Generalized Linear Mixed-Model Tree for Modeling Dengue Fever Cases

Erwan Setiawan*[1], Khairil Anwar Notodiputro[2], Bagus Sartono[3]
[1,2,3]IPB University; Jalan Meranti Wing 22 level 4. IPB Dramaga Campus, Bogor 16680
[1,2,3]Program Study of Statistics and Data Science, School of Data Science, Mathematics and Informatics, IPB University, Bogor
e-mail: ***[1]erwan85setiawan@apps.ipb.ac.id**, [2]khairil@apps.ipb.ac.id, [3]bagusco@apps.ipb.ac.id

Abstract

The GLMM tree demonstrates flexibility when applied to complex dataset structures such as multilevel and longitudinal data. However, there has been no assessment of the performance of GLMM trees on panel data structures. This study aims to assess the performance of the GLMM tree on a panel data structure using a case study of dengue fever cases in West Java. The performance evaluation focuses on the accuracy of the model. The dataset includes cross-sectional data from 27 regencies/cities in West Jawa, covering different regions at a single point in time, and time-series data from 2014 to 2022, tracking dengue fever cases over the years. The results of this study show that the GLMM tree model is suitable for panel data that exhibit nuanced or intricate variability unrelated to temporal effects. When developing the incidence rate of the dengue fever model, the GLMM tree separates into two submodels depending on a GRDP growth rate threshold of 5.5%. The GLMM tree model shows significant differences in the incidence rate of dengue fever between regencies/cities. However, the differences in the incidence rate of dengue fever from year to year between the regencies/cities are not significant. It indicates that local factors, such as research predictor variables, are more dominant in influencing the incidence rate than global factors.

Keywords— Panel Data, Generalized Linear Mixed-Model, GLMM tree, Dengue Fever

1. INTRODUCTION

Group-structured datasets are widely used in various research and practical applications. Many institutions provide access to this dataset for various users to contribute to social improvement. For instance, the BreizhCrops dataset leverages time series data from satellite imagery to map plant types in multiple locations within the Brittany region, France [1]. Another example is a dataset derived from a census or survey administered by the Central Statistics Agency (Badan Pusat Statistik – BPS), encompassing a wide range of data related to demographics, economics, environment, and social aspects across various regions in Indonesia. A group-structured dataset organizes data into natural or hierarchical groups and presents unique features and challenges in statistical analysis, such as interdependence among observations within groups, variation between groups, and specific effects within groups. The Generalized Linear mixed model (GLMM) is a popular statistical approach for analyzing datasets with a group structure [2].

GLMM, an expansion of the Generalized Linear Model (GLM), incorporates random effects. With the inclusion of random effects, GLMM can account for variations between different groups or subgroups in the data. This feature enables researchers to develop more precise and realistic models that accurately represent the complex data structures often encountered in research. Despite the numerous advantages of GLMM in group data analysis, it is important to acknowledge several weaknesses and challenges, including the complexity of GLMM models due to the presence of both random effects and fixed effects. This complexity can interpret GLMM models more challenging, especially for individuals without a strong statistical background [3][4]. To address this weakness, Fokkema, et al. [5] have introduced a GLMM tree as a potential solution.

The GLMM tree technique integrates the advantages of GLMM and decision tree models. Introduced by Fokkema et al. in 2018, the GLMM tree algorithm can detect interactions between treatments and subgroups within the data while also considering the inherent grouping structure present in the dataset [5]. By utilizing decision trees to divide the data into smaller subgroups, the GLMM tree simplifies interpretation by organizing the data hierarchically based on the most informative predictor variables. Notably, in the analysis of multilevel and longitudinal data, GLMM trees exhibit comparable performance to traditional GLMMs and Random Forest (RF) despite requiring the assessment of a smaller number of variables [6].

The GLMM trees have been identified as a versatile model suitable for handling complex data structures such as multilevel and longitudinal data [6][7]. Multilevel data exhibits a hierarchical format, while longitudinal data entails observations collected from the same unit over an extended period. However, certain research scenarios necessitate the combined analysis of both these dimensions. Panel data refers to datasets that combine elements of both multilevel and longitudinal structures. Specifically, panel data captures cross-sectional variation between units (such as regions, individuals, or companies) at a specific point in time, alongside temporal variation by observing these units over multiple periods. For instance, in this study, data from 27 regencies/cities in West Java were collected for multiple years (2014–2022), capturing both spatial differences (across regencies) and temporal changes (over the years). This integrated structure enables the analysis of how both local and time-based factors influence the outcome variable, such as the incidence rate of dengue fever. Given this complexity, further investigation into the effectiveness of GLMM trees for panel data is essential, as they can leverage both hierarchical and temporal components to enhance analytical techniques.

This study aims to evaluate the effectiveness of GLMM trees in modeling panel data by focusing on a case study that predicts the incidence rate of Dengue Fever (DF) cases in West Java. The dataset includes the number of DF cases in West Java at the regency/city level from 2014 to 2022, comprising panel data with regency/city level differences and temporal variations [8]. Modeling the number of DF cases is crucial for preventing and controlling the spread of this infectious disease, which is mainly transmitted by Aedes aegypti and Aedes albopictus mosquitoes and is widespread in tropical and subtropical regions, including Indonesia [9][10]. It's important to note that epidemiological data shows a consistent increase in DF cases during the rainy season in various Indonesian provinces. In 2023, the Indonesian Ministry of Health reported 114,720 DF cases nationwide, with the highest incidence recorded in West Java at 19,328 cases [11].

The GLMM tree model is particularly advantageous in this context due to its ability to handle complex hierarchical data structures and capture subtle variations that may not be fully explained by traditional GLMMs or other statistical approaches. Unlike conventional methods, GLMM trees integrate the strengths of regression trees and mixed-effects modeling, offering flexibility in identifying interaction effects and segmenting data based on key predictors. For instance, Hothorn and Zeileis [12] highlighted that GLMM trees can effectively uncover data-driven structures in hierarchical or panel datasets by combining recursive partitioning with random effects, making them well-suited for epidemiological analyses. Similarly, studies such as Fokkema et al. [6] emphasize the interpretability and practical value of GLMM trees in disentangling complex variable relationships, which is particularly relevant for DF modeling in regions with diverse sociodemographic and environmental conditions.

This research aimed to examine the application of the GLMM tree algorithm in panel data modeling using the incidence rate dataset of dengue fever cases in West Java. The objectives included assessing the performance of GLMM tree and GLMM in capturing the variance structure of regencies/cities and years. Additionally, the study sought to compare the accuracy of the GLMM tree and GLMM in predicting the incidence rate of DF. The article comprises four sections: (1) Introduction, (2) Research Methods, (3) Result and Discussion, and (4) Conclusion.

2. RESEARCH METHODS

2.1. Data

This study utilized secondary data from the West Java Government Open Data and the West Java Central Statistics Agency, accessible at opendata.jabarprov.go.id and jabar.bps.go.id. The data collection focused on 27 regencies/cities in West Java and spanned the years 2014 to 2022.

The response variable was the incidence rate (IR) of dengue fever. In GLMM tree modeling, three variables are independent: predictor variables, random effect variables, and partition variables. Predictor variables are employed in models for predicting responses and play a crucial role in shaping the GLMM model at each tree node. The predictor variables having a significant impact on dengue fever were population density, population growth rate, net enrollment rate for junior secondary education, and households with access to adequate sanitation [10][13][14]. Regarding the random effect variable, regency/city was considered a random intercept variable, while year was regarded as a random slope variable to explain the variance of the incidence rate in each distinct regency/city.

The partition variables in a decision tree play a crucial role in partitioning data into more coherent and smaller groups. These specific variables are selected with consideration of context and domain expertise to ensure the meaningfulness of the segmentation in the analysis [15]. In this study, the chosen partition variables encompass the growth rate of Gross Regional Domestic Product (GRDP) at constant prices, the human development index, the education index, the health index, and the impoverished population. These variables are used as partition factors rather than fixed effects to capture the heterogeneity between regions, allowing for more granular subgroup analyses and better interpretation of local variations in dengue incidence.

The GRDP growth rate measures economic performance, with studies indicating that regions with higher economic activity often have better infrastructure to combat vector-borne diseases [16]. The education index is critical as higher education levels correlate with increased health awareness and effective disease prevention strategies [17]. These variables are better suited as partition variables because they represent macro-level, structural characteristics that shape the broader context of dengue fever risk. Including them as fixed effects could oversimplify their influence, treating them as direct predictors rather than underlying conditions that differentiate regions. By using them for partitioning, the model accounts for hierarchical or nested relationships between socioeconomic conditions and dengue outcomes, enabling a more robust and context-sensitive analysis. See Table 1 for a comprehensive list of variables utilized in this study.

Table 1. List of variables utilized in the study

Types of Variables	Names of variables	Types of Data	Units
Response	1. Incidence rate (IR) of Dengue Fever (Y)	Numerical	People per 100,000 population
Fixed Effect (Predictor)	1. Population density $(X^{(1)})$	Numerical	People per 100 km^2
	2. Population growth rate $(X^{(2)})$	Numerical	%
	3. Net enrollment rate for junior secondary education $(X^{(3)})$	Numerical	%
	4. Households having access to adequate sanitation $(X^{(4)})$	Numerical	%
Random Effect	1. Regency/City as a random intercept	Categorical	-
	2. Year as random slope	Numerical	-
Partition	1. Human development index	Numerical	-
	2. GRDP growth rate at constant prices	Numerical	%
	3. Education index	Numerical	-
	4. Health index	Numerical	-
	5. Poor population	Numerical	%

Source: BPS, Open Data Jabar

2.2. Model

The GLMM model for this study is defined by the following specifications:

a) GLMM represents an expansion of the GLM, allowing for the incorporation of random effects alongside fixed effects. The GLMM components encompass: 1) fixed effect, which gauges the impact of the predictor variable on the response variable, 2) random effect, aimed at capturing inter-group variability, 3) link function, establishing the relationship between the predictor variable's linear combination and the mean of the response variable, and 4) error distribution, denoting the exponential family distribution of error/residual. The general form of GLMM is equation (1).

$$\begin{aligned} g(\mu_i) &= x_i^T\beta + z_i^T b \\ Y &\sim f(\mu_i, \theta) \end{aligned} \qquad (1)$$

where $g(\mu_i)$ is the link function, $x_i^T\beta$ is a fixed component, z_i^T is a design matrix of random effect, b is vector of random effect (assumed to follow a normal distribution, $b{\sim}N(0,\Sigma)$ where Σ is variance-covariance matrix of the random effects), and $f(\mu_i, \theta)$ is a distribution function of the response variable.

b) The response variable is the incidence rate of dengue fever cases, which generally has continuous and positive values. As a result, it is assumed that the response variable follows a lognormal distribution.
$Y_{ijt}{\sim}\text{Lognormal}(\mu_{ijt}, \sigma^2)$
Here, Y_{ijt} represents the value of the response variable in the i-th observation in the j-th region in the t-th year, and μ_{ijt} represents the mean, while σ^2 denotes the variance. The probability density function of a lognormal distribution can be described in equation (2).

$$f(y_i; \mu, \sigma) = \frac{1}{y_i\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right) \quad , \qquad y_i > 0 \qquad (2)$$

In terms of the link function, it is described by an identity function, $f(Y) = \log Y$.

c) The GLMM model in this study is equation (3).

$$\log(Y_{ijt}) = \beta_0 + \beta_1 X_{ijt}^{(1)} + \beta_2 X_{ijt}^{(2)} + \beta_3 X_{ijt}^{(3)} + \beta_4 X_{ijt}^{(4)} + b_{0j} Z_j^{(0)} + b_{1j} Z_j^{(1)} + \varepsilon_{ijt} \qquad (3)$$

where

- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are parameter of the predictor variable.
- $X_{ijt}^{(k)}$ is the k-th predictor variable in the i-th observation in the j-th region in the t-th year.
- b_{0j} is the random effect for intercept in the j-th region and $b_{0j}{\sim}N(0, \sigma_0^2)$
- b_{1j} is the random effect for slope of year in the j-th region and $b_{1j}{\sim}N(0, \sigma_1^2)$
- $Z_j = \left(Z_j^{(0)}, Z_j^{(1)}\right)$ is design matrix of random effect in the j-th region.
- ε_{ijt} is error in the i-th observation in the j-th region in the t-th year

The GLMM tree is a tree-based algorithm that can detect interactions in GLMM. The GLMM tree algorithm uses the GLM tree algorithm to estimate fixed effects and treats random effects as offsets. The algorithm of the GLM tree with the model's form $g(\mu) = x^T\beta$ is as follows [7][15].

1. Parameter estimation: Initially, the parameters of the GLM model $(\beta_j = \beta)$ are estimated for a single subgroup, assuming homogeneity within the node.
2. Testing instability: The instability of the estimated parameters is assessed across subgroups of partitioning variables $(Z_1, \dots, Z_j)$ using score contributions. The score contribution for a partitioning variable is defined as equation (4). This quantifies the sensitivity of the likelihood function $l((y,x)_i\beta)$ to change in $\beta_{(k)}$ for each subgroup.
$$s_{(k)}\big((y,x)_i\hat{\beta}\big) = \frac{\partial l((y,x)_i\beta)}{\partial \beta_{(k)}} \tag{4}$$
3. Statistical Testing Using M-Fluctuation: To determine the significance of parameter instability, the test statistic evaluates whether the score contributions fluctuate significantly from zero. The null hypothesis $H_0^{\beta(k),j}$ assumes independence (⊥) between the score contributions $S_{(k)}\big((Y,X)_i\hat{\beta}\big)$ and the partitioning variable Z_j. Formally
$$H_0^{\beta(k),j}\colon S_{(k)}\big((Y,X)_i\hat{\beta}\big) \perp Z_j \tag{5}$$
Here, ⊥ signifies that the score contributions are not influenced by the partitioning variable under the null hypothesis. Rejecting H_0 implies that Z_j introduces significant heterogeneity in the model parameters.
4. Variable Selection: If the test is significant for any partitioning variable, the variable Z_j with the lowest p-value is selected. The division point within Z_j that maximizes the likelihood is chosen to split the data into new subgroups.
5. Recursive Partitioning: Steps 1–4 are repeated iteratively until either the null hypothesis $H_0^{\beta(k),j} \forall k, j$ cannot be rejected or other criteria, such as the minimum size of subgroups, are met.

The GLMM tree algorithm [5] is as follows

1. Initialize the value of r and the whole value $\hat{b}_{(r)}$ with a value of 0
2. Update the $r = r + 1$. Estimate GLM tree with $z_i^T \hat{b}_{(r-1)}$ as an offset.
3. Estimate the mixed effect model $g(\mu_{ij}) = x_i^T \beta_j + z_i^T b$ with the terminal nodes $j(r)$ of the GLM tree estimated in step 2. Extract the estimated value $\hat{b}_{(r)}$ from the estimated model.
4. Repeat steps 2 and 3 until they are convergent.

2.3. Analysis Procedure

Analysis of the data was conducted using R software version 4.3.2 and R Studio 2023.09.1, along with the readxl, lme4 [18], glmertree [19], and panelr packages. We used the bobyqa optimizer, which excels in handling large datasets and complex random effects structures by providing efficient, derivative-free optimization that avoids local minima and dynamically adjusts the parameter search space for reliable convergence. The algorithm's convergence was tested using the default tolerance level of 1e-5 for both the fixed-effect coefficients and random effects. The data analysis procedure followed Figure 1.

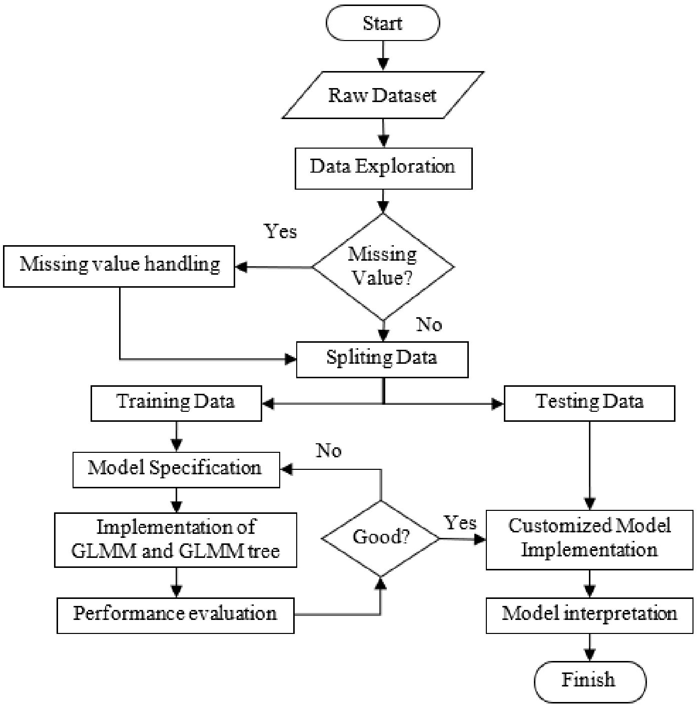

Figure 1. Research flow

The dataset in this study has a panel data structure, comprising annual observations of dengue fever incidence rates across multiple regencies and cities in West Java from 2014 to 2022. To address the issue of missing data, forward-filling was applied, wherein missing values were replaced with the most recent non-missing values from the same variable. This method was selected due to its simplicity and its ability to preserve temporal continuity within the panel structure.

Alternative imputation methods, such as mean imputation, regression-based imputation, or multiple imputation, were considered. However, these approaches were deemed less suitable for this dataset. Mean imputation risks oversmoothing the data and losing temporal variability, while regression and multiple imputation require assumptions about the underlying data distribution, which may not hold for epidemiological data with seasonal patterns. Forward-filling was chosen as it maintains the temporal trend and minimizes the introduction of biases that could distort the panel's temporal dynamics.

Following imputation, the complete dataset was divided into two parts: training data covering the years 2014 to 2021 and test data for the year 2022. This division ensures that the model can be validated on unseen data, simulating its performance in predicting future incidence rates. The training data was employed for constructing GLMM and GLMM tree models, while the test data was used for evaluating model performance. To mitigate overfitting, 10-fold cross-validation was implemented during model training, ensuring that the models were evaluated on different subsets of the training data. Model performance was assessed based on the Root Mean Squared Error (RMSE).

3. RESULT AND DISCUSSION

The research dataset includes information on the incidence rate of dengue fever cases and various predictor variables from 27 regencies/cities in West Java spanning from 2014 to 2022.

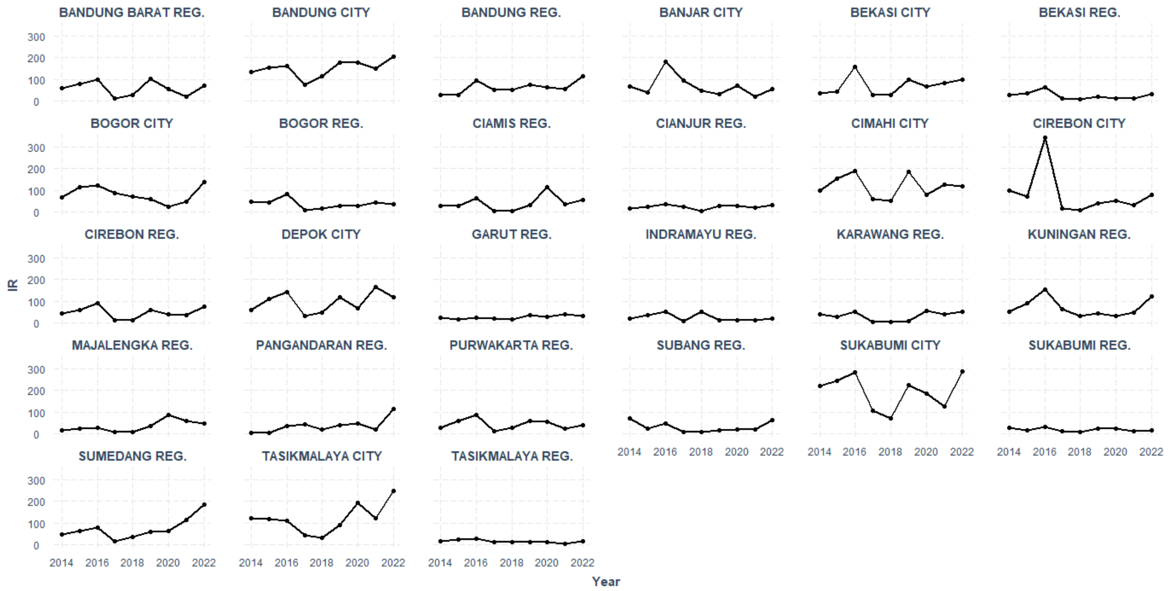

Figure 2. Trend patterns of the incidence rate (IR) of dengue fever in each regency/city

Each subplot in Figure 2 represents a specific regency or city, showcasing temporal trends in the DF incidence rate. The data reveal considerable variability in IR over the years, with some regions experiencing significant fluctuations (e.g., Cirebon City in 2017) and others showing relatively stable trends (e.g., Bandung Regency). Notably, a general increase in the IR is observed in several regions toward the later years of the dataset, particularly around 2022. These patterns suggest potential regional differences in DF risk factors and transmission dynamics, emphasizing the importance of localized public health interventions. These differences form the basis for the assumption that the model exhibits variations in intercepts and slopes at the regency/city level.

The utilization of GLMM and GLMM trees for modeling yields distinct parameter estimates. In contrast to GLMM, which generates a single model, the GLMM tree produces two models distinguished by the partition variable of the GRDP growth rate at constant prices. Figure 3 depicts the decision tree derived from the GLMM tree model, while Table 2 provides a comparison of the model parameter estimation.

Figure 3 represents the results of a GLMM tree model applied to panel data of dengue fever incidence rates in West Java. The tree splits based on the threshold of GRDP growth rate ($p < 0.001$) at 5.5%, dividing the dataset into two subgroups. Node 2 ($n = 121$) for regions with GRDP growth rate ≤ 5.5% and Node 3 ($n = 95$) for regions with GRDP growth rate > 5.5%. For Node 2, the fixed effects indicate a smaller intercept (0.371), alongside positive contributions from population density (0.012), population growth (0.050) and net enrollment ratio (NER) for junior education (0.039), while sanitation has a negative impact (-0.004). In Node 3, regions with higher GRDP growth exhibit a much larger intercept (4.253), with a smaller positive effect from population density (0.011) and a slight negative contribution from NER junior education (-0.003) and sanitation (-0.008). This split highlights the differential impact of socioeconomic factors on dengue fever incidence rates depending on economic growth levels, emphasizing the nuanced relationships captured by the GLMM tree model.

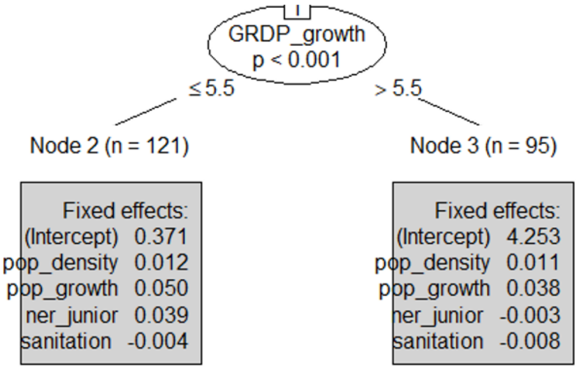

Figure 3. Decision tree of GLMM tree

The parameter estimation presented in Table 2 provides valuable insights into the impact of each predictor variable on the response variable. It is clear from all models that population density and population growth rate have a positive influence on the response variable [20]. Conversely, the household having access to adequate sanitation has been shown to have a negative effect on the response variable. Furthermore, the net enrollment rate for junior secondary education has a positive impact in the GLMM model, and its influence varies depending on the GRDP growth rate in the GLMM tree model

Table 2. Parameter estimation result

Predictor Variable	**GLMM tree**		**GLMM**
	Regency/city with GRDP growth rate ≤ 5.5%	**Regency/city with GRDP growth rate > 5.5 %**	
Intercept	0.371	4.253	**2.364***
Population density	0.012	0.011	**0.011***
Population growth rate	0.050	0.038	0.002
Net enrollment rate for junior secondary education	0.039	-0.003	0.016
Households having access to adequate sanitasion	-0.004	-0.008	-0.003

*Description: * significant at 95% confidence level*

The comparison between the GLMM tree and the GLMM model (in Tabel 2) reveals notable differences in parameter estimates. The intercept values in the GLMM tree vary significantly, with 0.371 for regions with GRDP growth rates ≤ 5.5% and 4.253 for those > 5.5%, while the GLMM model produces a significantly larger and statistically significant intercept of 2.364. Population density shows a positive association across all models, but only the GLMM model's coefficient (0.011) is statistically significant at the 95% confidence level. The population growth rate has higher coefficients in the GLMM tree (0.050 and 0.038 for ≤ 5.5% and > 5.5%, respectively) compared to the GLMM model (0.002), though none are statistically significant. Access to adequate sanitation has a consistently negative effect across models, though not statistically significant. The net enrollment rate for junior secondary education exhibits varying effects in the GLMM tree (0.039 for ≤ 5.5% and -0.003 for > 5.5%), whereas the GLMM model shows a smaller positive effect (0.016), all of which are statistically insignificant. Further understanding indicates that regencies/cities with GRDP growth rates below 5.5% may have inadequate school environments, potentially facilitating the spread of dengue fever. Conversely, regencies/cities with GRDP growth rates above 5.5% might possess better school environments, impeding the transmission of the dengue fever virus. A high-quality school environment can heighten students' awareness in stemming the spread of dengue fever, thereby negatively impacting the rise in dengue fever cases [21]. Contrarily, in the GLMM model, this predictor variable exerts a positive influence across all regencies/cities. These results highlight the GLMM tree's ability to capture regional variations based on GRDP growth rates, while the GLMM model identifies population density as a significant predictor.

In this study, the random effects in GLMM and GLMM tree are regency/city as the intercept random effect and year as the slope random effect for each district/city. The variance related to the impact of random effects on the response variable can be seen in Table 3. The variance of the intercept in the GLMM model is very high, indicating high variability between regencies/cities [19]. According to Gelman and Hill [2], an Intraclass Correlation Coefficient (ICC) close to 1 indicates that almost all the variability in the data is explained by the differences between regencies/cities. It suggests that the random effect for regencies/cities is dominant in explaining the data's variability. On the other hand, the variance for the year is 0.9989, indicating that annual variability is very small compared to the variance between regencies/cities [20]. The residual variance of 1340 indicates considerable variability not explained by random effects.

Table 3. Variance of the random effect

Model	Groups	Name	Variance	Std. Dev	ICC
GLMM	Regency/City	Intercept	4.131×10^{6}	2032.38	0,9997
		Year	0.9989	0.9994	
	Residual		1340	36.6115	
GLMM tree	Regency/City	Intercept	42.3020	6.5040	0,9769
		Year	1.156×10^{-5}	0.0034	
	Residual		1.0000	1.0000	

The random effects plot (see Figure 4) illustrates the variability in intercepts and year effects across regencies/cities, providing a clear validation of the random effects structure. The substantial differences in intercepts highlight the heterogeneity between regions, while the minimal variation in year effects aligns with the low variance estimates for temporal factors in Table 3. The narrow confidence intervals across most estimates indicate stability and reliability, with no apparent outliers that could undermine the model's assumptions. This visualization confirms that the random effects structure effectively captures spatial and temporal variability, supporting the model's validity in analyzing dengue fever incidence in West Java.

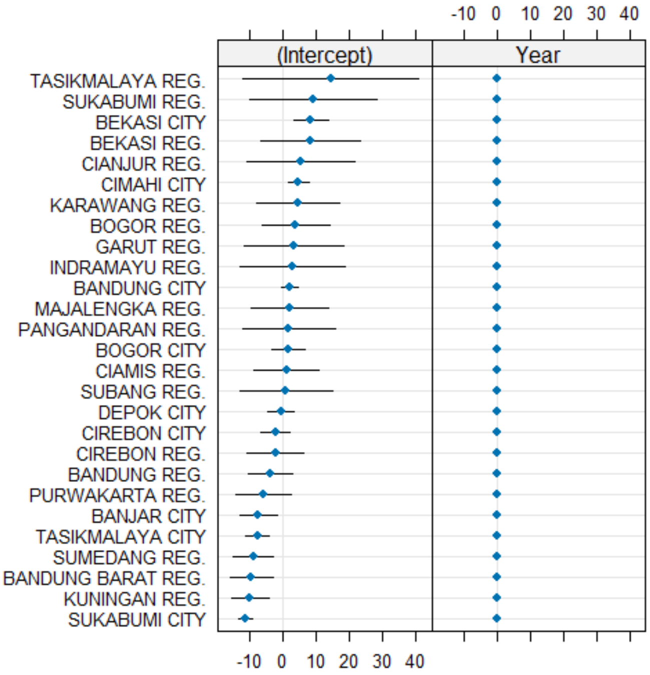

Figure 4. Random Effects Plot for Intercepts and Year Effects Across Regencies/Cities

In the GLMM tree, the variance of the intercept is lower (42.3020) compared to the GLMM model, but the ICC is still high (0.9769). Hothorn and Zeileis [12] explain that this shows the GLMM tree model is also effective in capturing variability between regencies/cities, though not as strongly as the GLMM model. The variance for the year in the GLMM tree is very small, indicating that the year effect is almost non-existent. Meanwhile, the residual variance is 1, indicating that the residual has been normalized or standardized. Based on these findings, it can be inferred that the GLMM model is well-suited for panel data with substantial variation between groups and some temporal fluctuations. This model effectively elucidates group differences [2]. On the other hand, the GLMM tree model is suitable for panel data that exhibit nuanced or intricate variability unrelated to temporal effects. This model offers greater flexibility in capturing unpatterned variations [12].

However, while the GLMM tree provides useful insights into the relationship between predictor variables, such as population density, and the dengue incidence rate, it is important to consider potential confounding factors and interactions. For instance, socioeconomic factors like income levels or healthcare access could interact with population density and growth rate, potentially influencing the incidence of dengue fever. These factors may not be directly accounted for in the model but could affect the observed relationships between predictors and the outcome. Future research should incorporate these interactions and confounders to refine the model and enhance its predictive accuracy. A more thorough understanding of these variables and their interdependencies will help ensure that policy interventions are based on a comprehensive view of the underlying factors driving dengue transmission.

The scatter plot in Figure 5 compares the predicted values from GLMM and GLMM Tree models across various clusters, with the diagonal red line representing perfect prediction (actual = predicted). While both models generally align closely with the actual values, slight differences in performance are observed across regions. The GLMM Tree model (blue points) demonstrates better alignment in clusters with more complex data structures, whereas the GLMM model (green points) occasionally shows deviations, indicating potential overfitting or underfitting in certain clusters. Overall, the GLMM Tree provides a more flexible approach for capturing regional variations, as reflected in its consistent performance near the identity line.

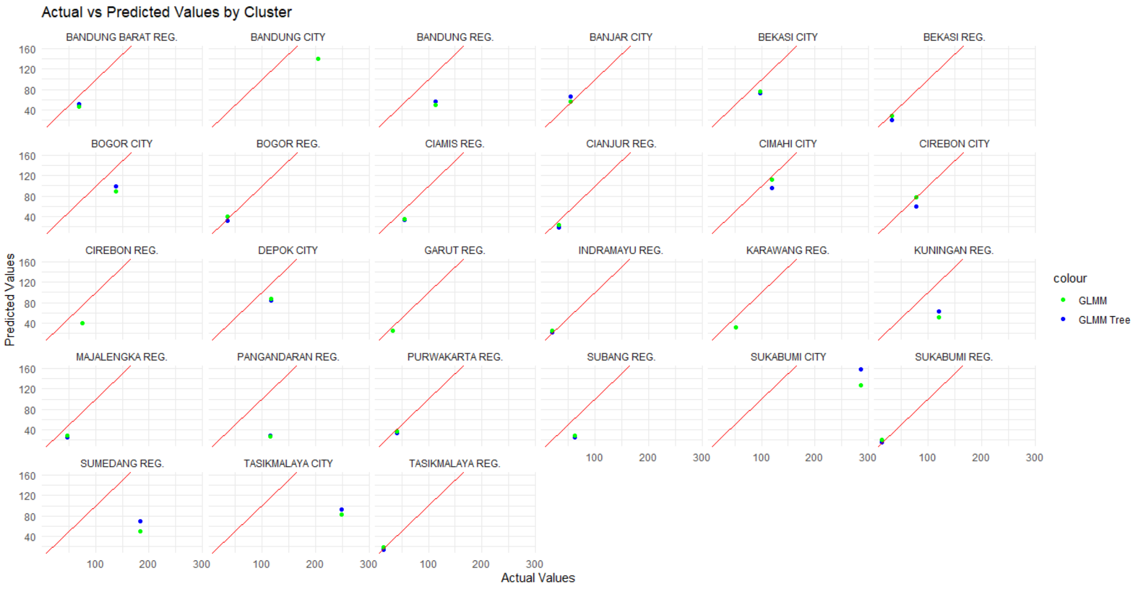

Figure 5. Comparisons between GLMM and GLMM tree

The following assessment of GLMM and GLMM tree models in analyzing panel data sets revolves around their accuracy level. The Root Mean Squared Error (RMSE) is utilized as the evaluation metric, and its values are detailed in Table 4.

Table 4. RMSE of the model

Model	Training Data	Testing Data

GLMM	36.6098	60.7298
GLMM tree	36.2199	54.6767

It can be seen in Table 4 that the GLMM model tends to overfit the training data because the RMSE for the testing data is higher than for the training data. It indicates that the model may be too complex and not generalize to new data. On the other hand, the GLMM Tree model tends to provide better results than the GLMM model, as the RMSE on the test data is lower. It indicates that the GLMM Tree may be better at generalizing and producing more accurate predictions for new data [12]. Therefore, in this case, the GLMM Tree performs slightly better in accuracy than the GLMM, based on RMSE values on test data.

4. CONCLUSION

The results of this study show that the GLMM tree model provides flexibility and expertise in handling panel data with complex or subtle variations that cannot be fully explained by temporal effects alone. Relating to accuracy, the GLMM tree model exhibits robust performance despite requiring significant adjustments when dealing with new data. When developing the incidence rate of the dengue fever model, the GLMM tree separates into two submodels depending on a GRDP growth rate threshold of 5.5%. Among the predictor variables, both population density and population growth rate have a positive influence on the dengue incidence rate, suggesting that regions with higher population density and growth are at greater risk. Conversely, households having access to adequate sanitation have a negative impact on the incidence rate, indicating that improvements in sanitation could be a key factor in controlling dengue fever. The net enrollment rate for junior secondary education is positively linked to regions below the 5.5% growth rate threshold and negatively associated with regencies/cities above the threshold, reflecting the complex role of education in influencing public health outcomes. The GLMM tree model shows significant differences in the incidence rate of dengue fever between regencies/cities in West Java. However, the difference in the incidence rate of dengue fever from year to year between regencies/cities is not significant. This indicates that local factors are more dominant in influencing the incidence rate than other factors outside the predictor variables.

The findings of this study have important implications for public health policy. Regions with higher population density and growth should prioritize interventions to mitigate the risk of dengue fever, including targeted vector control measures and public health campaigns. Additionally, enhancing sanitation infrastructure could be a cost-effective strategy for reducing the incidence of dengue fever.

For future research, it would be valuable to explore the interactions between socioeconomic factors and environmental variables that may influence dengue transmission. Further investigation into how other unobserved factors, such as climate data or local healthcare accessibility, affect dengue incidence could improve the model's predictive accuracy and lead to more tailored interventions.

5. ACKNOWLEDGMENTS

The author would like to thank Indonesian Education Scholarship - Ministry of Education, Culture, Research and Technology (BPI - Kemdikbudristek) for providing a scholarship so that the author was able to continue his studies and write this article.

REFERENCES

[1] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefˋevre, and M. K¨orner, “BREIZHCROPS: A TIME SERIES DATASET FOR CROP TYPE MAPPING,” *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, 2019.

[2] A. Gelman and J. L. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

[3] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith, *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media, 2009.

[4] B. M. Bolker *et al.*, “Generalized linear mixed models: a practical guide for ecology and evolution,” *Trends Ecol. Evol.*, vol. 24, no. 3, pp. 127–135, 2009, doi: 10.1016/j.tree.2008.10.008.

[5] M. Fokkema, N. Smits, A. Zeileis, T. Hothorn, and H. Kelderman, “Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees,” *Behav. Res. Methods*, vol. 50, no. 5, pp. 2016–2034, Oct. 2018, doi: 10.3758/s13428-017-0971-x.

[6] M. Fokkema, J. Edbrook-Childs, and M. Wolpert, “Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data,” *Psychother. Res.*, vol. 31, no. 3, pp. 329–341, 2021, doi: 10.1080/10503307.2020.1785037.

[7] B. Suseno, K. A. Notodiputro, and B. Sartono, “GLMM and GLMM Tree for Modeling Poverty in Indonesia,” 2023.

[8] Dinas Kesehatan, “Jumlah Kasus Demam Berdarah Dengue (DBD) Berdasarkan Jenis Kelamin di Jawa Barat,” *opendata.jabarprov.go.id*, 2024. https://opendata.jabarprov.go.id/id/dataset/jumlah-kasus-demam-berdarah-dengue-dbd-berdasarkan-jenis-kelamin-di-jawa-barat (accessed May 15, 2024).

[9] Z. Rafifah, R. Anisa, and Erfiani, “Penerapan Regresi Binomial Negatif untuk Mengatasi Overdispersi pada Regresi Poisson Kasus Demam Berdarah di Jawa Barat,” IPB University, 2022.

[10] Z. Martha, B. Susetyo, and M. N. Aidi, “Pemodelan Regresi Data Panel Pada Kasus Jumlah Penderita Demam Berdarah Dengue (Dbd) Di Kota Bogor.,” IPB University, 2015.

[11] M. Y. N. Prisie, “Kemenkes: Kasus DBD tahun 2023 turun 30 persen dari tahun sebelumnya,” *www.antaranews.com*, 2024. https://www.antaranews.com/berita/4021911/kemenkes-kasus-dbd-tahun-2023-turun-30-persen-dari-tahun-sebelumnya (accessed May 24, 2024).

[12] T. Hothorn and A. Zeileis, “Partykit: a modular toolkit for recursive partytioning in R,” *J. Mach. Learn. Res.*, vol. 16, pp. 3905–3909, 2015, doi: 10.5555/2789272.2912120.

[13] A. Suwandono, *Dengue Update: Menilik Perjalanan Dengue di Jawa Bara*. Jakarta: LIPI Press, 2019.

[14] H. P. Astuti, A. Adyas, and A. Djamil, “Analisis faktor yang berhubungan dengan kejadian demam berdarah dengue di kota Bandar Lampung tahun 2023,” *Sanitasi J. Kesehat. Lingkung.*, vol. 16, no. 2, 2023, doi: 10.29238/sanitasi.v16i2.1855.

[15] A. Zeileis, T. Hothorn, and K. Hornik, “Model-Based Recursive Partitioning,” *J. Comput.*

Graph. Stat., vol. 17, pp. 492–514, 2008, doi: 10.1198/106186008X319331.

[16] M. Basili and F. Belloc, “HOW TO MEASURE THE ECONOMIC IMPACT OF VECTOR-BORNE DISEASES AT COUNTRY LEVEL,” *J. Econ. Surv.*, vol. 29, no. 5, pp. 896–916, 2014, doi: doi:10.1111/joes.12075.

[17] C. E. Ross and C. Wu, “The Links Between Education and Health,” *Am. Sociol. Rev.*, vol. 60, no. 5, 1995, doi: https://doi.org/10.2307/2096319.

[18] D. M. Bates, M. Machler, B. M. Bolker, and S. C. Walker, “Fitting Linear Mixed-Effects Models Using lme4,” *J. Stat. Softw.*, vol. 67, pp. 1–48, 2014, doi: 10.18637/jss.v067.i01.

[19] M. Fokkema and A. Zeileis, “Fitting Generalized Linear Mixed-Effects Model Trees,” 2019.

[20] A. Ruliansyah, Y. Yuliasih, and S. Hasbullah, “Pemanfaatan Citra ASTER Dalam Penentuan Dan Verifikasi Daerah Rawan Demam Berdarah Dengue (DBD) Di Kota Banjar Provinsi Jawa Barat,” *ASPIRATOR - J. Vector-borne Dis. Stud.*, vol. 6, pp. 55–62, 2015, doi: 10.22435/ASPIRATOR.V6I2.3631.55-62.

[21] Y. Zhang *et al.*, “Knowledge, attitude and practice (KAP) and risk factors on dengue fever among children in Brazil, Fortaleza: A cross-sectional study,” *PLoS Negl. Trop. Dis.*, vol. 17, no. 9, 2023, doi: 10.1371/journal.pntd.0011110.