MODELING OPEN UNEMPLOYMENT RATE IN KALIMANTAN ISLAND USING NONPARAMETRIC REGRESSION WITH FOURIER SERIES ESTIMATOR

Rahmania¹, Sifriyani²*, Andrea Tri Rian Dani³

¹,²,³ Study Program of Statistics, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University
Barong Tongkok Street, Samarinda, 75123, Indonesia

Corresponding author’s e-mail: *sifriyani@fmipa.unmul.ac.id

ABSTRACT

Nonparametric regression is a regression approach used to determine the relationship between the response variable and the predictor variable if the shape of the regression curve is unknown. One of the popular estimators used in nonparametric regression is the Fourier series estimator. Fourier series nonparametric regression is generally used when the pattern of the investigated data is unknown and there is a tendency for the pattern to repeat. The purpose of this study is to estimate nonparametric regression using the Fourier series approach and to find out the factors that influence the open unemployment rate on the island of Borneo in 2021. The criteria for the goodness of the model used Generalized Cross Validation (GCV) and the coefficient of determination ($R^2$). Based on the results, it was found that the best nonparametric regression model for the Fourier series was the model with 5 oscillations, which indicated a minimum GCV of 10.47 and an $R^2$ of 74.22%. Furthermore, based on the results of parameter significance testing, either simultaneously or partially, it shows that all predictor variables have a significant effect on the open unemployment rate. The predictor variables include the labor force participation rate, the average length of schooling, the percentage of poor people, the economic growth rate, and the total population.

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:

1. INTRODUCTION

Regression analysis is a statistical method that is commonly used in analyzing and explaining the relationship between variables, namely the response variable and the predictor variable [1], [2]. There are three types of approaches used to estimate the regression curve, namely parametric, nonparametric, and semiparametric regression [3]. In the parametric regression approach, it is assumed that the shape of the regression curve is known and follows a certain pattern, for example, linear, quadratic, cubic, p-degree polynomial, and so on [4]-[6]. In contrast to parametric regression, the nonparametric regression approach is more flexible because it does not depend on a particular curve shape. The data are expected to be able to find their own form of estimation of the regression curve without being influenced by the subjectivity of the research [7], [8]. In addition to parametric and nonparametric regression approaches, there is also a semiparametric regression approach. Semiparametric regression is used if the regression model contains parametric and nonparametric regression components [9], [10]. Many nonparametric regression approaches have been developed; some of the most frequently used regression estimation methods include spline [11], [12], kernel [13], [14], and Fourier series [15], [16].

One of the most popular estimation methods used by researchers is the Fourier series estimator [17]. The Fourier series is a trigonometric polynomial that has flexibility, so that it can adapt effectively to the local nature of the data [18], [19]. The Fourier series is generally used when the data pattern being investigated is unknown, and there is a tendency for the pattern to repeat [20], [21]. The advantage of the Fourier series approach is that it can handle data if the curve shows cosine and sine waves. Research on the Fourier series was first carried out by Bilodeau [15] and then developed by several researchers, including [22]-[25]. One of the important things in nonparametric Fourier series regression modeling is the selection of smoothing parameters. The Fourier series uses the oscillation parameter as a smoothing parameter. Selection of the optimal oscillation needs to be done to get the best Fourier series estimator in nonparametric regression. One of the methods that can be used to determine the optimal oscillation parameters in Fourier series nonparametric regression is the Generalized Cross Validation (GCV) method. The GCV method theoretically has asymptotically optimal properties, the formula does not contain unknown population variance [26], [27].

The Fourier series nonparametric regression approach can be applied to various fields, one of which is the social field. One of the most common social problems is that the majority of the population has a low level of education, which will lead to high levels of unemployment. The Fourier series nonparametric regression approach can be applied to various fields, one of which is the social field. One of the most common social problems is that the majority of the population has a low level of education which will lead to high levels of unemployment. Not only that, there are many factors that affect the unemployment rate in a region, such as poverty, economic growth, and so on. Therefore, unemployment is one of the main problems that must be addressed immediately [28].

Unemployment is a problem faced by almost all regions in Indonesia, including Kalimantan Island. The main indicator used to measure the unemployment rate in the labor force is the open unemployment rate. The open unemployment rate is the percentage of the number of job seekers to the total workforce. The Central Statistics Agency recorded open unemployment rate in Kalimantan in February 2022 of 4.93%, a decrease of 0.23% points compared to February 2021 of 5.16%. Even though this figure has decreased, it is still relatively high compared to the national target of 4.8%. Based on the background that has been described, a nonparametric regression study will be carried out in this study using the Fourier series approach on the open unemployment rate data on the island of Kalimantan.

2. RESEARCH METHODS

2.1 Nonparametric Regression

Nonparametric regression is a regression method that is not bound by the assumption of a particular regression curve shape; a nonparametric regression approach is used to solve data patterns between response variables and predictor variables that do not form certain patterns. Nonparametric regression provides high flexibility in predicting a model. Therefore, there are many methods that can be used to estimate the nonparametric regression function. The methods that are often used include the Fourier series, spline, kernel, wavelet, k-nearest neighbor, etc.
2.2 Fourier Series Nonparametric Regression

According to Sudiarsa et al. in Adrianingsih et al. [29], the Fourier series is a trigonometric polynomial function that has high flexibility and its curve shows the sine and cosine functions. The Fourier series is one of the estimators that has been widely studied by researchers. The Fourier series estimator used is the sum of a linear function and a trigonometric polynomial function. The multivariable Fourier series nonparametric regression model is completely written in Equation (1).

\[ y_i = \frac{1}{2} \alpha_0 + \sum_{q=1}^{p} \left( \gamma_q (x_{iq}) + \sum_{q=1}^{Q} \alpha_{qj} \cos qx_j \right) + \varepsilon_i, \quad j = 1, 2, \ldots, p \]

Equation (1) can be expressed in the matrix form in Equation (2).

\[ \bar{y} = X(Q) \bar{\theta} + \bar{\varepsilon} \]

where \( \bar{y} \) is response vector with size \( n \times 1 \); \( X(Q) \) is matrix of size \( n \times (Q + 2) \); \( \bar{\theta} \) is parameter vector to be estimated with size \( (Q + 2) \times 1 \); and \( \bar{\varepsilon} \) is error vector of size \( n \times 1 \), which is assumed to be identical, independent, and normally distributed with a mean of zero and a variance \( \sigma^2 \).

2.3 Generalized Cross Validation

There are several methods that can be used to determine the optimal oscillation parameters in nonparametric Fourier series regression, one of which is the GCV method. According to Wang in Dani, the GCV method has several advantages compared to other methods, such as Cross Validation (CV) and/or the Unbiased Risk (UBR) method [3]. The GCV method theoretically has asymptotically optimal properties [4], [30], the formula does not contain the unknown population variance \( \sigma^2 \), and invariance to transformation. The GCV function for selecting optimal oscillations can be shown in Equation (3).

\[ \text{GCV}(Q) = \frac{\text{MSE}(Q)}{(n^{-1} \text{tr}(I - A(Q)))^2} \]  

\[ \text{MSE}(Q) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]  

where \( Q \) is oscillation parameters; \( I \) is identity matrix of size \( n \times n \); \( y_i \) is response variable to- \(\text{i} \); \( \hat{y}_i \) is estimator of the response variable and \( A(Q) \) is \( X(Q)X(Q)^T X(Q)X(Q)^T \). 

2.4 Model Parameter Testing

Model parameter testing was conducted to determine whether the predictor variable has a significant effect on the response variable. There are two stages in testing model parameters, namely simultaneous testing and then continuing with partial testing [27]. Simultaneous testing uses the F test, while partial testing uses the t-test.

Simultaneous testing is carried out to find out whether all predictor variables have an influence on the response variable. This test is carried out simultaneously or simultaneously with the parameters contained in the model [21]. The following are hypotheses for simultaneous testing:

\[ H_0: y_1 = y_2 = \ldots = y_p = \alpha_{11} = \ldots = \alpha_{pq} = 0; \]

\[ H_1: \text{there is at least one } y_j = \alpha_{jq} \neq 0; j = 1, 2, \ldots, p \quad q = 1, 2, \ldots, Q. \]

Test statistics used in parameter testing simultaneously are written in Equation (5).

\[ F = \frac{\text{MSR}}{\text{MSE}} \]

Reject \( H_0 \) if \( F > F_{\alpha(p + pQn - (p + pQ) - 1)} \) or \( p - \text{value} < \alpha \), which means that there is at least one significant parameter of the Fourier series regression model.
Partial testing is done to find out the parameters that are partially significant to the model. The hypotheses for partial testing are as follows:

\[ H_0 : \gamma_j, \alpha_{jq} \neq 0 \]
\[ H_1 : \gamma_j \neq 0, \alpha_{jq} \neq 0; j = 1,2,..,5; q = 1,2,..,5 \]

Test statistics used:

\[ t = \frac{\hat{\theta}_j}{SE(\hat{\theta}_j)} \]  

where \( SE(\hat{\theta}_j) \) is the standard error of \( \hat{\theta}_j \)

\[ SE(\hat{\theta}_j) = \sqrt{\text{var}(\hat{\theta}_j)} \]  

Reject \( H_0 \) if \(|t| \) greater than \( t_{\alpha/2; (n-(p+pq)-1)} \) or reject \( H_0 \) if \( p-value < \alpha \).

3. RESULTS AND DISCUSSION

The analysis includes descriptive and nonparametric regression analysis with the Fourier series approach. The research data used is the open unemployment rate on Kalimantan Island in 2021.

3.1 Data Description

The description of the data is expressed in descriptive statistics which are used to provide an overview of the data. Descriptive statistical analysis used in this study is the average, standard deviation, maximum, and minimum values of each variable used. These variables consist of the open unemployment rate (\( y \)), while the independent variables in this study are the labor force participation rate (\( x_1 \)), the average length of school (\( x_2 \)), the percentage of poor population (\( x_3 \)), the economic growth rate (\( x_4 \)), and the total population (\( x_5 \)). The results of the descriptive statistical analysis are shown in Table 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>4.95</td>
<td>2.06</td>
<td>12.38</td>
<td>2.30</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>68.82</td>
<td>3.97</td>
<td>78.4</td>
<td>60.86</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>8.41</td>
<td>1.22</td>
<td>11.53</td>
<td>6.02</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>6.30</td>
<td>2.30</td>
<td>12.01</td>
<td>2.89</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>3.66</td>
<td>1.38</td>
<td>5.61</td>
<td>-1.69</td>
</tr>
<tr>
<td>( x_5 )</td>
<td>298526.40</td>
<td>195069.10</td>
<td>859250</td>
<td>26000</td>
</tr>
</tbody>
</table>

Based on Table 1, it can be seen that the average for the open unemployment rate (\( y \)) on Kalimantan Island in 2021 is 4.95% with a standard deviation of 2.06%, which illustrates the distribution of data. The highest open unemployment rate is 12.34% in Pontianak and the lowest open unemployment rate is 2.30% in Lamandau.

After obtaining the average, standard deviation, maximum, and minimum values from the open unemployment rate data and the variables that are thought to have an effect, furthermore, visually the distribution of data from each variable is shown on the spatial mapping in Figure 1, where the grouping of each variable is presented in 5 color degradation categories.
The next step is to do before modeling using nonparametric Fourier series regression, namely the need to know the relationship pattern between the response variable and each predictor variable that is thought to influence it. The following is the pattern of the relationship that is formed between the response variable open unemployment rate and each predictor variable that is thought to have an influence, including the labor force participation rate, the average years of schooling, the percentage of poor population, the economic growth rate, and the total population.
Based on the explanation in Figure 2, the scatter plot shows that the pattern of relationship between the open unemployment rate (y) and the labor force participation rate (x1), the average length of school (x2), the percentage of poor population (x3), the economic growth rate (x4), and the total population (x5), do not form a specific pattern, so that the model used to model the open unemployment rate in Kalimantan Island is a nonparametric regression model. The estimator used in this study is the Fourier series.

3.2 Determining Optimal Oscillation

Fourier series nonparametric regression modeling is highly dependent on the oscillation parameter (Q). Selection of the optimal oscillation parameters needs to be done to obtain the best Fourier series estimator in nonparametric regression. The number of oscillations in this study will be limited to 1 to 5 oscillations. The selection of optimal oscillation parameters is carried out using the GCV method whose formula can be seen in Equation (2). The optimal oscillation value chosen is the oscillation value that has the minimum GCV. The minimum GCV value and the maximum R² are the optimal oscillation criteria that will be used [28]. The results of the analysis for oscillation values 1 to 5 are shown in Table 2.

<table>
<thead>
<tr>
<th>Oscillation (Q)</th>
<th>GCV</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47.75</td>
<td>63.49%</td>
</tr>
<tr>
<td>2</td>
<td>22.44</td>
<td>66.37%</td>
</tr>
<tr>
<td>3</td>
<td>13.11</td>
<td>67.53%</td>
</tr>
<tr>
<td>4</td>
<td>13.50</td>
<td>67.86%</td>
</tr>
<tr>
<td>5</td>
<td>10.47</td>
<td>74.22%</td>
</tr>
</tbody>
</table>

Table 2 shows that the minimum GCV value and the greatest coefficient of determination are obtained when the number of oscillations is 5 with a GCV of 10.47 and a coefficient of determination of 74.22%.

3.3 Fourier Series Nonparametric Regression Model

The best Fourier series nonparametric regression model is obtained by using optimal oscillations. Based on the selection of optimal oscillations that have been carried out, the best regression model is the model with 5 oscillations. The estimation of the model of the open unemployment rate in Kalimantan Island with the nonparametric Fourier series regression is as follows:

\[
\hat{y}_i = 20.12 - 0.12x_{1i} + 44.02\cos x_{1i} + 4.74\cos 2x_{1i} - 32.68\cos 3x_{1i} - 1.54\cos 4x_{1i} - 21.53\cos 5x_{1i} + 0.75x_{2i} + 74.27\cos x_{2i} + 96.98\cos 2x_{2i} - 5.77\cos 3x_{2i} - 10.56\cos 4x_{2i} + 13.71\cos 5x_{2i} + 0.14x_{3i} - 0.01\cos x_{3i} - 29.09\cos 2x_{3i} - 34.13\cos 3x_{3i} - 37.34\cos 4x_{3i} + 12.97\cos 5x_{3i} + 0.17x_{4i} + 59.67\cos x_{4i} - 59.47\cos 2x_{4i} + 1.25\cos 3x_{4i} - 89.46\cos 4x_{4i} - 10.13\cos 5x_{4i} + 0.00x_{5i} - 37.09\cos x_{5i} - 1.37\cos 2x_{5i} - 37.09\cos 3x_{5i} - 1.37\cos 4x_{5i} - 37.09\cos 5x_{5i}
\]

Based on Table 2 it can be seen that the optimal oscillation value (Q) is as many as 5 oscillations. This is because the value of Q=5 has the smallest GCV value compared to the oscillation values 1, 2, 3, 4, and 5. So, the above model is obtained as the best model. The number of oscillation values tested can actually be more than 5 oscillations, but in this study, it is limited to 5 oscillations because this will affect the number of parameters to be estimated, and the resulting model is not parsimony. The nonparametric regression model with the Fourier series approach can be used to predict the value of the response variable based on the predictor variable. Visually, the comparison between y and \( y^* \) from the Fourier series nonparametric regression model is shown in Figure 3.
Based on Figure 3, it is known that the prediction results using the best Fourier series nonparametric regression model tend to follow the actual data pattern, which means that the model obtained can be used as a consideration for making decisions in the future.

### 3.4 Model Parameter Significance Testing

After estimating the model parameters, the next step is to test the significance of the model parameters. This test was carried out in 2 stages, namely, simultaneous and partial testing. Testing the significance of model parameters simultaneously has the following hypotheses:

\[ H_0: \gamma_1 = \gamma_2 = \ldots = \gamma_p = \alpha_{11} = \ldots = \alpha_{pq} = 0 \]
\[ H_1: \text{there is at least one } \gamma_j = \alpha_{jq} \neq 0; j = 1, 2, \ldots, p \quad q = 1, 2, \ldots, Q \]

Based on the results of the analysis of variance in Table 4.4, it can be seen that the \( F = 3.86 \) and the p-value is 0.00 with \( F_{0.05; (30,25)} = 1.92 \) and a significance level of 0.05. This value indicates that the \( F \) value is more than the \( F_{0.05; (30,25)} \) value and the p-value is less than \( \alpha \) (0.05), so it can be concluded that \( H_0 \) is rejected, which means that simultaneously the labor force participation rate variable, average length of schooling, percentage of poor people, the rate of economic growth, and the total population have no effect on the open unemployment rate on Kalimantan Island in 2021.

Furthermore, testing the significance of model parameters partially has the following hypotheses:

\[ H_0: \gamma_j, \alpha_{jq} = 0 \]
\[ H_1: \gamma_j \neq 0, \alpha_{jq} \neq 0; j = 1, 2, \ldots, 5; q = 1, 2, \ldots, 5 \]

Based on the partial test results, it is known that after comparing the p-value with a significance level of 0.05, it is known that all parameter significant. These parameters are considered significant if the p-value is less than the significance level.

### 4. CONCLUSIONS

Based on the results and discussion that has been obtained, the following conclusions can be drawn:

1. Based on nonparametric regression modeling with the best Fourier series approach on the open unemployment rate data on Kalimantan Island, it can be concluded that the optimal (Q) oscillations obtained are as many as 5 oscillations with a minimum GCV of 10.47

2. The best nonparametric regression model with the Fourier series approach is obtained with \( Q = 5 \) as follows:

\[
\hat{y}_t = 20.12 - 0.12x_{1t} + 44.02 \cos x_{1t} + 4.74 \cos 2x_{1t} - 32.68 \cos 3x_{1t} - 1.54 \cos 4x_{1t} - 21.53 \cos 5x_{1t} + 0.75x_{2t} + 74.27 \cos x_{2t} + 96.98 \cos 2x_{2t} - 5.77 \cos 3x_{2t} - \]
10.56 \cos 4x_{2j} + 13.71 \cos 5x_{2j} + 0.14x_{3j} - 0.01 \cos x_{4j} - 29.09 \cos 2x_{5j} - 34.13 \cos 3x_{5j} - 37.34 \cos 4x_{5j} + 12.97 \cos 5x_{5j} + 0.17x_{4j} + 59.67 \cos x_{4j} - 59.47 \cos 2x_{4j} + 1.25 \cos 3x_{4j} - 89.46 \cos 4x_{4j} - 10.13 \cos 5x_{4j} + 0.00x_{5j} - 37.09 \cos x_{5j} - 1.37 \cos 2x_{5j} - 37.09 \cos x_{5j} - 1.37 \cos 4x_{5j} - 37.09 \cos 5x_{5j}

3. Based on the parameter significance test, it is found that all predictor variables have a significant effect on the open unemployment rate. These predictor variables include the labor force participation rate, the average length of schooling, the percentage of poor people, the rate of economic growth, and the total population.

REFERENCES


