



GLMMTree for Modelling Poverty in Indonesia

B Suseno^{1,*}, K A Notodiputro¹, B Sartono¹

¹ Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Indonesia

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

Abstract. GLMMTree is a tree-based algorithm that can detect interaction and find subgroups in the GLMM to improve fixed effect estimation. This study uses GLMMTree for the actual data applications of poverty in Indonesia and confirms that the GLMMTree algorithm method has better precision than GLMM. The significant predictors that affect poverty in Indonesia are the unemployment rate and the GRDP at a constant price. GLMMTree algorithm enriches the analysis by finding subgroups of provinces with electricity lighting access and clean drinking water sources variables.

1. Introduction

Poverty eradication is one of the national goals of the Republic of Indonesia, as stated in the fourth paragraph of the Preamble to the 1945 Constitution of the Republic of Indonesia. The government of the Republic of Indonesia has shown efforts to eradicate poverty through a commitment to achieve 2030's Sustainable Development Goals (SDGs), where the first goal is to eradicate poverty.

Based on the data from Statistics Indonesia (BPS), from 2010 to 2019, the percentage of poor people in Indonesia tended to decrease. The lowest poverty level was achieved in 2019, with the percentage of poor people being 9.41 percent, and unfortunately, it increased in the following years until 2021, when it became 10.14 percent. The percentages of poor people were diverse at the provincial level, with the highest occurring in Papua province, 26.86 percent, and the lowest occurring in Bali province, 4.53 percent. Many researchers have modeled the poverty panel data in Indonesia [2][5][6] using the panel regression model and Generalized Linear Mixed Models (GLMM).

The GLMMs are statistical models that handle the response variables with probability distribution from the exponential family, accommodate nonlinear models, and model correlated data [8]. GLMM can model covariates as fixed effects or random effects. A recent development of GLMM was combining the GLMM with a tree algorithm called the GLMMTree.

Model-based trees are used to find subgroups in data that differ concerning model parameters [9]. Model-based trees are beneficial because they can handle many potential predictor variables and automatically detect interactions between them [3]. The GLMMTree is a tree-based algorithm that allows for the detection of treatment-subgroup interactions while accounting for the clustered structure of a dataset. The algorithm uses model-based recursive partitioning to detect treatment-subgroup interactions and a GLMM to estimate the random-effects parameters [3]. We found that many studies used the GLMMTree with the health field dataset [3] [4] [13]. However, we have yet to find research that used the GLMMTree for socioeconomic datasets in their study. This study is filling this gap by



performing an empirical analysis of the GLMMTree to Indonesia's socioeconomic datasets applications of poverty in Indonesia.

The remainder of this paper is organized as follows. Section 2 presents some literature reviews of GLMM and GLMMTree. Section 3 provides the analysis step in the empirical study using poverty data in Indonesia. Section 4 analyzes and compares the GLMM and GLMMTree. Finally, Section 5 highlights the conclusions and states some points for future research.

2. Literature Review

2.1. Generalized Linear Mixed Models (GLMMs)

GLMMs are statistical models that handle the response variables with probability distribution from the exponential family, accommodate nonlinear models, and model correlated data [8]. The covariates in the GLMM can be treated as fixed and random effects. The fixed effect also called the regression coefficient or fixed effect parameter, explains the relationship between the response variable and the predictor variable for the entire population unit. A random effect is a random value realization corresponding to the level of a random factor that describes the random deviation of the relationship described by the fixed factor. Random effects describe the effect of a particular cluster or subject in a population, so it can be used to model the random diversity of the response variable for each different level of the data. A random effect can be a random intercept indicating a random deviation for a specific subject or cluster of a whole fixed intercept or in the form of a random coefficient indicating deviation random for a particular subject or cluster of an overall fixed effect.

Suppose that a random variable Y has an exponential family probability distribution as in equation (1) with an expectation value $E(Y) = \mu$ and has a variance $Var(Y) = \sigma^2$, so that the log of distribution of Y as in equation (2)

$$f(y|\theta) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right] \quad (1)$$

$$\log[f(y|\theta)] = \ell(\theta; y, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \quad (2)$$

The functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are an arbitrary known function with θ as the canonical parameter and ϕ as the scale parameter. The GLMM in matrix notation is as in equation (3).

$$g(\boldsymbol{\mu}_i|\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad (3)$$

where $\boldsymbol{\mu}_i|\mathbf{b}_i = E(\mathbf{Y}_i|\mathbf{b}_i)$, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$ and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$

The GLMMs structure consists of four components as follows

1. The linear predictors $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$
2. The distribution of response variable conditional on random effects $\mathbf{Y}_i|\mathbf{b}_i$
3. The link function $\boldsymbol{\eta} = g(\boldsymbol{\mu}_i|\mathbf{b}_i)$ where $\boldsymbol{\mu}_i|\mathbf{b}_i = E(\mathbf{Y}_i|\mathbf{b}_i)$
4. The distribution of random effects $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$

The response variable \mathbf{Y}_i in vector notation for the subject i is as in equation (4)

$$\mathbf{Y}_i^T = (Y_{1i} \ Y_{2i} \ \dots \ Y_{n_i i}) \quad (4)$$

The fixed effect design matrix \mathbf{X}_i with p covariate has a size of $n_i \times p$ as in equation (5)

$$\mathbf{X}_i = \begin{pmatrix} X_{1i}^{(1)} & X_{1i}^{(2)} & \dots & X_{1i}^{(p)} \\ X_{2i}^{(1)} & X_{2i}^{(2)} & \dots & X_{2i}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n_i i}^{(1)} & X_{n_i i}^{(2)} & \dots & X_{n_i i}^{(p)} \end{pmatrix} \quad (5)$$

The regression coefficient or fixed effect parameter $\boldsymbol{\beta}$ in vector notation as in equation (6)



$$\boldsymbol{\beta}^T = (\beta_1 \quad \beta_2 \quad \dots \quad \beta_p) \tag{6}$$

The random effect design matrix \mathbf{Z}_i with q covariate has a size of $n_i \times q$ as in equation (7)

$$\mathbf{Z}_i = \begin{pmatrix} Z_{1i}^{(1)} & Z_{1i}^{(2)} & \dots & Z_{1i}^{(q)} \\ Z_{2i}^{(1)} & Z_{2i}^{(2)} & \dots & Z_{2i}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n_i i}^{(1)} & Z_{n_i i}^{(2)} & \dots & Z_{n_i i}^{(q)} \end{pmatrix} \tag{7}$$

The random effect parameter vector \mathbf{b}_i for subject i is assumed to have a Multivariate Normal distribution with a mean vector of $\mathbf{0}$ and covariance matrix \mathbf{D} as in equation (8).

$$\mathbf{b}_i^T = (b_{1i} \quad b_{2i} \quad \dots \quad b_{qi}) \tag{8}$$

The elements of the main diagonal of matrix \mathbf{D} show the variance of each random effect \mathbf{b}_i , while the elements outside the main diagonal are the covariance between two random effects. The matrix \mathbf{D} is a symmetrical matrix of size $q \times q$ as in equation (9).

$$\mathbf{D} = Var(\mathbf{b}_i) = \begin{pmatrix} Var(b_{1i}) & cov(b_{1i}, b_{2i}) & \dots & cov(b_{1i}, b_{qi}) \\ cov(b_{1i}, b_{2i}) & Var(b_{2i}) & \dots & cov(b_{2i}, b_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(b_{1i}, b_{qi}) & cov(b_{2i}, b_{qi}) & \dots & Var(b_{qi}) \end{pmatrix} \tag{9}$$

The error vector $\boldsymbol{\varepsilon}_i$, whose individual element denotes the residual for the subject i , is assumed to have a Multivariate Normal distribution with a mean vector of $\mathbf{0}$ and a covariance matrix \mathbf{R}_i as in equations (10) and (11). These errors for the same subject can be correlated.

$$\boldsymbol{\varepsilon}_i^T = (\varepsilon_{1i} \quad \varepsilon_{2i} \quad \dots \quad \varepsilon_{n_i i}) \tag{10}$$

$$\mathbf{R}_i = Var(\boldsymbol{\varepsilon}_i) = \begin{pmatrix} Var(\varepsilon_{1i}) & cov(\varepsilon_{1i}, \varepsilon_{2i}) & \dots & cov(\varepsilon_{1i}, \varepsilon_{n_i i}) \\ cov(\varepsilon_{1i}, \varepsilon_{2i}) & Var(\varepsilon_{2i}) & \dots & cov(\varepsilon_{2i}, \varepsilon_{n_i i}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\varepsilon_{1i}, \varepsilon_{n_i i}) & cov(\varepsilon_{2i}, \varepsilon_{n_i i}) & \dots & Var(\varepsilon_{n_i i}) \end{pmatrix} \tag{11}$$

2.2. GLMM Estimation Methods

The GLMM parameter could be estimated using the likelihood function [10]. The likelihood function of GLMM is in equation (12).

$$L(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{D}|\mathbf{y}) = \int \prod_{i=1}^n f_{y_i|b}(\mathbf{y}|\mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\phi}) f_b(\mathbf{b}|\mathbf{D}) d\mathbf{b} \tag{12}$$

This function usually does not have a closed form except for the Normal distributed response variable because of the high dimensional integration for all levels of random effects \mathbf{b} [18]. This integration could be approximated by using Laplace approximation. Suppose that a form of function integration is in equation (13).

$$\int \exp\{-q(x)\} dx \tag{13}$$

With $q(\cdot)$ is a differentiable function with a minimum value for $x = x^*$ with $q'(x^*) = 0$ and $q''(x^*) > 0$. The Taylor series expansion of this function is as in equation (14).

$$q(x) = q(x^*) + \frac{1}{2}q''(x^*)(x - x^*)^2 + \dots, \tag{14}$$



So, the approximated value is as in equation (15).

$$\int \exp\{-q(x)\} dx \approx \sqrt{\frac{2\pi}{q''(x^*)}} \exp\{-q(x^*)\} \quad (15)$$

2.3. Generalized Linear Mixed Model Tree (GLMMTree)

The GLMMTree is a tree-based algorithm that can detect interactions in GLMM. The GLMM Tree algorithm was developed based on model-based recursive partitions capable of detecting subgroups in clustered data structures [3].

In the estimation process, the fixed effect on GLMM may be less than optimal even though the random effect component has accommodated the structure of clustered or nested data due to interactions with other variables outside the model. The GLMMTree's ability to detect interactions will improve GLMM estimation by partitioning the fixed effect component locally with a random effect globally.

$$g(\mu_{ij}) = \mathbf{x}_i^T \boldsymbol{\beta}_j + \mathbf{z}_i^T \mathbf{b}$$

The GLMMTree model treats the fixed effect as a local parameter that depends on the terminal node j . However, the random effect \mathbf{b} is treated as a global parameter. The GLMMtree 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 $g(\mu) = \mathbf{x}^T \boldsymbol{\beta}_j$ is as follows.

1. Estimate GLM model parameters on a single subgroup ($\boldsymbol{\beta}_j = \boldsymbol{\beta}$) in the node
2. Test the instability of the model parameters on each subgroup of the partition variable Z_1, \dots, Z_j
Calculate score contribution as in equation (16)

$$s_{(k)}((y, \mathbf{x})_i \hat{\boldsymbol{\beta}}) = \frac{\partial l((y, \mathbf{x})_i \hat{\boldsymbol{\beta}})}{\partial \beta_{(k)}} \Big|_{\hat{\boldsymbol{\beta}}} \quad (16)$$

Test scores fluctuate from 0 for each partition variable using M-fluctuation testing [14]

$$H_0^{\beta_{(k)},j} : S_{(k)}((Y, \mathbf{X})_i \hat{\boldsymbol{\beta}}) \perp Z_j$$

3. If the test is significant at a particular level, choose the variable Z_j with the least p-value.
4. Select a division point in the partition variable that maximizes the likelihood.
5. Iterate steps 1-4 until it cannot be rejected or other criteria $H_0^{\beta_{(k)},j} \forall k, j$ are met (such as the minimum size of subgroups)

The GLMMTree algorithm is as follows.

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

2.4. Percentage of Poor Population (P0)

Statistics Indonesia uses the ability to meet basic needs or the basic needs approach in measuring poverty. By using this approach, poverty could be seen as the inability from the economic side to meet the basic needs of food and non-food as measured in terms of expenditure. Residents are categorized as poor if they have an average per capita monthly expenditure below the poverty line.

The Poverty Line reflects the value of the minimum expenditure required for a person to meet the basic needs for a month, both food and non-food. The poverty line consists of the Food and Non-Food Poverty Line. The food poverty line is the minimum expenditure value for basic food needs, equivalent to 2100 kilocalories per capita per day. The commodity package of basic food needs is represented by



52 commodities, such as grains, tubers, fish, meat, eggs and milk, vegetables, nuts, fruits, oils, and fats. The Non-Food Poverty Line is the minimum expenditure value for non-food basic needs such as housing, clothing, education, and health. The commodity package for non-food basic needs is represented by 51 types of commodities in urban areas and 47 types of commodities in rural areas.

The formula to calculate the poverty line is as in equation (17)

$$GK = GKM + GKNM \quad (17)$$

GK = Poverty Line

GKM = Food Poverty Line

$GKNM$ = Non-Food Poverty Line

The formula to calculate the percentage of poor people is as in equation (18)

$$P_0 = \frac{1}{n} \sum_{i=1}^q \left(\frac{z - y_i}{z} \right)^0 \quad (18)$$

Where:

P_0 = percentage of poor population

z = Poverty line

y_i = Average per capita expenditure a month of population below the poverty line ($i=1, 2, 3, \dots, q$), $y_i < z$

q = The large number of people who are below the poverty line.

n = number of residents.

Research on factors affecting Indonesia's poverty has been carried out. Cassandra et al. found that the provincial minimum wage, open unemployment rate, gross regional domestic product (GRDP), and human development index (HDI) affect the poverty rate in Indonesia [2]. Hapsari et al. found that the open unemployment rate and provincial minimum wage positively affect the poverty rate in Indonesia. In contrast, the average length of schooling and life expectancy negatively affect the poverty rate in Indonesia [5].

2.5. Evaluation Measure

The evaluation measure used to assess the performance of the estimation method is the root mean squared error (RMSE) value as in equation (19).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2} \quad (19)$$

Where e_i is the residual of the i observations and n is the number of observations.

3. Methodology

3.1. Data

This study used longitudinal secondary data from Statistics Indonesia (BPS) and the Ministry of Investment (BKPM). In this case, the subject is the province that counts as 33 provinces in Indonesia that were observed annually from 2010 to 2021.

The response variable used is the percentage of the poor population as poverty variables. The predictor variables are seven variables based on previous research [2][5][6] that have a significant effect on poverty in Indonesia, namely the mean years of schooling, the expected years of schooling, the life expectancy rate at birth, the unemployment rate, the gross regional domestic product (GRDP) based on constant prices, and years that are assumed linearly as fixed effect and province variable are treated as a random effect. In addition, eight variables are used as partition variables based on previous research that are expected to interact with variables in the model. The partition variables are the percentage of



households with electricity lighting sources, the percentage of households that had access to proper sanitation, the percentage of households with the status of owning a self-owned house, the percentage of households with a clean source of drinking water, provincial minimum wage, gini ratio, the realization of foreign investment, the realization of domestic investment, construction cost index, and the number of working people.

The variables sourced from BPS were 14 socioeconomic indicators and two from BKPM administrative data, namely the realization of foreign and domestic investments. The list of variables used in this study can be seen in Table 1.

Table 1. List of variables used in the study

Types of Variables	Names of Variables	Types of Data	Units	Data sources
Response	1. Percentage of the poor population (Y)	Numerical	%	BPS
Fixed Effect	1. GRDP on a constant price basis 2010 (X ⁽¹⁾)	Numerical	Quadrillion rupiah	BPS
	2. Open Unemployment Rate (X ⁽²⁾)	Numerical	%	BPS
	3. Life Expectancy At Birth (X ⁽³⁾)	Numerical	Year	BPS
	4. Mean Years of Schooling (X ⁽⁴⁾)	Numerical	Year	BPS
	5. Expected Years of Schooling (X ⁽⁵⁾)	Numerical	Year	BPS
	6. Years (X ⁽⁶⁾)	Ordinal	-	BPS
Random Effect	1. Province (Z ⁽¹⁾)	Categorical	-	BPS
Partition	1. Realization of Foreign Investment	Numerical	Million US\$	BKPM
	2. Realization of Domestic Investment	Numerical	Billion Rupiah	BKPM
	3. Household percentage with electricity lighting source	Numerical	%	BPS
	4. Percentage of households having access to proper sanitation	Numerical	%	BPS
	5. Percentage of households with the status of owned house ownership	Numerical	%	BPS
	6. Percentage of households with clean drinking water sources	Numerical	%	BPS
	7. Provincial minimum wage	Numerical	Million Rupiah	BPS
	8. Gini Ratio	Numerical	-	BPS

Source: BPS and BKPM

3.2. Methods

The stages of the analysis, as shown in Fig. 1, are as follows:

1. At the beginning stage, the dataset exploration is carried out to explore the structure and to examine the problem of the dataset.
2. Next, the specification of the model is determined based on previous studies.

The response variable is assumed to have binomial probability distribution $y_{ti}|b_i \sim \text{Binomial}(n_{ti}, \pi_{ti}|b_i)$

Where y_{it} is the number of poor people of the i -th province and t -th year, n_{it} is the total population of the i -th province and t -th year, and b_i is the random effect of the i -th province that is assumed to be Normally distributed $b_i \sim N(\mathbf{0}, \mathbf{D})$.

The logit link function: $\eta_{ti} = \ln\left(\frac{\pi_{ti}|b_i}{1-\pi_{ti}|b_i}\right)$

The linear predictors component assumes the intercept and slope of the model vary in different provinces.

The GLMM model is as in equation (20).

$$\ln\left(\frac{\pi_{ti}|b_i}{1-\pi_{ti}|b_i}\right) = \beta_0 + \beta_1 X_{ti}^{(1)} + \beta_2 X_{ti}^{(2)} + \beta_3 X_{ti}^{(3)} + \beta_4 X_{ti}^{(4)} + \beta_5 X_{ti}^{(5)} + \beta_6 X_{ti}^{(6)} + b_{0i} Z_{ti}^{(0)} + b_{1i} Z_{ti}^{(1)} + \varepsilon_{ti} \quad (20)$$



The GLMMtree model is as in equation (21).

$$\ln\left(\frac{\pi_{ti}b_i}{1-\pi_{ti}b_i}\right) = \beta_{0j} + \beta_{1j}X_{tij}^{(1)} + \beta_{2j}X_{tij}^{(2)} + \beta_{3j}X_{tij}^{(3)} + \beta_{4j}X_{tij}^{(4)} + \beta_{5j}X_{tij}^{(5)} + \beta_{6j}X_{tij}^{(6)} + b_{0i}Z_{ti}^{(0)} + b_{1i}Z_{ti}^{(1)} + \varepsilon_{ti} \quad (21)$$

With j is the node and ε_{ti} is the error assumed to be normally distributed $\varepsilon_{ti} \sim N(\mathbf{0}, \mathbf{R}_i)$.

3. The implementation of the GLMM and the GLMMTree was carried out. The implementation is done by bootstrapping the data with the number of iterations 50 times.
4. The performance evaluation is based on the RMSE value.
5. The interpretation of the effect of the predictor variable is carried out on the response variable to the best estimation method.

The analysis in this study was carried out using Microsoft Excel software, R version 4.2.0, and R Studio 2022.02. 2 Build 485 using *package* lme4, glmertree, ggplot2, and readxl.

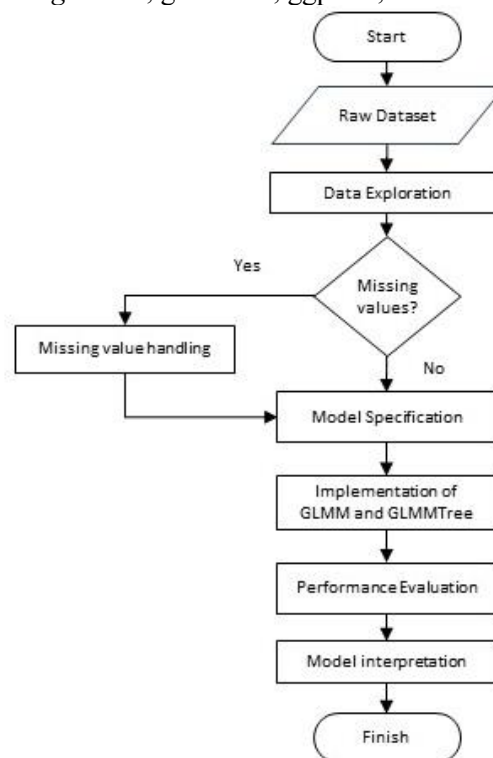


Figure 1. Research Flow

4. Results and Discussion

The dataset has complete variables observation for all provinces. The number of provinces is 33 observed from 2010 to 2021 (12 years). The time series figure of the percentage of poor people by provinces in Indonesia from 2010 to 2021 is shown in Figure 2. It showed that the variables tend to decrease with different trends and levels in each province. This pattern gives evidence to assume varying intercepts and slopes in the model.

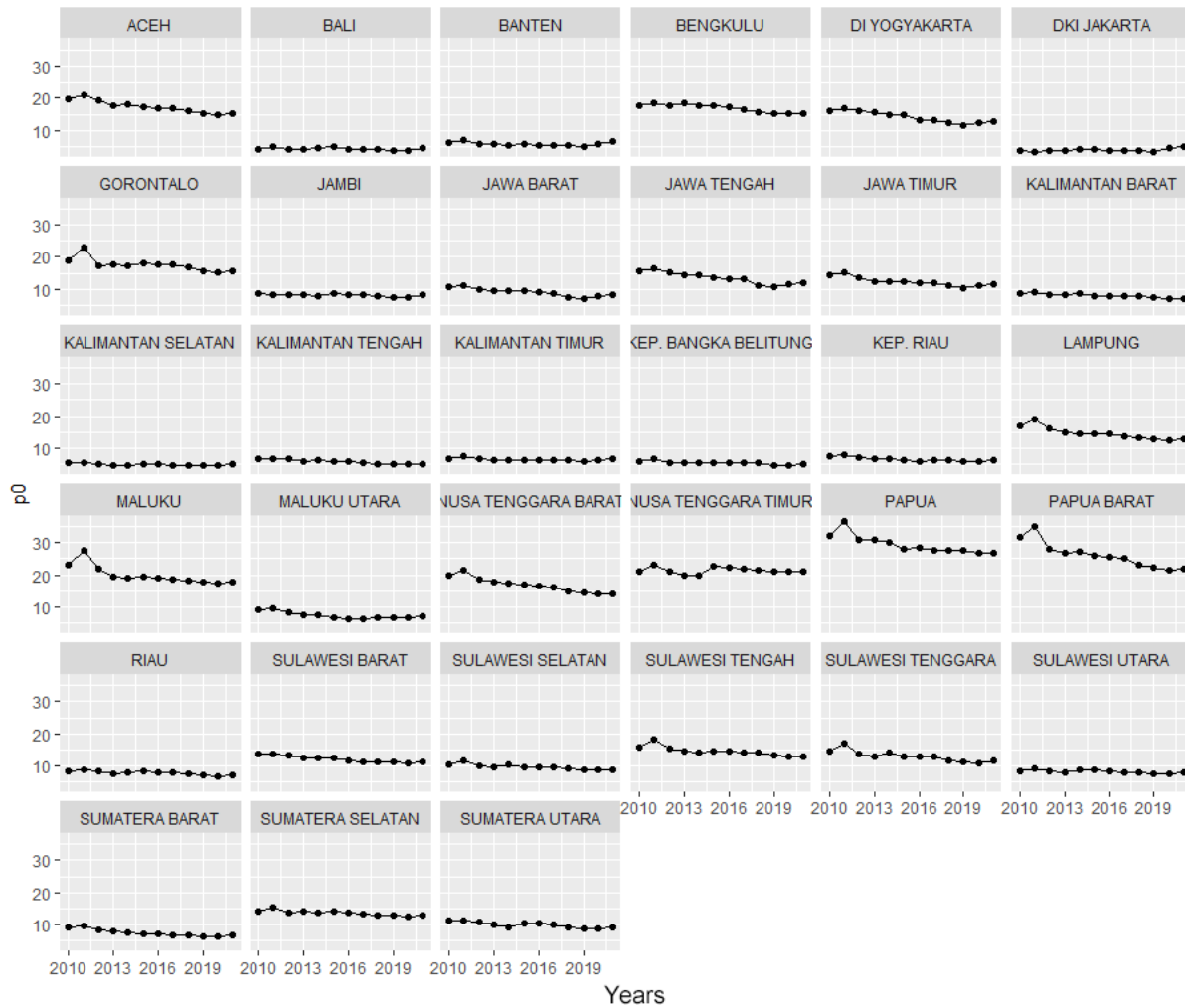


Figure 2. Percentage of poor people by province in the periods of 2010-2021 (%)

Using the GLMM and GLMMTree model specification in equations (20) and (21), the parameter estimations were carried out using the Laplace approximation and GLMMTree algorithm. The evaluation stage is carried out by using the RMSE value, as shown in Table 2.

Table 2. Evaluation of Estimation Methods

Estimation Methods	RMSE
GLMM	1.1576 ±0.0738
GLMMTree	1,0038 ±0.0956

Based on the RMSE value, the GLMMTree method gives a smaller RMSE value than GLMM, though not statistically significant. These results confirm that the GLMMTree could improve the GLMM estimation's precision by finding the variables' subgroups or interactions.

The results of parameter estimation can be seen in Table 3 and Figure 3. Estimating parameters in the GLMM model showed that two significant predictors with a 95 percent confidence level are the open unemployment rate and the GRDP at constant market prices. The open unemployment rate has an estimate of 0.0201, which indicates that an increase in the open unemployment rate by 1% will increase the percentage of the poor population by 0.0201 percent. The GRDP at constant market prices is estimated at -0.7168, indicating that an increase in GRDP at constant market prices by one quadrillion rupiah will reduce poverty by 0.7168 percent.



The significant parameters estimate coefficient of the GLMMTree is based on the GLMM estimation method. The parameter estimate of the open unemployment rate has the lowest value at 0.012 in the subgroup provinces, with the percentage of households with electricity lighting sources below 98.46 percent and those with clean drinking water sources below 55.06 percent. In these relatively poor infrastructure provinces, which is reflected by low electricity infrastructure and drinking water sources, the effect of the unemployment rate is lower than in the subgroup of provinces that have better infrastructure because they usually live in rural areas with lower unemployment rates and living costs and tends to work in any field.

The coefficient of the open unemployment rate variable has the highest value of 0.035 in the subgroup provinces, with the percentage of households with electricity lighting sources of more than 98.46 percent, indicating that the open unemployment rate has a higher effect on poverty in the subgroup provinces with relative better infrastructure conditions, like urban areas, because of limited job available to accommodate their skills. It is a problematic situation for unemployment in this subgroup because of the higher living cost in this subgroup province.

The effect of GRDP at constant market prices on poverty has the highest value at -0.2509 in the subgroup provinces, with the criteria of the percentage of households with electricity lighting sources of more than 98.46 percent, indicating that the change in GRDP at constant market prices has higher effect of reducing the percentage of poor people in the subgroup provinces that have relatively better infrastructure conditions. These are usually urban areas that have a higher living cost. Economic development will improve people's income and have a higher effect on reducing poverty.

Table 3. Parameter Estimation Results

Predictor Variables	GLMMTree			GLMM
	Node 3 The percentage of households with electricity lighting sources \leq 98.46%, and the percentage of households with clean drinking water sources \leq 55.06%	Node 4 The percentage of households with electricity lighting sources \leq 98.46%, and the percentage of households with clean drinking water sources $>$ 55.06%	Node 5 The percentage of households with electricity lighting sources $>$ 98.46%	
Intercept	2,3448	2,0185	-1.4340	0.6772
Mean Years of Schooling	0.0840	0.0171	-0.0873	-0.0362
Expected Years of Schooling	-0.0590	-0.0447	-0.0097	-0.0292
Life expectancy	-0.0629	-0.0534	0.0002	-0.0292
Unemployment Rate	0.0116	0.0247	0.0348	0.0201*
GRDP at constant market prices	-0.0986	-0.0567	-0.2509	-0.7168*
Year	-0.0069	-0.0113	0.0081	0.0042

Description: *Significant at 95% confidence level

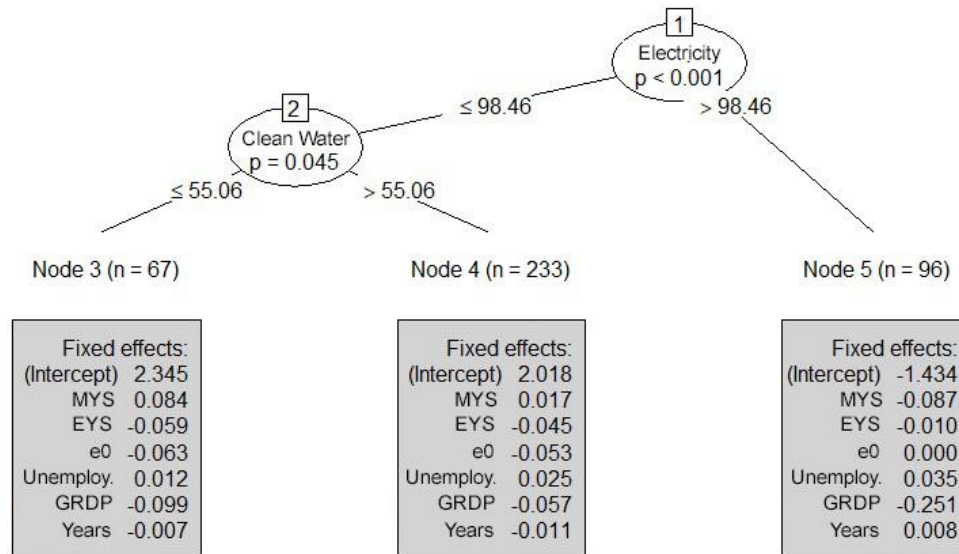


Figure 3. Parameter Estimation of the GLMMTree

5. Conclusion and Future Works

This study successfully applies the GLMMTree algorithm to the socioeconomic data in Indonesia. It confirms that the GLMMTree improves the GLMM estimation's precision by finding the variables' subgroups or interactions. The two significant predictor variables are the open unemployment rate and the GRDP at constant market prices. The GLMMTree can find the subgroup of provinces that interact with predictor variables, enriching the analysis.

As a suggestion for further development, the estimation method with the GLMMTree needs to be developed to give information on statistically significant predictor variables. This method can be improved using machine learning methods like Random Forest or Boosting algorithm.

Acknowledgment

The authors thank the reviewers for their comments and BPS-Statistics Indonesia for funding this research.

References

- [1] Boos, Dennis & Stefanski, Len. (2013). Essential Statistical Inference, Theory and Methods. 10.1007/978-1-4614-4818-1.
- [2] Cassandra. 2016. Analysis of Factors Affecting the Poverty Rate in Indonesia (Period 2008-2013) [Thesis]. Bogor (ID): IPB University. Cassandra. 2016. Analysis of Factors Affecting the Poverty Rate in Indonesia (Period 2008-2013) [Thesis]. Bogor (ID): IPB University.
- [3] Fokkema M, Smits N, Zeileis A, Hothorn T, Kelderman H. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav Res Methods*. 2016 Oct;50(5):2016-2034. doi: 10.3758/s13428-017-0971-x. PMID: 29071652.
- [4] Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. (2021). Generalized linear mixed model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy research: journal of the Society for Psychotherapy Research*, 31(3), 313–325. <https://doi.org/10.1080/10503307.2020.1785037>
- [5] Hapsari, AND. 2019. Factors Affecting the Poverty Rate in Indonesia (Period 2010-2017) [Thesis]. Bogor (ID): IPB University.
- [6] Khairi, A. 2017. Compressed Linear Mixed Model With Regional And Time Effects For Poverty Data Analysis In Aceh Province [Thesis]. Bogor (ID): IPB University.



- [7] McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Second Edition, Chapman & Hall.
- [8] McCulloch, C.E. & Neuhaus, J.M.. (2015). *Generalized Linear Mixed Models*. 10.1016/B978-0-08-097086-8.42017-9.
- [9] Seibold, Heidi & Hothorn, Torsten & Zeileis, Achim. (2018). *Generalised Linear Model Trees with Global Additive Effects*. *Advances in Data Analysis and Classification*. 13. 10.1007/s11634-018-0342-1.
- [10] Stroup, W.W. (2012). *Generalized Linear Mixed Models: Modern Concepts, Methods, and Applications* (1st ed.). CRC Press. <https://doi.org/10.1201/b13151>
- [11] Suhardin, John. (2012). The Role of the State and Law in Eradicating Poverty By Realizing General Welfare. *Journal of Law & Development*. 42. 302. 10.21143/jhp.vol42.no3.274.
- [12] Twisk, J. (2013). *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide* (2nd ed.). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139342834
- [13] Wei, Y., Liu, L., Su, X., Zhao, L., & Jiang, H. (2020). Precision medicine: Subgroup identification in longitudinal trajectories. *Statistical methods in medical research*, 29(9), 2603–2616. <https://doi.org/10.1177/0962280220904114>
- [14] Zeileis, A. and Hornik, K. (2007), Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, 61: 488-508. <https://doi.org/10.1111/j.1467-9574.2007.00371.x>
- [15] Zeileis, A. Torsten Hothorn & Kurt Hornik (2008) Model-Based Recursive Partitioning, *Journal of Computational and Graphical Statistics*, 17:2, 492-514, DOI: 10.1198/106186008X319331
- [16] Zeileis, A. (2005) A Unified Approach to Structural Change Tests Based on ML Scores, F Statistics, and OLS Residuals, *Econometric Reviews*, 24:4, 445-466, DOI: 10.1080/07474930500406053
- [17] West, B.T., Welch, K.B., & Galecki, A.T. (2014). *Linear Mixed Models: A Practical Guide Using Statistical Software*, Second Edition (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/b17198>
- [18] Jiang, J., & Nguyen, T. (2021). *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.