Java demonstrated that the GWRf method outperformed Random Forest based on RMSE, MAPE, and R^2 , indicating its ability to capture spatial variation [14]. Furthermore, other studies have demonstrated that the GWRf can model spatially varying relationships by integrating global and local approaches. The consistency of these findings further reinforces the view that the GWRf method is more robust in capturing spatial heterogeneity than global models, particularly in public health cases characterized by high inter-regional variation. These results are also supported by previous research on rainfall in East Java. The research showed that the GWRf is more effective at capturing spatial heterogeneity and local variations across regions. The study explained that the GWRf can adjust local weights at each observation site using adaptive Gaussian kernel weighting. This approach makes the model more flexible in representing complex spatial patterns. It also allows GWRf to capture local relationships between variables based on geographical proximity [26]. Similar findings were also supported by previous research on the prevalence of type 2 diabetes in the United States, which shows that the GW-RF model performs better in terms of prediction than both the global RF model and the GW-OLS model. The study reported that GW-RF yields a higher R^2 value and lower prediction error, indicating the models superior ability to account for spatial heterogeneity and capture spatially varying relationships among variables [15].

These findings can help local governments prioritize dengue prevention strategies based on region-specific risk factors for dengue. In particular, interventions such as improving sanitation, enhancing health facilities, and controlling population density in high-risk areas can be implemented more effectively by considering the local characteristics identified through the GWRf model.

IV. CONCLUSION

This study demonstrates that the Geographically Weighted Random Forest approach outperforms the global Random Forest Regression model in modeling dengue fever cases in West Java Province. This is evidenced by lower RMSE and MAPE values and a higher R^2 value in the GWRf model. The increase in the R^2 serves as empirical evidence that spatial factors play a significant role in explaining the variation in dengue fever cases, which global models cannot fully capture. Furthermore, the analysis results indicate spatial heterogeneity and non-stationarity, as reflected by variations in feature importance values across regions. This suggests that the factors influencing dengue fever incidence differ by location, making a local modeling approach more relevant than a global one. Methodologically, this study confirms that integrating spatial weighting into machine learning models can improve prediction accuracy while providing a deeper understanding of the patterns and relationships among variables. Thus, the GWRf excels not only in terms of prediction but also in revealing complex spatial dynamics. The implication of this study is the importance of applying a region-based approach to planning public health interventions. Dengue control policies should be

formulated specifically for each region's local characteristics, thereby improving the effectiveness of future management and prevention efforts in the region.

REFERENCES

- [1] Y. Chen, Y. Liao, K. Yuan, A. Wu, and L. Liu, "HLA-A, -B, -DRB1 Alleles as Genetic Predictive Factors for Dengue Disease: A Systematic Review and Meta-Analysis," *Viral Immunol.*, vol. 32, no. 3, pp. 121–130, 2019, doi: 10.1089/vim.2018.0151.
- [2] M. Kanan *et al.*, "Genetic variants associated with dengue hemorrhagic fever. A systematic review and meta-analysis," *J. Infect. Public Health*, vol. 17, no. 4, pp. 579–587, 2024, doi: 10.1016/j.jiph.2024.02.001.
- [3] Y. Liu *et al.*, "Trends and insights in dengue virus research globally : a bibliometric analysis (1995 – 2023)," *J. Transl. Med.*, vol. 22, no. 1, p. 818, 2024.
- [4] R. Yousuf, M. W. Salam, S. Akter, S. Sinha, and M. Haque, "Dengue Dynamics : A Global Update," *Advances in Human Biology*, vol. 14, no. December, pp. 5–10, 2023, doi: 10.4103/aihb.aihb_135_23.
- [5] A. M. Tejo, D. T. Hamasaki, L. M. Menezes, and Y. Ho, "Severe dengue in the intensive care unit," *Journal of Intensive Medicine*, vol. 4, no. September 2023, pp. 16–33, 2024, doi: 10.1016/j.jointm.2023.07.007.
- [6] W. C. Ng *et al.*, "A natural human cell-adapted dengue type 3 virus strain," *Microbiol. Spectr.*, vol. 12, no. 7, pp. 10–13, 2024, doi: 10.1128/spectrum.00701-24.
- [7] B. A. Ajlan, M. M. Alafif, M. M. Alawi, N. A. Akbar, E. K. Aldigs, and T. A. Madani, "Assessment of the new World Health Organization's dengue classification for predicting severity of illness and level of healthcare required," *PLoS Negl. Trop. Dis.*, vol. 13, no. 8, p. p.516229, 2019, doi: 10.1371/journal.pntd.0007144.
- [8] Ministry of Health of the Republic of Indonesia, "Indonesia Health Profile 2024," 2025.
- [9] WHO, "Dengue - Global situation," World Health Organization. Accessed: Apr. 21, 2026. [Online]. Available: <https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON498>
- [10] T. C. N. Monintja, A. A. Arsin, R. Amiruddin, and M. Syafar, "Analysis of temperature and humidity on dengue hemorrhagic fever in Manado Municipality &," *Gac. Sanit.*, vol. 35, no. 2, pp. S330–S333, 2021, doi: 10.1016/j.gaceta.2021.07.020.
- [11] H. Harapan *et al.*, "Dengue viruses circulating in Indonesia : A systematic review and phylogenetic analysis of data from five decades," *Rev. Med. Virol.*, no. December 2018, pp. 1–17, 2019, doi: 10.1002/rmv.2037.
- [12] I. Saiful Fauzi, N. Nuraini, R. Wahyudyah, R. W. S. Ayu, and B. W. Lestari, "Heliyon Temporal trend and spatial clustering of the dengue fever prevalence in West Java, Indonesia," *Heliyon*, vol. 8, no. 8, p. e10350, 2022, doi: 10.1016/j.heliyon.2022.e10350.
- [13] I. Suryani, H. Yasin, and P. Kartikasari, "Pemodelan jumlah kasus demam berdarah dengue (dbd) di jawa tengah dengan geographically weighted negative binomial regression (gwnbr)," *Gaussian*, vol. 10, 2021.
- [14] Y. S. Dewi, S. Hastuti, and M. Fatekurohman, "Analysis of stunting in East Java , Indonesia using random forest and geographically weighted random forest regression," vol. 224, pp. 213–224, 2024, doi: 10.28951/bjb.v42i3.679.
- [15] 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.*, pp. 1–13, 2021, doi: 10.1038/s41598-021-85381-5.
- [16] 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.
- [17] 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. Int.*, vol. 28, no. 6, pp. 6587–6599, 2021, doi: 10.1007/s11356-020-10962-2.
- [18] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J.*, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.
- [19] S. Astiti, A. W. Aranski, and Darmansah, "Prediction of Potential Fishing Zones Using K-Means Clustering and Random Forest in Batam Waters," *JUITA: Jurnal Informatika*, vol. 14, no. 1, pp. 221–230, Mar. 2026, doi: 10.30595/juita.v14i1.29679.
- [20] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wires Data Mining And Knowledge Discovery*, vol. 9, no. 3, 2019.
- [21] Y. Luo, J. Yan, S. C. McClure, and F. Li, "Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model," *Environ Sci Pollut Res*, vol. 29, pp. 33205–33217, 2022, doi: <https://doi.org/10.1007/s11356-021-17513-3>.
- [22] D. N. Gujarati, *Basic Econometrics*, 4th ed. New York: McGraw-Hill Companies, 2004.
- [23] L. Anselin, *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic, 1988. doi: 10.1007/978-94-015-7799-1.
- [24] U. Grömping, "Variable Importance Assessment in Regression: Linear Regression versus Random Forest," *Am. Stat.*, vol. 4, no. 63, pp. 308–319, 2012.

- [25] L. Feng, Y. Wang, Z. Zhang, and Q. Du, "Geographically and temporally weighted neural network for winter wheat yield prediction," *Remote Sens. Environ.*, vol. 262, 2021, doi: 10.1016/j.rse.2021.112514.
- [26] R. Hartanti, P. Yulianing, S. Astutik, and A. B. Astuti, "Geographically Weighted Random Forest Model for Addressing Spatial Heterogeneity of Monthly Rainfall with Small Sample Size," *Cauchy*, vol. 10, no. 1, pp. 442–456, 2025.