BAREKENG: Journal of Mathematics and Its Applications
March 2023 Volume 17 Issue 1 Page 0527–0536
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss1pp0527-0536

APPLICATION OF PENALIZED SPLINE-SPATIAL AUTOREGRESSIVE MODEL TO HIV CASE DATA IN INDONESIA

Nindi Pigitha [1*], Anik Djuraidah[2], Aji Hamim Wigena[3]

[1,2,3] *Department of Statistics, Faculty Mathematics and Natural Sciences, IPB University*
Jl. Raya Dramaga, Bogor, Jawa Barat, 16680, Indonesia

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

Article History:

Received: 24th November 2022
Revised: 6th February 2023
Accepted: 16th February 2023

Keywords:

Human immunodeficiency virus;
Nonlinear;
Penalized spline;
Penalized spline-spatial autoregressive model;
Spatial autoregressive model;
Semiparametrics.

ABSTRACT

Spatial regression analysis is a statistical method used to perform modeling by considering spatial effects. Spatial models generally use a parametric approach by assuming a linear relationship between explanatory and response variables. The nonparametric regression method is better suited for data with a nonlinear connection because it does not need linear assumptions. One of the nonparametric regression methods is penalized spline regression (P-Spline). The P-spline has a simple mathematical relationship with mixed linear model. The use of a mixed linear model allows the P-Spline to be combined with other statistical models. PS-SAR is a combination of the P-Spline and the SAR spatial model so that it can analyze spatial data with a semiparametric approach. Based on data from monitoring the development of the HIV situation in 2018, the number of HIV cases in Indonesia shows a clustered pattern that indicate spatial dependence. In addition, the relationship between the number of positive cases and the factors tends to be nonlinear. Therefore, this study aims to apply the PS-SAR model to HIV case data in Indonesia. The resulting model is evaluated based on the estimates of autoregressive spatial coefficient, MSE, MAPE, and Pseudo R2. Based on the results, the PS-SAR model has an autoregressive spatial coefficient similar to the SAR model and has smaller MSE and MAPE than the SAR model.

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

How to cite this article:

N. Pigitha, A. Djuraidah and A. H. Wigena., "APPLICATION OF PENALIZED SPLINE-SPATIAL AUTOREGRESSIVE MODEL TO HIV CASE DATA IN INDONESIA," *BAREKENG: J. Math. & App.*, vol. 17, iss. 1, pp. 0527-0536, March 2023.

Copyright © 2022 Author(s)
Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

Spatial regression analysis is a statistical method used to model spatial effects that consist of spatial dependencies and spatial heterogeneity [1]. Data with spatial dependencies can use spatial dependency models such as the spatial autoregressive model (SAR), spatial Durbin model (SDM), and spatial error model (SEM). On the other hand, data with spatial heterogeneity can be modeled by geographically weighted regression (GWR). Spatial regression analysis often used to model cases of diseases such as malaria [2][3], tuberculosis[4][5], dengue hemorragic fever [6].

Spatial models generally use a parametric approach by assuming a linear relationship between variables. However, the relationship between the response variable and the explanatory variable is not always linear but also nonlinear. If the model ignores nonlinear functional forms, it can lead to inconsistent parameter estimators and inaccurate conclusions [7]. The nonparametric regression method is better suited for data with a nonlinear connection because it does not need linear assumptions. One of the nonparametric regression methods is penalized spline regression (P-Spline). P-Spline is a combination of spline regression and spline smoothing [8][9]. P-spline has a simple mathematical relationship and equivalent to a mixed linear model. In this case, the parametric component is formulated as a fixed effect, while the smoothing function component is formulated as a random effect. By formulating the P-spline into a mixed linear model allows the P-Spline to be combined with other statistical models [10]. PS-SAR is a combination of the P-Spline and the SAR spatial model so that it can analyze spatial data with a semiparametric approach [11].

Indonesia is an epidemic area of HIV disease with the 3rd highest number of positive cases in Asia Pacific. One of the factors that influence HIV prevalence is key populations, namely groups of people who have a high level of risk of HIV transmission [12]. According to a World Health Organization (WHO) report, HIV-Tuberculosis (TB) fatalities contribute over 1.3 million deaths worldwide. HIV and TB weaken the patient's immune system, which reduce life expectancy [13]. External variables like unemployment and poverty also have an impact on HIV prevalence [14]. A person may decide to work as a sex worker due to the difficulty of finding a quality job and their need to make ends meet, which raises their risk of catching HIV. Based on data from monitoring the development of the HIV situation in 2018, the number of HIV cases in Indonesia shows a clustered pattern that indicate spatial dependence. In addition, the relationship between the number of positive cases and the factors tends to be nonlinear. Therefore, this study aims to apply the PS-SAR model to HIV case data in Indonesia.

2. RESEARCH METHODS

2.1 Linear Mixed Model

The mixed linear model can be written as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y}$ is a sample observation vector of size $n \times 1$, $\boldsymbol{X}$ is the design matrix of fixed effects $n \times p$ where p is the number of fixed variables, $\boldsymbol{\beta}$ is a fixed effect parameter vector of size $p \times 1$, $\boldsymbol{Z}$ is the design matrix of random effects of size $n \times k$ where k is the number of random variables, $\boldsymbol{u}$ is a random effect parameter vector of size $n \times k$, and ε is an error vector of size $n \times 1$. The assumption of the mixed linear model is that $\boldsymbol{u}$ and ε are independent with $\boldsymbol{u} \sim N(0, \boldsymbol{G})$ and $\sim N(0, R)$, where $\boldsymbol{G} = Var(\boldsymbol{u})$ and $\boldsymbol{R} = Var(\boldsymbol{\varepsilon})$ is variance-covariance matrix that involve unknown dispersion or an unknown variance components (σ). Based on **Equation (1)**, $\boldsymbol{y}$ is distributed as $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$ where $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{ZGZ'}$ [15].

The parameter estimation of the linear mixed model is carried out using the maximum likelihood method with the log likelihood function formulated as follows [16]:

$$\ln(L) = (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})'\boldsymbol{R^{-1}}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + \boldsymbol{u'Gu} \tag{2}$$

The resulting $\boldsymbol{\beta}$ and $\boldsymbol{u}$ estimators is as follows:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X'V^{-1}X}\right)^{-1}\boldsymbol{X'V^{-1}y}$$
$$\widehat{\boldsymbol{u}} = \boldsymbol{GZ'V^{-1}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \quad (3)$$

Estimation of the components of variance using the maximum likelihood method is as follows:

$$\ln(L)_{ML} = -\frac{1}{2}\left[\ln|\boldsymbol{V}| + \boldsymbol{y'V^{-1}}\left(\boldsymbol{I} - \boldsymbol{X}\left[\boldsymbol{X'V^{-1}X}\right]^{-1}\boldsymbol{X'V^{-1}}\right)\boldsymbol{y}\right] - \frac{n}{2}\ln(2\pi) \quad (4)$$

Estimating the variance component estimator using the maximum likelihood method produces a biased estimator so that the estimation of the variance component is better using the restricted maximum likelihood (REML) method. REML possibility function is as follows [12][13]:

$$\ln(L)_{REML} = \ln(L)_{ML} - \frac{1}{2}\ln\left|\boldsymbol{X'V^{-1}X}\right| + \frac{p}{2}\ln(2\pi) \quad (5)$$

2.2 P-Spline

The P-Spline model with one explanatory variable can be written as follows [9]:

$$f(\boldsymbol{X}, \boldsymbol{\beta}) = \boldsymbol{X\beta}, \text{ where}$$
$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p & (x_1-\kappa_1)_+^p & \cdots & (x_1-\kappa_K)_+^p \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^p & (x_n-\kappa_1)_+^p & \ldots & (x_1-\kappa_K)_+^p \end{bmatrix}_{n\times(p+K+1)}$$
$$\text{and } \boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_p \quad u_1 \quad \cdots \quad u_K]'_{(\mathrm{p+K+1})\times 1} \quad (6)$$

where p a positive integer as the exponent of the smoothing function, $(w)_+^p = w^p\, I(w \geq 0)$ is the p-degree truncated power function at a fixed node $\kappa_1 < \kappa_2 < \cdots < \kappa_K$. The recommended number of nodes is 5 to 40 nodes with the selection of κ_k being carried out based on the $k/(K+1)$ quantile of $\boldsymbol{X}$. Estimation of β is carried out using the least squares method, namely minimizing the sum of the squared penalized errors which are defined as follows:

$$J(f) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\boldsymbol{\beta'D\beta} \quad (7)$$

where $\lambda \geq 0$ and $\boldsymbol{D} = \begin{bmatrix} 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & 0 & \vdots & \vdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & 0 \\ 0 & \ldots & 0 & 0 & \ldots & 1 \end{bmatrix} = diag(\boldsymbol{0_{p+1}}, \boldsymbol{1_K})$

Minimizing $J(f)$ at a certain value of λ will compromise the good fit and the smoothness of the curve. The estimator based on the least squares method is as follows:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X'y} \quad (8)$$

2.3 P-Spline Using The Linear Mixed Model Approach

The P-Spline formulation using the linear mixed model approach becomes [10]:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \quad (9)$$

with $X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^p \end{bmatrix}$, $\beta = [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_p]'$,

$Z = \begin{bmatrix} (x_1-\kappa_1)_+^p & \ldots & (x_1-\kappa_K)_+^p \\ \vdots & \ddots & \vdots \\ (x_n-\kappa_1)_+^p & \ldots & (x_n-\kappa_K)_+^p \end{bmatrix}$, and $u = [u_1 \quad \cdots \quad u_K]'$

The sum of the squared penalized errors using a linear mixed model is:

$$\frac{1}{\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})'(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + \frac{\lambda}{\sigma_\varepsilon^2}\boldsymbol{u'u} \quad (10)$$

This function is the same as the BLUP criteria of the mixed linear model in **Equation (2)** by treating the covariance $\boldsymbol{u}$ as follows [15]:

$$cov(\boldsymbol{u}) = \sigma_u^2 \boldsymbol{I} \text{ where } \sigma_u^2 = \frac{\sigma_\varepsilon^2}{\lambda} \tag{11}$$

The P-Spline BLUP estimator using a linear mixed model method is written as follows:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}'(\sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma_\varepsilon^2 \boldsymbol{I})\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma_\varepsilon^2 \boldsymbol{I})^{-1}\boldsymbol{y} \\ \hat{u} &= \sigma_u^2 \boldsymbol{Z}'(\sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma_\varepsilon^2 \boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \end{aligned} \tag{12}$$

2.4 Spatial Autoregressive Model (SAR)

The SAR model equation can be written in matrix form as follows [19]:

$$y = \rho \boldsymbol{W} y + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \varepsilon \sim N(0, \sigma^2 \boldsymbol{I}) \tag{13}$$

where ρ is the parameter of the spatial lag coefficient of the response variable, $\boldsymbol{W}$ is the spatial weight matrix, $\boldsymbol{\beta}$ is a regression coeffisicient, and $\boldsymbol{\varepsilon}$ is a vector of normally distributed error SAR models with zero mean and σ^2 variance. Estimation is carried out using the maximum likelihood method and the β estimator is as follows:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \rho \boldsymbol{W})\boldsymbol{y} \tag{14}$$

If $\widehat{\boldsymbol{e_o}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}_o}$ obtained from the regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta_o} + \boldsymbol{e_o}$, $\widehat{\boldsymbol{e_l}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}_l}$ obtained from the regression model $\boldsymbol{W}\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta_l} + \boldsymbol{e_l}$, and $\widehat{\boldsymbol{\sigma^2}} = \frac{1}{n}(\widehat{\boldsymbol{e_o}} - \rho\widehat{\boldsymbol{e_l}})'(\widehat{\boldsymbol{e_o}} - \rho\widehat{\boldsymbol{e_l}})$, then the estimation of ρ can be simplified as follows:

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}_o} - \rho\widehat{\boldsymbol{\beta_l}} \tag{15}$$

2.5 PS-SAR

The SAR model with the mixed model P-Spline is formulated as follows [11]:

$$(\boldsymbol{I} - \rho \boldsymbol{W})\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \quad \boldsymbol{u} \sim Normal(0, \boldsymbol{G}), \quad \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{R}) \tag{16}$$

PS-SAR estimation is carried out using the maximum likelihood method with maximum function as follows:

$$\ln(L) = \ln c - \frac{1}{2}\boldsymbol{e}'\boldsymbol{R}^{-1}\boldsymbol{e} - \frac{1}{2}\boldsymbol{u}'\boldsymbol{G}^{-1}\boldsymbol{u} + \ln|\boldsymbol{I} - \rho \boldsymbol{W}| \tag{17}$$

The estimators obtained are as follows:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{I} - \rho \boldsymbol{W})\boldsymbol{y} \\ \widehat{\boldsymbol{u}} &= \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}((\boldsymbol{I} - \rho \boldsymbol{W})\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \end{aligned} \tag{18}$$

The estimation of the variance component and the ρ coefficient uses the REML method with the following criteria functions [11]:

$$\begin{aligned} \ln(L(\boldsymbol{V}, \rho))_{ML} = &-\frac{1}{2}\Big[\ln|\boldsymbol{V}| + \boldsymbol{y}'(\boldsymbol{I} - \rho \boldsymbol{W})\boldsymbol{V}^{-1}\left(\boldsymbol{I} - \boldsymbol{X}\left[\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\right)(\boldsymbol{I} - \\ &\rho \boldsymbol{W})\boldsymbol{y}\Big] - \frac{n}{2}\ln(2\pi) + \ln|\boldsymbol{I} - \rho \boldsymbol{W}| \\ \ln(L(\boldsymbol{V}, \rho))_{REML} = &\ln(L(\boldsymbol{V}, \rho))_{ML} - \frac{1}{2}\ln\left|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right| + \frac{p}{2}\ln(2\pi) \end{aligned} \tag{19}$$

2.6 Data

This study uses data from The Ministry of Health of the Republic of Indonesia, the Directorate General of Disease Prevention and Control, the Central Bureau of Statistics, and the Regional System Information and Basis Data Management of National Planning Development Agency. Data covers the number of HIV cases and the factors from 390 districts/cities in Indonesia in 2018. Table 1 shows the response variable and the explanatory variables used in this study [20].

Table 1. List of variables used in the study

Variable	Description
Y	The number of HIV cases (per 100,000 population)
X_1	The number of HIV cases in pregnant women (per 100,000 population)

X_2	General Population (per 100,000 population)
X_3	The number of Tuberculosis patients (per 100,000 population)
X_4	Percentage of poor people
X_5	Percentage of unemployment

2.7 Research Procedures

1. Data exporation
2. Form spatial weighting matrix and optimizing its parameters
3. Perform lagrange multiplier test and Breusch-Pagan test
4. Determine the best number of nodes for polynomial degree 1, 2, and 3. Optimization based on the minimum GCV value won't do because the resulting models tend to have singular matrices. Therefore, the number of nodes is determined using the Relative Deviation of GCV (RDGCV) value obtained from the smoothing spline between the Y variable and the nonlinear variable. The GCV value is calculated based on the equation as follow:

$$GCV(\lambda) = \frac{MSE(\lambda)}{\left[1 - \frac{tr(\boldsymbol{H}(\boldsymbol{\lambda}))}{n}\right]^2}, \quad (20)$$

$$with\ MSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(X, \hat{\beta})\right)^2\ and\ \boldsymbol{H}(\boldsymbol{\lambda}) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}'.$$

The Relative Deviation of GCV is calculated based on the equation as follows:

$$RDGCV_i = \frac{GCV_i - GCV_{i-1}}{GCV_{i-1}} \quad (21)$$

with GCV_i is the GCV value with i nodes, and GCV_{i-1} is the GCV value with $i-1$ nodes. If $RDGCV_i$ value is low, the addition of nodes does not significantly increase the GCV value, so the number of nodes that is better used is $i-1$.

5. Form PS-SAR models of degrees 1, 2 and 3 with the best number of spline nodes according to RDGCV.
6. Evaluate PS-SAR models based on ρ value, MSE, MAPE, and Pseudo R^2.

3. RESULTS AND DISCUSSION

3.1. Data Exploration

Indonesia is an epidemic area of HIV disease with the 3rd highest number of positive cases in Asia Pacific. Based on the data, the distribution of the number of HIV cases per 100,000 population is presented in **Figure 1**. Based on **Figure 1**, most districts/cities in Indonesia are in a low category, namely under 35 cases per 100,000 population. There are zero HIV cases in a total of 125 districts/cities. The province with the highest number of districts/cities with zero HIV cases is Papua Barat Daya Province which is 66.67% or 4 of 6 districts/cities have zero HIV cases. The district with the highest rate of HIV cases is Nabire District, Central Papua Province with 534.07 cases per 100,000 population. The district with the lowest rate of HIV cases is Langkat District, North Sumatera with 0.10 cases per 100,000 population. **Table 2** present descriptive statistics on the number of HIV cases per 100,000 population.

Table 2. Summary of HIV data

Statistics	Value
Minimum	0,10
Median	10,41
Mean	21,76
Maximum	534,07
Variance	167,66

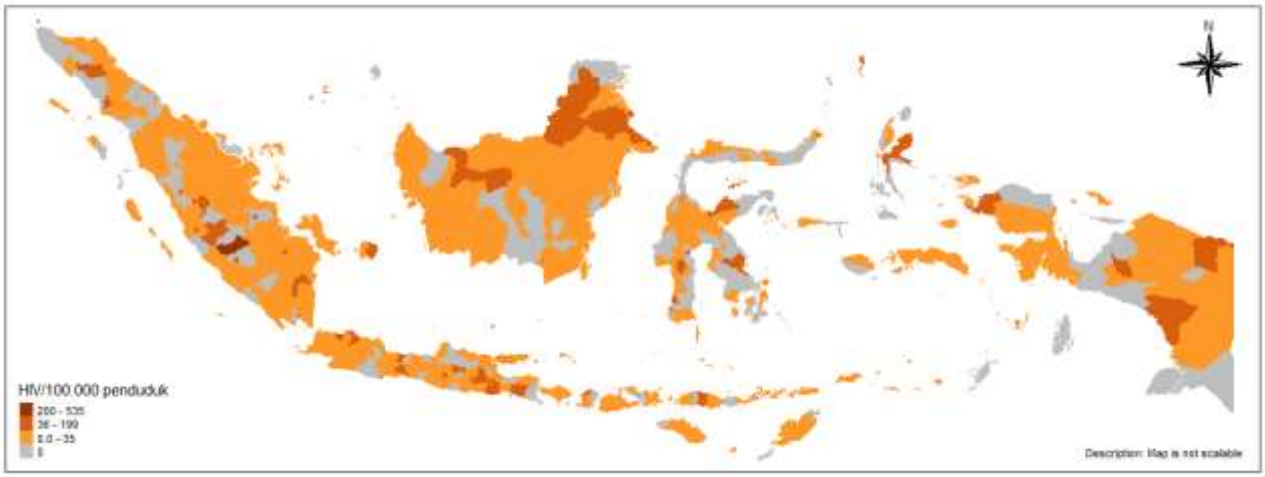

Figure 1. Thematic map of the number of HIV cases per 100,000 population

3.2 Spatial Weigth Matrix

This study applies the KNN matrix, the distance band matrix, the inverse distance matrix, and the negative exponential distance matrix. This study also optimizes the matrix parameters. The optimized parameters are the k parameter or number of neighbors in the KNN matrix, the optimum distance parameter on the distance band matrix, the α parameter or the inverse power on the inverse distance matrix, and the k parameter on the negative exponential distance matrix. Optimum parameter evaluation is carried out based on the statistical results of the Lagrange Multiplier (LMlag) test and its significance value (p-value). **Table 3** provides a summary of the evaluation outcomes of the spatial weighting matrix. Based on **Table 3**, the matrix with the highest LMlag test statistical value and the most significant p-value is the distance band matrix. Therefore, further analyses in this study use the distance band matrix.

Table 3. Summary of the evaluation outcomes of the spatial weighting matrix

Matrix	Parameter	LMlag	*p-value*
KNN	$k = 7$	4.94	0.03
DBM	$d_{max} = 327.824$	5.27	0.02
IDM	$\alpha = 1$	1.06	0.30
NEDM	$k = 0.5$	0.26	0.61

3.3 Spatial Effects Test

Spatial effects between locations can be caused by spatial dependencies and spatial heterogeneity **[1]**. This study uses the Lagrange Multiplier (LM) test to evaluate spatial dependency and the Breusch-Pagan (BP) test to evaluate spatial heterogeneity. **Table 4** shows the results of the LM test and BP test. Based on **Table 4**, the p-value of the LM and BP test is less than α = 5%. In conclusion, there is an effect of spatial dependency and spatial heterogeneity.

Table 4. The Results of the LM test and BP test

Test	Statistical result	*p-value*
LM	5.27	0.02
BP	133.98	$< 2.2 \times 10^{-16}$

3.4 Optimization of The Number of Spline Nodes

This study optimizes the number of spline nodes based on the Relative Deviation of GCV (RDGCV). The optimization is for models with spline degrees 1, 2, and 3. **Table 5** presents a summary of the RDGCV

optimization results for each spline degrees. Based on Table 5, the number of nodes used is the number of nodes that have the optimum RDGCV minus 1. For example, for a degree of spline 1, variable X1 is divided into six nodes while X2, X3, X4, and X5 are divided into five nodes. Figure 2 presents a spline smoothing plot with the optimal number of vertices for the degree of spline $m = 1$.

Table 5. Summary of the RDGCV Optimization Results for Each Spline Degrees

Model	Variable	Degree	Node	GCV	RDGCV
	X1	1	7	610.78	4.11×10^{-3}
	X2	1	6	1187.31	4.76×10^{-4}
Model 1	X3	1	6	823.51	8.67×10^{-3}
	X4	1	6	1637.90	5.26×10^{-5}
	X5	1	6	1588.70	9.57×10^{-4}
	X1	2	9	522.35	4.60×10^{-5}
	X2	2	8	1139.95	2.64×10^{-3}
Model 2	X3	2	6	596.61	1.42×10^{-4}
	X4	2	10	1625.10	5.32×10^{-4}
	X5	2	10	1581.30	9.57×10^{-5}
	X1	3	6	417.11	1.32×10^{-3}
	X2	3	8	1114.84	1.69×10^{-4}
Model 3	X3	3	6	538.33	1.60×10^{-4}
	X4	3	9	1621.30	2.35×10^{-5}
	X5	3	6	1578.40	7.06×10^{-5}

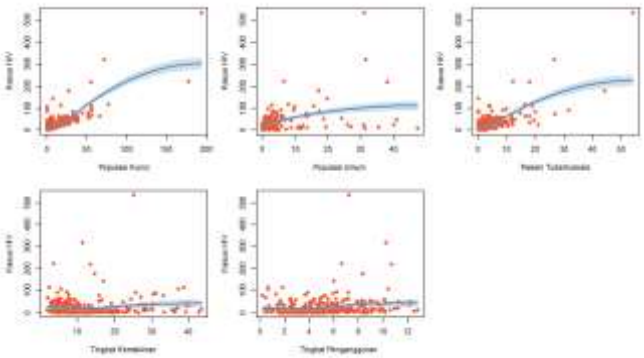

Figure 2. Smoothing spline plot of variable X with Y at spline degree m=1

3.5 Model Comparison

The PS-SAR model is formed with distance band matrices, number of nodes, and degrees according to the optimization in **Table 5**. Models are compared based on the estimated value of ρ, p-value of ρ, MSE, MAPE, and Pseudo R-square. **Table 6** provides the result of the comparison criteria. Based on **Table 6**, the estimated value of ρ in the PS-SAR model is close to the standard SAR model. In addition, the MSE and MAPE values of the PS-SAR model are already smaller than the classic regression and the SAR model. This indicates that the prediction results of the PS-SAR model are better. Based on the Pseudo R-Square value, the PS-SAR model has provided a higher Pseudo R-Square value than the classic regression model and SAR model so that the PS-SAR model better describes the diversity of the data. Based on the comparison of ρ values, the p-value of ρ, MSE, and R-Square, the best PS-SAR model is the PS-SAR model with a spline degree of 2.

Table 6. Model Comparison

Model	ρ	p-value ρ	MSE	MAPE	R-Square
Classic regression	-	-	243.52	144.02	85.36%
SAR	0.21	0.02	239.85	144.77	85.58%
Model 1	0.17	0.05	186.35	108.99	88.80%
Model 2	0.18	0.03	181.45	116.85	89.09%
Model 3	0.18	0.03	185.46	115.87	88.85%

4. CONCLUSIONS

The PS-SAR model is better at predicting the number the number of HIV cases than the classic regression model and the standard SAR model. The best spatial weight matrix used is the distance band matrix. The best PS-SAR model used is the PS-SAR model with a spline degree of 2.

REFERENCES

[1] L. Anselin, Spatial Econometrics: Method and Models. Dodrecht (NLD): Kluwer Academic Publishers, 1988.

[2] S. Sukmawati, A. Djuraidah, and A. H. Wigena, "Spatial Clustered Regression Analysis of 2017 Getis Score Indonesian Malaria Prevalence Data," *J. Phys. Conf. Ser.*, vol. 1863, no. 1, 2021, doi: 10.1088/1742-6596/1863/1/012043.

[3] A. Djuraidah, P. Silvianti, B. Djaafara, and S. N. Laila, "Modeling Annual Parasite Incidence of Malaria in Indonesia of 2017 using Spatial Regime," *Indones. J. Geogr.*, vol. 53, no. 2, pp. 185–191, 2021.

[4] A. Djuraidah, A. Alamudi, and A. Fadila, *Modeling the Number of Tuberculosis Patients in East Java with Geographically Weighted Negative Binomial Regression*. Bachelor [Undergraduate Theses]. Bogor, ID: IPB Univeristy., 2021. [Online]. Available: IPB Repository.

[5] A. Djuraidah, Z. Mar'ah, and R. Anisa, "a Bayesian Conditional Autoregressive With Inla: a Case Study of Tuberculosis in Java, Indonesia," *Commun. Math. Biol. Neurosci.*, vol. 2022, pp. 1–15, 2022, doi: 10.28919/cmbn/7709.

[6] Z. D. R, A. Saefuddin, and A. Djuraidah, "Modelling the Number of Cases of Dengue Hemorragic Fever with Mixed Geographically Negative Binomial Regression in West Java Province," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 18116, no. Kemenkes 2014, pp. 71–77, 2019, doi: 10.32628/ijsrset196124.

[7] Z. Chen and J. Chen, "Bayesian analysis of partially linear additive spatial autoregressive models with free-knot splines," *Symmetry (Basel).*, vol. 13, no. 9, pp. 1–20, 2021, doi: https://doi.org/10.3390/sym13091635.

[8] P. H. C. Eilers and B. D. Marx, "Flexible smoothing with B-splines and penalties," *Stat. Sci.*, vol. 11, no. 2, pp. 89–102, 1996, doi: 10.1214/ss/1038425655.

[9] D. Ruppert and R. J. Carroll, "*Penalized regression splines*," 1997.

[10] B. A. Brumback and J. A. Rice, "Smoothing spline models for the analysis of nested and crossed samples of curves," *J. Am. Stat. Assoc.*, vol. 93, no. 443, pp. 961–976, 1998, doi: https://doi.org/10.2307/2669837.

[11] J. Montero, R. Mínguez, and M. Durbán, "SAR models with nonparametric spatial trends: A P-spline approach," *Estadística Española*, vol. 54, no. 177, pp. 89–111, 2012.

[12] K. H. dan H. R. Dirjen P2 & PL (Direktorat Jendral Pengendalian Penyakit dan Penyehatan Lingkungan), Kementerian Kesehatan RI, Direktorat Jendral Permasyarakatan, *Pedoman Layanan Komprehensif HIV-AIDS & AIMS di Lapas, Rutan dan Bapas*. Jakarta: Dirjen P2 & PL, 2012.

[13] D. B. Lolong, O. S. Simarmata, Novianti, and F. P. Senewe, "Situasi Human Immunodeficiency Virus - Tuberkulosis di Kabupaten Merauke 2018: Ancaman pada umur produktif," *J. Kesehat. Reproduksi*, vol. 10, no. 1, pp. 1–9, 2019, doi: 10.22435/kespro.v10i1.1711.1-9.

[14] I. Bates *et al.*, "Vulnerability to malaria, tuberculosis, and HIV/AIDS infection and disease. Part 1: Determinants operating at individual and household level," *Lancet Infect. Dis.*, vol. 4, no. 5, pp. 267–277, 2004, doi: 10.1016/S1473-3099(04)01002-3.

[15] A. Djuraidah, *Model aditif spatio-temporal untuk pencemar udara PM10 dan ozon di Kota Surabaya dengan pendekatan model linear campuran*. PhD [Dissertation]. Bogor, ID: IPB Univeristy., 2007. [Online]. Available: IPB Repository.

[16] C. R. Henderson, O. Kempthorne, S. R. Searle, and C. M. Krosigk, "The estimation of environmental and genetic trends from records subject to culling," *Biometrics*, vol. 15, no. 2, pp. 192–218, 1959, [Online]. Available: http://www.jstor.org/stable/2527669.

[17] R. Christensen, *The Theory of Linear Models*. New Mexico (USA), 1987.

[18] S. R. Searle, G. Casella, and C. E. McCulloch, *Variance Components*. New Jersey (USA): John Wiley & Sons, Inc., 1992.

[19] J. LeSage and R. K. Pace, *Introduction to Spatial Econometrics*. Boca Raton (USA): CRC Press, 2009.

[20] M. B. El-Kautsar, A. Djuraidah, and Y. Angraini, *Identifikasi faktor-faktor yang memengaruhi kasus HIV di Indonesia tahun 2018 menggunakan regresi terboboti geografis campuran*. Bachelor [Undergraduate Theses]. Bogor, ID: IPB Univeristy., 2022. [Online]. Available: IPB Repository.