1st International Conference on Multidisciplinary Engagement

Analysis of the Percentage of the Poor Population in Regencies and Cities on the Island of Java

Elfrida Arroudhatun Nadya[1*], Dewi Retno Sari Saputro[2], Purnami Widyaningsih[3]
[1,2,3] *Department of Mathematics, Faculty of Mathematics and Natural Science, Sebelas Maret University*
Ir. Sutami No. 36, Surakarta, 57126, Indonesia

Corresponding Author.
**Email*: elfrida_nadya11@student.uns.ac.id

Abstract: *Poverty remains a significant issue in development on the island of Java, even though this region is the center of national economic activity. The relationship between socioeconomic factors and the percentage of the poor population is not always linear, necessitating an approach that better accommodates the data patterns. This study aims to analyze the percentage of the poor population using a B-Spline semiparametric regression approach. The variables used are the Human Development Index (HDI), Labor Force Participation Rate (LFPR), Open Unemployment Rate (OUR), and percentage of the population. The research data used consists of data from 119 regencies/cities in the Java region in 2024, which was then divided into 80% training data and 20% testing data. The selection of the order and knot points was performed using cross-validation (CV). The best combination was obtained at the second order for the variables z_1 and z_2, and a third-order model for the variable z_3 which yields a minimum CV value of 8.891438. Model evaluation yields a Mean Absolute Percentage (MAPE) of 20.97% on the training data and 30.04% on the testing data, which falls into the fairly accurate category. Overall, the model is able to comprehensively describe the relationship between the predictor variables and the percentage of the poor population. The HDI has a consistent negative linear effect in reducing poverty, while the LFPR, OUR, and percentage of the population show nonlinear relationships that vary across each value interval. This indicates that the B-Spline semiparametric regression approach is effective in capturing the complex patterns of relationships between socioeconomic factors and poverty levels.*
Keywords: *Semiparametric Regression, B-Spline, Poverty, Cross Validation*

1. INTRODUCTION

Poverty is a condition in which individuals or groups of people are unable to adequately meet their basic needs, such as food, clothing, shelter, education, and health care [1]. This condition is not only related to low income levels but also reflects limitations in accessing the various resources needed to improve living standards. Meanwhile, the percentage of the poor population indicates the proportion of the population whose average monthly per capita expenditure falls below the established poverty line [2].

To this day, the issue of poverty remains a barrier to social and economic development in various regions of Indonesia, including Java. Java, as the center of national economic activity, still faces the problem of regional disparities in welfare. Based on data from the Central Statistics Agency [3], the percentage of the poor population in 2024 in every province on Java Island remains quite high. Central Java is recorded as having a poverty rate of 10.47%, East Java at 9.79%, and West Java at 7.46%.

The relationship between predictor variables and the percentage of the poor population is generally analyzed using parametric regression, such as linear regression. This type of regression has several advantages, namely its simple structure, easily interpretable results, and efficiency when the specified conditions are met [4], [5], [6]. However, parametric regression also has limitations, particularly when the relationship between variables. The predictors and response do not follow a specific functional form assumed in parametric regression approaches. This condition causes parameter estimates to be less accurate or even biased [7]. As a solution, nonparametric regression is used, which does not specify a particular functional form from the outset. This type of regression better fits the data because it directly captures nonlinear patterns. One widely used method is *B-Spline*, which has several advantages, such as consistent calculation results, the ability to adjust data patterns in specific sections without affecting the

overall data, and a faster calculation process, thereby capturing complex relationships [8], [9]. Semiparametric regression utilizes *splines* to fit the shape of nonlinear relationships based on data characteristics. In this regression, the relationship is divided into simple segments that are arranged sequentially to form a *spline* [10].

Research on poverty modeling has been extensively conducted using various statistical approaches. Nonparametric approaches have also begun to be widely applied, such as the study by Li et al. [11] in 2007, which used spline regression to identify nonlinear relationships between socioeconomic indicators and poverty levels. A more flexible approach was used by Fotheringham et al. [12] in 2014 through the Geographically Weighted Regression (GWR) method to analyze regional poverty, which is capable of capturing the spatial variation in the relationships between variables in each region. In 2016, Ravallion [13] analyzed the determinants of global poverty using a cross-country data-based parametric regression approach, which showed that economic growth and inequality have a significant impact on poverty rates. Furthermore, in 2024, Huang [14] modeled poverty rates in the Asian region using panel data regression with the Fixed Effects Model (FEM) and Random Effects Model (REM) approaches, which assume a linear relationship between the response variable and the predictors.

Semiparametric regression has been applied in various fields of research. In 2003, Ruppert et al. [15] developed a basic framework for semiparametric regression to accommodate flexible relationships between variables. Furthermore, Wood [16] in 2017 demonstrated that the spline approach in semiparametric models is highly effective in capturing nonlinear relationships through Generalized Additive Models (GAM). In 2024, Zou and Zhu [17] developed a spline-based approach to improve estimation accuracy through more optimal knot selection. In contrast to these studies, this research employs a B-Spline semiparametric regression approach with knot point selection based on the minimum Cross-Validation (CV) value to model the percentage of the poor population on Java Island, thereby aiming to provide more optimal estimation results in capturing complex relationship patterns.

2. METHOD

This section describes the data sources, research steps, and methods applied in semiparametric regression modeling using the B-Spline function.

2.1 Research Data

The data sources used in this study consist of secondary data obtained from the Central Statistics Agency (BPS). The data covers 119 regencies/cities on the island of Java in 2024, with the percentage of the poor population Y as the response variable and the following as predictor variables X_1, X_2, X_3, and X_4. Next, the data was divided into 80% as training data to develop the B-Spline semiparametric regression approach and 20% as testing data to evaluate the performance of the B-Spline semiparametric regression approach. Details of the research variables are shown in Table 1.

Table 1. Details of the research data variables

Variable	Research data
Y	Percentage of the poor population
X_1	Human Development Index (HDI)
X_2	Labor Force Participation Rate (LFPR)
X_3	Open Unemployment Rate (OUR)
X_4	Percentage of the population

2.2 Analysis Steps

The data analysis process was conducted through several stages, as described below.

1. Preprocessing data consists of collecting data on the percentage of the poor population, the Human Development Index (HDI), the Labor Force Participation Rate (LFPR), the Open Unemployment Rate (OUR), and the percentage of the population. Next, cleaning data is performed.
2. Exploratory data analysis (EDA) involves determining descriptive statistics.
3. Split the data into training data (80%) and testing data (20%).
4. Create a scatter plot visualizing the relationship between the response variable and the predictors

using the training data.

5. Identify linear or nonlinear relationship patterns to determine parametric and nonparametric components.
6. Select the order and combination of knot points, then calculate the Cross-Validation (CV) score for each combination.
7. Determine the optimal knot points based on the minimum Cross-Validation (CV) value.
8. Construct the B-Spline basis function according to the optimal order and knot points.
9. Forming a B-Spline semiparametric regression model.
10. Estimating parameters using the Ordinary Least Squares method.
11. Evaluation and validation include calculating the Mean Absolute Percentage Error and testing the data.
12. Analysis of results.

2.3 Parametric Regression

Parametric regression is a regression method used to analyze the relationship between a response variable and one or more predictor variables, assuming that the form of the relationship function has been determined in advance [18]. This approach is easy to interpret and efficient as long as the conditions are met [19]. The parametric regression approach is written as

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + \varepsilon_i \text{ for } i = 1,2,\dots,n \quad (1)$$

where Y_i is the nth response variable i, $x_{1i}, x_{2i}, \dots, x_{ni}$ are the nth predictor variables i, β_0 is the constant term, $\beta_1, \beta_2, \dots, \beta_n$ are the regression coefficients, and ε_i is the nth error term i, which is assumed $\varepsilon_i \sim N(0, \sigma^2)$.

2.4 Nonparametric B-Spline Regression

Nonparametric regression is a regression method that allows the form of the regression function to be determined based on data patterns so that it captures complex or nonlinear relationships [20]. This regression reduces bias due to specification errors and provides more accurate predictions when the data has unstructured patterns [21]. The nonparametric regression formula is written as

$$Y_i = g(z_i) + \varepsilon_i \text{ for } i = 1,2,\dots,n \quad (2)$$

where $g(z_i)$ is a nonparametric regression function used to describe nonlinear relationships and z_i is the i th predictor variable.

A spline is a function used in nonparametric regression, where one of the commonly used basis functions
is the B-Spline. According to [22], a B-Spline is a basis function constructed in stages and is bounded within a specific range, making it highly efficient for computation. If the function $g(z_i)$ in Equation (2) is approximated using a B-Spline, the general form of the B-Spline is written as

$$g(z_i) = \sum_{j=1}^{m+k} \gamma_j B_{j-m,m}(z_i) \text{ for } i = 1,2,\dots,n \quad (3)$$

where γ_j is a parameter for the B-spline and $B_{j-m,m}(z_i)$ is the B-spline basis.

How to construct *a B-spline* of order m and k knot points with $a < u_1 < \cdots < u_k < b$, requiring an additional $2m$ is $u_{-(m-1)}, \dots, u_{-1}, u_0, u_1, u_2, \dots, u_k, u_{k+1}, \dots, u_{k+m}$. These additional knot points are defined by $u_{-(m-1)} = \cdots = u_{-1} = u_0 = a$ and $u_{k+1} = \cdots = u_{k+m} = b$ where a is the minimum value of the data z and b is the maximum value of the data z. According to [23], the basis of the *B-Spline* function of order m and the knot points $u_1, u_2, \dots, u_k$ are written as

$$B_{j,m}(z) = \frac{z - u_j}{u_{j+m-1} - u_j} B_{j,m-1}(z) + \frac{u_{j+m} - z}{u_{j+m} - u_{j+1}} B_{j+1,m-1}(z)$$

where $j = -(m-1), \dots, k$ and the initial condition (order $m = 1$) is written as

$$B_{j,1}(z) = \begin{cases} 1, & u_j \le z < u_{j+1} \\ 0, & \text{otherwise} \end{cases}$$

According to [9], the order m is the degree of the polynomial plus one. Based on their order, B-spline forms are classified into three categories, namely

1. Order $m = 2$, A B-spline is linear, so the basis is written as
$$B_{j,2}(z) = \frac{z - u_j}{u_{j+1} - u_j} B_{j,1}(z) + \frac{u_{j+2} - z}{u_{j+2} - u_{j+1}} B_{j+1,1}(z) \text{ for } j = -1, \dots, k$$
2. Order $m = 3$, A B-spline is quadratic, so the basis is written as
$$B_{j,3}(z) = \frac{z - u_j}{u_{j+2} - u_j} B_{j,2}(z) + \frac{u_{j+3} - z}{u_{j+3} - u_{j+1}} B_{j+1,2}(z) \text{ for } j = -2, \dots, k$$
3. Order $m = 4$, A B-spline is cubic, so the basis is written as
$$B_{j,4}(z) = \frac{z - u_j}{u_{j+3} - u_j} B_{j,3}(z) + \frac{u_{j+4} - z}{u_{j+4} - u_{j+1}} B_{j+1,3}(z) \text{ for } j = -3, \dots, k$$

2.5 B-Spline Semiparametric Regression

B-Spline semiparametric regression is a regression method that combines parametric regression and nonparametric regression approximated by B-Splines. This approach provides a balance between the ability to fit data patterns and ease of interpretation, while addressing the limitations of parametric regression, which is too restrictive, and nonparametric regression, which is too free-form [15]. The B-Spline semiparametric regression approach is written as

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + g(z_i) + \varepsilon_i$$

The function $g(z_i)$ is written as in Equation (3). Therefore, the general formula for B-spline semiparametric regression is written as

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + \sum_{j=1}^{m+k} \gamma_j B_{j-m,m}(z_i) + \varepsilon_i \quad (4)$$

Equation (4), when written in matrix form, is expressed as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{B\gamma} + \boldsymbol{\varepsilon} \quad (5)$$

Next, let $\boldsymbol{C} = [\boldsymbol{XB}]$ and $\boldsymbol{\omega} = [\boldsymbol{\beta\gamma}]$ so that Equation (5) is written as

$$\boldsymbol{Y} = \boldsymbol{C\omega} + \boldsymbol{\varepsilon} \quad (6)$$

The vectors in matrix (5) are written in full as

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{i1} \\ 1 & x_{12} & \cdots & x_{i2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{in} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, \boldsymbol{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{m+k} \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} B_{-(m-1),m}(z_1) & B_{-(m-2),m}(z_1) & \cdots & B_{k,m}(z_1) \\ B_{-(m-1),m}(z_2) & B_{-(m-2),m}(z_2) & \cdots & B_{k,m}(z_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_{-(m-1),m}(z_n) & B_{-(m-2),m}(z_n) & \cdots & B_{k,m}(z_n) \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where $\boldsymbol{Y}$ is an $n \times 1$ response vector, X is an $n \times (k+1)$ matrix of predictor variables in the parametric component, $\boldsymbol{\beta}$ is a $(k+1) \times 1$ parametric regression parameter vector, B is the B -Spline basis functions of size $n \times (m+k)$, $\boldsymbol{\gamma}$ is the B-Spline parameter vector of size $(m+k) \times 1$, and $\boldsymbol{\varepsilon}$ is the random error vector assumed to be normally distributed with a mean of zero and variance $\sigma^2 \boldsymbol{I}$.

The parameter estimates were obtained using the Ordinary Least Squares (OLS) method, which minimizes the sum of the squares of the errors. The error matrix for Equation (6) is written as

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{C\omega}$$

by minimizing the sum of squared residuals (SSR), which is expressed as

$$SSR = \varepsilon^T \varepsilon = (\boldsymbol{Y} - \boldsymbol{C\omega})^T (\boldsymbol{Y} - \boldsymbol{C\omega})$$

$$= (\boldsymbol{Y}^T - \boldsymbol{\omega}^T\boldsymbol{C}^T)(\boldsymbol{Y} - \boldsymbol{C\omega})$$
$$= \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\omega}^T\boldsymbol{C}^T\boldsymbol{Y} + \boldsymbol{\omega}^T\boldsymbol{C}^T\boldsymbol{C\omega}$$

to minimize the sum of squared errors, the SSR is first differentiated with respect to $\boldsymbol{\omega}$ and set equal to zero, so it is written as

$$-2\boldsymbol{C}^T\boldsymbol{Y} + 2\boldsymbol{C}^T\boldsymbol{C}\widehat{\boldsymbol{\omega}} = 0$$
$$\boldsymbol{C}^T\boldsymbol{C}\widehat{\boldsymbol{\omega}} = \boldsymbol{C}^T\boldsymbol{Y}$$
$$\widehat{\boldsymbol{\omega}} = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{Y}$$

The predicted value can be written as

$$\hat{Y} = \boldsymbol{C}\widehat{\boldsymbol{\omega}} = \boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{Y}$$

or written as

$$\hat{Y} = HY$$

with

$$H = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{Y}$$

H is a hat matrix that is symmetric and idempotent.

2.6 Selection of Knot Points

Cross Validation (CV) is a method used to assess how well a spline predicts data, including cases where the knot points are not specified in advance [24]. The CV formula is written as

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{Y}_i}{1 - a_{ii}}\right)^2$$

This formula represents a form of leave-one-out Cross Validation (LOOCV). For each data point i, the value $\hat{Y}_i$ is the spline prediction obtained using the entire dataset, while the correction factor $(1 - a_{ii})$ adjusts for the influence of data point i, which contributes to the formation of the predicted value. The element a_{ii} is a diagonal element of the smoothing matrix or hat matrix from the nonparametric B-Spline regression parameter estimation, namely $\hat{Y}_i = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{Y}$. The value of a_{ii} indicates the magnitude of the influence of the i-th data point on the estimated value $\hat{Y}_i$. The larger the value of a_{ii}, s the greater the contribution of the data to the formation of the estimate.

2.7 Mean Absolute Percentage Error (MAPE)

MAPE is used to measure the magnitude of prediction error, expressed as a percentage of the actual value. The MAPE formula is written as

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{Y}_i}{y_i}\right| \times 100\%$$

where $\hat{Y}_i$ is the predicted value from the parameter estimation. In general, a MAPE value of less than 10% indicates that the model has a very high level of accuracy, 10%–20% indicates that the model has an accurate level of accuracy, 20%–50% indicates that the model has a fairly accurate level of accuracy, and a value greater than 50% indicates that the model has an inaccurate level of accuracy [25].

3. RESULT AND DISCUSSION

This section describes the data, presents a scatter plot to examine the relationship patterns among variables, determines the optimal semiparametric B-Spline regression approach based on a combination of order and knot points using Cross Validation, assesses the fit between actual data and predicted results, and analyzes the outcomes.

3.1 Description Data

The data used in this study are from the Java Island region in 2024, with the percentage of the poor population as the response variable (Y) and the predictor variables being the Human Development Index (X_1), Labor Force Participation Rate (X_2), Open Unemployment Rate (X_3), and percentage of the population (X_4). Descriptive statistics for each response and predictor variable are shown in Table 2.

Table 2. Descriptive statistics for each response and predictor variables

Variables	Max	Min	Median	Mean
Y	20.83	2.34	8.98	9.07
X_1	88.77	64.65	74.10	74.95
X_2	86.62	58.13	71.92	71.58
X_3	9.18	1.56	4.91	5.03
X_4	31.08	0.27	2.90	5.04

Table 2 shows that the percentage of the population (Y) has a highest value of 20.83 and a lowest value of 2.34, with an average of 9.07 and a median of 8.98. The Human Development Index (X_1) variable has a highest value of 88.77 and a lowest value of 64.65, with a mean of 74.95 and a median of 74.10. The variable Labor Force Participation Rate (X_2) has a highest value of 86.62 and a lowest value of 58.13, with an average of 71.58 and a median of 71.92. The Open Unemployment Rate variable (X_3) has a highest value of 9.18% and a lowest value of 1.56%, with a mean of 5.03 and a median of 4.91. The percentage of the population variable (X_4) has a highest value of 31.08 and a lowest value of 0.27, with a mean of 5.04 and a median of 2.90.

3.2 Scatter Plot

In this study, a scatter plot was used to observe the relationship between the response variable and the predictor variables so that both parametric and nonparametric components could be identified. If the scatter plot shows a regular pattern, such as a straight line or a specific pattern, then a parametric regression approach is generally more appropriate. Conversely, if the distribution of points on the scatter plot does not form a clear pattern or appear random, then a nonparametric regression approach is more appropriate to apply. The scatter plot illustrating the relationship between the response variable the percentage of the poor population and each predictor variable is shown in Figure 1.

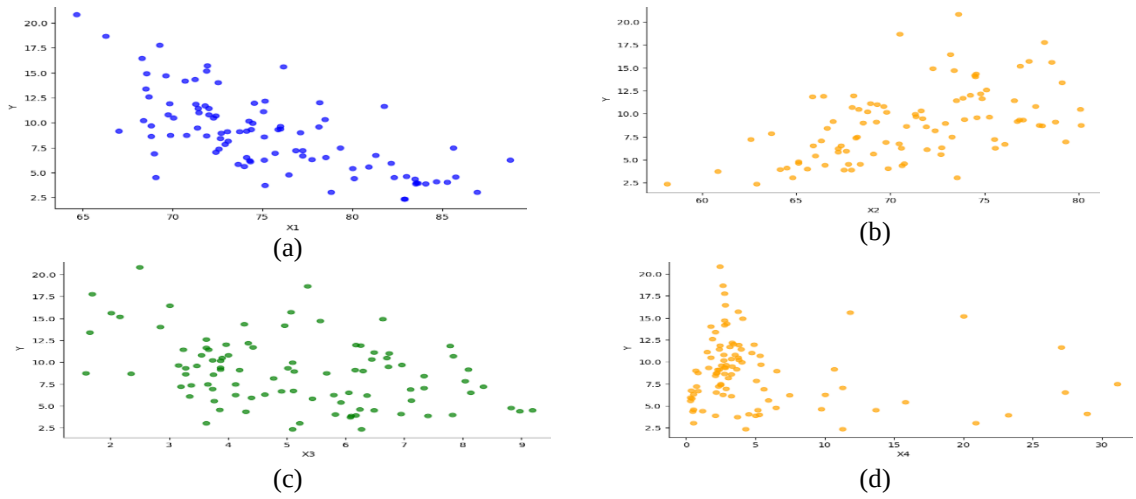

Figure 1. Scatter plot: (a) Relationship between X_1 and Y; (b) Relationship between X_2 and Y; (c) Relationship between X_3 and Y; (d) Relationship between X_4 and Y

Figure 1(a) shows that the scatter plot of the relationship between Y and X_1 forms a pattern that tends to be a straight line or linear with a negative correlation; therefore, the variable X_1 uses a parametric component denoted by x. Meanwhile, Figure 1(b) shows the scatter plot of the relationship between Y and X_2, Figure 1(c) shows the scatter plot of the relationship between Y and X_3, and Figure 1(d) shows the scatter plot of the relationship between Y and X_4 which displays a relationship pattern that does not clearly

form a straight line and tends to scatter irregularly, thus indicating a nonlinear pattern. Therefore, variables X_2, X_3 and X_4 are better suited to using non-parametric components denoted by z_1, z_2, z_3.

3.3 Determining the Best B-Spline Approach

The best B-spline approach was determined by selecting a combination of second-order (linear) and third-order (quadratic) terms, along with one and two knot points for each predictor variable. Since each combination yielded different results, an evaluation was necessary to select the best combination. The optimal knot points are determined using the CV method by calculating the CV value for each combination of order and number of knot points. The combination with the smallest CV value is selected as the best approach because it indicates a lower prediction error rate compared to other combinations. The best CV values from various combinations of order and number of knot points for each variable z_1, z_2, z_3 are shown in Table 3.

Table 3. Results of experiments on the combination of orders and number of knot points z_1, z_2, z_3

Orders			Knot points			CV scores
z_1	z_2	z_3	z_1	z_2	z_3	
2	2	2	71.610000	6.280000	3.020000	8.896512
2	**2**	**3**	**71.610000**	**6.280000**	**2.255000; 3.020000**	**8.891438**
2	2	2	71.610000	5.070000; 6.280000	3.020000	9.176437
2	2	3	71.610000	5.070000; 6.280000	2.255000; 3.020000	9.152965
2	2	2	67.860000; 71.610000	6.280000	3.020000	9.173522
2	2	3	67.860000; 71.610000	6.280000	2.255000; 3.020000	9.234882
2	2	2	67.860000; 71.610000	5.070000; 6.280000	3.020000	9.491913
2	2	3	67.860000; 71.610000	5.070000; 6.280000	2.255000; 3.020000	9.660251

Table 3 shows that the best fit is obtained with a second-order combination for the variables z_1, z_2 and a third-order combination for the variable z_3 with one knot point each at z_1 (71.610000), one knot at z_2 (6.280000), and two knots at z_3 (2.255000 and 3.020000). This combination yields the minimum CV value of 8.891438 compared to other order combinations and numbers of knots, and is therefore selected as the best B-Spline approximation.

The parameter estimates of the semiparametric B-spline regression model obtained using a combination of second-order B-splines for z_1 and z_2, third-order B-splines for z_3 and one knot point each for z_1 and z_2, and two knot points for z_3 are expressed as

$$Y_i = \beta_0 + \beta_1 x + \gamma_{11} B_{-1,2}(z_1) + \gamma_{12} B_{0,2}(z_1) + \gamma_{13} B_{1,2}(z_1) + \gamma_{21} B_{-1,2}(z_2) + \gamma_{22} B_{0,2}(z_2) + \gamma_{23} B_{1,2}(z_2) + \gamma_{31} B_{-2,3}(z_3) + \gamma_{32} B_{-1,3}(z_3) + \gamma_{33} B_{0,3}(z_3) + \gamma_{34} B_{1,3}(z_3) + \gamma_{35} B_{2,3}(z_3)$$

Based on the results of data processing using Google Colab Python, the estimated parameter values for the best B-Spline semiparametric regression approach are shown in Table 4.

Table 4. Variables, parameters, and parameter estimations

Variables	Parameters	Parameter estimations
	β_0	38.592272
x_1	β_1	-0.365666
z_1	γ_{11}	0.399895
	γ_{12}	5.226114
	γ_{13}	0.140288
z_2	γ_{21}	-6.905872
	γ_{22}	-2.721704
	γ_{23}	-6.929847
z_3	γ_{31}	-2.448788
	γ_{32}	0.817829
	γ_{33}	-5.754671
	γ_{34}	2.924526

γ_{35}	-0.606696

Table 4 shows the parameter estimates obtained from the best B-Spline semiparametric regression model, such that the resulting model is expressed as

$$Y_i = 38.592272 - 0.365666x + 0.399895B_{-1,2}(z_1) + 5.226114B_{0,2}(z_1) + 0.140288B_{1,2}(z_1) - 6.905872B_{-1,2}(z_2) - 2.721704B_{0,2}(z_2) - 6.929847B_{1,2}(z_2) - 2.448788B_{-2,3}(z_3) + 0.817829B_{-1,3}(z_3) - 5.754671B_{0,3}(z_3) + 2.924526B_{1,3}(z_3) - 0.606696B_{2,3}(z_3)$$

The resulting B-spline function is expressed as

$$B_{-1,2}(z_1) = \begin{cases} \frac{71.61 - z_1}{13.48}, & 58.13 \le z_1 < 71.61 \\ 0, & \text{otherwise} \end{cases}$$

$$B_{0,2}(z_1) = \begin{cases} \frac{z_1 - 58.13}{13{,}48}, & 58.13 \le z_1 < 71.61 \\ \frac{80.12 - z_1}{8.51}, & 71.61 \le z_1 < 80.12 \end{cases}$$

$$B_{1,2}(z_1) = \begin{cases} \frac{z_1 - 71.61}{8.51}, & 71.61 \le z_1 < 80.12 \\ 0, & \text{otherwise} \end{cases}$$

$$B_{-1,2}(z_2) = \begin{cases} \frac{5.07 - z_2}{3.49}, & 1.58 \le z_2 < 5.07 \\ 0, & \text{otherwise} \end{cases}$$

$$B_{0,2}(z_2) = \begin{cases} \frac{z_2 - 1.58}{3.49}, & 1.58 \le z_2 < 5.07 \\ \frac{9.18 - z_2}{4.11}, & 5.07 \le z_2 < 9.18 \end{cases}$$

$$B_{1,2}(z_2) = \begin{cases} \frac{z_2 - 5.07}{4.11}, & 5.07 \le z_2 < 9.18 \\ 0, & \text{otherwise} \end{cases}$$

$$B_{-2,3}(z_3) = \begin{cases} \left(\frac{3.02 - z_3}{2.75}\right)^2, & 0.27 \le z_3 < 3.02 \\ 0, & \text{otherwise} \end{cases}$$

$$B_{-1,3}(z_3) = \begin{cases} \frac{(z_3 - 0.27)(3.02 - z_3)}{7.56} + \frac{(31.08 - z_3)(z_3 - 0.27)}{77.17}, & 0.27 \le z_3 < 3.02 \\ \left(\frac{31.08 - z_3}{28.06}\right)^2, & 3.02 \le z_3 < 31.08 \end{cases}$$

$$B_{0,3}(z_3) = \begin{cases} \frac{(z_3 - 0.27)^2}{77{,}17}, & 0.27 \le z_3 < 3.02 \\ \frac{(z_3 - 0.27)(31.08 - z_3)}{864.53} + \frac{(31.08 - z_3)(z_3 - 3.02)}{787.36}, & 3.02 \le z_3 < 31.08 \end{cases}$$

$$B_{1,3}(z_3) = \begin{cases} \left(\frac{z_3 - 3.02}{28.06}\right)^2, & 3.02 \le z_3 < 31.08 \\ 0, & \text{otherwise} \end{cases}$$

3.4 Model Accuracy Evaluation

Based on the model evaluation results using training and testing data, a MAPE value of 20.97% was obtained for the training data and 30.04% for the testing data. The MAPE value for the training data indicates that the prediction accuracy falls into the "fairly accurate" category. Meanwhile, the MAPE value on the test data, which is higher than that on the training data by approximately 9.07%, indicates a decline in model performance on new data, although it remains in the fairly accurate category.

3.5 Analysis of Results

The next step is to interpret the influence of each predictor variable on the response variable.

3.5.1 The Effect of the Human Development Index (HDI) on the Percentage of the Poor Population

A coefficient of −0.365666 indicates that a 1-unit increase in the Human Development Index (HDI) reduces the percentage of the poor population by 0.365666%, assuming all other variables remain constant. This suggests that as the level of human development encompassing education, health, and living standards increases, poverty rates tend to decrease. This relationship is linear and negative, meaning that the Human Development Index (HDI) acts as a consistent factor in reducing poverty.

3.5.2 The Effect of Labor Force Participation Rate (LFPR) on the Percentage of the Poor Population

The effect of the Labor Force Participation Rate (LFPR) is modeled using a B-Spline function written as

$$0.399895B_{-1,2}(z_1) + 5.226114B_{0,2}(z_1) + 0.140288B_{1,2}(z_1)$$

This indicates that the relationship between the Labor Force Participation Rate (LFPR) and the percentage of the poor population is not linear, but rather depends on the intervals of the Labor Force Participation Rate (LFPR) defined by the knot points. In general, the varying coefficients across each basis function indicate that changes in the Labor Force Participation Rate (LFPR) can have different effects on poverty at specific levels. In some value ranges, an increase in the Labor Force Participation Rate (LFPR) may correlate with a decrease in poverty, but in other ranges, the effect may weaken or change. Thus, the Labor Force Participation Rate (LFPR) has a flexible and non- constant relationship with the poverty rate.

3.5.3 The Effect of the Open Unemployment Rate (OUR) on the Percentage of the Poor Population

The effect of the Open Unemployment Rate (OUR) is written as

$$-6.905872B_{-1,2}(z_2) - 2.721704B_{0,2}(z_2) - 6.929847B_{1,2}(z_2)$$

The fact that the dominant coefficient is negative indicates that, within certain ranges of values, changes the Open Unemployment Rate (OUR) tends to reduce or increase poverty unevenly. This indicates that the relationship between the Open Unemployment Rate (OUR) and poverty is nonlinear and complex, so it cannot be explained by a single direction of relationship. In this context, the unemployment rate does not always directly increase poverty, as it can be influenced by other factors such as economic structure and the informal sector.

3.5.4 The Effect of the Percentage of the Population on the Percentage of the Poor Population

The effect of the percentage of the population variable is modeled using a third-order spline, written as

$$-2.448788B_{-2,3}(z_3) + 0.817829B_{-1,3}(z_3) - 5.754671B_{0,3}(z_3) + 2.924526B_{1,3}(z_3) - 0.606696B_{2,3}(z_3)$$

The use of a third-order spline indicates that the relationship between population percentage and poverty exhibits a more complex pattern and can change subtly across various value intervals. The variation in the signs and magnitudes of the coefficients suggests that at certain population density levels, the impact on poverty may differ. Under some conditions, an increase in population may heighten economic pressure and poverty; however, under other conditions, it may be associated with higher economic activity, so the impact is not always negative.

4. CONCLUSION

Based on the analysis results, it was concluded that the best semiparametric B-Spline regression approach for analyzing the percentage of the poor population in regencies and cities on the island of Java was obtained using a combination of a second-order spline with one knot point at the Labor Force Participation Rate (LFPR) variable at 71.61, second-order with one knot point on the Open

Unemployment Rate (OUR) variable at 6.28, and third-order with two knot points on the percentage of the total population variable at 2.255 and 3.020, resulting in a minimum Cross Validation value of 8.891438. Model evaluation shows a Mean Absolute Percentage Error (MAPE) of 20.97% on the training data and 30.04% on the testing data, indicating that the predictive capability falls into the "fairly accurate" category. In general, this model is capable of comprehensively describing the relationship between the predictor variables and the percentage of the poor population, where the Human Development Index (HDI) has a consistent negative linear effect in reducing poverty, while Labor Force Participation Rate (LFPR), Open Unemployment Rate (OUR), and the percentage of the population show nonlinear relationships that vary across each value interval, meaning their effects are not constant and depend on specific conditions.

REFERENCES

[1] E. Muti'ah, M. Listiani, and N. N. Sukma, "Poverty measurement: Income thresholds and multidimensional approaches," *Bina Bangsa International Journal of Business and Management*, vol. 4, no. 3, pp. 1–10, 2024, doi: 10.46306/bbijbm.v4i3.104.

[2] D. S. Dedduwakumara, L. A. Prendergast, and R. G. Staudte, "Insights and inference for the proportion below the relative poverty line," *arXiv preprint arXiv:1908.08133*, 2019.

[3] Badan Pusat Statistik, *Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen), 2024*. Jakarta, Indonesia: Badan Pusat Statistik, 2024. [Online]. Available: https://www.bps.go.id

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 2nd ed. New York, NY, USA: Springer, 2021.

[5] J. M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th ed. Boston, MA, USA: Cengage Learning, 2020.

[6] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2008.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2017.

[8] L. L. Schumaker, *Spline Functions: Basic Theory*, 3rd ed. Cambridge, U.K.: Cambridge University Press, 2007.

[9] P. H. C. Eilers and B. D. Marx, "Practical smoothing: The joy of P-splines," *Statistical Science*, vol. 36, no. 3, pp. 417–438, 2021, doi: 10.1214/20-STS792.

[10] A. Perperoglou, W. Sauerbrei, M. Abrahamowicz, and M. Schmid, "A review of spline function procedures in R," *BMC Medical Research Methodology*, vol. 19, pp. 1–16, 2019.

[11] Q. Li and J. S. Racine, *Nonparametric Econometrics: Theory and Practice*. Princeton, NJ, USA: Princeton University Press, 2007.

[12] B. Lu, M. Charlton, P. Harris, and A. S. Fotheringham, "Geographically weighted regression with a non-Euclidean distance metric: A case study using hedonic house price data," *International Journal of Geographical Information Science*, vol. 28, no. 4, pp. 660–681, 2014, doi: 10.1080/13658816.2013.865739.

[13] M. Ravallion, *The Economics of Poverty: History, Measurement, and Policy*. Oxford, U.K.: Oxford University Press, 2016.

[14] Y. Huang, "The poverty reduction effect of economic growth: How does China's economic growth impact the population size of the urban livings of minimum living allowance?," in *Proceedings of the 2nd International Conference on Management Research and Economic Development*, 2024, doi: 10.54254/2754-1169/72/20240695.

[15] D. Ruppert, M. P. Wand, and R. J. Carroll, "Semiparametric regression during 2003–2007," *Electronic Journal of Statistics*, vol. 3, pp. 1193–1256, 2009.

[16] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Boca Raton, FL, USA: Chapman & Hall/CRC, 2017.

[17] Q. Zou and L. Zhu, "SplineGen: A generative model for B-spline approximation of unorganized points," *Computer-Aided Design*, 2024.

[18] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ, USA: John Wiley & Sons, 2013.

[19] W. Hardle and O. Linton, *Applied Nonparametric Methods*. Cambridge, U.K.: Cambridge University Press, 1994.

[20] H. F. F. Mahmoud, "Parametric versus semi and nonparametric regression models," *International Journal of Statistics and Probability*, vol. 10, no. 2, 2021.

[21] J. Fan and I. Gijbels, *Local Polynomial Modelling and Its Applications*. Boca Raton, FL, USA: Chapman & Hall/CRC, 1996.

[22] C. de Boor, *A Practical Guide to Splines*, Rev. ed. New York, NY, USA: Springer, 2001.

[23] R. L. Eubank, *Nonparametric Regression and Spline Smoothing*, 2nd ed. New York, NY, USA: Marcel Dekker, 1999.

[24] S. W. Keith and D. B. Allison, "A free-knot spline modeling framework for piecewise linear logistic regression in complex samples with body mass index and mortality as an example," *National Library of Medicine*, 2014.

[25] P. C. Chang, Y. W. Wang, and C. H. Liu, "The development of a weighted evolving fuzzy neural network for PCB sales forecasting," *Expert Systems with Applications*, vol. 32, pp. 86–96, 2007.