



Available online at <http://scik.org>

Commun. Math. Biol. Neurosci. 2024, 2024:18

<https://doi.org/10.28919/cmbn/8281>

ISSN: 2052-2541

BAYESIAN CONDITIONAL NEGATIVE BINOMIAL AUTOREGRESSIVE MODEL: A CASE STUDY OF STUNTING ON JAVA ISLAND IN 2021

DINA JUMIATUL FITRI*, ANIK DJURAIDAH, HARI WIJAYANTO

Department of Statistika, IPB University, Bogor, Indonesia

Copyright © 2024 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: Sustainable Development Goals (SDGs) include stunting as one of its worldwide objectives, which requires an average annual reduction in the stunting rate of about 2.7%. Therefore, mapping stunting cases along with the underlying risk factors is crucial in Indonesia, especially on the island of Java, which has the largest population. The result can then be taken into consideration for policymaking to address stunting cases, particularly in Java and in Indonesia in general. This research aims to analyze stunting cases on the island of Java using the best models among GLM (Generalized Linear Model), GLMM (Generalised Linear Mixed Models), ICAR (Intrinsic Autoregressive), and CAR BYM (Besag York Mollie) with a negative binomial distribution. The data used in this study were obtained from the 2021 Health Profile of all provinces in Java, with predictors including the percentage of infants with exclusive breastfeeding, the percentage of complete basic immunization, the percentage of families with access to adequate sanitation, and the percentage of poverty rates. The study's findings reveal that the CAR BYM model has the best forecasts since it has the lowest DIC (Deviance Information Criterion) and MAD (Mean Absolute Deviance) values. The significant predictors are the percentage of families with access to adequate sanitation (X_3) and the percentage of poverty rates (X_4). A relative risk greater than one exists in 31.1% of Java's districts and cities, the majority of

*Corresponding author

E-mail address: dinajumiatulfitri@apps.ipb.ac.id

Received October 13, 2023

which are found in West Java and Central Java. Bogor district/city has the highest relative risk (RR) of stunting (RR = 5.833), followed by Bandung district/city (RR = 3.721), Tegal (RR = 3.291), and Brebes (RR = 2.252). Meanwhile, areas with low risk are found in districts and cities in the Special Region of Yogyakarta (RR = 0.610).

Keywords: negative binomial; GLM; GLMM; conditional autoregressive; spatial; stunting; mapping disease.

2020 AMS Subject Classification: 62M30, 62H11.

1. INTRODUCTION

Poisson regression analysis is used to examine the relationship between a dependent variable representing count data that follows a Poisson distribution. In Poisson regression analysis, there is an assumption that must be met, which is equidispersion. In real-world cases, count data often exhibit variability that is greater than its mean (overdispersion) or variability that is less than its mean (underdispersion) [1]. To handle overdispersion in Poisson regression, a negative binomial regression is employed, which can be analyzed using the Generalized Linear Model (GLM) method. Negative binomial regression assumes that observations are independent. In reality, geographic, sociocultural, and economic conditions will vary between different regions. This implies the presence of spatial effects among regions. A common approach to addressing this is to use Generalized Linear Mixed Models (GLMM), which incorporate structured random spatial effects and are often modeled using conditional autoregressive methods.

The conditional autoregressive (CAR) model is a spatially conditional model where observations at one location are influenced by other locations [2]. CAR operates on spatial autocorrelation by adding spatially correlated random components. Spatial correlation in the CAR model is generated through neighborhood weighting [3]. The CAR-Besag and Besag York Mollie (CAR-BYM) model is used for negative binomial observations [4]. Bayesian spatial models, such as the Besag, York, and Mollié Poisson models (BYM), have been implemented for mapping and identifying areas with a high risk of tuberculosis (TB) cases in Portugal [5]. In the study by Djuraidah *et.al.* [6], the Besag York Mollie (BYM) model was found to be the best in predicting TB on the island of Java.

In CAR model estimation, the Bayesian method is commonly employed. A method that can provide a better and more efficient posterior distribution using Bayesian is the Integrated Nested Laplace Approximation (INLA) introduced by Rue *et.al.* [7]. INLA focuses on marginal inference using Laplace and numerical integration approaches. Bivand *et.al.* [8] employed Bayesian inference with the INLA approach, which offers faster computation compared to Markov Chain Monte Carlo (MCMC) and can accommodate large datasets.

The negative binomial conditional autoregressive (CAR) model with the integrated nested laplace approximation (INLA) approach was applied to analyze the prevalence of stunting cases, involving count data, and is thus considered discrete. Stunting is a significant issue faced by Indonesia and has serious implications for human resource quality (HRQ). According to the 2021 Indonesia Nutritional Status Survey (SSGBI), the prevalence of stunting in Indonesia is 24.4%. This means that approximately one out of four children under five years old in Indonesia is experiencing stunting. This figure is high compared to the threshold set by the World Health Organization (WHO), which is 20%. The Sustainable Development Goals (SDGs) include stunting as one of its worldwide objectives. With the right policies, the government aims to reduce stunting in children by 40% by 2025 and eradicate all forms of malnutrition by 2030. To achieve this, the prevalence of stunting in Indonesian children has to decrease from 37.2% in 2013 to 14.9% by 2025, which requires an average annual reduction of 2.7%. Therefore, mapping stunting cases and the underlying risk factors is crucial in Indonesia, especially on the island of Java, where approximately 56.10% of Indonesia's population lives. Apart from that, infrastructure in Java, particularly in terms of clean water and sanitation, is better than any other island in Indonesia. This has been reflected in the relatively low prevalence of stunting among children in Java's provinces [9]. However, the local government still has to take stunting into account. While its prevalence may not be high, due to its large population, there will probably be a sizable number of cases of stunting to deal with.

Therefore, this research aims to analyze cases of stunting and determine the influencing factors. Moreover, it is aimed at mapping stunting cases on the island of Java using the best-fitting

models from the Generalised Linear Model (GLM), Generalised Linear Mixed Models (GLMM), Intrinsic Conditional Autoregressive (ICAR), and Conditional Autoregressive Bessag York Mollie (CAR BYM) with a negative binomial distribution.

2. PRELIMINARIES

2.1 Generalized Linear Model (GLM)

2.2.1 Poisson Regression

Poisson regression is one of the generalized linear models (GLMs) used to simulate the relationship between a response variable following a Poisson distribution and predictor variables. The Poisson distribution is employed to handle data that represents the count of events in a random process with equal mean and variance [10]. The general equation for GLM with a Poisson distribution is as follows:

$$\begin{aligned} y_i &\sim \text{Poisson}(\mu_i) \\ \log(\mu_i) &= \mathbf{X}_i^T \boldsymbol{\beta} \end{aligned} \quad (1)$$

In the equation above, where y_i is the i -th response variable following a Poisson distribution, μ_i is the mean of the Poisson distribution, \mathbf{X}_i^T is the matrix of predictors, $\boldsymbol{\beta}$ is the parameter vector to be estimated, and $\log(\cdot)$ to bridge the Poisson distribution. When the variance of the response variable $\text{Var}(y_i)$ is greater than the mean $E(y_i)$, overdispersion occurs in GLM with a Poisson distribution [10].

2.2.2 Overdispersion Test

Overdispersion can be detected partly by using the Deviance (D) value. The deviance is a likelihood ratio test that compares the current model to its saturated model, divided by its degrees of freedom, and can be expressed as follows [11] :

$$D = 2 \ln \left[\frac{L(y_i, y_i)}{L(\hat{\mu}_i, y_i)} \right] \quad (2)$$

$$\phi = \frac{D}{db} \quad (3)$$

y_i is the value of the response variable, and $\hat{\mu}_i$ is the estimated result from the Poisson

distribution. The value of db is calculated as n minus k , where n is the number of observations, and k is the number of parameters. If the dispersion test statistic is greater than one, then overdispersion is present in the response variable [10]. To address the issue of overdispersion, several techniques can be used, such as employing more flexible distributions like the negative binomial distribution.

2.2.3 Negative Binomial Regression

GLM (Generalized Linear Model) with a negative binomial distribution is a statistical method used to mimic the relationship between a response variable following a negative binomial distribution and predictor variables. The general equation for GLM with a negative binomial distribution is as follows [12] :

$$\begin{aligned} y_i &\sim NB(\mu_i, \alpha) \\ \log(\mu_i) &= \mathbf{X}_i^T \boldsymbol{\beta} \end{aligned} \quad (4)$$

with $\log(\cdot)$ functions as the link of the negative binomial distribution.

2.2 Negative Binomial Conditional Autoregressive Model

Generalized Linear Mixed Models (GLMM) are an extension of GLM (Generalized Linear Models) with the addition of random components in the predictor part of the model. The general equation for GLMM can be formulated by modifying the GLM model and adding random effects. θ_i represents independent random effects (unstructured random effect). Another development in GLMM is the consideration of structured random spatial effects in the predictors through the Conditional Autoregressive (CAR) prior, hence often referred to as the Conditional Autoregressive (CAR) model.

If the response variable $Y = (y_1, \dots, y_n)$ is a univariate vector representing the count of cases at location i originating from a negative binomial distribution, the predictor vector at location i is $X_i^T = (1, x_{1i}, x_{2i}, \dots, x_{pi})$ with $i = 1, 2, 3, \dots, n$ and there are p independent variables. The structured spatial random effect components are represented by $\phi = (\phi_1, \phi_2, \dots, \phi_n)$. The CAR model can be formulated as follows:

$$y_i | \mu_i \sim NB(y_i | \mu_i, \gamma) \quad (5)$$

$$\log(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \phi_i \quad (6)$$

The random effect ϕ_i is modeled using the class of conditional autoregressive (CAR) prior distributions. Various CAR prior distributions have been introduced by various experts, especially in the context of disease mapping. These include the Intrinsic CAR (ICAR), Besag York Mollié (BYM), Stern-Cressie, and Leroux priors, among others. However, this study will only discuss the ICAR and BYM priors.

2.2.1 ICAR

Currently, the simplest CAR prior is the Intrinsic Autoregressive (IAR) prior, which was proposed by Besag *et.al.* [13] and has a full conditional distribution, namely:

$$\phi_i = \psi_i \quad (7)$$

$$\psi_i \mid \boldsymbol{\psi}_{-i}, \mathbf{W}, \tau^2 \sim N\left(\frac{\sum_{j \sim i} w_{ji} \psi_j}{\sum_{j \sim i} w_{ji}}, \frac{\tau^2}{\sum_{j \sim i} w_{ji}}\right) \quad (8)$$

$$\tau^2 \sim \text{Inverse-Gamma}(a, b) \quad (9)$$

The conditional expectation of ψ_i is equal to the average of the random effects in neighboring areas, while the conditional variance is inversely proportional to the number of neighbors. Thus, the variance structure of ψ_i will have a strong spatial correlation when an area has more neighbors. The variance parameter τ^2 is used to control the magnitude of variation among random effects, typically following an inverse gamma distribution process.

2.2.2 BYM

The Besag-York-Mollie (BYM) model was first described by Besag *et.al.* [13] and incorporates random effects from two components, namely, combining structured random spatial effects with independent random effects. The model takes the following form:

$$\phi_i = \psi_i + \theta_i \quad (10)$$

$$\psi_i \mid \boldsymbol{\psi}_{-i}, \mathbf{W}, \tau^2 \sim N\left(\frac{\sum_{j \sim i} w_{ji} \psi_j}{\sum_{j \sim i} w_{ji}}, \frac{\tau^2}{\sum_{j \sim i} w_{ji}}\right) \quad (11)$$

$$\theta_i \sim N(0, \sigma^2) \quad (12)$$

$$\tau^2, \sigma^2 \sim \text{Inverse-Gamma}(a, b) \quad (13)$$

The random effects $\theta = (\theta_1, \theta_2, \dots, \theta_n)$ are independent with a mean of zero and constant variance, while the spatial autocorrelation random effects are modeled through ψ . This is the ubiquitous CAR model in practice.

2.3 Integrated Nested Laplace Approximation (INLA)

The Integrated Nested Laplace Approximation (INLA) method developed by Rue *et.al.* [7], is more effective than MCMC simulations since it uses direct numerical computations on the latent Gaussian Markov random fields (GMRF) density [14].

For each component of the parameter vector, the INLA technique aims to estimate the marginal posterior distribution. The marginal posterior distribution for the parameter, according to Blangiardo and Cameletti [15], is as follows:

$$\pi(\theta_i | \mathbf{Y}) = \int \pi(\theta_i | \boldsymbol{\psi}, \mathbf{Y}) \times \pi(\boldsymbol{\psi} | \mathbf{Y}) d\boldsymbol{\psi} \quad (14)$$

Then, for the hyperparameter ψ , the marginal posterior distribution is collected as follows:

$$\pi(\psi_i | \mathbf{Y}) = \int \pi(\boldsymbol{\psi} | \mathbf{Y}) d\psi_{-i} \quad (15)$$

Once the marginal posterior distributions are obtained, the calculation process for $\pi(\boldsymbol{\psi} | \mathbf{Y})$ and $\pi(\theta_i | \boldsymbol{\psi}, \mathbf{Y})$ through the Laplace approximation, as described by Blangiardo and Cameletti [15], is as follows:

1. Approaching $\pi(\boldsymbol{\psi} | \mathbf{Y})$ through the marginal distribution $\pi(\psi_i | \mathbf{Y})$ obtained from:

$$\begin{aligned} \pi(\boldsymbol{\psi} | \mathbf{Y}) &= \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{Y})}{\pi(\boldsymbol{\theta} | \boldsymbol{\psi}, \mathbf{Y})} \\ &= \frac{\pi(\mathbf{Y} | \boldsymbol{\theta}, \boldsymbol{\psi}) \times \pi(\boldsymbol{\theta}, \boldsymbol{\psi})}{\pi(\mathbf{Y})} \frac{1}{\pi(\boldsymbol{\theta} | \boldsymbol{\psi}, \mathbf{Y})} \\ &= \frac{\pi(\mathbf{Y} | \boldsymbol{\theta}, \boldsymbol{\psi}) \times \pi(\boldsymbol{\theta}, \boldsymbol{\psi}) \times \pi(\boldsymbol{\psi})}{\pi(\mathbf{Y})} \frac{1}{\pi(\boldsymbol{\theta} | \boldsymbol{\psi}, \mathbf{Y})} \\ &\propto \frac{\pi(\mathbf{Y} | \boldsymbol{\theta}, \boldsymbol{\psi}) \times \pi(\boldsymbol{\theta}, \boldsymbol{\psi}) \times \pi(\boldsymbol{\psi})}{\pi(\boldsymbol{\theta} | \boldsymbol{\psi}, \mathbf{Y})} \\ &\approx \left. \frac{\pi(\mathbf{Y} | \boldsymbol{\theta}, \boldsymbol{\psi}) \times \pi(\boldsymbol{\theta}, \boldsymbol{\psi}) \times \pi(\boldsymbol{\psi})}{\pi(\boldsymbol{\theta} | \boldsymbol{\psi}, \mathbf{Y})} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*(\boldsymbol{\psi})} = \tilde{\pi}(\boldsymbol{\psi} | \mathbf{Y}) \end{aligned} \quad (16)$$

where $\tilde{\pi}(\boldsymbol{\psi} | \mathbf{Y})$ is an approximation of $\pi(\boldsymbol{\psi} | \mathbf{Y})$.

2. Approaching the conditional posterior distribution $\pi(\theta_i | \boldsymbol{\psi}, \mathbf{Y})$ to obtain $\pi(\theta_i | \mathbf{Y})$, let $\boldsymbol{\theta}$

be a parameter vector that can be expressed as $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$, through the Laplace approximation, it can be elaborated as follows:

$$\begin{aligned}
\pi(\theta_i|\boldsymbol{\psi}, \mathbf{Y}) &= \frac{\pi((\theta_i, \boldsymbol{\theta}_{-i}), \boldsymbol{\psi}|\mathbf{Y})}{\pi((\boldsymbol{\theta}_{-i}, \theta_i)|\boldsymbol{\psi}, \mathbf{Y})} \\
&= \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{Y})}{\pi(\boldsymbol{\psi}|\mathbf{Y}) \pi(\boldsymbol{\theta}_{-i}, \theta_i|\boldsymbol{\psi}, \mathbf{Y})} \frac{1}{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{Y})} \\
&\propto \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{Y})}{\pi(\boldsymbol{\theta}_{-i}, \theta_i|\boldsymbol{\psi}, \mathbf{Y})} \\
&\approx \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{Y})}{\pi(\boldsymbol{\theta}_{-i}, \theta_i|\boldsymbol{\psi}, \mathbf{Y})} \Big|_{\theta_i=\theta_{-i}^*(\theta_i, \boldsymbol{\psi})} = \tilde{\pi}(\theta_i|\boldsymbol{\psi}, \mathbf{Y}) \tag{17}
\end{aligned}$$

with $\tilde{\pi}(\boldsymbol{\theta}_{-i}, \theta_i|\boldsymbol{\psi}, \mathbf{Y})$ is the Laplace approximation of $\pi(\theta_i|\boldsymbol{\psi}, \mathbf{Y})$

After having $\tilde{\pi}(\boldsymbol{\psi}|\mathbf{Y})$ through equation (16) and $\tilde{\pi}(\theta_i|\boldsymbol{\psi}, \mathbf{Y})$ through equation (17), the marginal posterior distribution of $\pi(\theta_i|Y)$ can be estimated by calculating the following :

$$\tilde{\pi}(\theta_i|\mathbf{Y}) \approx \int \tilde{\pi}(\theta_i|\boldsymbol{\psi}, \mathbf{Y}) \times \tilde{\pi}(\boldsymbol{\psi}|\mathbf{Y}) d\boldsymbol{\psi} \tag{18}$$

The integral form of equation (18) can be solved numerically using this formula:

$$\tilde{\pi}(\theta_i|\mathbf{Y}) \approx \sum_j \tilde{\pi}(\theta_i|\boldsymbol{\psi}^{(j)}, \mathbf{Y}) \times \tilde{\pi}(\boldsymbol{\psi}^{(j)}|\mathbf{Y}) d\boldsymbol{\psi} \tag{19}$$

3. MAIN RESULTS

3.1 Data

The 2021 Health Profile of all the provinces on the island of Java, specifically Banten, DKI Jakarta, West Java, Yogyakarta Special Region, Central Java, and East Java, provided the data for this study. The 119 districts and cities that make up the entire island of Java served as the study's observational units. Table 1 lists the variables used in this investigation.

Tabel 1 The variables used

| The variables | Description | Factors | Literature |
|----------------|---|----------------------------------|---------------------------|
| Y | The prevalence of stunting per 100,000 toddlers | | |
| X ₁ | The percentage of infants with exclusive breastfeeding | The factor of nutritional intake | Hasiru <i>et.al.</i> [16] |
| X ₂ | The percentage of infants with complete basic immunization | The factor of nutritional intake | Manaf <i>et.al.</i> [17] |
| X ₃ | The percentage of families with access to adequate sanitation | The factor of parenting style | Fadliana and Drajat [18] |
| X ₄ | The percentage of the impoverished population | The economic factor | Bele <i>et al.</i> [19] |

3.2 Analysis Method

The method used to analyze stunting data is the Negative Binomial Conditional Autoregressive (CAR) method with the INLA approach. Data processing was performed using R Studio software. The data analysis steps conducted as follows:

1. Conducting data exploration for the research
2. Utilizing the Variance Inflation Factor (VIF) values to identify multicollinearity among explanatory factors
3. Checking for overdispersion by examining the deviance and Pearson chi-squared values divided by their degrees of freedom. If these values are greater than one, it indicates overdispersion in the data, and more flexible distribution like the negative binomial distribution can be used.
4. Calculating the Spatial Weight Matrix (W), with the details of each weighting performed [6] :
 - a. For queen contiguity, the weighting is done with the following formula:

$$W_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are adjacent.} \\ 0 & \text{if } i \text{ and } j \text{ are not adjacent} \end{cases}$$

- b. For exponential weight, the weighting is done using the following formula: $W_{ij} = \exp(-d_{ij})$, where d_{ij} is the distance between area i and area j .
 - c. For inverse distance weight, the weighting is done as : $W_{ij} = d_{ij}^{-1}$, where d_{ij} is the distance between area i and area j .
 - d. For KNN weighting :
 - i. Calculate the distance from the center of unit i to all other units $j \neq i$.

- ii. Rank them as follows: $d_{ij(1)} \leq d_{ij(2)} \leq \dots \leq d_{ij(n-1)}$.
- iii. Then, for each $k = 1, \dots, n - 1$, set $N_k(i) = \{j(1), j(2), \dots, j(k)\}$ containing the k nearest units to i .
- iv. For each k , the weight matrix W had elements w_{ij} equal to 1 if area i was adjacent to area j , while the main diagonal elements would always be zero.

5. Conducting spatial autocorrelation tests by calculating the Moran's index value on the response variable [20], with the hypothesis below :

H_0 : There is no spatial autocorrelation in the data

H_1 : There is spatial autocorrelation in the data

Test statistics: $z = \frac{I - E(I)}{\sqrt{\text{var}(I)}}$

where: $I = \frac{n \sum_{i=1}^n \sum_{j=1}^n W_{ij} (x_i - \bar{x})(x_j - \bar{x})}{s_0 \sum_{i=1}^n (x_i - \bar{x})^2}$

The testing criterion: H_0 is rejected if the $|Z|$ value is greater than $Z_{\alpha/2}$

6. Conducting estimation of the Negative Binomial CAR model using the INLA approach,

6.1 Creating four models to be examined [21] as follows :

- a. GLM model: $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$
- b. GLMM model: $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta} + \theta_i$
- c. ICAR model: $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta} + \psi_i$
- d. BYM model: $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta} + \psi_i + \theta_i$

6.2 Determining the prior distribution for hyperparameters in the ICAR and BYM models with the specified prior values stated below:

- a. Structured Random Effect, $\psi_i \mid \boldsymbol{\psi}_{-i} \sim N\left(\frac{\sum_{j \sim i} w_{ji} \psi_j}{\sum_{j \sim i} w_{ji}}, \frac{\tau^2}{\sum_{j \sim i} w_{ji}}\right)$
- b. Unstructured Random Effect, $\theta_i \sim N(0, \sigma^2)$
- c. Ragam ψ_i , $\tau^2 \sim \text{Inverse-Gamma}(1, 0.0005)$
- d. Ragam θ_i , $\sigma^2 \sim \text{Inverse-Gamma}(1, 0.0005)$

7. Selecting the best model requires considering several criteria, as follows:

- a. Deviance Information Criterion (DIC)

$$DIC = \bar{D} + 2PD$$

where $PD = E(D(\theta)) - D(E(\theta)) = \bar{D} - D(\bar{\theta})$ and $D(\theta) = -2\log(p(y|\theta))$. The quality of the built model is inversely proportional to AIC values.

b. Mean Absolute Deviance (MAD)

$$MAD = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{N}$$

where y_i is the response data at location i , \hat{y}_i is the prediction result at location i , and N represents the number of observation locations. The smaller the MAD value, the smaller the prediction error will be.

8. Utilizing the goodness-of-fit criteria to determine the optimal model's relative risk. The relative risk was determined by exponentiating the structured random effect component ψ_i from the best model using the following formula [15]: $RR_i = e^{\psi_i}$

9. Interpreting the research results and drawing conclusions

3.3 Data Exploration

The distribution of stunting cases on Java Island varied significantly, with a range of 1-368 cases per 100.000 population and an average of 56 cases per 100.000 population. The areas with the lowest stunting cases were in Blitar and Mojokerto in East Java Province, with 4 and 1 case per 100.000 population, respectively. Meanwhile, Bogor, Garut, and Bandung in West Java Province had the highest stunting cases, with 368, 292, and 227 cases per 100.000 population, respectively.

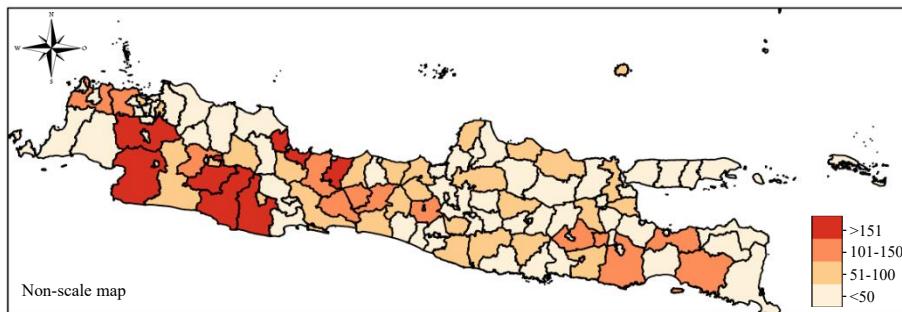


Figure 1 Number of Stunting Cases per 100.000 on Java Island in 2021

The distribution map of stunting cases in Figure 1, shows that the areas with the lowest

stunting cases were located in East Java Province, while the areas with the highest stunting cases were mostly in West Java Province. The spatial reliance in the statistics on cases of stunting on Java Island is indicated by this color grouping.

Reducing the stunting rate can involve identifying factors that influence it. One of these factors is nutritional intake. To measure the intake, data on the percentage of infants with exclusive breastfeeding and the percentage of infants with complete basic immunization are used. Poor maternal factors and inadequate childcare practices, especially in terms of proper sanitation within the family, can also contribute to child stunting. To assess good childcare practices, data on the percentage of families with access to proper sanitation is implemented. Stunting is not only influenced by factors directly related to health but also by socioeconomic issues such as poverty levels [19].

Figure 2 shows that there are still areas with low percentages of infants with exclusive breastfeeding (below 50%), including Blitar, Bangkalan, and Sumenep Regencies in East Java Province. Areas with low percentages of infants with complete basic immunization, which all were below 20%, included Tasikmalaya Regency in West Java Province, as well as Pandeglang Regency and South Tangerang City in Banten Province. Regions with low percentages (below 30%) of proper sanitation were Serang and Tangerang Regencies in Banten Province. Areas with a poverty rate above 19% were Bangkalan, Sumenep, and Sumpang Regencies in East Java Province.

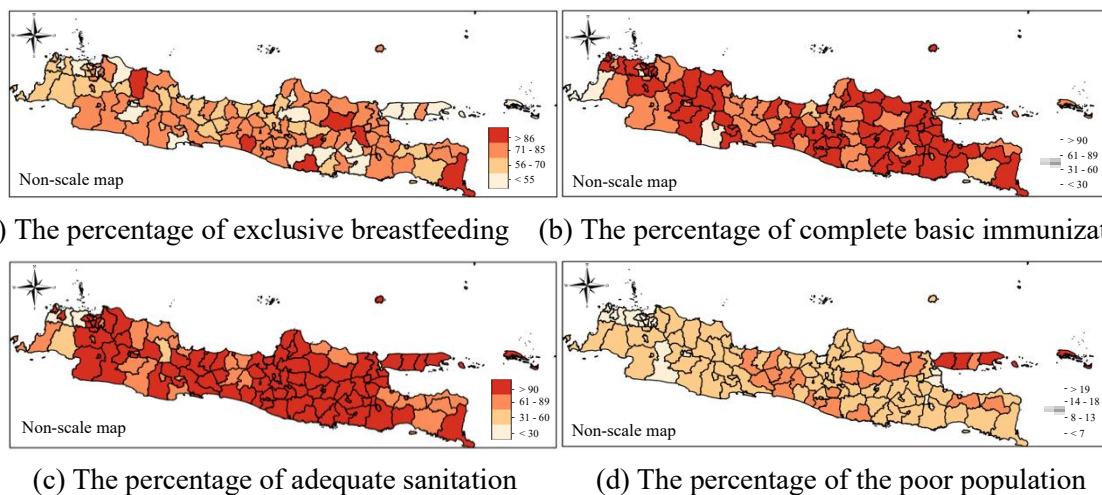


Figure 2 Percentage of Explanatory Variables

BAYESIAN CONDITIONAL NEGATIVE BINOMIAL AUTOREGRESSIVE MODEL

The VIF (variance inflation factor) values were used to examine the multicollinearity among the explanatory variables. Based on Table 2, the VIF values for all explanatory variables are less than 10, indicating the absence of multicollinearity in the data. Therefore, all explanatory variables are included in the model.

Table 2 shows the VIF values for each variable

| variable | X ₁ | X ₂ | X ₃ | X ₄ |
|----------|----------------|----------------|----------------|----------------|
| VIF | 1.019 | 1.021 | 1.001 | 1.002 |

The testing of overdispersion was done by examining the Pearson chi-squared value and the deviance divided by its degrees of freedom. Table 3 shows that the deviation divided by degrees of freedom is 41.101 and the Pearson chi-squared value divided by degrees of freedom is 57.668. Since both of these numbers are more than 1, the data is overdispersed. The negative binomial regression model is one approach to dealing with overdispersion in Poisson regression. Consequently, the data was analyzed using the negative binomial regression model.

Table 3 Overdispersion Testing Results

| Test | Statistics test | df | Statistics test/ df |
|------------|-----------------|-----|---------------------|
| Chi-square | 6574.215 | 114 | 57.688 |
| Devians | 4685.603 | 114 | 41.101 |

Spatial dependence can be depicted by calculating Moran's index value. Based on Table 4, there is a significant Moran's index value using the k-Nearest Neighbour Weight matrix with a value of $k = 4$, which is $I = 0.1052$ with a p-value of 0.02 at a significant level of 5%. This indicates that there is spatial autocorrelation or similarity in values between neighboring areas, and their distribution tends to be clustered. Therefore, in this study, only the k-Nearest Neighbour Weight matrix with a value of $k = 4$ is used.

Table 4 Moran's Index value

| Weight Matrix | Moran's Index | E(I) | Var(I) | p-value |
|--------------------------|---------------|--------|--------|---------|
| queen contiguity | 0.0591 | -0.008 | 0.0045 | 0.15 |
| exponential alpha=1 | 0.0228 | -0.008 | 0.0001 | 0.01 |
| inverse distance alpha=1 | 0.0115 | -0.008 | 0.0006 | 0.21 |
| 4-Nearest Neighbor | 0.1052 | -0.008 | 0.0032 | 0.02 |

3.4 Negative Binomial CAR with INLA in The Case of Stunting

The modeling was performed with four models, including GLM, GLMM, ICAR, and BYM.

Table 5 presents the parameter estimates of each model using explanatory variables.

Table 5 Parameter estimates in the models and their 95% credible intervals

| Model | Intersept | X_1 | X_2 | X_3 | X_4 |
|-------|--------------------------|---------------------------|---------------------------|-----------------------------------|--------------------------------|
| GLM | 1.109 (-0.247, 2.468) | -0.005 (-0.018, 0.007) | -0.003 (-0.011, 0.007) | -0.009 (-0.017, -0.002) | 0.037 (-0.003, 0.077) |
| GLMM | 0.533 (-0.851, 1.931) | -0.004 (-0.013, 0.003) | -0.005 (-0.017, 0.009) | -0.008 (-0.016, 0.000) | 0.056 (0.017, 0.095) |
| ICAR | 0.999 (-0.382, 2.364) | -0.005 (-0.010, 0.005) | -0.002 (-0.018, 0.007) | -0.009 (-0.017, -0.002) | 0.042 (0.001, 0.083) |
| CAR | 0.377 | -0.004 | -0.003 | -0.009 | 0.062 |
| BYM | (-0.978, 1.737) | (-0.017, 0.008) | (-0.011, 0.005) | (-0.017, -0.001) | (0.024, 0.100) |

The four models were estimated with negative signs for the exclusive breastfeeding (X_1), Immunity (X_2), and Sanitation variables (X_3), indicating that higher values of these explanatory variables are highly likely to reduce the incidence of stunting cases in Java Island. On the other hand, the Poverty variable (X_4) has a positive sign, suggesting that higher levels of poverty are associated with a higher incidence of stunting on Java Island. In the GLM model, the explanatory variable Sanitation (X_3) is statistically significant based on the 95% confidence interval. In the GLMM model, the significant explanatory variable is Poverty (X_4). In both the ICAR and CAR BYM models, the significant explanatory variables are Sanitation (X_3) and Poverty (X_4). The goodness-of-fit measure used for model comparison is the Deviance Information Criterion (DIC). Additionally, the Mean Absolute Deviance (MAD) values are compared for each model. Based on Table 6, the CAR BYM model with four explanatory variables yields the lowest DIC and MAD values. Therefore, the CAR BYM model, which makes the best predictions, is more likely to be the best-fitting.

Table 6 Comparison of Model Goodness-of-Fit Values

| Model | DIC | MAD |
|---------|---------|-------|
| GLM | 1196.11 | 35.73 |
| GLMM | 997.69 | 3.39 |
| ICAR | 1190.22 | 35.17 |
| CAR BYM | 985.71 | 2.44 |

The proportion of spatial variability in the CAR BYM model can be calculated by dividing the variance of the spatial structured component by the total variance of the random components. In this case, the proportion was 99.66%, indicating that the proportion of variability in the structured spatial component (u_i) is larger than the unstructured spatial component (v_i).

3.5 Mapping Relative Risk Using The CAR BYM Model

The spatial patterns of diseases are visualized by disease mapping, which also identifies regions with a relative risk greater or lower than one. Figure 3 shows that 31.1% of the regions had a high relative risk, including 11.76% in Central Java, 9.24% in West Java, 7.56% in East Java, 1.68% in Banten, and 0.84% in DKI Jakarta provinces. In Figure 1, it is shown that Cianjur and Bandung Barat districts had 78 and 106 cases per 100,000 population, respectively, which are classified as low and moderate areas. However, in Figure 3, Bandung Barat and Cianjur districts had high relative risk values of 1.32 and 1.79, respectively. This is because of their proximity to Bogor, Sukabumi, Bandung, and Garut districts, which are classified as high-risk areas. Meanwhile, districts and cities in the Special Region of Yogyakarta had low-risk areas.

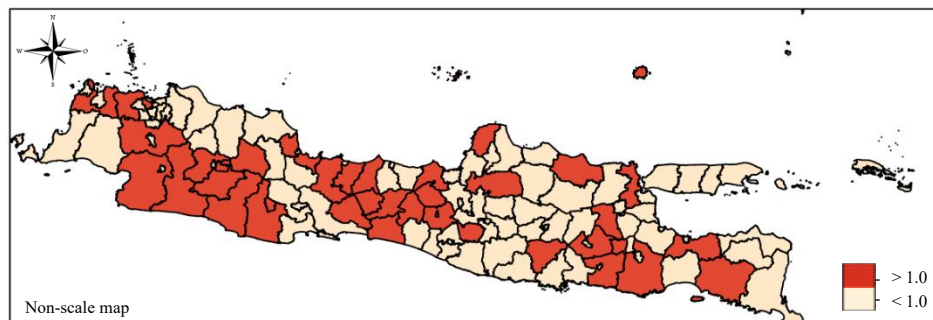


Figure 3 Relative Risk Map of Stunting Cases on Java Island 2021

4. CONCLUSION

The CAR BYM model effectively predicted the relative risk levels of stunting cases on Java Island in 2021. Significant factors affecting stunting cases were the percentage of families with access to proper sanitation (X_3), and the percentage of poverty levels (X_4). The proportion of districts and cities on Java Island at high risk is 31.1%, with 11.7% in Central Java and 9.24% in West Java. Meanwhile, the province of DI Yogyakarta had areas with low risk. Stunting cases on

Java Island can be addressed by estimating the relative risk levels of those cases using the BYM model, which produces precise predictions and useful information for policymaking.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

- [1] K.F. Sellers, B. Premeaux, Conway-Maxwell-Poisson regression models for dispersed count data, *WIREs Comput. Stat.* 13 (2020), e1533. <https://doi.org/10.1002/wics.1533>.
- [2] M.M. Wall, A close look at the spatial structure implied by the CAR and SAR models, *J. Stat. Plan. Inference.* 121 (2004), 311-324. [https://doi.org/10.1016/s0378-3758\(03\)00111-3](https://doi.org/10.1016/s0378-3758(03)00111-3).
- [3] D. Lee, A comparison of conditional autoregressive models used in Bayesian disease mapping, *Spatial Spatio-Temp. Epidemiol.* 2 (2011), 79-89. <https://doi.org/10.1016/j.sste.2011.03.001>.
- [4] O. Tufvesson, J. Lindström, E. Lindström, Spatial statistical modelling of insurance risk: a spatial epidemiological approach to car insurance, *Scandinavian Actuarial J.* 2019 (2019), 508-522. <https://doi.org/10.1080/03461238.2019.1576146>.
- [5] O. Oliveira, A.I. Ribeiro, E.T. Krainski, et al. Using Bayesian spatial models to map and to identify geographical hotspots of multidrug-resistant tuberculosis in Portugal between 2000 and 2016, *Sci. Rep.* 10 (2020), 16646. <https://doi.org/10.1038/s41598-020-73759-w>.
- [6] A. Djuraidah, Z. Mar'ah, R. Anisa, A Bayesian conditional autoregressive with INLA: A case study of tuberculosis in Java, Indonesia, *Commun. Math. Biol. Neurosci.* 2022 (2022), 102. <https://doi.org/10.28919/cmbn/7709>.
- [7] H. Rue, S. Martino, N. Chopin, Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations, *J. R. Stat. Soc. Ser. B: Stat. Methodol.* 71 (2009), 319-392. <https://doi.org/10.1111/j.1467-9868.2008.00700.x>.
- [8] R.S. Bivand, V. Gómez-Rubio, H. Rue, Approximate Bayesian inference for spatial econometrics models, *Spatial Stat.* 9 (2014), 146-165. <https://doi.org/10.1016/j.spasta.2014.01.002>.

BAYESIAN CONDITIONAL NEGATIVE BINOMIAL AUTOREGRESSIVE MODEL

- [9] DINKES Pulau Jawa, Dinas Kesehatan Provinsi di Pulau Jawa, Profil Kesehatan Provinsi di Pulau Jawa 2021, Dinas Kesehatan Provinsi di Pulau Jawa, Jawa, 2021.
- [10] J.M. Hilbe, Negative binomial regression, Cambridge University Press, (2011).
- [11] P. McCullagh, Generalized linear models (2nd ed.), Routledge, (1983).
- [12] J. Stoklosa, R.V. Blakey, F.K.C. Hui, An overview of modern applications of negative binomial modelling in ecology and biodiversity, *Diversity*. 14 (2022), 320. <https://doi.org/10.3390/d14050320>.
- [13] J. Besag, J. York, A. Mollie, Bayesian image restoration, with two applications in spatial statistics, *Ann. Inst. Stat. Math.* 43 (1991), 1-20. <https://doi.org/10.1007/bf00116466>.
- [14] F. Lindgren, H. Rue, Bayesian spatial modelling with R-INLA, *J. Stat. Softw.* 63 (2015), 1-25.
- [15] M. Blangiardo, M. Cameletti, Spatial and spatio-temporal Bayesian models with R-INLA, John Wiley & Sons, Hoboken, (2015).
- [16] L.S. Hasiru, I. Djakaria, I.K. Hasan, Penerapan Model Spasial Durbin dengan Uji Lanjutan Local Indicator of Spatial Autocorrelation untuk Melihat Penyebaran Stunting di Kabupaten Bone Bolango, *Jambura J. Prob. Stat.* 3 (2022), 19-28.
- [17] S.A.R. Manaf, A. Fitrianto, R. Amelia, Faktor–Faktor yang Memengaruhi Permasalahan Stunting di Jawa Barat Menggunakan Regresi Logistik Biner, *J. Statistika: J. Ilmiah Teori Apl. Stat.* 15 (2022), 265-274.
- [18] A. Fadliana, P.P. Drajat, Pemetaan faktor risiko stunting berbasis sistem informasi geografis menggunakan metode geographically weighted regression, *J. Ikraith- Inform.* 5 (2021), 91-102.
- [19] M.G.L. Bele, E.M.P. Hermanto, F. Fitriani, Pemodelan Geographically Weighted Regression pada Kasus Stunting di Provinsi Nusa Tenggara Timur Tahun 2020, *J. Stat. Apl.* 6 (2022), 179-191.
- [20] L. Anselin, *Spatial econometrics: methods and models*, Kluwer Academic Publisher, (1988).
- [21] J. Stoklosa, R.V. Blakey, F.K.C. Hui, An overview of modern applications of negative binomial modelling in ecology and biodiversity, *Diversity*, 14 (2022), 320. <https://doi.org/10.3390/d14050320>.