STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–13.
Published online in International Academic Press (www.IAPress.org)

Advanced statistical methods for analyzing spatially varying relationships in overdispersed HIV case counts in East Java Province, Indonesia: GWRF vs. GWNBR

Yuliani Setia Dewi*, Renata Wijayanti, Mohammad Fatekurrohman

Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Jember, Indonesia

Abstract This study investigates the efficacy of Geographically Weighted Random Forest (GWRF) compared to Negative Binomial Regression (NBR) and Geographically Weighted Negative Binomial Regression (GWNBR) in modeling spatially varying, overdispersed count data using HIV cases from East Java Province, Indonesia. The dataset covers 38 regencies/cities and examines the relationship between HIV cases and five independent variables. GWNBR incorporates spatial weighting based on adaptive bisquare kernel function and Euclidean distance, while GWRF combines random forests with geographical weighting. GWRF emerges as the superior model based on RMSE, MAPE, and R² values, outperforming NBR and GWNBR. GWRF identifies five groups based on the three most important predictor variables. In approximately 60% of the region, the percentage of drug users (X_2), the percentage of individuals living in poverty (X_4), and the open unemployment rate (X_5) are identified as important variables. Notably, the percentage of drug users and the open unemployment rate are consistently associated with HIV cases across nearly all regions. This study offers valuable insights into HIV transmission patterns and associated risk factors across the province, contributing to a better understanding of the spatial distribution of HIV cases and informing targeted interventions and resource allocations.

Keywords GWRF, NBR, GWNBR, HIV cases, East Java Province, Indonesia, Spatial heterogeneity, Overdispersion

DOI: 10.19139/soic-2310-5070-2827

1. Introduction

Studies on overdispersed count data have gained significant attention due to their prevalence in various fields, such as epidemiology [1, 2], ecology [3, 4, 5] and social sciences [6, 7]. Understanding and accurately modeling overdispersed count data is crucial for making informed decisions and predictions in life science research. Researchers have noted that ignoring overdispersion can lead to biased parameter estimates, erroneous conclusions, and an overestimation of the precision of the model parameters [1, 8, 9, 10].

Previous studies have used various statistical methods to address overdispersion, including negative binomial [11, 12, 13, 14] and the development of Poisson regression models [15, 16, 17]. Negative binomial modeling has emerged as a popular tool for analyzing such data, offering advantages over traditional Poisson models in dealing with overdispersion [4]. Interestingly, while negative binomial models are commonly used, recent studies have explored alternative approaches for handling overdispersion. For instance, generalized linear mixed models (GLMMs) have been recognized for their effectiveness in managing overdispersed count data [18], particularly in single-case experimental designs [19, 20]. The zero-inflated negative binomial (ZINB) model delivers precise estimates of treatment effects in single-case experimental designs characterized by zero-inflated and overdispersed data [20], while some researchers have discovered that both negative binomial models and negative binomial

*Correspondence to: Yuliani Setia Dewi (Email: yulidewi.fmipa@unej.ac.id). Department of Mathematics, University of Jember. 37 Kalimantan Road, Jember, Eas Java Province, Indonesia (68121).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

generalized linear mixed models (NB-GLMM) are effective in managing overdispersion from different sources [4, 8, 21].

Despite advancements in statistical modeling, there remains a lack of effective methods for spatially varying relationships in overdispersed count data. To address this gap, our study focuses on the efficacy of GWRF against other methods, such as NBR and GWNBR, specifically in the context of spatially varying relationships observed in HIV case data in East Java. Geographically Weighted Negative Binomial Regression has been introduced to address the challenges of modeling count data that are overdispersed and non-stationary [22]. This method permits model parameters to change according to spatial variations, proving more effective than global regressions and Geographically Weighted Poisson Regression when dealing with non-stationary overdispersed count data [14].The Geographically Weighted Random Forest method offers a promising approach to capture spatial heterogeneity and improve the accuracy of predictions in overdispersed count data. GWRF combines the strengths of random forests with the ability to account for spatial non-stationarity, making it particularly suitable for modeling spatially varying relationships [23, 24]. This strategy enables the model to adjust to local data variations, which is essential when handling overdispersed count data frequently exhibiting spatial heterogeneity. This technique can capture non-linear relationships and spatial heterogeneity [25, 26]. This research evaluates how well the Geographically Weighted Random Forest technique performs compared to other approaches for modeling count data that exhibit overdispersion and spatial variability using the number of HIV cases in the East Java Province, Indonesia. This research examines various modeling strategies, such as Geographically Weighted Random Forest, Negative Binomial Regression, and Geographically Weighted Negative Binomial Regression, employing R and GeoDa software. Additionally, this study provides a comprehensive framework for diagnostic checking, model estimation, prediction, finding the optimum bandwidth and weight on spatial models, finding important variables, mapping location groups, and comparing methods.

2. Basic algorithm and extensions

The distribution characteristics of the data guide the choice of the analytical method for count data. The analysis is straightforward when dealing with Poisson count data, which assumes equal mean and variance. However, a transition to a negative binomial distribution is necessary when the data exhibits overdispersion (variance exceeding the mean). This shift allows for greater variability in count data [27].

Negative Binomial Regression is commonly employed to model negative binomial data, incorporating a dispersion parameter for extra variability [3, 4]. The algorithm estimates parameters using maximum likelihood estimation [13]. The model establishes a relationship between the logarithm of the expected count and a linear combination of predictor variables. The parameters are iteratively refined during the estimation process to maximize the likelihood function, typically using numerical optimization methods. The algorithm accounts for overdispersion by including an additional parameter that allows the variance to exceed the mean, making it more flexible than the Poisson regression for datasets with extra variability.

In cases where spatial heterogeneity is present [28], more advanced techniques are required. Geographically Weighted Negative Binomial Regression extends NBR to account for spatial non-stationarity, allowing model parameters to vary across geographical spaces. Unlike simplistic models, GWNBR allows for spatially varying parameters by employing kernel functions, enabling superior modeling of non-stationary spatial data [14, 22]. GWNBR involves an iterative process to estimate local parameters at each location.The regression coefficients and dispersion parameters are initially set to their initial values. Subsequently, a spatial weighting method is employed to assign greater importance to observations that are in close proximity. The log-likelihood is computed using these weights, and the parameter estimates are adjusted to optimize the local log-likelihood. This procedure is iteratively applied to each location within the study area to derive coefficient estimates that are specific to each location and iterates until convergence, typically using numerical optimization techniques.

Alternatively, Geographically Weighted Random Forests offer a machine learning approach, combining Random Forests with geographical weighting to capture complex, nonlinear relationships in spatial data [29, 30, 31]. Initially, a spatial weighting method is established to provide greater significance to closer observations. A local

Random Forest model is developed using the weighted observations for each distinct location. This technique involves randomly selecting subsets of features and observations, constructing decision trees, and combining their predictions.This procedure is carried out for every location in the study area, leading to customized Random Forest models for each location. The importance of variables can be calculated locally, allowing the analysis of spatial non-stationarity in feature importance. The final output provides spatially varying predictions and variable importance measures, enabling the exploration of local relationships between predictors and response variables across geographical space.

3. Materials and methods

3.1. Negative Binomial Regression

Different statistical approaches have been employed to handle overdispersion. One method involves adjusting Poisson models using techniques such as negative binomial regression, which adds a random effect typically modeled with a gamma distribution. This approach is beneficial because it accommodates extra variability beyond that explained by the standard Poisson model [4, 32, 33]. The negative binomial regression model for a response variable Y, characterized by a mean μ and a dispersion parameter θ, can be represented as $Y \sim (\mu, \theta)$. The probability mass function is

$$\mathrm{P}(Y = y) = \left(\frac{\Gamma(\mathrm{y} + 1/\theta)}{(\Gamma(\mathrm{y}+1)\Gamma(1/\theta))}\right)\left(\frac{\theta\mu}{(1+\theta\mu)}\right)^{y}\left(\frac{1}{(1+\theta\mu)}\right)^{(1/\theta)} \tag{1}$$

The expected value and variance are given by E(Y) = μ and Var(Y) = $\mu + \theta\mu^2$, respectively. For parameter estimation, the log-likelihood function (Equation 2 is maximized:

$$\mathrm{L}(\beta, \theta) = \sum \left[\log\left(\Gamma\left(\mathrm{y_i} + 1/\theta\right)\right) - \log\left(\Gamma\left(\mathrm{y}_i + 1\right)\right) - \log(\Gamma(1/\theta)) + \mathrm{y}_i \log\left(\theta\mu_i\right) - \left(\mathrm{y}_i + 1/\theta\right)\log\left(1+\theta\mu_i\right)\right] \tag{2}$$

Where β denotes the vector of regression coefficients, $\mu_i = \exp(X_i'\beta)$ signifies the predicted mean for observation i, X_i represents the vector of predictor variables for observation i, and y_i denotes the observed count for observation i. The maximum likelihood estimates for β and θ are obtained by solving $\partial L/\partial\beta = 0$ and $\partial L/\partial\theta = 0$.These equations are typically solved using numerical optimization methods due to their complexity. These methods iteratively update parameter estimates until convergence, which is usually defined by a slight change in the log-likelihood or parameter values between iterations. Following the discussion on NBR, we now explore the GWNBR and GWRF methods, which introduce spatial heterogeneity into the modeling process.

3.2. Geographically Weighted Negative Binomial

The Geographically Weighted Negative Binomial model broadens the application of spatial heterogeneity to the analysis of count data. It incorporates local parameter estimates for each location, allowing for varying relationships between the predictors and response variable across space. This approach is beneficial when dealing with overdispersed count data that exhibit spatial non-stationarity, providing more accurate and locally tailored predictions than global models.

GWNBR weighting in this research is obtained from the adaptive bisquare kernel function. Euclidean distance and bandwidth are required to compute the weights $w\left(u_i, v_i\right)$, as shown in Equation 3.

$$w\left(u_i, v_i\right) = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h_i}\right)^2\right)^2, \text{ if } d_{ij} \leq h_i \\ 0, \text{ if } d_{ij} > h_i \end{cases} \tag{3}$$

Euclidean d_{ij} represents the distance between location i and location j, where u denotes the latitude and v signifies the longitude. The Euclidean distance can be determined using Equation 4.

$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \tag{4}$$

The bandwidth (h_i) determines the range around each point where observations are considered for the local regression process. Optimal bandwidth is determined by cross-validation (CV) [14] according to Equation 5.

$$\text{CV}(h_i) = \sum_{i=1}^{n} \left[y_i - \hat{y}_{\neq i}(h_i)\right]^2 \tag{5}$$

where n represents the number of observation sites, y_i signifies the ith observation, $i = 1, 2, \ldots, n$, and $\hat{y}_{\neq i}(h_i)$ denotes the estimated value of the ith observation, excluding the ith observation.

Hypothesis testing in Geographically Weighted Negative Binomial Regression encompasses two main approaches, simultaneous and partial testing. Simultaneous testing evaluates the overall model significance by examining whether any predictor variables influence the response variable. The null hypothesis is rejected when the p-value falls below the predetermined significance level (α). Conversely, partial testing evaluates the importance of the individual predictors at each specific location. The null hypothesis for each predictor β_k suggests that there is no effect at a particular area, and it is rejected if the absolute t-value surpasses the critical value $t_{(\alpha/2,df)}$ or if the p-value is below α. It enables researchers to assess both the overall model significance and the spatial variability of relationships, facilitating the detection of local patterns and heterogeneity in the data. Simultaneous and partial testing approaches complement each other, comprehensively assessing the geographically weighted regression model. By combining these methods, researchers can identify globally significant predictors and locally influential variables across spatial regions. This multifaceted analysis allows for a nuanced understanding of spatial relationships, helping to uncover complex patterns that may not be apparent through traditional global regression techniques.

3.3. *Geographically Weighted Random Forest*

Geographically Weighted Random Forest is an advanced machine-learning method that integrates the concepts of Random Forest with geographical weighting. This method is designed to capture spatial heterogeneity and improve the prediction accuracy in spatially varying data, such as overdispersed count data. The GWRF uses a spatial weighting function to assign importance to observations based on their proximity to the location of interest [31, 34]. This research uses the adaptive kernel for the process. The process evaluate candidate $K \in \{K_1, K_2, \ldots\}$ by training and picking the K nearest neighbor that maximizes OOB R^2. A local Random Forest model is constructed using the weighted observations for each location. Predictions for new locations are made by aggregating the results from the nearby local models. The steps of GWRF at Figure 1are given as follows:

1. For each location $s_i = (u_i, v_i)$:

   a. The spatial weights for all observations are calculated based on their distance to location s_i. Starting from location s_i, the process involves analyzing all other points in the dataset. Distances are calculated from location s_i to each point, and weights are assigned based on proximity. Points closer to location s_i receive higher weights, whereas more distant points are assigned lower weights. This method reveals the spatial relationships and patterns within the data, allowing for the identification of local trends and influences. Considering the weighted contributions of nearby points provides a nuanced understanding of how geographical features or phenomena vary across the study area. This step is crucial for modeling spatial relationships accurately [35]. The process identifies the K nearest observations, and letting h_i be the distance to the K-th nearest one, it computes bi-square weights w_{ij} for all j using the formula 3.

   b. Training a local random forest model using weighted observations: The local random forest model is trained using the weighted observations derived from the spatial weighting process. Observations with higher case weights are more likely to be included in each bootstrap sample. The integration of bootstrap selection probability with spatial sampling in the GWRF process allows it to manage spatial heterogeneity effectively, improving the model's ability to predict and capture patterns in geographically diverse datasets [25, 26, 31]. The resulting model can adapt to varying spatial dependencies and provide more accurate predictions for different regions within the study area.

c. The $F_i = \{\text{trees at } s_i\}$ or local model parameters and variable importance measures are stored. Local model parameters and variable importance measures offer granular insights into the behavior of machine-learning models. Storing this information allows researchers to understand the feature contributions for specific instances or data subsets, aiding in model interpretation and bias detection. It enables the examination of model behavior across different feature space regions, revealing non-linear relationships and potential unreliability. Preserving local variable importance measures enhances the understanding of feature relevance and guides feature selection. The Increase in Mean Squared Error (%IncMSE) is a feature selection technique in random forest models to assess the importance of individual variables. This method examines how each predictor influences the model's accuracy by observing the increase in the mean squared error when the predictor values are shuffled randomly. The %IncMSE is computed for each tree within the forest and then averaged across all trees. A higher %IncMSE value signifies greater importance of the variable, as it indicates that randomizing the predictor's values leads to a notable rise in prediction error.The formula for %IncMSE can be expressed as

$$\% \text{ incMSE } = \frac{\text{MSE}_{\text{ permuted }} - \text{ MSE}}{\text{MSE}} \times 100\% \tag{6}$$

MSE permuted refers to the mean squared error calculated after the values of a particular predictor are randomly shuffled, and MSE represents the mean squared error of the model in its original form [36].

2. For prediction at a new location s^*:
Once the GWRF model has been fitted, it can be applied to predict the response variable at new, unseen locations. The prediction process at a new site $s^* = (u^*, v^*)$ involves several steps: identifying its spatial neighbors, computing spatial weights, selecting or interpolating between local random forest models, and combining predictions from trees and forests according to the weights. Unlike a standard Random Forest, where prediction depends only on feature values X^*, GWRF predictions additionally depend on the spatial position of the new site, as written in the Equation below:

$$\hat{y}(s^*) = \frac{\sum_{i \in \mathcal{N}(s^*)} w(s^*, s_i)\, \hat{f}_i(X^*)}{\sum_{i \in \mathcal{N}(s^*)} w(s^*, s_i)} \tag{7}$$

where $\mathcal{N}(s^*)$ is set of nearest neighbors of s^* from the training data, $w(s^*, s_i)$ is spatial kernel weight between the new site and training location s_i, and $\hat{f}_i(X^*)$ is prediction from the local random forest model fitted at location s_i. This ensures that predictions at s^* are spatially adaptive, nearby training sites exert more decisive influence than distant ones. The global prediction can also be blended with the local one [31].

3.4. Best Model Selection

This research uses three model evaluations: RMSE, MAPE, and R^2 [37]. The root mean square error is derived from the square root of the Mean Squared Error. It indicates how much the prediction error deviates from observed outcomes [38]. The RMSE calculation formula is given by Equation 8.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{8}$$

where y_i and $\hat{y}_i$ represent observation and estimated value respectively, and n denotes the number of data points. MAPE refers to the mean percentage discrepancy between predicted and actual values [38]. It expresses the average

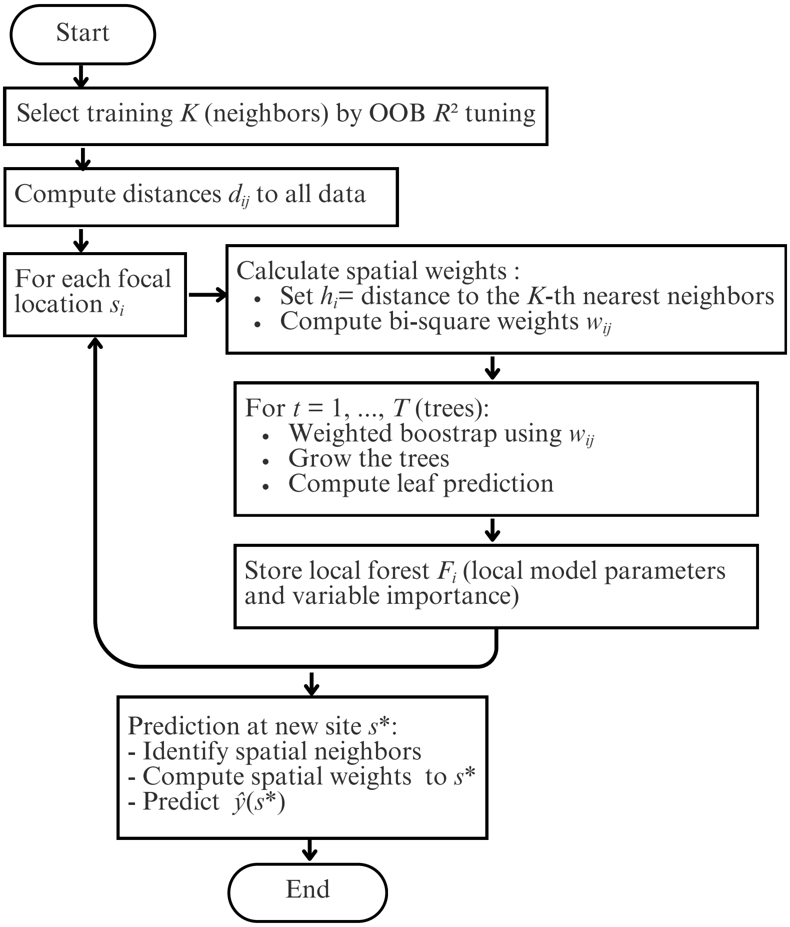
Start
Select training K (neighbors) by OOB R^2 tuning
Compute distances d_{ij} to all data
For each focal location s_i
Calculate spatial weights :
• Set h_i= distance to the K-th nearest neighbors
• Compute bi-square weights w_{ij}
For t = 1, ..., T (trees):
• Weighted boostrap using w_{ij}
• Grow the trees
• Compute leaf prediction
Store local forest F_i (local model parameters and variable importance)
Prediction at new site s*:
- Identify spatial neighbors
- Compute spatial weights to s*
- Predict $\hat{y}(s^*)$
End

Figure 1. GWRF process with adaptive bandwidth.

absolute percentage difference between predicted and actual values.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{9}$$

The coefficient of determination (R^2) is regarded as a valuable metric in comparison to other regression analysis measures, such as the Mean Squared Error (MSE) or Root Mean Squared Error (RMSE), as it offers a clear depiction of the proportion of variance explained by the model. It indicates the model's effectiveness, with values ranging from zero to one. If the value is close to 1, the model will be stronger and deliver better results [39].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

where $\bar{y}$ represents the average value. A specific categorization of the R^2 values is provided as follows:
•$R^2 \geq 0.75$ often considered a strong indicator of model fit [38].
• $0.5 \leq R^2 < 0.75$ indicates a moderate fit, suggesting that, while there is substantial explanatory power, the model still has significant variance unexplained.
•$R^2 < 0.5$ indicates a poor fit and suggests that the model explains only a small portion of the variability in the response data [40].

3.5. Software and Package

In this research, we utilize R software to analyze our data comprehensively. We employ various statistical methods and packages available in R to manage our dataset effectively. The flexibility and robustness of R enable us to execute complex tasks such as data visualization, hypothesis testing, and regression modeling. The MASS package provides the glm.nb() function for negative binomial regression, where the dependent and independent variables and the dataset are specified. For GWNBR, the Gwmodel package offers the gw.dist() function for calculating distance matrices that are needed in geographically weighted (GW) models. The spgwr package includes the gwr.bisquare() function for the bisquare kernel function, necessitating additional parameters such as spatial distance and bandwidth. We use an adaptive bisquare kernel function in the process. The SpatialML package includes the grf.bw() and grf() functions for GWRF analysis, which incorporate spatial information and allow for the specification of the number of trees. The grf.bw() function contributes to determining the optimal bandwidth for the geographically weighted random forest model. Once the bandwidth is selected, the grf() function fits the GWRF model with the specified parameters. We use parameters ntree = 500, and five options for mtry 1,2,3,4,5. These functions facilitate the examination of how relationships between variables vary across different locations. We employ GeoDa software to visualize spatial data. GeoDa provides robust tools for exploring spatial data, including interactive mapping. Users can import various data formats and create thematic maps to identify patterns and clusters within their spatial data.

4. Experimental results

4.1. Descriptive study

The research utilizes secondary data from the East Java Provincial Health Office and Badan Pusat Statistik (BPS) publications in 2021, covering 38 Regencies/Cities in East Java as observational units. The research explores the correlation between the number of HIV cases, which serves as the dependent variable (Y), and five independent variables, including the percentage of family planning program users (X_1), percentage of drug users (X_2), percentage of health screening (X_3), percentage of poor people (X_4), and open unemployment rate (X_5).

The connection between independent variables and the risk of HIV transmission can be examined through various socio-economic and health-related factors that impact HIV epidemiology. Combining family planning initiatives with sexual and reproductive health services can reduce HIV transmission by promoting safe sexual practices, improving access to contraceptives, and encouraging regular health check-ups. Drug use, especially injection drug use, poses a significant risk for HIV transmission, as sharing needles can directly spread the virus among users. Regular health screenings can aid in the early detection and management of HIV, thereby lowering transmission risk. These screenings often include education on HIV transmission and prevention, promoting safer behaviors. Poverty is a crucial social determinant of HIV risk, as economic vulnerability can lead to high-risk behaviors such as sex work or transactional sex for survival. Additionally, poverty often restricts access to healthcare services, including HIV prevention and treatment programs, thereby increasing the risk of HIV transmission. Unemployment is linked to an increased risk of HIV transmission due to the economic stress and instability it causes. Unemployed individuals may turn to drug use or engage in high-risk sexual behaviors as coping mechanisms. This dataset enables the analysis of potential correlations between HIV prevalence and various socio-economic and health-related factors across different regions in East Java, providing insights into the complex dynamics of HIV transmission and its associated risk factors in the province.

Figure 2 presents the distribution of HIV cases in the 38 regencies/cities of East Java Province. The map categorizes the distribution of HIV case numbers into three groups according to quartile values. In urban areas, the number of HIV cases surpasses that in rural areas. The top 25% group, which has the highest number of HIV cases, includes Malang, Batu, and Surabaya City, along with the regencies of Malang, Lumajang, Jember, Banyuwangi, Pasuruan, Sidoarjo, and Bojonegoro. Surabaya City records the most HIV cases, totaling 694. The middle 50% group consists of regions with moderate HIV case numbers, such as the cities of Probolinggo and Mojokerto, and the regencies of Tulungagung, Blitar, and Lamongan. The bottom 25% group, representing areas with the fewest

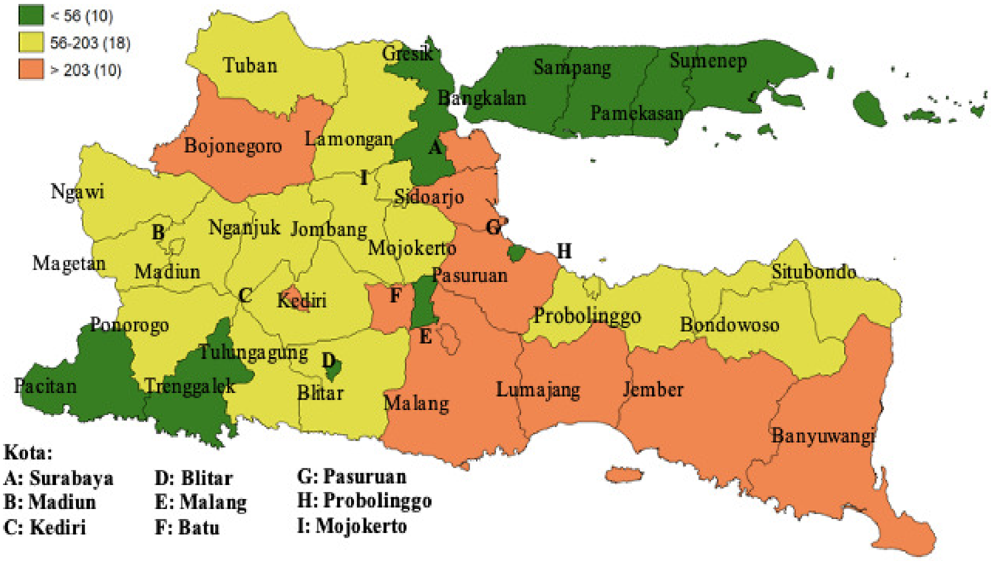

Figure 2. Map of HIV cases in East Java in 2021.

HIV cases, includes smaller regencies and cities throughout East Java. Trenggalek has the lowest number of HIV cases, with 18. This distribution underscores the concentration of HIV cases in urban centers and more densely populated regions, indicating a need for focused interventions and resource distribution in these high-prevalence areas.

4.2. Diagnostic Checking

The HIV case count data demonstrates substantial overdispersion, as evidenced by the considerable disparity between the mean and variance (Table 1). The dispersion value 74.825 further corroborates this observation (Table 2), indicating that the data deviates from a Poisson distribution. This excessive variability may be attributed to the regional heterogeneity in HIV risk, case clustering, or observational dependence. Consequently, standard Poisson regression is deemed inappropriate for this dataset, necessitating alternative modeling approaches that adequately account for the observed overdispersion. Negative binomial regression is an appropriate technique for analyzing count data that exhibit overdispersion, where the variance surpasses the mean, thus breaching the Poisson regression assumption of an equal mean and variance. This method includes a dispersion parameter to address additional variability, resulting in more precise parameter estimates and standard errors. It is often effective in managing overdispersion.

Table 1. Descriptive statistics of variable Y.

Characteristics	Value
Variance	20038.63
Mean	145.70

Table 2. Poisson regression dispersion.

Deviance Value	Df	Dispersion
2394.4	32	74.825

Table 3 presents the Variance Inflation Factor (VIF) to identify multicollinearity among the predictor variables in the regression model. VIF measures the extent to which the variance of the estimated regression coefficients is inflated because of collinearity among the predictors. A VIF of 1 signifies no correlation, whereas higher values indicate greater multicollinearity. Typically, VIF values greater than five or ten are deemed problematic. The VIF values in (Table 3) suggest no multicollinearity issue in the data.

Table 3. The VIF value.

Variable	VIF
X_1	1.487608
X_2	2.499138
X_3	1.041342
X_4	1.812299
X_5	2.078656

The Breusch–Pagan test serves as a tool for detecting data heterogeneity, and the p-value of 0.02172, as shown in Table 4, highlights significant spatial heterogeneity, thereby rejecting the null hypothesis of homoscedasticity. These findings have several implications. It confirms heteroscedasticity in the regression model, suggests spatial variation in the error term variance, and may lead to biased standard errors and unreliable hypothesis tests. We consider the spatial regression techniques GWNBR and GWRF to address this issue.

Table 4. The Breusch–Pagan test value.

BP	Df	p-value
13.183	5	0.02172

4.3. Geographically Weighted Negative Binomial Regression

GWNBR is an extension of Negative Binomial regression that incorporates spatial weighting. This research calculates the weights using an adaptive bisquare kernel function based on the distance between the regencies/cities. We use Euclidean distance obtained from the latitude and longitude of each location (Table 5). This weighting matrix is then used to estimate the GWNBR parameters for each regency/city. GWNBR has an optimal proportion of 0.9 to include in the weighting scheme obtained from the smallest CV value from the cross-validation process, 731968.7.

Table 5. Euclidean distance between regencies/cities in east java province.

	1	2	3	4	5	6	7	8	$\cdots$	38
1	0.000	0.419	0.480	0.824	1.079	1.092	1.578	2.003	$\cdots$	1.413
2	0.419	0.000	0.350	0.544	0.796	0.722	1.301	1.690	$\cdots$	1.064
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
38	1.413	1.064	0.963	0.606	0.400	0.348	0.402	0.646	$\cdots$	0.000

The parameters estimated by GWNBR are evaluated individually at each site to identify those significantly correlated with the number of HIV cases. Partial testing uses test statistics for each parameter in all regencies/cities, which are compared with the t-table values to determine the significant variables. The level of significance used is 5%. The locations are grouped based on the same significant variables. The estimated results of the partial test of the parameters yield significant GWNBR variables in each region. The variables affecting HIV infection in each location, regency, or city within the same group of significant variables are mapped in Figure 3. The locations are divided into two groups based on significant predictor variables. The open unemployment rate (X_5) is a significant variable in one group, including fewer locations, whereas it is not significant in the other group. The spatial heterogeneity observed in the significance of predictor variables across different regions highlights the complexity of HIV transmission patterns. This variability suggests that targeted intervention strategies should be tailored to address the specific risk factors prevalent in each location. Further research is needed to explore the underlying reasons for the differential impact of the open unemployment rate on HIV cases between the two groups of locations.

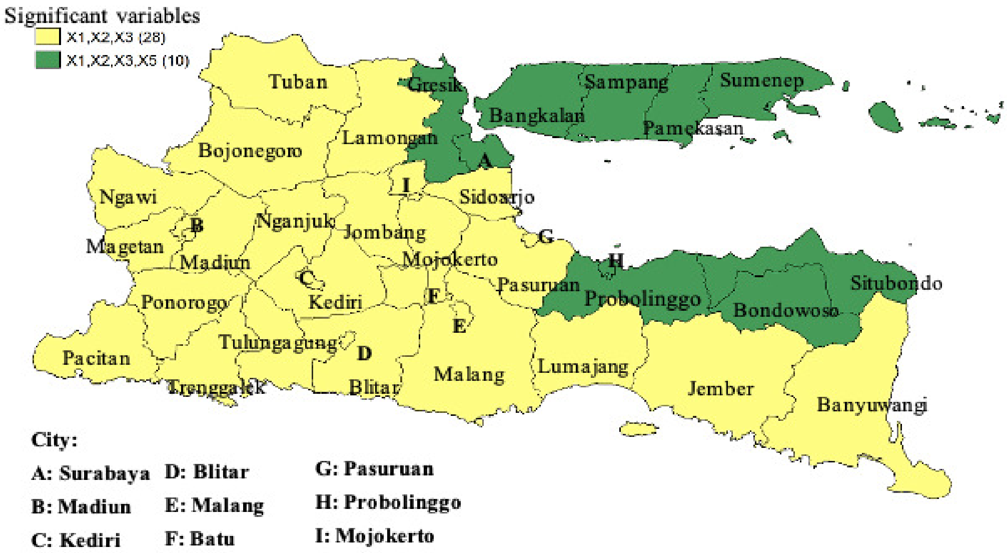

Figure 3. GWNBR grouping map.

4.4. Geographically Weighted Random Forest

Geographically weighted random forest (GWRF) incorporates spatial heterogeneity by assigning weights to observations based on geographical proximity. We determine the weight using the adaptive bisquare kernel, and determine the optimum bandwidth value by looking at the largest R^2 OOB value obtained from the local RF model for each regency/city. In the adaptive Kernel of GWRF, determining the weight of neighboring data is based on the number of nearest neighbors (K), not a fixed physical distance. Figure 4 graphically depicts the R^2 values for each bandwidth. The optimal bandwidth (highest R^2 value) is 24. Building a random forest in the regency or city involves 24 regions.

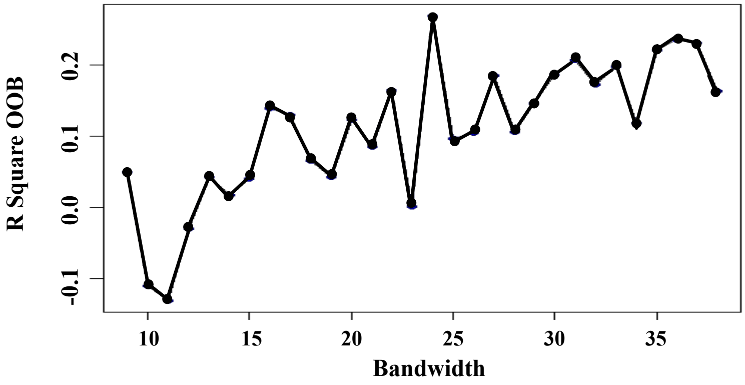

Figure 4. R^2 of OOB value.

Percent Increase in Mean Squared Error (%IncMSE) is a crucial metric used to assess variable importance in Geographically Weighted Random Forests. It measures the reduction in the model accuracy when a particular

Table 6. %IncMSE value of explanatory variable.

Regency/City	X_1	X_2	X_3	X_4	X_5
Probolinggo Regency	43788.86	134889.80	35882.18	85854.64	76287.57
Situbondo Regency	47803.70	202892.95	41044.43	114207.44	76624.54
⋮	⋮	⋮	⋮	⋮	⋮
Blitar Regency	35513.81	76809.50	41060.92	31418.98	72885.65

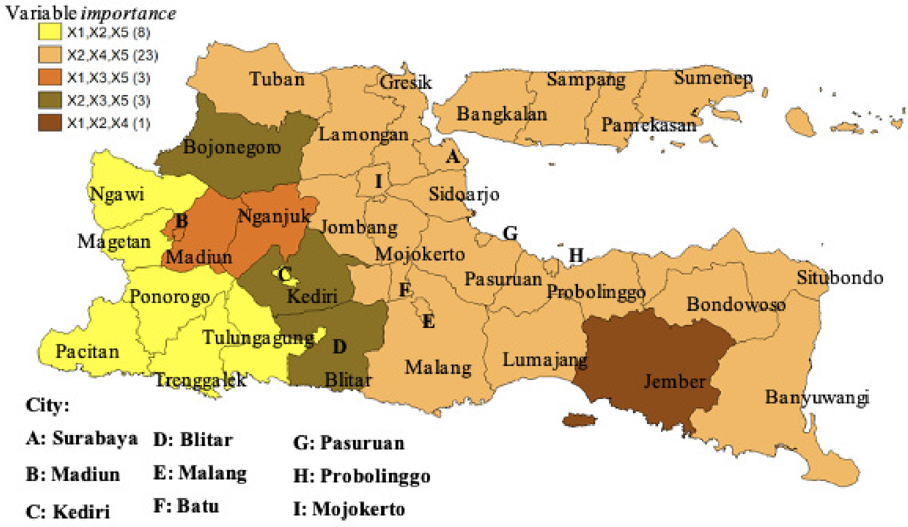

Figure 5. GWRF Grouping Map.

predictor is randomly shuffled, with larger values signifying greater significance. We observe that %IncMSE differs across various geographical areas, offering insights into the spatial variability of variable importance (Tabel 6). This research categorizes the locations based on the three most important variables locally. As illustrations, in Probolinggo and Situbondo Regencies, the three key variables are the percentage of drug users (X_2), the percentage of poor people (X_4), and the open unemployment rate (X_5). Conversely, in Blitar Regency, the three most important variables include the percentage of drug users (X_2), the percentage of health screening (X_3), and the open unemployment rate (X_5). Figure 5 illustrates the locations grouped according to the same top three important variables. These locations are divided into five categories based on the similarity of the three most important predictor variables of GWRF. As illustrated in Figure 5, these locations tend to cluster with their proximate neighbors. The majority of locations are characterized by the important variables X_2, X_4, and X_5. Variables X_2 and X_5 are notably important in nearly every location.

In the central and eastern regions of East Java, variables X_2, X_4, and X_5 are the three most important, except in Jember Regency, X_5 (open unemployment rate) is not inclued. The open unemployment rate in Jember Regency decreased by 1.38% from 2018 to 2022. Jember has effective local policies or community initiatives that support employment and economic participation. Local government efforts in vocational training or job placement help bridge the gap between job seekers and employment opportunities, reducing unemployment [41]. Sociocultural factors and community cohesion can contribute to employment levels [42]. Additionally, regional economic and labor market structure mobility can also influence the open unemployment in the area.

The variation in key variables across different regencies highlights the complex and region-specific nature of socio-economic factors influencing each area. It underscores the need for tailored interventions and policies that address the unique challenges faced by each location. By identifying and focusing on these critical factors, local governments and organizations can more effectively allocate resources and implement targeted strategies to improve overall community well-being.

4.5. *Comparison of the methods*

The goodness of fit of the model is evaluated by analyzing RMSE, MAPE and other metrics derived, with the outcomes presented in Table 7.

Table 7. Goodness of fit of the methods.

Criteria	Negative Binomial	GWNBR	GWRF
RMSE	110.554	102.814	25.4114
MAPE	105.296	94.827	16.9077
R^2	0.374	0.458	0.9669

Table 7 shows that the RMSE values of the Negative Binomial regression and GWNBR are larger than those of the GWRF. The larger the RMSE value, the higher the prediction error, indicating that the GWRF model performs better than the other two. The RMSE value of GWRF is the smallest among the models, indicating that the predicted values are closer to the observed results. The MAPE values of the Negative Binomial regression and GWNBR, each with MAPE percentages exceeding 50%, are higher than those of the GWRF. Based on the MAPE criteria, the GWRF delivers the best performance. The R^2 value of the GWRF model is close to 1, indicating that the predictor variables collectively have a strong ability to explain the response variable. This value is higher than that of the Negative Binomial regression and GWNBR. Therefore, based on all the criteria, RMSE, MAPE, and R^2 for each model, it can be concluded that the GWRF method is the best model for analyzing HIV cases in East Java Province in 2021, where the data exhibit overdispersion and spatial heterogeneity.

The accuracy vs interpretability trade-off for GWRF and GWNBR presents a classic dilemma in spatial modeling. GWRF (Geographically Weighted Random Forest) demonstrates superior accuracy, outperforming GWNBR regarding RMSE, MAPE, and R^2 values. It captures complex non-linear relationships and spatial heterogeneity, providing spatially varying variable importance measures. It allows for grouping locations based on the most important predictor variables. However, as a machine learning method, GWRF sacrifices interpretability, making it challenging to explain its decision-making process. On the other hand, GWNBR offers higher interpretability, with coefficients that have clear statistical interpretations. It accounts for spatial heterogeneity by allowing coefficients to vary across locations and enables the identification of significant variables for each location through partial testing. It facilitates grouping locations based on significant variables. However, based on model fit metrics, GWNBR's accuracy is lower than that of GWRF. The choice between these methods ultimately depends on the specific research goals, with GWRF favoring predictive power and GWNBR prioritizing interpretability and statistical inference.

5. Conclusions and discussion

This research investigates the effectiveness of Geographically Weighted Random Forest (GWRF) compared to Negative Binomial Regression (NBR) and Geographically Weighted Negative Binomial Regression (GWNBR) in modeling overdispersed count data with spatial variations using HIV case data from East Java Province, Indonesia. HIV case count data demonstrate substantial overdispersion, and the Breusch-Pagan test indicates significant spatial heterogeneity.

The Negative Binomial Distribution fit the HIV data in East Java better than the Poisson distribution. The Negative Binomial model accounts for this overdispersion by incorporating an additional parameter to capture the extra variability in the data. Spatial heterogeneity suggests that HIV cases are not uniformly distributed across East Java, with some areas experiencing higher concentrations of cases than others. This spatial variation in HIV incidence highlights the need for targeted intervention strategies that address local risk factors.

There are five groups of locations according to the similarity of the three most important predictor variables of GWRF. In most areas, the primary variables, the percentage of drug users (X_2), the percentage of poor people (X_4), and the open unemployment rate (X_5) are prevalent, making them crucial for the local government to prioritize when developing policies concerning factors that affect HIV cases. Notably, variables X_2 and X_5 are linked to

HIV cases in nearly every region. Several steps can be taken to validate and implement the findings of the GWRF model further. Validating the model using data from another year or province would assess its robustness and generalizability. Public health officials should utilize the GWRF prediction map to identify high-risk areas for targeted investigations and interventions. Focus should be placed on addressing the primary variables influencing HIV cases, percentage of drug users (X_2), percentage of poor people (X_4), and open unemployment rate (X_5). Interventions targeting drug use and unemployment should be prioritized, as variables X_2 and X_5 are linked to HIV cases in nearly every region. Further research is needed to understand the underlying mechanisms connecting these variables to HIV prevalence in different areas. Regular monitoring and evaluation of interventions should be implemented to assess their effectiveness in reducing HIV cases across the identified high-risk areas.

The research indicates that GWRF surpasses NBR and GWNBR in terms of RMSE, MAPE, and R² metrics, establishing it as the most effective model for examining HIV cases in East Java Province in 2021, especially when the data exhibit overdispersion and spatial heterogeneity. Nonetheless, the study has limitations, such as the small sample size of 38 regencies/cities, which limits statistical power and generalizability while heightening the risk of overfitting, particularly with the complex GWRF model. The cross-sectional nature of the 2021 data hinders causal inference, the evaluation of temporal trends, and the analysis of long-term patterns. Potential omitted variable bias might have resulted in skewed estimates of predictor effects and overlooked significant spatial or socioeconomic factors. The computational complexity of the GWRF model restricts extensive sensitivity analyses and bootstrap resampling, making it less interpretable than parametric alternatives. These limitations necessitate careful interpretation and generalization of the results. Future studies should address these issues by using larger sample sizes, longitudinal data, and a more comprehensive set of variables to enhance the robustness and applicability of the findings.

6. Acknowledgement

The authors thank the East Java Provincial Health Office and Badan Pusat Statistik (BPS) for providing access to HIV data and related variables in this study. The authors also gratefully acknowledge the anonymous referee for valuable comments and suggestions.

REFERENCES

1. X. A. Harrison, *Using observation-level random effects to model overdispersion in count data in ecology and evolution*, PeerJ, vol. 2, p. e616, 2014, doi: 10.7717/peerj.616.
2. H. Jin, A. Restar, and C. Beyrer, *Overview of the epidemiological conditions of HIV among key populations in Africa*, J Int AIDS Soc, vol. Suppl 24 3, no. S3, 2021, doi: 10.1002/jia2.25716.
3. A. Lindén and S. Mäntyniemi, *Using the negative binomial distribution to model overdispersion in ecological count data*, Ecology, vol. 92, no. 7, pp. 1414–1421, 2011, doi: 10.1890/10-1831.1.
4. J. Stoklosa, F. K. C. Hui, and R. V Blakey, *An Overview of Modern Applications of Negative Binomial Modelling in Ecology and Biodiversity*, Diversity (Basel), vol. 14, no. 5, p. 320, 2022, doi: 10.3390/d14050320.
5. H. Campbell, *The consequences of checking for zero-inflation and overdispersion in the analysis of count data*, Methods Ecol Evol, vol. 12, no. 4, pp. 665–680, 2021, doi: 10.1111/2041-210x.13559.
6. T. Y. Jang, *Count Data Models for Trip Generation*, J Transp Eng, vol. 131, no. 6, pp. 444–450, 2005, doi: 10.1061/(asce)0733-947x(2005)131:6(444).
7. L. Tan et al., *The selection of statistical models for reporting count outcomes and intervention effects in brief alcohol intervention trials: A review and recommendations*, Alcohol, clinical & experimental research, vol. 48, no. 1, pp. 16–28, 2023, doi: 10.1111/acer.15232.
8. E. H. Payne, M. Gebregziabher, J. W. Hardin, V. Ramakrishnan, L. E. Egede, and A. Selassie, *Approaches for dealing with various sources of overdispersion in modeling count data: Scale adjustment versus modeling*, Stat Methods Med Res, vol. 26, no. 4, pp. 1802–1823, 2015, doi: 10.1177/0962280215588569.
9. E. H. Payne, M. Gebregziabher, J. W. Hardin, V. Ramakrishnan, and L. E. Egede, *An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data*, Commun Stat Simul Comput, vol. 47, no. 6, pp. 1722–1738, 2017, doi: 10.1080/03610918.2017.1323223.
10. G. Popovic et al., *Four principles for improved statistical ecology*, Methods Ecol Evol, vol. 15, no. 2, pp. 266–281, 2024, doi: 10.1111/2041-210x.14270.
11. H. Zhu and H. Lakkis, *Sample size calculation for comparing two negative binomial rates*, Stat Med, vol. 33, no. 3, pp. 376–387, 2013, doi: 10.1002/sim.5947.

12. W. H. Aeberhard, S. Heritier, and E. Cantoni, *Robust inference in the negative binomial regression model with an application to falls data*, Biometrics, vol. 70, no. 4, pp. 920–931, 2014, doi: 10.1111/biom.12212.
13. Yee, Thomas W, *The VGAM package for negative binomial regression*, Australian & New Zealand journal of statistics, vol. 62, no. 1, pp. 116–131, 2020, doi: 10.1111/anzs.12283.
14. J. Chen, L. Liu, C. Xu, D. Long, and L. Xiao, *Integrative Analysis of Spatial Heterogeneity and Overdispersion of Crime with a Geographically Weighted Negative Binomial Model*, ISPRS Int J Geoinf, vol. 9, no. 1, p. 60, 2020, doi: 10.3390/ijgi9010060.
15. D. P. Blevins, S. M. Spain, and E. W. K. Tsang, *Count-Based Research in Management*, Organ Res Methods, vol. 18, no. 1, pp. 47–69, 2014, doi: 10.1177/1094428114549601.
16. A. Huang, *Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts*, Stat Modelling, vol. 17, no. 6, pp. 359–380, 2017, doi: 10.1177/1471082x17697749.
17. F. Sami, M. M. Butt, and M. Amin, *On the ridge estimation of the Conway-Maxwell Poisson regression model with multicollinearity: Methods and applications*, Concurrency and Computation: Practice and Experience, vol. 34, no. 1, 2021, doi: 10.1002/cpe.6477.
18. M. Brooks E et al., *glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling*, R J, vol. 9, no. 2, p. 378, 2017, doi: 10.32614/rj-2017-066.
19. H. Li, W. Luo, E. Baek, K. H. Lam, and C. G. Thompson, *Multilevel modeling in single-case studies with count and proportion data: A demonstration and evaluation*, Psychol Methods, 2023, doi: 10.1037/met0000607.
20. H. Li, W. Luo, and E. Baek, *Multilevel modeling in single-case studies with zero-inflated and overdispersed count data*, Behav Res Methods, vol. 56, no. 4, pp. 2765–2781, 2024, doi: 10.3758/s13428-024-02359-7.
21. M. R. Rahman Shaon and X. Qin, *Use of Mixed Distribution Generalized Linear Models to Quantify Safety Effects of Rural Roadway Features*, Transportation Research Record: Journal of the Transportation Research Board, vol. 2583, no. 1, pp. 134–141, 2016, doi: 10.3141/2583-17.
22. A. R. Da Silva and T. C. V. Rodrigues, “Geographically Weighted Negative Binomial Regression—incorporating overdispersion,” Stat Comput, vol. 24, no. 5, 2013, doi: 10.1007/s11222-013-9401-9.
23. G. Nduwayezu et al., *Understanding the spatial non-stationarity in the relationships between malaria incidence and environmental risk factors using Geographically Weighted Random Forest: A case study in Rwanda*, Geospat Health, vol. 18, no. 1, 2023, doi: 10.4081/gh.2023.1184.
24. 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*, Accident Analysis & Prevention, vol. 199, p. 107528, 2024, doi: 10.1016/j.aap.2024.107528.
25. S. N. Khan, M. Maimaitijiang, and D. Li, *A Geographically Weighted Random Forest Approach to Predict Corn Yield in the US Corn Belt*, Remote Sens (Basel), vol. 14, no. 12, p. 2843, 2022, doi: 10.3390/rs14122843.
26. Z. Wu, F. Yao, J. Zhang, and H. Liu, *Estimating Forest Aboveground Biomass Using a Combination of Geographical Random Forest and Empirical Bayesian Kriging Models*, Remote Sens (Basel), vol. 16, no. 11, p. 1859, 2024, doi: 10.3390/rs16111859.
27. Hilbe, Joseph M, *Negative Binomial Regression Second Edition-Negative Binomial Regression: Second Edition Joseph M. Hilbe Frontmatter More information*, Cambridge University Press, 2011, [Online]. Available: www.cambridge.org.
28. T. S. Breusch and A. R. Pagan, *A Simple Test for Heteroscedasticity and Random Coefficient Variation*, Econometrica: Journal of the econometric society, pp. 1287–1294, 1979.
29. F. Santos, V. Graw, and S. Bonilla, *A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon*, PloS one vol. 14, no. 12. 2019. doi: 10.1371/journal.pone.0226224.
30. 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/Published.
31. S. Georganos and S. Kalogirou, *A Forest of Forests: A Spatially Weighted and Computationally Efficient Formulation of Geographical Random Forests*, ISPRS Int J Geoinf, vol. 11, no. 9, 2022, doi: 10.3390/ijgi11090471.
32. P. Hougaard, G. A. Whitmore, and M.-L. T. Lee, *Analysis of Overdispersed Count Data by Mixtures of Poisson Variables and Poisson Processes*, Biometrics, vol. 53, no. 4, p. 1225, 1997, doi: 10.2307/2533492.
33. S. Gschlößl and C. Czado, *Modelling count data with overdispersion and spatial effects*, Statistical Papers, vol. 49, no. 3, pp. 531–552, 2006, doi: 10.1007/s00362-006-0031-6.
34. 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, 2021, doi: 10.1038/s41598-021-85381-5.
35. 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*, Environmental Science and Pollution Research, vol. 29, no. 22, pp. 33205–33217, 2022, doi: 10.1007/s11356-021-17513-3.
36. H. Liang, Z. Guo, J. Wu, and Z. Chen, *GDP Spatialization in Ningbo City based on NPP/VIIRS Night-Time Light and Auxiliary Data Using Random Forest Regression*, Advances in Space Research, vol. 65, no. 1, pp. 481–493, 2020, doi: 10.1016/j.asr.2019.09.035.
37. Feng, Luwei and Wang, Yumiao and Zhang, Zhou and Du, Qingyun, *Geographically and temporally weighted neural network for winter wheat yield prediction*, Remote Sensing of Environment, vol. 262, p.112514, 2021.
38. D. Chicco, M. J. Warrens, and G. Jurman, *The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation*, PeerJ Comput Sci, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.
39. 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.
40. C. Onyutha, *From R-squared to coefficient of model accuracy for assessing " goodness-of-fits"*, Geoscientific Model Development Discussions, vol. 2020, pp. 1–25, 2020.
41. V. V. , Maulana and A. Ilyasi, *Strategi Dinas Tenaga Kerja dan Transmigrasi dalam Upaya Mengurangi Pengangguran di Kabupaten Jember*, formula Jurnal Administrasi Publik, vol. 1, no. 2, pp. 156–166, 2024.
42. M. Asensio and C. Ferreira, *Labor-market Reforms in Southern Europe: From Protection to Flexibility*, intechopen, 2024.