1st International Conference on Multidisciplinary Engagement

Modeling the Open Unemployment Rate on Java Island Using Based on Influencing Factors

Muthia Hashna Firdausi S.[1*] , Dewi Retno Sari S.[2] , Nughtoh Artafawi Kurdhi[3]
[123] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Sebelas Maret University*
Jl. Ir. Sutami, Surakarta City, Indonesia

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

Abstract: *The open unemployment rate is a key indicator reflecting the labor market conditions of a region. Java Island, as the economic hub of Indonesia with a large population, is at risk of facing unemployment issues. This study aims to model the open unemployment rate in Java Island using truncated spline semiparametric regression with optimal knot selection based on the minimum Cross Validation (CV) value. The data used are secondary data from the Central Statistics Agency (BPS) for the year 2024. The data cover 119 regencies/cities on Java and are divided into 80% training data and 20% testing data. The variables used in this study include the minimum wage, the percentage of the poor population, Gross Regional Domestic Product (GRDP), population growth rate, and Labor Force Participation Rate (LFPR). The results show that the best model was obtained using a first-order spline with two knot points, yielding a minimum CV value of 0.994062. Model evaluation shows that the training data achieved a Mean Absolute Error (MAE) of 0.84332 and 0.964139 on the testing data, as well as a Mean Absolute Percentage Error (MAPE) of 19.63% on the training data and 21.29% on the testing data. The coefficient of determination ((R^2)) was 63.66% for the training data and 48.88% for the testing data. The difference in values between the training and testing data indicates that the model is capable of modeling the open unemployment rate with a reasonably good level of prediction accuracy.*
Keywords: *Open Unemployment Rate , Semiparametric Regression , Truncated Spline , Cross-Validation , Optimal Knot, Java Island*

1. INTRODUCTION

Unemployment refers to the condition of individuals who are not currently working or are actively seeking employment[1] . According to Satrio and Utomo[2] , unemployment can be measured by the open unemployment rate. The open unemployment rate represents the proportion of the labor force that is not employed. The island of Java is home to approximately 56% of Indonesia's total population of 270 million[3] . Although its land area accounts for only 7% of Indonesia's total landmass, the island serves as a central hub, causing various community activities to be concentrated in this region [4] . This situation leads residents living outside Java to choose cities or regencies on Java as their place of residence, hoping to secure more stable employment and improve their quality of life. On the other hand, this creates competition for jobs and drives up the open unemployment rate[5] .

Economic growth on Java is reflected in the increase in Regional Gross Domestic Product (RGDP). RGDP has the potential to create added value in the production of goods and services across various sectors, thereby creating job opportunities[6] . However, this has not yet fully succeeded in reducing unemployment because the available job opportunities do not match the size of the labor force[7] .

Wages serve as a source of income for workers and as a production cost for companies[8] . Changes in wage levels can influence individuals' decisions when seeking employment as well as companies' policies regarding the management of production costs. Higher wages can increase income, but on the other hand , they can raise production costs, leading to higher product prices. If product prices continue to rise, demand for goods decreases and companies may potentially incur losses. In such circumstances, companies may implement efficiency measures to address the resulting losses, such as reducing the workforce or implementing layoffs

(PHK)[9] . High unemployment rates can lead to various issues, such as reduced purchasing power and increased criminal activity[10] . Therefore, an analysis of the relationship between the unemployment rate and its influencing factors is necessary.

Regression analysis is a model used to explain the relationship between the dependent variable and the independent variables[11] . Parametric regression is often the preferred choice in this analysis because the method is relatively easy to interpret[12] . Parametric regression assumes that the form of the relationship between variables is known, such as linear[13] . However, this assumption is not always met, so nonparametric regression is used as an alternative because this method does not assume a specific curve shape[14] . In some cases, some variables still exhibit a linear relationship, while others show relationship patterns of unknown form. One method that can be used is semiparametric regression[15] . The nonparametric component in semiparametric regression can be estimated using various methods, one of which is *the spline*[16] . However, *splines* have limitations, leading to the development of *truncated* basis functions. These functions aim to divide the data into several segments so that the model can adapt to changes in the data pattern[17] .

Numerous studies on modeling the open unemployment rate have been conducted using various approaches. In 2022, Santi *et al.*[10] modeled the open unemployment rate on the island of Java using *penalized spline* nonparametric regression, which falls under nonlinear modeling. Then in 2023, researchers Satrio and Utomo[2] analyzed the open unemployment rate on Java Island for the period 2012–2021 using a panel data regression method based on *the Random Effects Model* (REM). This model is a parametric approach assuming a linear relationship between the dependent and independent variables. A linear approach was also used by researchers Amarta and Setiawati [1] in 2025. The method employed was multiple linear regression. Unlike previous studies that used parametric or nonparametric approaches separately, this study utilized semiparametric regression, which is capable of capturing both linear and nonlinear relationships within a single model.

Semi-parametric regression has been applied in various fields of research. Marbun *et al.* ([18] , 2020) used this model to analyze movements in the Jakarta Composite Index (JCI) using the R GUI application. Parameter estimation was performed using *the Ordinary Least Squares method*, with knot selection based on the minimum MSE value. The best model was obtained at the second order with three optimal knot points, yielding a coefficient of determination ((R^2)) of 90.75% and a MAPE of 3.77%, indicating excellent predictive capability. In 2024, researchers Setyawati *et al.*[19] applied the *truncated spline* semiparametric regression method to model the effects *of Time* (parametric) as well as *Temperature* and *Sunlight Intensity* (nonparametric) on COVID-19 CIR and CFR in Indonesia by determining knots based on the minimum *Generalized Cross Validation* (GCV). Unlike previous studies, the researchers in this study used knot selection based on minimum *Cross Validation* (CV) to model the Open Unemployment Rate on the island of Java.

2. THE COMPREHENSIVE THEORETICAL BASIS

2.1. *Truncated Spline*

Truncated spline is a method capable of handling data with different patterns in each sub-interval[20] . Its advantage lies in the *truncated* function that forms knot points[22] . A knot point is the point where two curves meet, indicating a change in the pattern's shape across different intervals[21] . Generally, an nth-order *spline* m with knot points$T_1,\ T_2, \dots ,\ T_j$ is written as

$$g(z_i) = \sum_{v=1}^{s}\sum_{u=1}^{m} \gamma_{vu}\, {z_{vi}}^{u} + \sum_{v=1}^{s}\sum_{r=1}^{j} \gamma_{v,m+r}\, {(z_{vi} - T_{rv})_+}^{m}, \qquad i = 1,2,\dots,n \tag{1}$$

whereγ_{vu} is the coefficient of the variablez_{vi} ,$\gamma_{v,m+r}$ is the coefficient in the *truncated* function,$v = 1,2,\dots,s$ where$s$ is the number of nonparametric independent variables,z_{vi} is the independent variable,$u = 1,2,\dots,m$ where$m$ is the degree of the polynomial,T_{rv} is the knot point,$r = 1,2,\dots,j$ where$j$ is the number of knot points, and$(z_i - t_j)_+^{\ m}$ is the *truncated* function defined as

$${(z_{vi} - t_{ru})_+}^{m} = \begin{cases} {(z_{vi} - t_{rv})_+}^{m}, & x_i \geq t_{rv} \\ 0, & x_i < t_{rv} \end{cases}.$$

2.2. *Truncated Spline* Semiparametric Regression

According to Eubank[22] , semiparametric regression is a regression method that combines parametric and nonparametric components into a single model. This approach is used when the form of some of the

relationships between variables is known, and the form of others is not yet known with certainty, resulting in both linear and nonlinear relationships[7] . In general, semiparametric regression can be written as

$$y_i = f(x_i) + g(z_i) + \varepsilon_i, \qquad i = 1,2,\dots,n \tag{2}$$

In equation (2), the parametric component $f(x_i)$ assumes a linear relationship between variables, allowing it to be modeled using parametric regression. Parametric regression itself can be written as

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1,2,\dots,n$$

If there is only one independent variable that has a linear relationship with the dependent variable, then the parametric component uses the $f(x_i)$ component in simple linear regression. This component can be written as

$$f(x_i) = \beta_0 + \beta_1 x_{i1}, \qquad i = 1,2,\dots,n \tag{3}$$

However, if there is more than one independent variable with a linear relationship, then the parametric component uses the $f(x_i)$ component in multiple linear regression, which can be written as

$$f(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{il}, i = 1,2,\dots,n \tag{4}$$

Therefore, the *truncated spline* semiparametric regression model can be expressed by substituting equation (1) as the nonparametric component and equation (3) or (4) as the parametric component into equation (2), resulting in the *truncated spline* semiparametric regression model as

$$y_i = f(x_i) + \sum_{v=1}^{s}\sum_{u=1}^{m} \gamma_{vu} {z_{vi}}^{u} + \sum_{v=1}^{s}\sum_{r=1}^{j} \gamma_{v,u+r} {(z_{vi} - T_{rv})_+}^{m} + \varepsilon_i, \qquad i = 1,2,\dots,n.$$

2.3. Selection of Optimal Knot

In the *truncated spline* semiparametric regression method, the selection of optimal knots significantly influences the creation of the best model[23] . One of the methods is *Cross Validation* (CV). *Cross Validation* (CV) is a method used to determine knots. The minimum CV value yields the best model. The CV formula is written as

$$CV(T) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{Y}}{1 - L_{ii}(T)}\right],$$

where n is the number of data points, T is the number of knot points, y_i is the th dependent variable, i and $\hat{Y}$ are the estimated values from the model, and $L_{ii}(T)$ is the diagonal element of *the* **hat** matrix or *smoothing matrix* for the *truncated spline* model.

2.4. Model Accuracy Evaluation

Model accuracy can be evaluated using prediction error metrics, including *Mean Absolute Error* (MAE) and *Mean Absolute Percentage Error* (MAPE). These two metrics are used to determine the magnitude of the difference between actual values and the model's predicted values. The first prediction error metric described is MAE. *Mean Absolute Error*, or MAE, is defined as the average of the absolute values of the differences between the actual values and the predicted values. Mathematically, MAE is expressed as

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

where y_i is the actual value, $\hat{y}_i$ is the predicted value, n is the number of data points, and $|y_i - \hat{y}_i|$ is the absolute difference between the actual and predicted values. The smaller the MAE value, the better the model is at making predictions[24] .

Furthermore, the measure of prediction error can also be expressed as a percentage using the *Mean Absolute Percentage* Error (MAPE). *The Mean Absolute Percentage Error* (MAPE) is a metric used to assess the accuracy of predictions made against actual data. The MAPE formula is written

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

where y_i is the actual value, $\hat{Y}_i$ is the predicted value, and n is the number of data points. The smaller the MAPE value, the more accurate the model. Generally, a MAPE value below 10% indicates a model with very high accuracy, the 10%–20% range indicates an accurate model, the 20%–50% range indicates moderate accuracy, and above 50% indicates a model with low accuracy[25] .

2.5. Coefficient of Determination (R^2)

The coefficient of determination(R^2) is a method used to determine the extent to which the dependent variable is explained by the independent variables. The formulaR^2 is written as

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

wherey_i is the actual value,$\hat{y}_i$ is the predicted value from the model, and$\bar{y}$ is the average value of$y$. The value ofR^2 ranges from 0 to 1[24] . If the value ofR^2 approaches one, the independent variables are better at explaining the dependent variable . However, if the value ofR^2 approaches zero, the resulting regression model is less suitable for use.

3. METHOD

This section explains the data sources, research steps, and methods used in *truncated spline* semiparametric regression modeling.

3.1. Research Data Sources

The data in this study are secondary data obtained from the Central Statistics Agency (BPS) in 2024. The data were accessed through official BPS publications at both the central and regional levels. The data used cover 119 regencies/cities on the island of Java. Next, the data was divided into two groups: *training* data and *testing* data, based on an 80:20 ratio. Eighty percent of the *training* data was used to build and estimate the model, while 20% was used as *testing data* to evaluate the model's predictive capability. The variables used consist of one dependent variable ((Y)) and five independent variables (($X_1 - X_5$)), as shown in Table 1.

Table 1. Research Variables

Variable	Description
Y	Open unemployment rate (percent)
X_1	Minimum wage (rupiah)
X_2	Percentage of the poor population (percent)
X_3	Gross regional domestic product (billion rupiah)
X_4	Population growth rate (thousands)
X_5	Labor force participation rate (percent)

3.2. Research Steps

The research steps were carried out systematically following the flow below:

1. Collected data on the open unemployment rate in Java along with its independent variables through access to the Central Bureau of Statistics (BPS), both from the central and regional BPS offices.
2. Conducting *Exploratory Data Analysis* (EDA) to understand the data and identify variables through descriptive statistics.
3. Performing the data *preprocessing* stage, which includes *data cleaning* and *data scaling*, to obtain data ready for modeling.
4. Split the data into two groups: *training data* and *testing data*, with a ratio of 80% to 20%.
5. Identify patterns of relationships between the dependent variable and independent variables using *scatter plots* to determine parametric and nonparametric components.
6. Model the parametric components using linear regression, while the nonparametric components are approximated using *truncated splines* with various combinations of order and number of knot points
7. Determine the optimal knot points for each model combination using the minimum *Cross-Validation* (CV) value to obtain the best model.
8. Estimating the parameters of the *truncated spline* semiparametric regression model using the *Ordinary Least Squares* (OLS) method.
9. Evaluate model performance using prediction error measures, namely *Mean Absolute Error (MAE), Mean Absolute Percentage Error* (MAPE), and the coefficient of determination ((R^2)) on *training* and *testing* data.
10. Analysis of results and conclusions of the *truncated spline* semiparametric regression model.

4. RESULTS AND DISCUSSION

This chapter discusses the results and discussion regarding the modeling of the Open Unemployment Rate on Java Island using *truncated spline* semiparametric regression. The Results and Discussion begin with a description of the data, followed by the identification of patterns of relationships between variables, model formulation, evaluation of the obtained model, and interpretation of the resulting model.

4.1. Description of Research Data

The characteristics of the variables in this study were analyzed using descriptive statistics, including the mean, standard deviation, minimum value, and maximum value. The purpose of this analysis is to provide an initial overview of the condition and distribution of the data. The results of the descriptive statistics are shown in Table 2.

Table 2. Descriptive statistics

Variable	Mean	Standard Deviation	Minimum	Maximum
Y	5.03	1.80	1.56	9.18
X_1	2,961,742.21	1,028,504.30	2,038,005.00	5,343,430.00
X_2	9.07	3.72	2.34	20.83
X_3	64,614.02	100,110.46	3,268.50	532,606.90
X_4	1,318.73	880.80	28.80	5,682.30
X_5	71.58	4.67	58.13	86.62

Based on Table 2, the average open unemployment rate(Y) on Java Island in 2024 is 5.03% with a standard deviation of 1.80%. The maximum value is 9.18% in Serang Regency, while the minimum value is 1.56% in Pacitan Regency. These differences indicate variations in unemployment rates among regencies and cities on Java Island. The average minimum wage(X_1) is Rp2,961,742.21 with a standard deviation of Rp1,028,504.30. The highest value is in Bekasi City at Rp5,343,430.00 and the lowest in Banjarnegara Regency at Rp2,038,005.00. These wage differences reflect variations in economic conditions and wage-setting policies across regions on the island of Java.

Next, the percentage of the poor population variable ((X_2).). This variable has a mean of 9.07% and a standard deviation of 3.72%. The maximum value for this variable is 20.83% in Sampang Regency, and the minimum value is 2.34% in Depok City. The significant difference between the maximum and minimum values indicates that poverty levels across regencies/cities on Java Island are uneven, meaning some areas have a higher percentage of the poor population compared to others. The GRDP variable(X_3) has a mean of Rp64,614.02 billion with a standard deviation of Rp100,110.46 billion. The highest value was recorded in Central Jakarta at Rp532,606.90 billion, and the lowest in the Thousand Islands at Rp3,268.50 billion. The vast disparity in GRDP values indicates significant economic activity gaps between regions, where areas with high economic activity have significantly larger GRDP values compared to other regions.

The population variable(X_4) has an average of 1,318,730 people and a standard deviation of 880,800 people. The maximum value is in Bogor Regency at 5,682,300 people, and the minimum value is in the Thousand Islands at 28,800 people. The difference between these two values indicates a considerable variation in population density.Meanwhile, the Labor Force Participation Rate (LFPR) (X_5) averages 71.58% with a standard deviation of 4.67%. The highest value is found in Pacitan Regency at around 86.62%, and the lowest in South Tangerang City at around 58.13%. These two values indicate that labor force participation is more evenly distributed compared to other variables.

4.2. **Research** Data *Preprocessing*

Before modeling, the data first undergoes a *preprocessing* stage. During this stage, the data is checked to ensure there are no missing or duplicate values, so that the data used is clean and ready for analysis. Next, *scaling* is performed. In this study, the variables for the open unemployment rate(Y) , the percentage of the poor population(X_2) , and the Labor Force Participation Rate (LFPR)(X_5) all have the same unit, namely percent. Meanwhile, the minimum wage variable(X_1) is measured in rupiah, GRDP(X_3) in billions, and the population growth rate (X4) is measured in thousands of people. This indicates that these three variables have different units and a very wide range of values, so a scaling process is necessary to ensure the scale differences between variables are not too large. The method used is min-max scaling. This method transforms the value of each variable into a range between 0 and 1.

4.3. Identifying the Relationship Pattern Between the Dependent and Independent Variables

After *scaling*, the next step is to determine the appropriate regression method by first identifying the relationship patterns between the dependent and independent variables through *a scatter plot*. If the *scatter plot* results show a specific pattern, such as a straight line, a parametric approach is more appropriate; whereas if the *scatter plot* results show a random distribution pattern, the data is better suited for a nonparametric approach. Meanwhile, if some data points exhibit a specific relationship pattern while others do not, a semiparametric

regression is more appropriate to use. Therefore, five *scatter plots* are presented for each variable to identify the relationship patterns as the basis for selecting a semiparametric regression, as shown in Figure 1 and Figure 2.

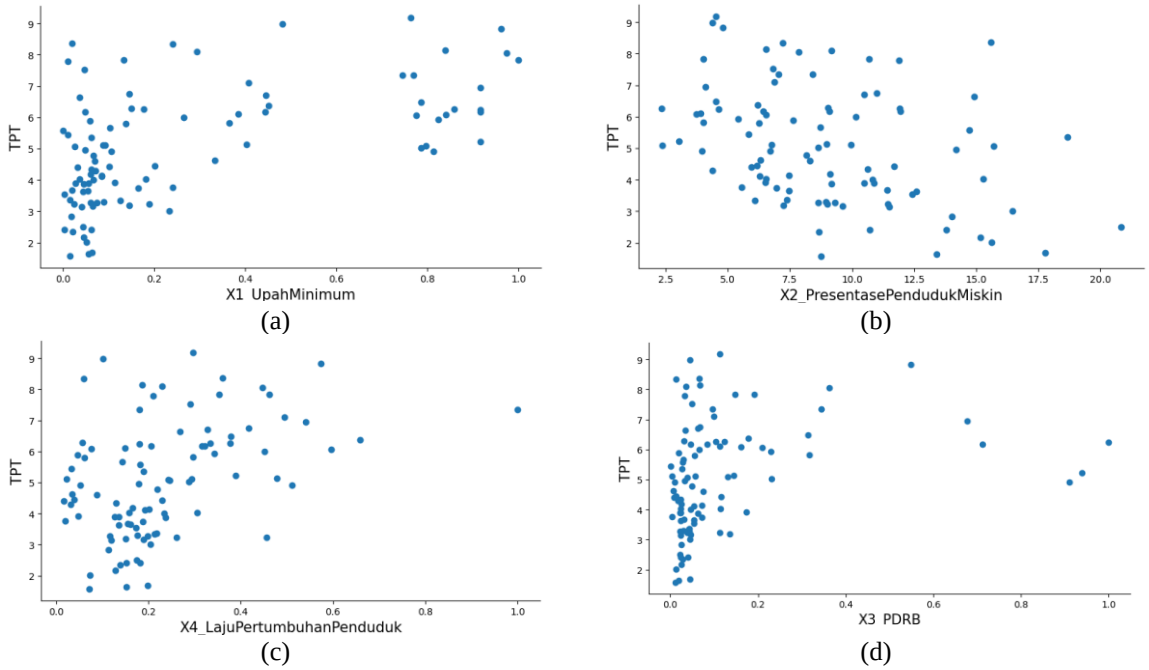

Figure 1. Relationship between nonparametric independent variables and the Open Unemployment Rate: (a) Minimum Wage (X_1), (b) Percentage of the Poor Population (X_2), (c) Population Growth Rate (X_4), and (d) GRDP (X_3). The Minimum Wage (X_1), Population Growth Rate (X_2), and GRDP (X_3) variables use *scaled* data, while the Percentage of Poor Population (X_2) variable uses raw data.

Figure 1 shows that the relationship between the dependent variable and the unknown independent variable is unclear. The data points are scattered randomly and do not form a linear pattern or any other specific pattern. This criterion can be categorized as a nonparametric component within a semiparametric model. Based on Figure 1(a), the Minimum Wage variable is denoted asz_1 ; Figure 1(b) shows the Percentage of the Poor Population denoted asz_2 ; Figure 1(c) shows the Population Growth Rate asz_4, ; and Figure 1(d) shows the GRDP denoted asz_3 .

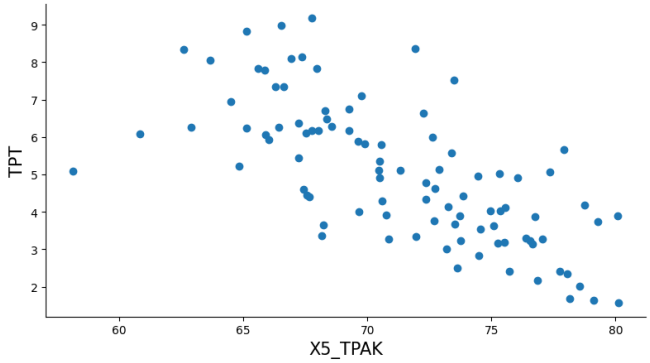

Figure 2. *Scatter plot* between the dependent variable and the parametric independent variables

Figure 2 shows that the relationship between the open unemployment rate and the Labor Force Participation Rate (LFPR) follows a linear pattern. Thus, the Labor Force Participation Rate (LFPR) can be categorized as a parametric component with a linear curve. This variable is denoted by x_1 .

4.4. Forming a *Truncated Spline* Semiparametric Regression Model

Modeling of the Open Unemployment Rate on Java was conducted using *truncated spline* semiparametric regression by determining the location and number of knot points in several orders. The following is the model of the *truncated spline* semiparametric regression.

$$y_i = \beta_0 + \beta_1 x_{i1} + \sum_{v=1}^{s}\sum_{u=1}^{m} \gamma_{vu}\, z_{vi}{}^{u} + \sum_{v=1}^{s}\sum_{r=1}^{j} \gamma_{v,m+r}\,(z_{vi} - T_{rv})^{m}{}_{+} + \varepsilon_i.$$

In modeling the Open Unemployment Rate, *splines* of order 1 (linear) and order 2 (quadratic) were used. The selection of these orders was based on the consideration that models with lower orders produce simpler function forms that are easier to interpret. The number of knots used was 1, 2, and 3 to observe the differences. The number of knots used is limited so that the model is not too complex, remains stable, and does not overfit the random variations in the data.

4.5. Selection of Optimal Knot Points Based on Minimum CV

Optimal knots are determined based on the minimum *Cross-Validation* (CV) value. Knot selection is performed using a *grid search* approach based on the data range of each variable. This data range is then divided into several candidate knot points, which serve as potential positions for the " " knots in the truncated spline model. Subsequently, each combination of order and number of knot points is tested iteratively to generate multiple models with different knot positions. After obtaining the minimum CV value for each combination of order and number of knot points, a comparison of the CV values was performed to select the best model with the minimum CV value, as shown in Table 3.

Table 3. Optimal knot point results

Order	Number of Knots	Knot				CV
		z_1	z_2	z_3	z_4	
1.	1	0.100747	6.300000	0.113850	0.301197	1.008026
	2	**0.082727** **0.160**	**3.869627** **5.29840**	**0.083794** **0.160977**	**0.097919** **0.173912**	**0.994062**
	3	0.082727 0.160 0.225385	3.869627 5.298400 6.507362	0.083794 0.160977 0.226285	0.097919 0.173912 0.238213	1.010657
2.	1	0.082727	3.869627	0.083794	0.097919	1.041779
	2	0.082727 0.160 0.175455	3.869627 5.298400 5.584155	0.083794 0.160977 0.176413	0.097919 0.173912 0.189110	1.106848
	3	0.245 0.303846	6.870050 7.958115	0.245878 0.304656	0.257504 0.315375	1.051871

Table 3 shows that the optimal knot has a minimum CV value of 0.994062. This CV value is located in the 7th column and the 2nd row. The best model was obtained at the first order with two knots. The optimal knot values for each variable are 0.082727 and 0.160 for the minimum wage variable, 3.869627 and 5.298400 for the percentage of the poor population variable, 0.083794 and 0.160977 for the GRDP variable, and 0.097919 and 0.173912 for the population growth rate variable. Based on the optimal knots in Table 3, a first-order *truncated spline* semiparametric regression model with two knots is obtained, which can be written as

$$\begin{aligned}\hat{y}_i = {} & \beta_0 + \beta_1 x_{i1} + \gamma_{11} z_{1i}{}^{1} + \gamma_{1,2}(z_{1i} - 0{,}082727)^{1}{}_{+} + \gamma_{1,3}(z_{1i} - 0{,}160)^{1}{}_{+} + \gamma_{21} z_{2i}{}^{1} \\ & + \gamma_{2,2}(z_{2i} - 3{,}869627)^{1}{}_{+} + \gamma_{2,3}(z_{2i} - 5{,}29840)^{1}{}_{+} + \gamma_{31} z_{3i}{}^{1} + \gamma_{3,2}(z_{3i} - 0{,}083794)^{1}{}_{+} \\ & + \gamma_{3,3}(z_{3i} - 0{,}160977)^{1}{}_{+} + \gamma_{41} z_{4i}{}^{1} + \gamma_{4,2}(z_{4i} - 0{,}097919)^{1}{}_{+} + \gamma_{4,3}(z_{4i} - 0{,}173912)^{1}{}_{+}\end{aligned}$$

4.6. Parameter Estimation

After obtaining the truncated spline semiparametric regression model based on the knot points, parameter estimation was then performed to obtain the coefficient values for each variable in the model. Model parameter estimation was obtained using the OLS method. The parameter estimation results are shown in Table 4.

Table 4. Parameter estimation results

Variable	Parameter	Estimation	Variable	Parameter	Estimates
-	β_0	12,78326	z_2	$\gamma_{2,3}$	0,17263
x_1	β_1	−0,24386		γ_{31}	4,42145
	γ_{11}	−1,22101	z_3	$\gamma_{3,2}$	−24,13899
z_1	$\gamma_{1,2}$	8,54109		$\gamma_{3,3}$	18,73397
	$\gamma_{1,3}$	−5,72423		γ_{41}	4,03033
	γ_{21}	2,34664	z_4	$\gamma_{4,2}$	−1,42821
z_2	$\gamma_{2,2}$	−2,53803		$\gamma_{4,3}$	−0,13461

Table 4 shows the parameter estimates for the best *truncated spline* semiparametric regression model. Based on these estimates, the model can be written as

$$\begin{aligned}\widehat{y_i} = {} & 12{,}78326 - 0{,}24386x_{i1} - 1{,}22101{z_{1i}}^1 + 8{,}54109(z_{1i} - 0{,}082727)^1{}_+ \\ & - 5{,}72423(z_{1i} - 0{,}160)^1{}_+ + 2{,}34664{z_{2i}}^1 - 2{,}53803(z_{2i} - 3{,}869627)^1{}_+ \\ & + 0{,}17263(z_{2i} - 5{,}29840)^1{}_+ + 4{,}42145{z_{3i}}^1 - 24{,}13899(z_{3i} - 0{,}083794)^1{}_+ \\ & + 18{,}73397(z_{3i} - 0{,}160977)^1{}_+ + 4{,}03033{z_{4i}}^1 - 1{,}42821(z_{4i} - 0{,}097919)^1{}_+ \\ & - 0{,}13461(z_{4i} - 0{,}173912)^1{}_+, \qquad i = 1{,}2, \dots, n\end{aligned} \tag{5}$$

4.7. Model Evaluation Based on MAE, MAPE, and Coefficient of Determination

After obtaining the best truncated spline semiparametric regression model, the model was evaluated using the *Mean Absolute Error* (MAE), *Mean Absolute Percentage Error* (MAPE), and the coefficient of determination ((R^2)). The tests were conducted on the *training* data to assess the model's fit to the data, and on the *testing* data to measure its predictive ability for new data. Additionally, a comparison was made with the multiple linear regression model as a benchmark to assess the performance of the truncated semiparametric spline regression model ( ) compared to a simpler regression model. The model evaluation results are shown in Table 5.

able 5. Model evaluation results

Model	Dataset	MAE	MAPE	R^2
Multiple Linear Regression	*Training* Data	0.966729	22.20%	55%
	Test Data	0.942104	18.89%	48.41
Truncated spline semiparametric regression	*Training* Data	0.84332	19.63%	63.66
	Test Data	0.964139	21.29%	48.88

Based on Table 5, the multiple linear regression model and the truncated spline semiparametric regression model produce relatively similar performance. On the *training* data, the truncated spline semiparametric regression model produced smaller MAE and MAPE values and a higher coefficient of determination ((R^2)) compared to multiple linear regression. Meanwhile, on the *testing* data, multiple linear regression produced slightly smaller MAE and MAPE values, while the truncated spline semiparametric regression model produced slightly higher values.

For the truncated spline semiparametric regression model, the MAE value on the *training* data was 0.84332 and on the *testing* data was 0.964139. These values indicate that the model's average prediction error is relatively small and does not show a significant increase on the *testing* data. The MAPE value on the *training* data, 19.63%, indicates that the model's predictive capability is considered good as it is below 20%, while the MAPE value on the *testing* data, 21.29%, indicates that the model's predictive capability is still considered fairly good.

The coefficient of determination ((R^2)) for the *training* data is 63.66%, indicating that the model is capable of explaining most of the variation in the Open Unemployment Rate. Meanwhile, the R-squared value ((R^2)) for the *testing* data is 48.88%, indicating that the model can still explain the variation in the testing data, although its explanatory power is lower compared to the *training* data. The relatively small difference in evaluation values between the *training* and *testing* data suggests that the model does not exhibit excessive *overfitting* and can still be used for predictive purposes.

4.8. Visual Evaluation of the Model

To gain a deeper understanding of the model's performance evaluation, a visual analysis was conducted by comparing the actual and predicted values of the open unemployment rate using the testing data. This approach aims to provide a clearer picture of the model's ability to capture the observed data patterns. Figure 3 shows a visual comparison of the actual and predicted values.

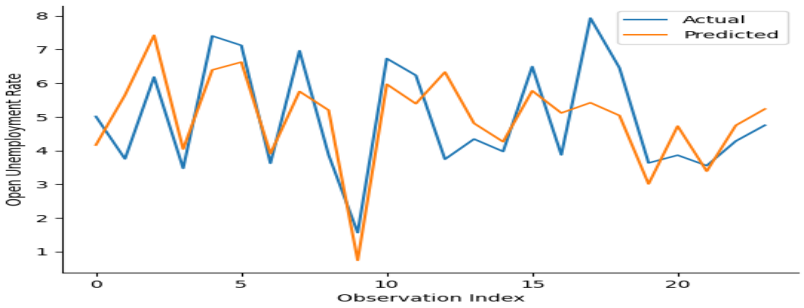

Figure 3. Actual vs. Predicted Values of the Open Unemployment Rate (Testing Data)

Figure 3 shows that the predicted values generally follow the pattern of the actual values at each observation. This indicates that the truncated spline semiparametric regression model is capable of capturing the main trends in the data. Although there are some differences between the actual and predicted values at certain observation points, the resulting discrepancies are relatively small and do not form a specific pattern. This indicates that the model does not exhibit significant bias and possesses adequate predictive capability for data not previously used in the modeling process. Furthermore, the alignment between actual and predicted values is supported by the results of the previous model evaluation, which showed relatively small error values. Therefore, the obtained model can be considered sufficiently capable of representing the relationship between the open unemployment rate and the variables influencing it.

Model Interpretation

The next step is to interpret the influence of each independent variable on the dependent variable based on the obtained truncated spline semiparametric regression model. The interpretation is based on the *scaled* data so that the intervals formed for each variable correspond to the intervals in the transformed data.

3.10.1. The Effect of Labor Force Participation Rate (LFPR) on the Open Unemployment Rate

Based on the parameter estimation results, the effect of the Labor Force Participation Rate (LFPR) on the open unemployment rate can be expressed through the equation

$$\hat{y}_i = 12{,}78326 - 0{,}24386x_{i1}. \quad (6)$$

The results of equation (6) yield a coefficient for the Labor Force Participation Rate (LFPR) of −0.24386, indicating that the relationship between the Labor Force Participation Rate (LFPR) and the Open Unemployment Rate is negative. This value indicates that every one-percent increase in the Labor Force Participation Rate (LFPR) tends to be followed by a 0.24386, 0.24386, 0.24386 percent decrease in the Open Unemployment Rate. Thus, the higher the Labor Force Participation Rate (LFPR), the lower the Open Unemployment Rate tends to be.

3.10.2. The Effect of the Minimum Wage on the Open Unemployment Rate

Based on the obtained *truncated spline* semiparametric regression model, the effect of the Minimum Wage variable on the open unemployment rate is modeled through a nonparametric component using a truncated spline function. The form of this relationship can be expressed through the equation

$$\hat{y}_i = \begin{cases} -1{,}22101\, z_{i1}\,, & z_{i1} < -0{,}08727 \\ -0{,}70657875243 + 7{,}3208 z_{i1}, & -0{,}08727 \le z_{i1} < -0{,}160 \\ 0{,}20929804757 + 1{,}59585 z_{i1}\,, & z_{i1} \ge -0{,}160. \end{cases}$$

The model interpretation results indicate that when the Minimum Wage value is less than −0.08727, the relationship between the Minimum Wage and the Open Unemployment Rate is negative. This indicates that every one-unit increase in the Minimum Wage tends to be followed by a decrease in the Open Unemployment Rate of 1.22101. Furthermore, when the Minimum Wage falls within the intervals $-0{,}08727$ and $-0{,}160$, the relationship formed is positive. This condition indicates that every one-unit increase in the Minimum Wage tends to be followed by an increase in the Open Unemployment Rate of 7.3208. Finally, when the minimum wage is greater than or equal to $-0{,}160$, the relationship between the Minimum Wage and the Open Unemployment Rate remains

positive. This indicates that every one-unit increase in the Minimum Wage tends to be followed by an increase in the Open Unemployment Rate of 1.59585.

4.8.3. The Effect of the Percentage of the Poor Population on the Open Unemployment Rate

Based on the obtained *truncated spline* semiparametric regression model, the effect of the Percentage of the Poor Population variable on the Open Unemployment Rate is modeled through a nonparametric component using a *truncated spline* function. The form of this relationship can be expressed through the equation

$$\hat{y}_i = \begin{cases} -3{,}869627\, z_{i2}, & z_{i2} < -3{,}869627 \\ 9{,}821294148 - 0{,}19139\, z_{i2}, & -3{,}869627 \le z_{i2} < -5{,}29840 \\ 8{,}90656132451 - 0{,}018759\, z_{i2}, & z_{i2} \ge -5{,}29840. \end{cases}$$

The model interpretation results indicate that the relationship between the Percentage of the Poor Population and the Open Unemployment Rate varies across intervals. When the value of the Percentage of the Poor Population is less than -3.869627, the relationship is negative. This suggests that an increase in the Percentage of the Poor Population tends to be followed by a decrease in the Open Unemployment Rate. Meanwhile, when the value of the Percentage of the Poor Population is between -5.29840 and less than or equal to -3.869627, the relationship between the two variables remains negative. This condition indicates that an increase in the Percentage of the Poor Population is still followed by a decrease in the Open Unemployment Rate, but with a smaller effect compared to the previous interval. Finally, when the value of the Percentage of the Poor Population is greater than or equal to -5.29840, the relationship formed is also negative. However, the effect is smaller than in the previous two intervals, so changes in the Percentage of the Poor Population within that range tend to be followed by smaller changes in the Open Unemployment Rate.

3.10.3. The Effect of GRDP on the Open Unemployment Rate

Based on the obtained *truncated spline* semiparametric regression model, the effect of the GRDP variable on the open unemployment rate is modeled through a nonparametric component using a *truncated spline* function. The form of this relationship can be expressed through the equation

$$\hat{y}_i = \begin{cases} -4{,}42145 z_{i3}, & z_{i3} < -0{,}083794 \\ 2{,}02270252806 - 19{,}71754\, z_{i3}, & -0{,}083794 \le z_{i3} < -0{,}169077 \\ -0{,}99303576063 - 0{,}98357\, z_{i3}, & z_{i3} \ge -0{,}169077. \end{cases}$$

When the GRDP value is less than −0.083794, the relationship between GRDP and the Open Unemployment Rate is negative. This indicates that every increase in GRDP tends to be followed by a decrease in the Open Unemployment Rate. In the next interval, when the PDRB value is between −0.169077 and −0.083794, the relationship between PDRB and the Open Unemployment Rate is also negative. This condition indicates that an increase in PDRB is still followed by a decrease in the Open Unemployment Rate, with a greater effect compared to the previous interval. Finally, when the GRDP value is greater than or equal to −0.169077, the relationship remains negative. These three intervals collectively indicate that an increase in GRDP tends to be followed by a decrease in the Open Unemployment Rate.

3.10.4. The Effect of Population Growth Rate on the Open Unemployment Rate

Based on the *truncated spline* semiparametric regression model obtained, the effect of the Population Growth Rate variable on the Open Unemployment Rate is modeled through a nonparametric component using a *truncated spline* function. The form of this relationship can be expressed by the equation

$$\hat{y}_i = \begin{cases} 4{,}03033 z_{i4}, & z_{i4} < -0{,}097919 \\ 0{,}1398449037811 + 2{,}060212\, z_{i4}, & -0{,}097919 \le z_{i4} < -0{,}173912 \\ 0{,}163257593011 + 2{,}4866\, z_{i4}, & z_{i4} \ge -0{,}173912. \end{cases}$$

When the Population Growth Rate is less than −0.097919, the relationship between the Population Growth Rate and the Open Unemployment Rate is positive with a coefficient of 4.03033. This indicates that every one-unit increase in the Population Growth Rate tends to be followed by an increase in the Open Unemployment Rate of 4.03033. Furthermore, when the Population Growth Rate is between −0.173912 and −0.097919, the relationship formed is also positive. This condition indicates that an increase in the Population Growth Rate is still followed by an increase in the Open Unemployment Rate, but with a smaller effect compared to the previous interval.

Finally, when the Population Growth Rate is greater than or equal to −0.173912, the relationship remains positive. This indicates that an increase in the Population Growth Rate tends to be followed by an increase in the Open Unemployment Rate.

5. CONCLUSION

Based on the research results, the open unemployment rate on Java Island can be modeled using truncated spline semiparametric regression as shown in equation (5). The best model was obtained at the first order with two knot points selected based on the minimum *Cross Validation* (CV) value. The CV value obtained was 0.994062. In this model, the Labor Force Participation Rate (LFPR) variable was modeled parametrically (linearly), while the Minimum Wage, Percentage of the Poor Population, GRDP, and Population Growth Rate variables were modeled nonparametrically using a truncated spline. The evaluation results indicate that the model possesses fairly good predictive capability, as reflected by relatively low error values and consistent performance between the *training* and *testing* datasets. Additionally, visual evaluation shows that the predicted values follow the pattern of the observed data, indicating that the model captures the underlying data structure fairly well.

The research findings also indicate that the open unemployment rate on Java Island is associated with several socioeconomic factors, such as Labor Force Participation Rate (LFPR), Minimum Wage, Poverty Rate, GRDP, and Population Growth Rate. These results indicate that efforts to reduce the unemployment rate are not only related to job creation but also require policies that support increased labor force participation, improved economic conditions for the community, and more equitable regional economic growth. Therefore, integrated development policies that combine increased employment opportunities, poverty alleviation, and the strengthening of regional economic activities are crucial in efforts to reduce the open unemployment rate on Java Island.

This study still has several limitations, including the absence of *outlier* detection in the data and further exploration of higher-order polynomials or a greater number of knot points. The study also did not conduct further examination of residual assumptions, such as heteroscedasticity or the influence of extreme observations; thus, these aspects could serve as areas for development in future research. Therefore, future research may consider using other methods, such as *spline smoothing*, to compare model performance in capturing nonlinear relationship patterns in the data. Additionally, further examination of residual assumptions is necessary to ensure that the resulting model provides better estimation and prediction results.

REFERENCES

[1] M. A. A. Pratiwi, R. S. Wijaya, and P. Perdana, "Analysis of Factors Affecting the Open Unemployment Rate," *Formosa Journal of Multidisciplinary Research (FJMR)*, vol. 4, no. 7, pp. 3069–3082, July 2025, doi: 10.55927/fjmr.v4i7.309.

[2] B. R. Satrio, and Y. P. Utomo, "Analysis of the Influence of Education, Economic Growth, Technological Development, and Wages on Unemployment on Java Island," *in Proc. Int. Conf. Sustainable Innovation*, Aug. 2023, pp. 67–72, doi: 10.18196/icosi.v3i1.116.

[3] A. E. Pravitasari, *et al.*, "Spatiotemporal Distribution Patterns and Local Driving Factors of Regional Development in Java," *ISPRS Int. J. Geo-Inf*, vol. 10, no. 12, pp. 1–13, Nov. 2021, doi: 10.3390/ijgi10120812.

[4] A. Putri, N. Murialti, and M. Hidayat, "Analysis of the Factors Driving Economic Growth in Central Java," *" Journal of Islamic Accounting, Management, and Economics (JAM-EKIS)*, vol. 8, no. 2, pp. 746–755, May 2025, doi: 10.36085/jamekis.v8i2.7789.

[5] P. S. Kinasih, and D. M. Nihayah, "Determinants of Unemployment Among University Graduates on Java Island," *Indonesian Journal of Development Economics*, vol. 5, no. 1, pp. 1505–1519, Jan.2022 , doi:10.15294/efficient.v5i1.49150.

[6] S. Akbar, A. S. Herisma, and T. Suharto, "Analysis of Factors Influencing Employment Opportunities in West Java Province in 2015–2020," *Journal of Economic Development and Village Building*, vol. 1, no. 1, pp. 53–65, Mar. 2023, doi: 10.59261/jedvb.v1i1.4.

[7] S. Purwandari, and N. Hanifa, "The Effect of Labor Force, Education, and Minimum Wage on Open Unemployment in East Java Province: A Spatial Disparity *Approach*2020-2024,", vol. 6, no. 1, pp. 44–57, Dec. 2025, doi: 10.38142/jtep.v6i1.1779.

[8] M. E. Rayhan, and P. M. A. Saputra, "An Analysis of the Determinants of the Unemployment Rate in West Java," *Journal of Development Economics and Social Studies*, vol. 4, no. 1, pp. 77–90, Jan. 2025, doi: 10.21776/jdess.2025.04.1.07.

[9] P. C. Amarta, and R. I. S. Setiawati, "Analysis of the Effect of Population, Education Level, and Minimum Wage on the Open Unemployment Rate in West Java Province," *International Journal of Business and Applied Economics (IJBAE)*, vol. 4, no. 5, pp. 2567–2576, Sep. 2025, doi: 10.55927/ijbae.v4i5.214.

[10] V. M. Santi, N. R. Ningsih, and F. Ladayya, "Analyzing the in Java Using Penalized Spline Nonparametric Regression," *International Journal of Applied Science and Sustainable Development (IJASSD)*, vol. 4, no. 2, pp. 70–83, Sep. 2022.

[11] S. S. Handajani, H. Pratiwi, Respatiwulan, Y. Susanti, M. B. Nirwana, and L. P. Nareswari, "Truncated Spline Semiparametric Regression to Handle Mixed-Pattern Data in Modeling Rice Production in East Java Province," *BAREKENG: Journal of Mathematics and Its Application*, vol. 19, no. 4, pp. 2597–2608, Dec. 2025, doi: 10.30598/barekengvol19iss4pp2597-2608.

[12] H. Nurcahayani, I. N. Budiantara, and I. Zain, "Nonparametric Truncated Spline Regression for Modeling," *IP Conf. Proc.: The 2nd International Conference on Science, Mathematics, Environment, and Education*, , Oct020073-1–020073-8, doi: 10.1063/1.5139805.

[13] D. Fitriana, I. N. Budiantara, and V. Ratnasari, "Semiparametric Truncated Spline Regression for Modeling AHH in Indonesia," *in*

Proc.3rd Int. Seminar Sci. Technol., Surabaya, Indonesia, vol. 4, no. 1, pp. 26, Aug. 2018, doi: 10.12962/j23546026.y2018i1.3502.

[14] R. D. Fadlirhohim, Sifriyani, and A. T. R. Dani, "Modeling Stunting Prevalence in Indonesia Using Spline Truncated Semiparametric *Regression*," , vol. 18, no. 3, pp. 2015–2028, Sep. 2026, doi: 10.30598/barekengvol18iss3pp2015-2026.

[15] Y. V. T. Saputra, Moh. Hafiyusholeh, H. Khaulasari, Y. Farida, and P. K. Intan, "Modeling the Number of Crimes in East Java Using a Truncated Spline Semiparametric Regression Approach," *BAREKENG: Journal of Mathematics and Its Application*, vol. 20, no. 2, pp. 1627–1642, Sep. 2025, doi: 10.30598/barekengvol20iss2pp1627-1642.

[16] A. Achmad, R. Fernandes, and B. Hutahayan, "Comparison of Curve Estimation of the Smoothing Spline Nonparametric Function Path Based on PLS and PWLS atLevels of Heteroscedasticity," IOP Conf. Series: Materials Science and Engineering 546, June 2019, pp. 1–8, doi: 10.1088/1757-899X/546/5/052024.

[17] F. Fadri, K. A. Santoso, and S. N. I. Karsan, "A Semiparametric Regression for Modeling Life Expectancy: Spline Truncated and Fourier Series Estimators," *in Proc. 9th Int. Conf. on Research, Implementation, and Education in Mathematical Sciences (ICRIEMS)*, Advances in Social Science, *Education and Humanities Research*, vol. 957, Nov. 2025, doi: 10.2991/978-2-38476-481-5_18.

[18] W. Marbun, Suparti, and D. A. I. Maruddani, "Modeling of the Composite Stock Price Index (CSPI) Using Semiparametric Regression with Truncated Splines Based on R GUI," *J. Phys. Conference Series.*, vol. 1524, no. 1, June 2020, doi: 10.1088/1742-6596/1524/1/012096.

[19] M. Setyawati, N. Chamidah, A. Kurniawan, and D. Aydin, "Confidence Interval for Semiparametric Regression Model Parameters Based on Truncated Spline with Application to COVID-19 Dataset in Indonesia," *Data and Metadata*, vol. 3, pp. 1–13, Dec. 2024, doi: 10.56294/dm2024.609.

[20] A. P. Anisar, Sifriyani, and A. T. R. Dani, "Estimation of a Bivariate Truncated Spline Nonparametric Regression Model on Life Expectancy and Prevalence of Underweight Children in *Indonesia*," , vol. 17, no. 4, pp. 2011–2022, Dec. 2023, doi: 10.30598/barekengvol17iss4pp2011-2022.

[21] F. Ubaidillah, A. A. R. Fernandes, A. Iriany, N. W. S. Wardhani, and S. Solimun "Truncated Spline Path Analysis Modeling in Company X with the Government's Role as a Mediating Variable," *Journal of Statistics Application & Probability*, vol. 794, no. 3, pp. 781–794, Sep. 2022, doi:10.18576/jsap/110303.

[22] R. L. Eubank, ""Smoothing" in *, 2ed. Boca Raton, FL, USA: CRC Press, 1999.

[23] A. A. D. Lestari, A. T. Damaliana, and D. A. Prasetya, "Implementing GCV and mGCV to Determine Optimal Knot in Spline Regression for East Java Life Expectancy," *International Journal of Advances in Data and Information Systems*, vol. 6, no. 2, pp. 242–258, Aug. 2025, doi: 10.59395/ijadis.v6i2.1379.

[24] A. Jierula, S. Wang, T. Oh, and P. Wang, "Applied Sciences Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data," *Appl. Sci*, vol. 11, no. 5, Mar. 2021, doi: 10.3390/app11052314.

[25] A. Fayyazbakhsh, T. Kienberger, and J. Vopara-Wrienz, "Comparative Analysis of Load Profile Forecasting: LSTM, SVR, and Ensemble Approaches for Singular and Cumulative Load Categories," *Smart Cities*, vol. 8, no. 2, Apr. 2025, doi: 10.3390/smartcities8020065.