PARAMETER
JURNAL MATEMATIKA, STATISTIKA DAN TERAPANNYA

e-ISSN. 2829-6303
Vol. 05 No. 3, December 2026
Pages 471-488

Check for updates

Spatial Autoregressive Generalized Poisson and Negative Binomial Models for Leprosy Case in East Java

Dewi Indra Setiawan

Data Science Department, Faculty of Science, Institut Teknologi Sumatera
Lampung, 35365, Indonesia

E-mail Correspondence Author: *dewi.setiawan@sd.itera.ac.id

Abstract

Leprosy remains a public health concern in East Java Province, with 2,225 new cases recorded in 2025. Poisson regression with covariates and a population offset indicated substantial conditional overdispersion (Pearson $\chi^2/df = 19.61$, $p < 0.001$), while Moran's I test on the raw response variable indicated significant spatial autocorrelation ($I = 0.600$, $p < 0.001$). This study compares the Spatial Autoregressive Generalized Poisson (SAR-GP) and Spatial Autoregressive Negative Binomial (SAR-NB) models for leprosy cases across 38 regencies and cities in East Java. The main contribution is a direct empirical comparison of these two spatial count models within a common analytical framework, complemented by an assessment of alternative spatial weight matrix specifications, including the Suramadu Bridge connection. Both models were estimated using the Berndt–Hall–Hall–Hausman (BHHH) method, incorporating poverty rate, population density, access to proper sanitation, and the ratio of health workers as predictors, with population size as the exposure variable. The estimated dispersion parameters and spatial lag parameters were statistically significant in both models (SAR-GP: $\rho = 0.740$, $p = 0.003$; SAR-NB: $\rho = 0.650$, $p = 0.029$). Based on RMSE and MAE, SAR-NB showed better in-sample performance (RMSE = 26.81; MAE = 19.83) than SAR-GP (RMSE = 57.00; MAE = 34.22). Poverty rate was the only statistically significant predictor in the SAR-NB model. These findings indicate that SAR-NB provided a better fit to the observed data among the two models considered.

Keywords: *Leprosy; overdispersion, spatial autocorrelation, spatial autoregressive generalized poisson, spatial autoregressive negative binomial*

: https://doi.org/10.30598/parameterv5i3pp471-488

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.

Copyright © 2026 The Author(s).
Research Ariticle ***Open Acces***
Email: jurnalparameter@gmail.com
Homepage: https://ojs3.unpatti.ac.id/index.php/parameter

Submitted: July 2026
Accepted: December 2026

1. INTRODUCTION

Leprosy is a chronic infectious disease caused by the bacterium *Mycobacterium leprae*. Transmission occurs primarily through respiratory droplets from infected individuals during coughing or sneezing, particularly in cases of close and prolonged contact. Because *Mycobacterium leprae* multiplies very slowly, leprosy has a long incubation period, averaging around five years and potentially extending to 20 years or more. This characteristic makes transmission difficult to detect directly. In 2025, 2,225 new leprosy cases were recorded in East Java Province, an increase of 27 cases compared to the previous year, with the majority (95.5%) classified as the Multibacillary (MB) type [1]. Densely populated environments and poor sanitation conditions are also among the factors associated with the spread of leprosy in the population [2].

Leprosy case data in East Java Province constitute count data, which can be analyzed using Poisson regression. Poisson regression assumes equidispersion, namely that the mean and variance are equal. However, the 2025 leprosy case data in East Java Province exhibit overdispersion. A dispersion test on a Poisson regression including the predictors and a log-population offset confirmed significant overdispersion conditional on covariates and exposure (Pearson $\chi^2/df = 19.61$, $p < 0.001$). This condition renders Poisson regression less suitable, necessitating a model capable of accommodating overdispersion. Two approaches commonly used to address this issue are Generalized Poisson (GP) regression, which can accommodate both overdispersion and underdispersion [3] , and Negative Binomial (NB) regression, which is specifically designed to handle overdispersion [4].

In addition to overdispersion, leprosy cases across regions in East Java may exhibit spatial dependence, as conditions in one region can be influenced by neighboring regions. Spatial dependence tests indicate that the 2025 leprosy case data in East Java Province show evidence of spatial dependence. This finding suggests that global regression modeling alone is insufficient to explain the variation in leprosy cases, as the interrelationships between regions must also be taken into account. One approach that can accommodate this is the Spatial Autoregressive (SAR) model, which incorporates spatial dependence through a spatial autoregressive parameter (ϱ) and has been applied in various count-data health studies, such as dengue fever cases on Lombok Island [5].

Several previous studies have applied spatial regression models to health data exhibiting overdispersion. The Spatial Autoregressive-Generalized Poisson (SAR-Generalized Poisson) model has been used to model pneumonia cases in Tuban Regency [6], while the Spatial Autoregressive-Negative Binomial (SAR-Negative Binomial) model has been applied to stunting cases in West Java Province [7]. A comparison of several spatial models within the Negative Binomial family (SAR, SEM, and SARMA) has also been conducted for malnutrition cases in Central Java [8]. However, these studies have not addressed a direct comparison between SAR-Generalized Poisson and SAR-Negative Binomial models for leprosy count data within a common analytical framework. Thus, the research gap addressed in this study is the lack of a direct empirical comparison of these two models for leprosy cases using a consistent estimation and evaluation framework.

To address this gap, this study compares SAR-Generalized Poisson and SAR-Negative Binomial models using leprosy data from 38 regencies and cities in East Java in 2025. Beyond comparing model fit, this study examines the estimated spatial lag parameters and the discrepancy between Poisson-based Lagrange Multiplier diagnostics and the estimated spatial dependence in both models. Furthermore, a sensitivity analysis is conducted using alternative spatial weight matrix specifications, including one that

incorporates the Suramadu Bridge connection between Madura Island and mainland Java, to assess whether the model comparison results vary with the representation of regional connectivity.

2. METHOD

2.1 Morans I Test

The Moran's I statistic is used to identify the presence of global spatial autocorrelation in a dataset [9] . The hypotheses tested are H_0: I = 0 (no spatial dependence) versus H_1: I ≠ 0 (spatial dependence exists) [10] . The Moran's I statistic is formulated as:

$$I = \frac{n}{S_0}\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \quad (1)$$

where n is the number of observations, x_i and x_j are the observed values at location i and j, $\bar{x}$ is the sample mean, w_{ij} is the spatial weight between location i and j, and $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ is the sum of all spatial weights. Under the null hypothesis, the expected value of Moran's I is :

$$E(I) = \frac{-1}{(n-1)} \quad (2)$$

Significance is assessed using the standardized test statistic :

$$Z(I) = \frac{[I - E(I)]}{\sqrt{Var(I)}} \quad (3)$$

H_0 is rejected if $|Z(I)| > Z_{\alpha/2}$ or the p-value is smaller than the significance level α, indicating the presence of significant spatial autocorrelation between locations [10].

2.2 Generalized Poisson – Spatial Autoregressive (SAR-GP)

Unlike the SAR model for continuous data, modeling SAR for count data requires a distinct spatial autoregressive link function, as originally developed by Lambert [11] . The construction of this model follows the framework applied in [6].

Let the linear predictor be defined as an $n \times 1$ vector, $\eta = X\beta$. The spatial component is obtained through the transformation $A^{-1}\eta$, where $A = (I - \rho W)$. The link function for the $i-th$ observation, incorporating the exposure component, is formulated as:

$$\mu_i = q_i exp((A^{-1}\eta)_i) \quad (4)$$

which can be written in full as $\mu_i = q_i exp(((I - \rho W)^{-1}X\beta)_i)$. In this equation, q_i denotes the exposure variable, A^{-1} is the inverse of matrix A, and $(A^{-1}\eta)_i$ represents the i-th element of the resulting $n \times 1$ transformed vector. The parameter ρ is the spatial lag coefficient, W is the row-standardized Queen Contiguity spatial weight matrix, X is the $n\,x\,(p+1)$ matrix of predictor variables, and β is the $(p+1)\,x\,1$ vector of regression coefficients. In this specification, spatial dependence is incorporated through the spatial transformation of the linear predictor, $A^{-1}\eta = (I - \varrho W)^{-1}X\beta$, which determines the conditional mean of the count response. Thus, spatial dependence is introduced through the transformed linear predictor rather than through a direct spatial lag of the observed response variable. By combining the structure of the Generalized Poisson Regression (GPR) model with the SAR link function in Equation (4), the SAR-Generalized Poisson model is obtained and can be expressed as follows:

$$P(Y = y_i|x_i) = \left[\frac{\mu_i{}^{SAR}}{1+\varphi\mu_i{}^{SAR}}\right]^{y_i} \frac{(1+\varphi y_i)^{y_i-1}}{y_i!} exp\left(-\frac{\mu_i^{SAR}(1+\varphi y_i)}{1+\varphi\mu_i{}^{SAR}}\right),\ y_i = 0,1,2,\ldots \quad (5)$$

with $\mu_i^{SAR} = q_i exp((A^{-1}X\beta)_i)$ as defined in Equation (4), where y_i is the response variable and φ is the dispersion parameter of y.

2.2.1 Parameter Estimation of SAR-GP

Parameters β, ρ, and φ of the SAR-GP model are estimated using the Maximum Likelihood Estimation (MLE) method. The spatial autoregressive parameter ρ is restricted to the interval $|\rho| < 1/\lambda_{max}(W)$, where $\lambda_{max}(W)$ is the spectral radius of matrix W; for a row-standardized weight matrix, this constraint simplifies to $|\rho| < 1$ **[6]** . This restriction ensures that matrix A = (I– ρW) remains invertible throughout the estimation process. Given n independent observations $(y_i, x_i), i = 1,2,\ldots,n,$ the likelihood function of the SAR-GP model is given by:

$$L(\theta) = \prod_{i=1}^{n} P(Y = y_i|x_i) \tag{6}$$

where $P(Y = y_i|x_i)$ follows the form in Equation (5). Taking the natural logarithm of Equation (6) yields the log-likelihood function:

$$lnL(\theta) = \sum_{i=1}^{n}\left[y_i ln\frac{{\mu_i}^{SAR}}{1+\varphi{\mu_i}^{SAR}} + (y_i - 1)ln(1+\varphi y_i) - lny_i! - \frac{\mu_i^{SAR}(1+\varphi y_i)}{1+\varphi{\mu_i}^{SAR}}\right] \tag{7}$$

The likelihood is constructed under the assumption that, conditional on the covariates, the exposure, the spatial weight matrix, and the parameters, the responses (y_i) are mutually independent. Spatial dependence is incorporated through the conditional mean: the spatially filtered linear predictor $(I - \rho W)^{-1}X\beta$ is a deterministic function of the covariates and model parameters. Consequently, the conditional means μ_i^{SAR} are fixed given these quantities. The joint likelihood therefore takes the product form of the marginal GP probability mass functions. This specification does not imply direct stochastic dependence among the observed responses; rather, spatial dependence is introduced through the spatial structure of the conditional mean. Any residual spatial autocorrelation not captured by this mean structure is not explicitly modeled and is assessed through residual diagnostics. with parameter vector $\theta = [\beta^T \rho\varphi]^T$. Since Equation (7) does not yield a closed-form solution, parameter estimates are obtained iteratively using the Berndt–Hall–Hall–Hausman (BHHH) algorithm **[12]** , which approximates the Hessian matrix using the outer product of the gradient vectors, thereby avoiding the need for second-derivative computation **[6]** . The iteration proceeds until the change between successive parameter estimates falls below a predetermined convergence threshold. The covariance matrix of the parameter estimates was estimated using the inverse of the outer product of gradients (OPG) information matrix. The standard errors were calculated as the square roots of the diagonal elements of the estimated covariance matrix.

2.2.2 Hypothesis Testing of SAR-GP

Simultaneous testing of the predictor parameters is conducted using the Maximum Likelihood Ratio Test (MLRT), with the hypotheses:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \text{ vs. } H_1: \text{at least one } \beta_k \neq 0$$

the test statistics is defined as,

$$G^2 = 2\left[lnL(\widehat{\Omega}) - lnL(\widehat{\omega})\right] \quad (8)$$

where $\widehat{\Omega}$ and $\widehat{\omega}$ denote the parameter estimates under the full and reduced (null) models, respectively. Under H_0, G^2 asymptotically follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models; H_0 is rejected if $G^2 > \chi^2_{(\alpha,df)}$. Partial significance testing for each parameter (ρ β_k and φ) is conducted using the Wald test:

$$W_{Wald} = \left(\frac{\hat{\theta}}{se(\hat{\theta})}\right)^2 \sim \chi^2_{(\alpha,1)} \quad (9)$$

where θ denotes each parameter of the SAR-GP model tested individually, namely the spatial lag coefficient (ρ), each regression coefficient (β_k), and the dispersion parameter (φ). The null hypothesis H_0: $\theta = 0$ is rejected if $W_{Wald} > \chi^2_{(\alpha,1)}$

2.3 Spatial Autoregressive – Negative Binomial (SAR-NB)

Model. The general spatial autoregressive equation for count data can be extended to accommodate the Negative Binomial (NB) distribution to address overdispersion. Following the same spatial autoregressive link function used in the SAR-Generalized Poisson model, the SAR-Negative Binomial model incorporates spatial dependence through the transformation $A^{-1}\eta$, where $A = (I - \rho W)$. The mean structure for the i-th observation is defined as:

$$\mu_i = q_i exp((A^{-1}X\beta)_i) \quad (10)$$

where q_i is the exposure variable (population size), ρ is the spatial lag coefficient, W is the row-standardized Queen Contiguity spatial weight matrix, X is the $n\ x(p+1)$ matrix of predictor variables, and β is the $(p+1)\ x\ 1$ vector of regression coefficients. The response variable is assumed to follow a Negative Binomial (NB) distribution, with probability mass function for the i-th observation given by:

$$P(Y = y_i|x_i) = \frac{\Gamma(y_i + r)}{y_i!\,\Gamma(r)}\left(\frac{r}{r+\mu_i}\right)^r \left(\frac{\mu_i}{r+\mu_i}\right)^{y_i}, y_i = 0,1,2,\ldots \quad (11)$$

where $r > 0$ is the dispersion parameter. Under this parameterization, the variance of Y_i is given by $Var(Y_i) = \mu_i + \mu_i^2/r$, so that smaller values of r correspond to greater overdispersion, and the model reduces to the standard Poisson distribution as r $\rightarrow\infty$.

2.3.1 Parameter Estimation of SAR-NB

Parameters β, ρ, and r of the SAR-NB model are estimated using the Maximum Likelihood Estimation (MLE) method. Given n observations, the likelihood function is formed as:

$$L(\beta, \rho, r) = \prod_{i=1}^{n} P(Y = y_i|x_i) \quad (12)$$

Taking the natural logarithm of Equation (12) yields the log-likelihood function:

$$lnL(\beta, \rho, r) = \sum_{i=1}^{n}\left[ln\Gamma(y_i + r) - lny_i! - ln\Gamma(r) + rln\left(\frac{r}{r+\mu_i}\right) + y_i ln\left(\frac{\mu_i}{r+\mu_i}\right)\right] \quad (13)$$

with μ_i as defined in **Equation (10).** Since **Equation (13)** does not admit a closed-form solution, parameter estimates are obtained iteratively using the Berndt–Hall–Hall–

Hausman (BHHH) algorithm, following the same estimation procedure described in Section 2.2 for the SAR-GP model. To ensure that (r>0) throughout the optimization process, estimation is performed using the parameter log(r), which is unrestricted in sign; the estimate of r is subsequently obtained as $\hat{r} = exp(\widehat{log(r)})$. The BHHH algorithm approximates the information matrix using the outer product of individual score vectors (OPG). Upon convergence, the covariance matrix is estimated as the inverse of the OPG information matrix evaluated at the final parameter estimates. The standard errors are calculated as the square roots of the diagonal elements of this covariance matrix. Therefore, the standard errors are based on the OPG approximation rather than a separately computed numerical Hessian.

2.3.2 Hypothesis Testing of SAR-NB

Simultaneous testing of the predictor parameters is conducted using the Maximum Likelihood Ratio Test (MLRT), with the hypotheses:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \text{ vs. } H_1: \text{at least one } \beta_k \neq 0$$

using the same test statistic form as in **Equation (8)**:

$$G^2 = 2\left[lnL(\hat{\Omega}) - lnL(\hat{\omega})\right] \quad (14)$$

where $\hat{\Omega}$ and $\hat{\omega}$ denote the parameter estimates under the full and reduced (null) models, respectively. Under H_0, G^2 asymptotically follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models where H_0 is rejected if $G^2 > \chi^2_{(\alpha,df)}$. Partial significance testing for each parameter (ρ, β_k and r) is conducted using the Wald test:

$$W_{wald} = \left(\frac{\hat{\theta}}{se(\hat{\theta})}\right)^2 \sim \chi^2_{(\alpha,1)} \quad (15)$$

Where θ denotes each parameter of the SAR-NB model tested individually, namely the spatial lag coefficient (ρ), each regression coefficient (β_k), and the dispersion parameter (r). The null hypothesis $H_0: \theta = 0$ is rejected if $W_{wald} > \chi^2_{(\alpha,1)}$

2.4 Spatial Weight Matrix

The spatial weight matrix (W) represents the spatial neighborhood structure among the regencies and cities in East Java. In this study, Queen contiguity is used to define spatial neighbors, where two regions are considered neighbors if their boundaries share either an edge or a vertex **[13]**. The binary spatial weight matrix is defined as follows:

$$w_{ij} = \begin{cases} 1, & \text{if i and j share an edge or corner} \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

The initial Queen contiguity matrix was constructed from the administrative polygons of the 38 regencies and cities in East Java using the GADM dataset. Based on the polygon-based Queen contiguity definition, Madura Island forms a separate connected component from mainland East Java. To represent the transportation connection between Bangkalan Regency and Surabaya City via the Suramadu Bridge, an additional symmetric link was introduced between these two areas. This modification connects the Madura component to mainland East Java and represents an extension of the original Queen contiguity structure. The resulting spatial weight matrix was then row-standardized before being incorporated into the SAR-GP and SAR-NB models. To assess the sensitivity of spatial count model results to the specification of the spatial weight

matrix, following a similar approach in previous research [6], the SAR-GP and SAR-NB models were re-estimated using alternative spatial weight matrices. Specifically, the modified Queen contiguity matrix incorporating the Suramadu link was compared with the Queen contiguity matrix without the additional link and the Rook contiguity matrix. Parameter estimates and model selection criteria were examined to evaluate the sensitivity of the findings to alternative spatial neighborhood specifications.

2.5 Model Selection Criteria

The relative fit of the SAR-Generalized Poisson and SAR-Negative Binomial models was assessed using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) [14] . These criteria are defined respectively as:

$$AIC = -2ln\hat{L} + 2p \quad (16)$$

$$BIC = -2ln\hat{L} + p\ ln\ n \quad (17)$$

where $ln\hat{L}$ is the maximized value of the log-likelihood function, p is the number of estimated parameters in the model, and n is the number of observations. Lower AIC and BIC values indicate a better relative fit, accounting for model complexity. Model performance was additionally evaluated using the root mean square error (RMSE) [6] and mean absolute error (MAE), calculated as:

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

and

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

where y_i and $\hat{y}_i$ denote the observed and fitted numbers of leprosy cases for region (i), respectively, and (n) is the number of regencies and cities. Lower RMSE and MAE values indicate smaller prediction errors. In this study, both metrics were calculated using the same data employed for model estimation and therefore represent in-sample prediction errors.

2.5 Data Source and Process Analysis

This study uses secondary data on leprosy cases and socio-demographic indicators in East Java Province for the year 2025, obtained from the *Health Profile of East Java Province 2025* published by the East Java Provincial Health Office [1] , and from the official website of the Statistics of East Java Province (*Badan Pusat Statistik Provinsi Jawa Timur*) [15], [16], [17] . The unit of analysis is the regency/city (*kabupaten/kota*) level, comprising 38 regencies and cities in East Java Province. All analyses were conducted using RStudio software.The response variable (Y) is the number of new leprosy cases recorded in each regency/city in 2025. The predictor variables (X) used in this study are presented in **Table 1**.

Table 1. Research Variable

Variable	Data	Source
Y	Number of Leprosy Cases	Health Profile of East Java Province 2025
q	Population size (exposure variable)	BPS Provinsi Jawa Timur
X_1	Poverty Rate (%)	BPS Provinsi Jawa Timur
X_2	Population Density	BPS Provinsi Jawa Timur
X_3	Access to proper sanitation	BPS Provinsi Jawa Timur
X_4	Ratio of Health Workers	BPS Provinsi Jawa Timur

The data analysis in this study is conducted through the following stages:

1. Data preparation. Leprosy case data and predictor variables obtained from the East Java Provincial Health Office and BPS Provinsi Jawa Timur are merged with the administrative boundary shapefile of East Java Province based on standardized regency/city names. The resulting dataset is verified to ensure all 38 regencies/cities are matched correctly with no missing values.
2. Data exploration. Descriptive statistics of the response variable (number of leprosy cases) are summarized, and a thematic map is generated to illustrate the spatial distribution of leprosy cases across regencies/cities in East Java.
3. Overdispersion test. The presence of overdispersion in the response variable is examined using the variance-to-mean ratio (VMR). A VMR substantially greater than 1 indicates a violation of the equidispersion assumption required for standard Poisson regression.
4. Spatial weight matrix construction. A first-order Queen Contiguity spatial weight matrix (W) is constructed based on the administrative boundary map of East Java Province, where two regions are considered neighbors if they share a common boundary (an edge or a vertex). A manual adjacency correction is applied to account for the Suramadu Bridge connection between Bangkalan Regency and Surabaya City, which are not physically contiguous but are connected via this bridge. The matrix is row-standardized.
5. Spatial dependence test on the response variable. Moran's I test is conducted on the raw response variable to provide an initial assessment of spatial autocorrelation in leprosy case distribution.
6. Multicollinearity test. The correlation matrix among predictor variables is examined, followed by the calculation of the Variance Inflation Factor (VIF) for each predictor, with VIF ≥ 10 indicating a multicollinearity problem.
7. Baseline Poisson regression. A standard Poisson regression model is fitted with the number of leprosy cases as the response variable, the predictor variables as covariates, and the logarithm of population size as an offset to account for exposure.
8. Spatial dependence test on model residuals. Lagrange Multiplier (LM) tests (LMlag, LMerr, and their robust variants) and Moran's I test on the Pearson residuals of the baseline Poisson model are conducted to formally assess the presence of spatial dependence after accounting for the predictor variables.
9. SAR-Generalized Poisson model estimation. The SAR-GP model is fitted using the Berndt–Hall–Hall–Hausman (BHHH) iterative algorithm, following the model specification, parameter estimation, and hypothesis testing procedures described in Section 2.2.

10. SAR-Negative Binomial model estimation. The SAR-NB model is fitted using the BHHH iterative algorithm, following the model specification, parameter estimation, and hypothesis testing procedures described in Section 2.3.
11. Model comparison. The SAR-GP and SAR-NB models are compared based on their AIC and BIC values, as described in Section 2.4, to determine the best-fitting model for the leprosy case data in East Java Province.
12. Interpretation of the best model. The coefficients of the selected best-fitting model are interpreted, including the significant predictor variables and the spatial lag parameter (ρ), to explain the factors associated with the number of leprosy cases across regencies/cities in East Java Province.

3. RESULTS AND DISCUSSION

This section presents the results of the descriptive analysis, spatial dependence testing, and parameter estimation for the SAR-Generalized Poisson and SAR-Negative Binomial models applied to leprosy case data in East Java Province. The discussion is organized into several subsections, covering data exploration, model diagnostics, parameter estimation and significance testing for both models, and model comparison based on the AIC and BIC criteria.

3.1. Data Exploration

The distribution of leprosy cases across East Java Province is presented in **Figure 1**. Sampang and Sumenep Regencies recorded the highest number of leprosy cases in East Java Province, with 309 and 217 cases, respectively, both located on Madura Island. Kediri City recorded no leprosy cases in 2025.

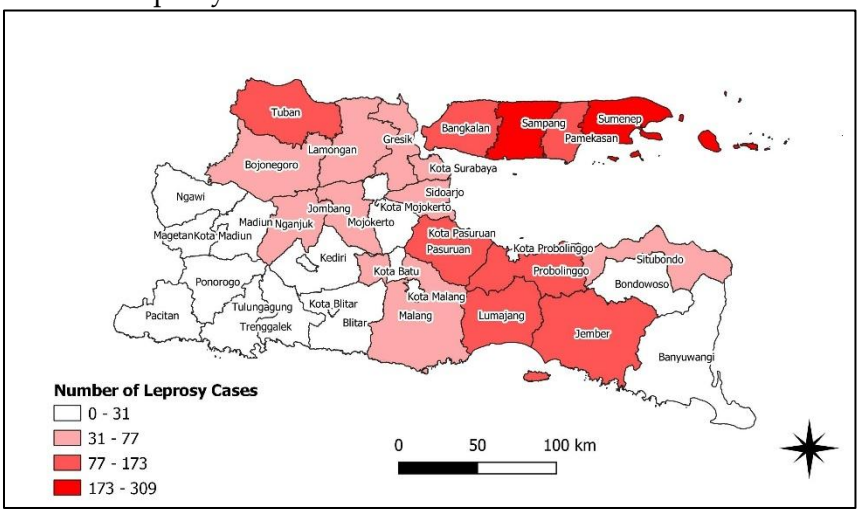


Figure 1. Distribution Map of Leprosy Cases in East Java Province, 2025

Table 2 presents the descriptive statistics of the response variable and the four predictor variables used in this study. The number of leprosy cases (Y) ranges from 0 to 309 cases, with a mean of 58.55 and a standard deviation of 68.95. The standard deviation exceeds the mean, indicating a highly skewed distribution and substantial variability in leprosy case counts across regencies/cities. The percentage of the poor population (X_1) ranges from 2.86 to 20.61 percent, with a mean of 9.48. The standard deviation of 4.21 appears relatively low, indicating that the variation in the percentage of the poor population across regencies/cities in East Java Province is relatively small. The highest population density (X_2), 8,727, is found in Surabaya City. This is fairly common, given that Surabaya is a large city with economic centers and complete facilities. The variability of population

density, based on the standard deviation value of 2,299.08, indicates considerable variation. This condition reflects high inequality, where some areas are very densely populated while others are sparsely populated.

Table 2. Descriptive Statistics

Variable	Min	Max	Mean	Standard Deviation	VMR (Variance to Mean Ratio)
Number of Leprosy Cases (Y)	0	309	58.55	68.95	81.18
Poverty Rate (X_1)	2.86	20.61	9.48	4.21	
Population Density (X_2)	411	8727	1969.97	2299.08	
Access to Proper Sanitation (X_3)	65.34	99.81	88.79	9.07	
Ratio of Health Workers (X_4)	0.20	1.96	0.54	0.36	

Access to proper sanitation (X_3) ranges from 65.34% to 99.81%, with a mean of 88.79%. The standard deviation of 9.07 is relatively small, indicating that sanitation access is fairly similar across regencies/cities in East Java Province. The ratio of health workers (X_4) ranges from 0.20 to 1.96, with a mean of 0.54. The standard deviation of 0.36 indicates considerable variation, meaning the availability of health workers differs quite widely between regencies/cities. The response variable (number of leprosy cases) exhibits a variance-to-mean ratio (VMR) of 81.18, as shown in **Table 2**. This value substantially exceeds 1, indicating strong overdispersion and confirming that the equidispersion assumption of standard Poisson regression is violated. Prior to model estimation, a multicollinearity test was conducted to examine the relationships among the predictor variables.

Table 3. Multicollinearity Test Result

Variable	VIF
Poverty Rate (X_1)	1.863
Population Density (X_2)	2.384
Access to Proper Sanitation (X_3)	1.527
Ratio of Health Workers (X_4)	2.043

The results of the multicollinearity test, based on the Variance Inflation Factor (VIF), are presented in **Table 3**. All VIF values range from 1.527 to 2.384, well below the commonly used threshold of 10. This indicates that there is no multicollinearity among the predictor variables, and all four predictors can be included simultaneously in the regression model.

3.2. Spatial Dependence Test

3.2.1 Moran's I Test

The hypotheses for Moran's I test are formulated as follows:

H_0: There is no spatial autocorrelation in the number of leprosy cases among regencies/cities.

H_1: There is spatial autocorrelation in the number of leprosy cases among regencies/cities.

The results of the Moran's I test are presented in **Table 4.**

Table 4. Moran's I Test Result

Parameter	Value
I	0.600
E(I)	-0.027
Z(I)	5.074
p-value	1.94×10^{-7}

Based on **Table 4**, the p-value (1.94×10^{-7}) is smaller than the significance level of $\alpha = 0.05$, so H_0 is rejected. This indicates the presence of significant positive spatial autocorrelation in the number of leprosy cases ($I = 0.600$, $Z = 5.074$, $p < 0.001$). This result suggests that leprosy cases tend to cluster spatially, with regencies/cities having high case counts located near other regencies/cities with similarly high case counts, as also reflected in the distribution pattern shown in **Figure 1.**

3.2.2 Lagrange Multiplier Test

A baseline Poisson regression model was fitted as a preliminary step to obtain residuals for spatial dependence testing. A dispersion test on a Poisson regression including the predictors and a log-population offset confirmed significant overdispersion conditional on covariates and exposure (Pearson $\chi^2/df = 19.61$, $p < 0.001$). Given the extreme overdispersion identified earlier, the coefficient estimates from this baseline model were not further interpreted, as standard errors under the Poisson assumption are known to be underestimated in the presence of overdispersion. Lagrange Multiplier (LM) tests and Moran's I test were subsequently conducted on the Pearson residuals of this baseline model to formally assess the presence of remaining spatial dependence after accounting for the predictor variables.

Table 5. Lagrange Multiplier Test Result

Test	Statistics	Df	P-value
LMlag	1.380	1	0.240
LMerr	0.141	1	0.707
Robust LMlag	1.348	1	0.246
Robust LMerr	0.109	1	0.742

To formally assess the presence of spatial dependence after accounting for the predictor variables, Lagrange Multiplier (LM) tests were conducted on the residuals of the baseline Poisson regression model. As shown in Table 5, none of the LM test statistics are statistically significant at $\alpha = 0.05$ (all p-values > 0.05), indicating no significant spatial dependence in the model residuals based on this diagnostic test. This result is further supported by Moran's I test on the Pearson residuals of the baseline model, which also indicates no significant spatial autocorrelation ($I = 0.026$, $Z = 0.421$, $p = 0.337$). Although the LM tests and the residual-based Moran's I test did not indicate significant spatial dependence in the baseline Poisson model, the observed leprosy case counts exhibited strong positive spatial autocorrelation ($I = 0.600$, $p < 0.001$). This spatial pattern motivated further investigation using spatial count regression models that accommodate overdispersion and incorporate spatial dependence. The spatial autoregressive parameters in both the SAR-GP and SAR-NB models were statistically significant, indicating evidence of spatial association under the respective model specifications. However, these findings do not establish residual spatial autocorrelation in the baseline model or demonstrate direct disease transmission between neighboring

regions. The SAR-GP and SAR-NB models were subsequently compared using the specified model selection criteria to assess their relative suitability for the observed leprosy case counts.

3.3. Spatial Autoregressive Generalized Poisson (SAR-GP) Model to Leprosy Cases in East Java

The parameter estimates of the SAR-Generalized Poisson model, obtained using the BHHH iterative algorithm, are presented in **Table 6.** Both structural parameters of the model were found to be highly significant: the spatial lag parameter ($\rho = 0.740$, p.value=0.003) and the dispersion parameter (φ = 0.061, p.value < 0.001).

Table 6. Parameter Estimation of SAR-GP Model

Parameter	Estimate	Std. Error	Wald (χ^2)	p-value	95% CI Lower	95% CI Upper
Intercept	-2.879	2.239	1.654	0.198	-7.267	1.509
Poverty Rate (X1)	0.089	0.049	3.335	0.068*	-0.006	0.185
Population Density (X2)	0.00014	0.00011	1.747	0.186	-0.075	0.000
Sanitation Access (X3)	-0.0071	0.0102	0.480	0.489	-0.027	0.013
Health Workers Ratio (X4)	-0.478	0.378	1.596	0.206	-1.218	0.263
ρ (Spatial Lag)	0.740	0.247	8.953	0.003	0.255	1.224
φ (Dispersion)	0.061	0.015	15.924	<0.001	0.031	0.091

*)Significance at α = 0.10

The significant ρ confirms the presence of spatial dependence among leprosy cases in East Java Province, while the significant φ confirms the presence of overdispersion, supporting the appropriateness of the Generalized Poisson distribution over the standard Poisson. Among the four predictor variables, none were found to be statistically significant at α = 0.05. Poverty rate showed the closest association with leprosy cases (p.value = 0.068), approaching significance at the 10% level, while population density, sanitation access, and the ratio of health workers were not significant predictors in this model. Based on the estimated parameters in Table 6, the fitted SAR-Generalized Poisson model can be expressed as follows:

$$\hat{\eta}_i = -2.879 + 0.089X_{1i} + 0.00014X_{2i} - 0.0071X_{3i} - 0.478X_{4i}$$
$$\hat{\mu}_i^{SAR} = q_i exp[(I - 0.740W)^{-1}\hat{\eta}_i]$$

with $\hat{\varphi} = 0.061$, where q_i denotes the population size of regency/city i.

3.4. Spatial Autoregressive Negative Binomial (SAR-NB) Model to Leprosy Cases in East Java

The parameter estimates of the SAR-Negative Binomial model, obtained using the BHHH iterative algorithm, are presented in **Table 7.** Both structural parameters of the model were found to be statistically significant: the spatial lag parameter (ρ=0.650, p.value=0.029) and the dispersion parameter ($\widehat{log(r)} = 1.304, p.value < 0.001$, corresponding to $\hat{r}$ = exp(1.304)≈3.684). The significant ρ confirms the presence of spatial dependence among leprosy cases, while the significant dispersion parameter confirms the presence of overdispersion, supporting the use of the Negative Binomial distribution over the standard Poisson.

Table 7. Parameter Estimation of SAR-NB Model

Parameter	Estimate	Std. Error	Wald (χ^2)	p-value	95% CI Lower	95% CI Upper
Intercept	-3.586	2.480	2.090	0.148	-8.448	1.276
Poverty Rate (X1)	0.089	0.043	4.361	0.037*	0.005	0.173
Population Density (X2)	0.00012	0.00009	1.507	0.220	-0.000	0.0003
Sanitation Access (X3)	-0.0088	0.0125	0.490	0.484	-0.033	0.016
Health Workers Ratio (X4)	-0.562	0.393	2.042	0.153	-1.333	0.209
ρ (Spatial Lag)	0.650	0.298	4.751	0.029	0.065	1.234
r (Dispersion)	1.304	0.344	14.358	<0.001	1.877	7.232

*)Significance at α = 0.05

Among the four predictor variables, poverty rate was found to be a statistically significant predictor of leprosy cases $(\widehat{\beta_1} = 0.089,\ p.value = 0.037)$, indicating that regencies/cities with higher poverty rates tend to have a higher number of leprosy cases. Population density, sanitation access, and the ratio of health workers were not statistically significant at α =0.05. Based on the estimated parameters in Table 7, the fitted SAR-Negative Binomial model can be expressed as follows:

$$\hat{\eta}_i = -3.586 + 0.089X_{1i} + 0.00012X_{2i} - 0.0088X_{3i} - 0.562X_{4i}$$
$$\hat{\mu}_i^{SAR} = q_i exp[(I - 0.650W)^{-1}\hat{\eta}_i]$$

with $\hat{r} \approx 3.684$, where q_i denotes the population size of regency/city i

3.4. Sensitivity Analysis of Spatial Weight Matrices

To assess the sensitivity of the model estimates to the specification of the spatial weight matrix, three configurations were considered: Queen contiguity with the Suramadu link, Queen contiguity without the Suramadu link, and Rook contiguity. The estimated spatial lag parameters and model selection criteria are presented in **Table 8.** The results show that the estimated spatial lag parameter remained positive under all specifications, although its magnitude decreased when the Suramadu link was excluded.

Table 8. Sensitivity Analysis of Spatial Weight Matrices

Spatial Weight Matrix	Model	ρ	AIC	BIC
Queen + Suramadu	SAR-GP	0.740	344.627	356.090
Queen + Suramadu	**SAR-NB**	**0.650**	**339.147**	**350.610**
Queen without Suramadu	SAR-GP	0.677	345.078	356.541
Queen without Suramadu	SAR-NB	0.601	340.048	351.511
Rook	SAR-GP	0.677	345.078	356.541
Rook	SAR-NB	0.601	340.048	351.511

The AIC and BIC values were also slightly lower when the Suramadu link was included. Furthermore, SAR-NB consistently yielded lower AIC and BIC values than SAR-GP across the specifications considered, supporting the relative preference for SAR-NB in this dataset. The Queen matrix without the Suramadu link and the Rook matrix were identical for the study area; therefore, these two configurations produced identical estimation results. The Negative Binomial-SAR model had lower AIC and BIC values than the Generalized Poisson-SAR model. This result is consistent with Rosilala and Hasanah **[8]** , who found that the negative binomial model was more suitable for highly

overdispersed data in their study of child malnutrition in Central Java. In this study, the overdispersion was reflected by the dispersion parameter of 3.684. The significant spatial parameter ($\hat{\varrho}$ = 0.650) also indicates a spatial association in the number of leprosy cases across regions. This finding is consistent with the studies of Husnaeni et al. **[5]** and Jayani et al. [18] on dengue fever, as well as Anselin **[13]** , who noted that spatial dependence should be considered in disease modeling.

3.5. Model Selection Criteria

Model performance was further evaluated by comparing the root mean square error (RMSE) and mean absolute error (MAE) of the SAR-Generalized Poisson and SAR-Negative Binomial models using the same data employed for model estimation (in-sample evaluation). The results are presented in **Table 9**.

Table 9. Comparison of RMSE and MAE Values

Model	RMSE	MAE
SAR – Generalized Poisson	57.003	34.222
SAR – Negative Binomial	**26.807**	**19.834**

The SAR-Negative Binomial model yielded lower RMSE (26.807 vs. 57.003) and MAE (19.834 vs. 34.222) than the SAR-Generalized Poisson model, indicating smaller prediction errors on the data used for model estimation. Moran's I tests were also conducted to assess spatial autocorrelation in the model residuals. The SAR-GP residuals exhibited significant positive spatial autocorrelation (I = 0.6622, p < 0.001), whereas the SAR-NB residuals showed no statistically significant spatial autocorrelation (I = 0.0426, p = 0.290). These results provide additional diagnostic information for comparing the two models. However, because RMSE and MAE were calculated using the same observations employed for model estimation, they reflect in-sample fit rather than predictive performance on unseen data. Therefore, the results do not establish that the SAR-NB model would necessarily perform better on new observations.

$$\hat{\eta}_i = -3.586 + 0.089X_{1i} + 0.00012X_{2i} - 0.0088X_{3i} - 0.562X_{4i}$$
$$\hat{\mu}_i^{SAR} = q_i exp[(I - 0.650W)^{-1}\hat{\eta}_i]$$

This model shows that the estimated spatial lag parameter ($\hat{\varrho}$ = 0.650) indicates positive spatial dependence in the modeled leprosy case counts across regencies and cities in East Java. A one-percentage-point increase in the poverty rate ((X_1)) is associated with a multiplicative factor of (exp(0.089) = 1.093) in the expected number of leprosy cases, holding the other variables constant. This corresponds to an approximately 9.3% increase in the expected case count, and the association is statistically significant (p = 0.037). This finding highlights the relevance of socioeconomic conditions in understanding the spatial distribution of leprosy in East Java. Previous spatial research in the province has also identified poverty and other socioeconomic conditions as relevant factors in the geographic distribution of leprosy cases **[19]** . In addition, research conducted in Bangkalan, East Java, has examined the relationship between poverty-related dietary conditions and leprosy, suggesting that socioeconomic vulnerability may be relevant to leprosy susceptibility **[20]** . Possible pathways include constraints on access to healthcare, living conditions, and timely case detection; however, these mechanisms were not directly assessed in the present study. Therefore, the significant poverty coefficient should be interpreted as an ecological association rather than

evidence of a causal effect. Differences in case detection, healthcare accessibility, and reporting practices may also influence the observed case counts.

A one-percentage-point increase in access to proper sanitation (X_3) is associated with a multiplicative factor of exp(−0.0088) ≈ 0.991 in the expected number of leprosy cases, holding the other variables constant. This corresponds to an approximately 0.9% decrease in the expected case count. However, this association was not statistically significant (p = 0.484). For every one-unit increase in the ratio of health workers (X_4), the expected number of leprosy cases is associated with a multiplicative factor of exp(−0.562) ≈ 0.570, holding the other variables constant. This corresponds to an approximately 43.0% decrease in the expected case count, although the association was not statistically significant (p = 0.153). The estimated dispersion parameter was approximately 3.684, indicating that the variance of the leprosy case counts exceeds the mean and suggesting overdispersion in the data, which supports the use of the Negative Binomial model. Poverty level was also found to have a significant effect on the number of leprosy cases. This finding is consistent with Muharry [2] , who reported an association between socioeconomic conditions and leprosy incidence in Indonesia. It is also in line with the policies of the Indonesian Ministry of Health [21] , [22] , which include poverty reduction as part of the leprosy elimination strategy.

4. CONCLUSION

This study compared the SAR Generalized Poisson (SAR-GP) and SAR Negative Binomial (SAR-NB) models for leprosy cases across 38 regencies and cities in East Java in 2025. The findings indicate substantial overdispersion and significant spatial autocorrelation in the observed case counts. Although the Lagrange Multiplier tests based on the baseline Poisson model were not significant, the estimated spatial lag parameters were significant in both models. Overall, the SAR-NB model provided a more favorable fit to the analyzed data than the SAR-GP model, based on the information criteria and in-sample RMSE and MAE. Residual diagnostics also indicated significant spatial autocorrelation in the SAR-GP residuals but no statistically significant spatial autocorrelation in the SAR-NB residuals. Among the four predictors considered, only the poverty rate was statistically significant in the SAR-NB model, suggesting an association between socioeconomic conditions and the spatial distribution of leprosy cases in East Java. The findings highlight the importance of considering spatial dependence when analyzing leprosy case counts and suggest that public health planning should account for potential connections between neighboring regencies and cities, particularly in areas with high case concentrations, such as Madura Island. The significant association with poverty also suggests that socioeconomic vulnerability deserves consideration when identifying areas for further investigation and planning leprosy control efforts. This is relevant to Indonesia's target of achieving leprosy elimination in 111 regencies and cities by 2030 [21] , [22]. However, the poverty association should not be interpreted as causal, and the apparent advantage of SAR-NB does not establish superior predictive performance on new data because RMSE and MAE were calculated using the same observations employed for model estimation. Other limitations include the relatively small number of observations (n = 38) and the use of cross-sectional data from a single year. Future research could use data from multiple years, finer spatial units such as the sub-district (*kecamatan*) level, and alternative spatial models, including SEM and SARMA, to assess the robustness of the findings and investigate spatio-temporal patterns of leprosy.

Acknowledgments

During the preparation of this manuscript, the authors used Claude (Anthropic) and ChatGPT for assistance with statistical script development, literature searching, and language refinement. The authors reviewed and take full responsibility for the content of this publication.

Funding Information

Authors state there is no funding involved.

Author Contributions Statement

Dewi Indra Setiawan : Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Authors state no conflict of interest.

Data Availability

The data used in this study are publicly available from the East Java Provincial Health Office (Dinas Kesehatan Provinsi Jawa Timur) and the Statistics of East Java Province (Badan Pusat Statistik Provinsi Jawa Timur), as cited in the references. Derived data and analysis scripts supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Dinas Kesehatan Provinsi Jawa Timur, "Profil Kesehatan Provinsi Jawa Timur Tahun 2025," 2025. Accessed: Sep. 22, 2026. [Online]. Available: https://dinkes.jatimprov.go.id/

[2] A. Muharry, "Faktor Risiko Kejadian Kusta," *KEMAS*, vol. 9, no. 2, pp. 174–182, 2014, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/kemas

[3] F. Famoye, J. T. Wulu, and K. P. Singh, "On the Generalized Poisson Regression Model with an Application to Accident Data," 2004.

[4] J. M. Hilbe, *Negative Binomial Regression*. Cambridge, U.K: Cambridge University Press, 2011.

[5] R. R. Husnaeni, S. Hauliati, I. Sholihah, B. T. A. Lasmiani, S. H. Hastuti, and M. Gazali, "Spatial Autoregressive (SAR) Poisson Modeling In Dengue Fever Cases On Lombok Island In 2021," *VARIANCE: Journal of Statistics and Its Applications*, vol. 6, no. 2, pp. 143–154, Oct. 2024, doi: 10.30598/variancevol6iss2page143-154.

[6] J. C. G. Sihombing, S. Sutikno, and A. Choiruddin, "Generalized Poisson Spatial Autoregressive Modeling of Pneumonia Cases in Children Under Five in Tuban," *CAUCHY: Jurnal Matematika Murni dan Aplikasi*, vol. 10, no. 2, pp. 1219–1232, Nov. 2025, doi: 10.18860/cauchy.v10i2.37288.

[7] A. Djuraidah, M. Amelia, and R. Anisa, "Estimation of Factors Affecting Stunting Cases in West Java in 2021 Using Negative Binomial Spatial Regression," *Jurnal Matematika, Statistika dan Komputasi*, vol. 20, no. 1, pp. 41–51, Sep. 2023, doi: 10.20956/j.v20i1.26984.

[8] A. Rosilala and S. H. Hasanah, "Comparison Of Negative Binomial SAR, SEM, and SARMA Methods In Modeling The Number Of Malnutrition Cases Among Toddlers In Central Java," *Jurnal Gaussian*, vol. 14, no. 1, pp. 42–53, Feb. 2025, doi: 10.14710/j.gauss.14.1.42-53.

[9] P A P Moran, "Notes on continuous stochastic phenomena.," *Biometrika*, vol. 37, pp. 17–23, Jun. 1950.

[10] S. Wahyuni, N. Amalita, F. Fitri, J. Hamka, K. Padang, and S. Barat, "Implementation of Spatial Autoregressive Analysis to Determine Factors Affecting the Population Dependency Ratio in West Sumatera," 2025. [Online]. Available: https://siperindu.online/,

[11] D. M. Lambert, J. P. Brown, and R. J. G. M. Florax, "A Two-Step Estimator For A Spatial Lag Model Of Counts: Theory, Small Sample Performace and an Application," 2010.

[12] E. K. Berndt, B. H. Hall, R. E. Hall, and J. A. Hausman, "Estimation and Inference in Nonlinear Structural Models," 1974.

[13] Luc Anselin, *Spatial Econometrics: Methods and Models*. Dordrecht, Netherlands: Kluwer Academic Publishers, 1988.

[14] C. Agiakloglou and A. Tsimpanos, "Evaluating the performance of AIC and BIC for selecting spatial econometric models," *Journal of Spatial Econometrics*, vol. 4, no. 1, Dec. 2023, doi: 10.1007/s43071-022-00030-x.

[15] BPS Provinsi Jawa Timur, "Persentase Penduduk Miskin Menurut Kabupaten/Kota di Provinsi Jawa Timur." Accessed: Aug. 10, 2026. [Online]. Available: jatim.bps.go.id

[16] BPS Provinsi Jawa Timur, "Kepadatan Penduduk Menurut Kabupaten/Kota di Provinsi Jawa Timur." Accessed: Aug. 10, 2026. [Online]. Available: jatim.bps.go.id

[17] BPS Provinsi Jawa Timur, "Statistik Perumahan dan Permukiman Provinsi Jawa Timur 2025." Accessed: Aug. 10, 2026. [Online]. Available: jatim.bps.go.id

[18] M. R. Makful and Y. Handayani, "Spatial Model of Geographic Distribution of Leprosy Cases in East Java Province, Indonesia," *Unnes Journal of Public Health*, vol. 13, no. 2, pp. 93–104, Jul. 2024, doi: 10.15294/ujph.v13i2.4188.

[19] S. Oktaria, N. S. Hurif, W. Naim, H. B. Thio, T. E. C. Nijsten, and J. H. Richardus, "Dietary diversity and poverty as risk factors for leprosy in Indonesia: A case-control study," *PLoS Negl. Trop. Dis.*, vol. 12, no. 3, Mar. 2018, doi: 10.1371/journal.pntd.0006317.

[20] N. Made Surya Jayanti, I. Wayan Sumarjaya, and M. Susilawati, "Pemodelan Penyebaran Kasus Demam Berdarah Dengue (DBD) Di Kota Denpasar Dengan Metode Spatial Autoregressive (SAR)," vol. 6, no. 1, pp. 37–46, 2017.

[21] Kementerian Kesehatan RI, "Indonesia percepat eliminasi kusta dan filariasis, target bebas NTDs pada 2030." Accessed: Sep. 22, 2026. [Online]. Available: kemkes.go.id

[22] Kementerian Kesehatan Republik Indonesia, "Kemenkes RI dan Sasakawa Health Foundation Perkuat Komitmen Eliminasi Kusta," Jul. 2025. Accessed: Aug. 14,

2026. [Online]. Available: https://kemkes.go.id/id/kemenkes-ri-dan-sasakawa-health-foundation-perkuat-komitmen-eliminasi-kusta.