Fourier series estimator in semiparametric regression to predict criminal rate in Indonesia

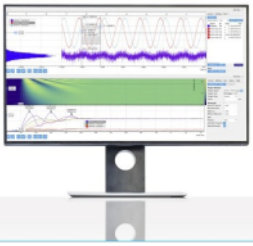
by Aang Darmawan

Submission date: 24-Jul-2023 05:34PM (UTC+0700)
Submission ID: 2136027699
File name: 5.0042123.pdf (528.24K)
Word count: 4050
Character count: 20055

Fourier series estimator in semiparametric regression to predict criminal rate in Indonesia

Cite as: AIP Conference Proceedings **2329**, 060023 (2021); https://doi.org/10.1063/5.0042123
Published Online: 26 February 2021

Rini Kustianingsih, M. Fariz Fadillah Mardianto, Belindha Ayu Ardhani, Kuzairi, Amin Thohari, Raka Andriawan, and Tony Yulianto



Fourier Series Estimator in Semiparametric Regression to Predict Criminal Rate in Indonesia

Rini Kustianingsih[1], M. Fariz Fadillah Mardianto[2, a)], Belindha Ayu Ardhani[2, b)], Kuzairi [1, c)], Amin Thohari[2], Raka Andriawan[2] and Tony Yulianto[1]

[1]*Department of Mathematics, Universitas Islam Madura, Pamekasan, Indonesia*
[2]*Department of Mathematics, Universitas Airlangga, Surabaya, Indonesia*

[a)] Corresponding author: m.fariz.fadillah.m@fst.unair.ac.id
[b)] abelindha58@gmail.com
[c)] kuzairi81@gmail.com

Abstract. Regression is an analysis for determining relationship between response variables and predictor variables. There are three approaches to estimate the regression curve. Those are parametric regression, nonparametric regression, and semiparametric regression. This study focused on the estimator form of semiparametric regression curve using Fourier series approach with sine and cosine base (general); sine base; and cosine base. The best estimator, which is obtained using ordinary least square optimization was applied to model the percentage of criminal incidents in Indonesia. The goodness-of-fit criteria of a model used are high coefficient of determination, minimum Generalized Cross Validation (GCV) and Mean Square Error (MSE) value with determining parsimony model. In this study, the authors obtained the best fourier estimator for predicting percentage of criminal incidents based on cosine fourier series that had minimun GCV and MSE values, of 2.471 and of 0.0006, respectively, and determination coefficient of 77.545%. So, the estimator (cosine-fourier series) was used for predicting the out-sample data and it met Mean Absolute Error (MAE) of 0.02.

INTRODUCTION

Regression analysis is a method to analyze the pattern of functional relationships between two or more variables. The main purpose of this method is to know the form of regression curve estimation. There are three approaches to estimate regression curves, such as parametric regression, nonparametric regression, and semiparametric regression [1]. This study focused on semiparametric regression, which used combination of linear estimator for parametric component and Fourier series estimator for nonparametric component. It also used percentage of criminal incidents data in the application of model. It was related to predict the percentage of criminal incidents in Indonesia based on factors that influence it using Fourier series estimators in semiparametric regression.

The determination of the use of this method was based on relationships between percentage of criminal incidents and its influenced factors. The relationship between percentage of criminal incidents and percentage of population showed parametric form. It was related to fact that the increase of population contributes to the increase of crime rate [2]. Both of them actually have upward trend every year. That was why they had parametric pattern. In another hand, the relationship between percentage of criminal incidents and percentage of unemployed population; also relationship between percentage of criminal incidents and percentage of poor people showed nonparametric form. Both unemployed population and poor people have fluctuating percentage every year. That was why the relationships had nonparametric pattern. Then, the combination of parametric and nonparametric one encouraged the author to use semiparametric regression.

Fourier series is a trigonometric polynomial function that often be used in Mathematical and Statistical modeling. Fourier series estimators are used to describe the two curves of sine and cosine waves. There are several forms of Fourier series in Mathematical study, such as cosine and sine base (general Fourier series); Fourier series of sine base; and Fourier series of cosine base. This research compared three forms of Fourier series in semiparametric

International Conference on Mathematics, Computational Sciences and Statistics 2020
AIP Conf. Proc. 2329, 060023-1–060023-9; https://doi.org/10.1063/5.0042123
Published by AIP Publishing. 978-0-7354-4073-9/$30.00

regression to get a selected model. This regression actually combines two kind of approaches that are parametric and nonparametric regression.

Previous research using Fourier series in nonparametric regression has been developed. One of them is Bilodeau [3], who used Fourier cosine series as an estimator to make regression curve smoother. Then, Bilodeau's research was developed by Mardianto et. al. [4], whom used specific cases in nonparametric regression. The optimal design for obtaining constraints in Fourier series estimator examined by Biedermann et. al. [5], and Dette et. al. [6]. Both of them use Fourier series with sine and cosine basis. Meanwhile, previous research related to the use of Fourier series in semiparametric regression has also been carried out by many researchers. One of them is Asrini and Budiantara [1], who applied research of Pane et. al., [7] in modelling of rice production in Central Java. This study is different from the previous ones because it used three kind of Fourier series to analyze the criminal rate as one of big problems in Indonesia [8].

LITERATURE REVIEW

Considering data $(x_1, x_2, \dots, x_p, t_1, t_2, \dots, t_q, y)$ with x and t are predictor variables and y is response variable. The relationship between x and y is known to form a pattern, whereas the relationship between t and y has unknown pattern forms. Therefore, the relationship between x_i, t_i and y_i that is assumed to follow a semiparametric regression model. In this study, the semiparametric regression model is assumed with p predictors of parametric components $x_1, x_2, \dots, x_p$ and q predictors of nonparametric components $t_1, t_2, \dots, t_q$ as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \sum_{l=1}^{r} f(t_{il}) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), \quad (1)$$

with $i = 1, 2, \dots n,\ j = 1,2, \dots, p$, and $l = 1,2, \dots, r$

If $f(t)$ is a function, which can be integrated and diferensiable at the interval $[a, a + 2L]$, then the Fourier series representation at the interval relating to $f(t)$, which contains the components of trigonometric sine and cosine is as follows [9]:

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos k^* t + b_n \sin k^* t) \quad (2)$$

with $k^* \approx \frac{n\pi}{L} : n = 1, 2, 3, \dots$

The Fourier coefficient is determined by the following formula:

$$a_0 = \frac{1}{L}\int_a^{a+2L} f(t)dt$$

$$a_n = \frac{1}{L}\int_a^{a+2L} f(t)\cos k^* t\, dt$$

$$b_n = \frac{1}{L}\int_a^{a+2L} f(t)\sin k^* t\, dt$$

If the Fourier series using the same way only contains the cosine element, it is called the cosine Fourier series. While the Fourier series which only contains the sin element is called the sine Fourier series.

The equation (1) can be written in the form of a matrix below:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{f} + \boldsymbol{\varepsilon} \quad (3)$$

with:

$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, the response variable

$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$, regression parameters for parametric components

$\boldsymbol{f} = \begin{bmatrix} f(t_{11}) + f(t_{12}) + \cdots + f(t_{1q}) \\ f(t_{21}) + f(t_{22}) + \cdots + f(t_{2q}) \\ \vdots \\ f(t_{n1}) + f(t_{n2}) + \cdots + f(t_{nq}) \end{bmatrix}$, nonparametric components

$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$, random error vector

Then the matrix is given below, which contains the parametric component of predictor variables.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

In equation (3), the function $\boldsymbol{f}$ is assumed to be an unknown regression curve pattern, so it is approximated by a Fourier series function by using sine and cosine bases as follows [9]:

$$f(t_i) = \frac{a_0}{2} + \gamma t_i + \sum_{k=1}^{K} (a_k \cos kt_i + b_k \sin kt_i) \quad (4)$$

with k is an oscillation parameter.

For q predictor, equation (4) becomes as follows:

$$f(t_{ij}) = \sum_{j=1}^{q} \left(\frac{a_{0j}}{2} + \gamma_j t_{ij} + \sum_{k=1}^{K} (a_{kj} \cos kt_{ij} + b_{kj} \sin kt_{ij}) \right)$$

Then, it continues to find out the nonparametric component matrix of semiparametric regression model using Fourier series approach. The equation (4) is substituted into equation (1), so it becomes a vector equation as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{T\eta} + \boldsymbol{\varepsilon} \quad (5)$$

where:

$$\boldsymbol{T} = [\boldsymbol{T_1} \quad \boldsymbol{T_2} \quad \cdots \quad \boldsymbol{T_q}]$$

$$\boldsymbol{T_q} = \begin{bmatrix} 1 & t_{1q} & \cos t_{1q} & \cdots & \cos kt_{1q} & \sin t_{1q} & \cdots & \sin kt_{1q} \\ 1 & t_{2q} & \cos t_{2q} & \cdots & \cos kt_{2q} & \sin t_{2q} & \cdots & \sin kt_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{nq} & \cos t_{nq} & \cdots & \cos kt_{nq} & \sin t_{nq} & \cdots & \sin kt_{nq} \end{bmatrix}$$

$$\boldsymbol{\eta} = (\boldsymbol{\eta_1} \quad \boldsymbol{\eta_2} \quad \cdots \quad \boldsymbol{\eta_q})^T$$

$$\boldsymbol{\eta_q} = \begin{pmatrix} a_{0q} & \gamma_q & a_{1q} & \cdots & a_{kq} & b_{1q} & \cdots & b_{kq} \end{pmatrix}^T$$

The error is minimized using Ordinary Least Square (OLS) method through the following equation:

$$\min(\boldsymbol{\beta}, \boldsymbol{\eta}) = \min \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = \min(\boldsymbol{y} - \mathbf{X\beta} - \boldsymbol{T\eta})^T (\boldsymbol{y} - \mathbf{X\beta} - \boldsymbol{T\eta}) \quad (6)$$

Then by elaborating on optimization equation (6), this equation is obtained:

$$\boldsymbol{R(\beta, \eta)} = \boldsymbol{y^T y - 2y^T X\beta - 2\eta^T T^T y + 2\beta^T X^T T\eta + \beta^T X^T X\beta + \eta^T T^T T\eta} \quad (7)$$

Estimator of parameter $\boldsymbol{\beta}$ is done by doing partial derivatives $(\boldsymbol{\beta}, \boldsymbol{\eta})$ to $\boldsymbol{\beta}$, so it is obtained:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \{\boldsymbol{X^T y - X^T T\widehat{\eta}}\} \quad (8)$$

Meanwhile, estimator of parameter $\boldsymbol{\eta}$ is done by doing partial derivatives $(\boldsymbol{\beta}, \boldsymbol{\eta})$ to $\boldsymbol{\eta}$, so it is obtained:

$$\widehat{\boldsymbol{\eta}} = (\boldsymbol{T^T T})^{-1} \{\boldsymbol{T^T y - T^T X\widehat{\beta}}\} \quad (9)$$

The estimators in equation (8) and (9) still have parameter. One of good estimator criteria is not included parameter. Therefore, substitution of equation (9) to equation (8) aims to obtain $\widehat{\boldsymbol{\beta}}$ that is free of parameters as follows:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{M(X^T X)^{-1} \{X^T - X^T T(T^T T)^{-1} T^T\} y} = \boldsymbol{B}(k)\boldsymbol{y} \quad (10)$$

with $\boldsymbol{M} = (\boldsymbol{I} - (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T T (T^T T)^{-1} T^T X})^{-1}$

while substitution of equation (8) to equation (9) aims to get $\widehat{\boldsymbol{\eta}}$ that is not included of parameters as follows:

$$\widehat{\boldsymbol{\eta}} = \boldsymbol{N(T^T T)^{-1} \{T^T - T^T X (X^T X)^{-1} X^T\} y} = \boldsymbol{C}(k)\boldsymbol{y} \quad (11)$$

with $\boldsymbol{N} = (\boldsymbol{I} - (\boldsymbol{T^T T})^{-1} \boldsymbol{T^T X (X^T X)^{-1} X^T T})^{-1}$

After obtaining estimators for parametric and nonparametric components, it continues to determine the estimator of semiparametric regression model using Fourier series approach below.

$$\widehat{\boldsymbol{y}} = \boldsymbol{X\widehat{\beta} + T\widehat{\eta}} = \boldsymbol{D}(k)\boldsymbol{y}$$

with $\boldsymbol{D}(k) = \boldsymbol{XB}(k) + \boldsymbol{TC}(k)$. $\boldsymbol{B}(k)$ is the hat matrix for parametric components. $\boldsymbol{C}(k)$ is the hat matrix for nonparametric components. Hat matrix for semiparametric regression models with the Fourier series approach is symbolized as $\boldsymbol{D}(k)$. In this case, k shows the oscillation parameter that is contained in the matrix $\boldsymbol{D}(k)$. The optimal k value in the model is obtained using Generalized Cross Validation (GCV) method. GCV is often used because it has asymptotically optimal properties [10]. For determining an optimal oscillation parameter, it can be seen based on the minimum GCV value, the formula is given as follows:

$$GCV(k) = \frac{MSE(k)}{(n^{-1} trace(\mathbf{I} - \mathbf{D}[k]))^2} \quad (12)$$

where:

k : oscillation parameter
n : the number of observation
$\boldsymbol{I}$: identity matrix
$\boldsymbol{D}[k]$: hat matrix

and $MSE[k] = \frac{1}{n}\mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{D}[k])^{\mathrm{T}}(\mathbf{I} - \mathbf{D}[k])\mathbf{y}$

In Fourier series estimator in semiparametric regression, a measure of the goodness-of-fit of the model can be determined based on the minimum of GCV and MSE values, and the high value of R^2. The following equation is a formula for R^2:

$$R^2 = \frac{(\hat{y}-\bar{y})^T(\hat{y}-\bar{y})}{(y-\bar{y})^T(y-\bar{y})} \quad (13)$$

with $\bar{\mathbf{y}}$ is a vector that contains the average response data.

The Fourier series estimator for the semiparametric regression curve with cosine and sine bases is as follows:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} + \sum_{j=1}^{q}\left(\frac{\hat{a}_{0j}}{2} + \hat{\gamma}_j t_{ij} + \sum_{k=1}^{K}(\hat{a}_{kj}\cos kt_{ij} + \hat{b}_{kj}\sin kt_{ij})\right) \quad (14)$$

DATA AND PROCEDURES

Fourier series estimator in semiparametric regression used in this study is consists of one response variable, and three predictor variables. The response variable represents the percentage of criminal incidents from 34 provinces in Indonesia that is denoted by y; the first predictor represents the percentage of population that is denoted by x as parametric component; the second predictor represents the percentage of unemployed population that is denoted by t_1 as nonparametric component; the third predictor represents the percentage of poor people that is denoted by t_2 as nonparametric component. The data was taken from Badan Pusat Statistik (BPS) or Central Bureau of Statistics with data in 2018 used as training data for estimating curve regression, and data in 2019 used as testing data for predicting based on the selected Fourier series estimator. The steps of analysis using Fourier series estimator in semiparametric regression are given as follows:

1. Inputting pairs of data (y_i, x_{ij}, t_{il}); $i = 1,2,\ldots,34$, $j = 1$, and $l = 1,2$.
2. Identifying predictors that are included in parametric predictor and nonparametric predictor based on training data.
3. Determining GCV value using equation (12) for $k = 1,2,3,4,5$ for Fourier series estimator based on cosine and sine (general Fourier estimator); cosine and sine.
4. Selecting an oscillation parameter based on the smallest GCV for $k = 1,2,3,4,5$ (parsimonious model).
5. Determining the other goodness-of-fit criteria, such as MSE and R^2.
6. Repeating step 3 and step 5 for the other Fourier series estimator.
7. Selecting the best Fourier series estimator in semiparametric regression after finishing step 3 until step 6 for all of Fourier series estimator.
8. Calculating a prediction based on testing data.
9. Presenting plot between response data and estimator data for testing data.

Description, which is based on training data is useful to know the general description about the response variables and the predictor variables used. The general description discussed is the maximum, minimum, and mean of data.

TABLE 1. Descriptive Statistics of Research Variables

Variable	Minimum Value	Province	Maximum Value	Province	Mean
y	0.003	North Maluku	0.123	DKI Jakarta	0.0312
x	0.003	North Maluku and Island of Bangka Belitung	0.168	East Java	0.0313
t_1	0.028	West Java	0.035	Papua dan Bali	0.0314
t_2	0.003	West Papua	0.185	West Java	0.031

Based on Table 1, it can be seen that province with the lowest percentage of criminal incidents in 2018 is North Maluku, and province with the highest percentage of criminal incidents is DKI Jakarta with the average value of 0.0312. Meanwhile, province with the lowest percentage of population in 2018 are North Maluku and Bangka Belitung Island, and province with the highest percentage of population is East Java with the average value of 0.0313. In addition, province with the lowest percentage of unemployed population in 2018 is West Java, and province with the highest percentage of unemployed population are Papua and Bali with the average value of

0.0314. Then province with the lowest percentage number of poor people in 2018 is West Papua, and province with the highest percentage number of poor people is West Java with the average value of 0.031.

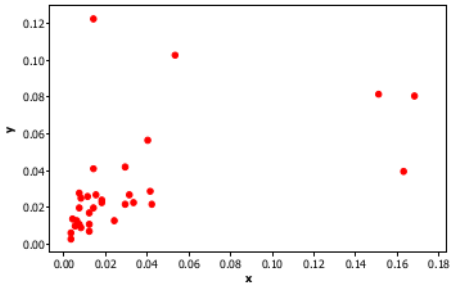


FIGURE 1. Scatter Plot between Response and Predictor Variables for Parametric Component

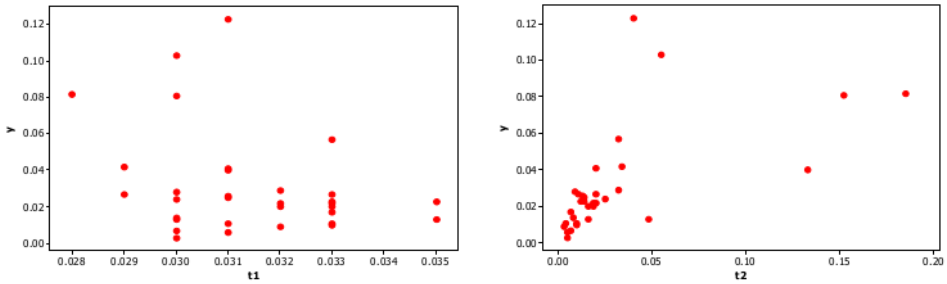


FIGURE 2. Scatter Plot between Response and Predictor Variables for Nonparametric Components

Figure 1 presents scatter plot between y and x. Based on Fig. 1, the scatter plot shows linear pattern. It can be identified that the relationship between the percentage of population and the percentage of criminal incidents in Indonesia had parametric component because of the fact that the total population had upward trend every time. It was also based on the result of correlation test between them, which shows that correlation between y and x has p-value of 0.047. This identification was supported by linearity test. While Fig. 2 presents scatter plot between y and t_1, and also between y and t_2. Based on Fig. 2, it can be identified that the pattern had nonparametric component. This identification was supported by linearity test using Open Source Software (OSS) R with package "lmtest".

Besides that, OSS R was also used to analyze data based on Fourier series estimator in semiparametric regression. It has an oscillation parameter k. Then, GCV method was used to determine the optimal k value using parsimony model with $k = 1,2,3,4,5$ for inputting.

RESULT AND DISCUSSION

The Data Analysis based on Cosine Fourier Series

Cosine Fourier series in semiparametric regression can be presented based on equation (14) by eliminating $\widehat{b}_{kj} \sin kt_{ij}$. The GCV value based on cosine Fourier series estimator is presented in Table 2 as follows:

TABLE 2. GCV Value for Cosine Base

k	GCV Value
1	**2.471**
2	159.615
3	238.456
4	723.24
5	113.501

Based on Table 2, the minimum GCV value was 2.471 with k equals to 1. The goodness-of-fit criteria of this model were the value of k equals to 1; GCV value of 2.471; MSE value of 0.0006; and R^2 value of 77.545%.

The Data Analysis based on Sine Fourier Series

Sine Fourier series in semiparametric regression can be presented based on equation (14) by eliminating $\hat{a}_{kj} \cos k t_{ij}$. The GCV value based on sine Fourier series estimator is presented in Table 3 as follows:

TABLE 3. GCV Value for Sine Base

k	GCV Value
1	8.521
2	**4.785**
3	44.411
4	11.722
5	14.963

Based on Table 3, the minimum GCV value was 4.785 with k equals to 2. The goodness-of-fit criteria of this model were the value of k equals to 2; GCV value of 4.785; MSE value of 0.0003; and R^2 value of 50.033%.

The Data Analysis based on General Fourier Series

The general Fourier series in semiparametric regression can be presented based on equation (14) without elimination. The GCV result based on general Fourier series estimator is presented in Table 4 as follows:

TABLE 4. GCV Value with Sine and Cosine Base

k	GCV Value
1	7.045
2	**6.720**
3	6.749
4	8.056
5	6.761

Based on Table 4, the minimum GCV value was 6.720 with k equals to 2. The goodness-of-fit criteria of this model were the value of k equals to 2; GCV value of 6.720; MSE value of 0.0003; and R^2 value of 57.981%.

The Best Fourier Series Estimator

Based on the three estimation above, the comparison between GCV, MSE and R^2 is as follows:

TABLE 5. Comparison with Sine, Cosine and General Base

Estimator	Optimum k	GCV	MSE	R^2
Cosine Fourier	1	**2.471**	0.0006	77.545%
Sine Fourier	2	4.785	0.0003	50.033%
General Fourier	2	6.720	0.0003	57.981%

Based on Table 5, it can be concluded that the Fourier cosine series estimator was the best estimated model with GCV value of 2.471, MSE value of 0.0006, and R^2 value of 77.545%. Refer to goodness-of-fit criteria, which are minimum GCV and MSE value; and high value of R^2, these values of Fourier cosine series estimator shows that it was good. The estimator of cosine Fourier series was the most parsimony estimator. Based on the optimum oscillation parameter value, the estimator obtained can be presented as follows:

$$\hat{y}_i = -342.145 - 0.564x_{i1} + 10.704t_{i1} + 330.487 \cos t_{i1} + 1.982t_{i2} + 12.113 \cos t_{i2} \quad (15)$$

Prediction of Percentage of Criminal Incidents in Indonesia in 2019

After obtaining semiparametric regression model with Fourier series approach, the prediction of percentage of criminal incidents in Indonesia can be presented based on testing data of percentage of criminal incidents in 2019. It can be seen in Table 6. In addition, Fig. 3 shows the prediction result visually to make the interpretion of prediction result easily to do based on equation (15). Based on Fig. 3 the predicted value of the percentage of criminal incidents was getting closer to the actual value. This indicates the model used to predict the percentage of criminal incidents in Indonesia was suitable.

TABLE 6. Comparison between Predicted Value and Actual Value of the Percentage of Criminal Incidents in 2019

Number	Province	Actual Value (Testing Data)	Predicted Value
1	DKI Jakarta	0.027	0.038
2	North Sumatera	0.103	0.080
3	West Java	0.041	0.049
4	East Java	0.024	0.055
5	South Sumatera	0.026	0.039
6	South Sulawesi	0.057	0.055
7	West Sumatera	0.017	0.027
8	Central Java	0.029	0.055
9	Lampung	0.006	0.028
10	North Sulawesi	0.014	0.033
11	Aceh	0.123	0.082
12	Central Sulawesi	0.082	0.092
13	Jambi	0.04	0.085
14	East Kalimantan	0.023	0.037
15	Riau	0.081	0.086
16	DI Yogyakarta	0.013	0.087
17	Papua	0.013	0.046
18	East Nusa Tenggara	0.022	0.040
19	West Nusa Tenggara	0.022	0.034
20	West Kalimantan	0.02	0.048
21	South Kalimantan	0.01	0.037
22	Bengkulu	0.02	0.047
23	Riau Islands	0.025	0.042
24	Bali	0.028	0.033
25	Banten	0.027	0.033
26	Gorontalo	0.042	0.063
27	Southeast Sulawesi	0.011	0.033
28	Central Kalimantan	0.011	0.024
29	West Papua	0.007	0.026
30	Maluku	0.003	0.028
31	Island of Bangka Belitung	0.009	0.022
32	North Maluku	0.023	0.024

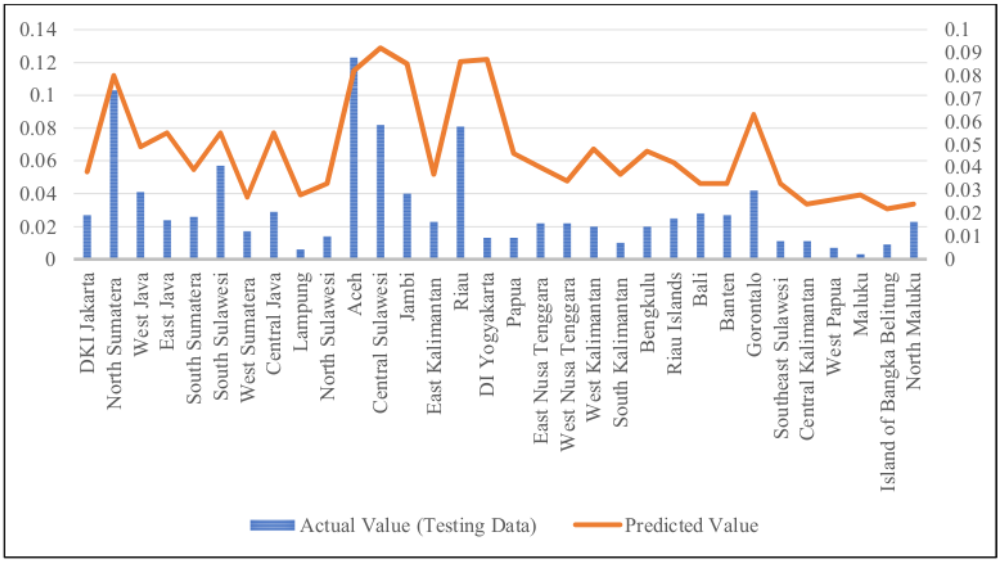


FIGURE 3. Comparison of Predicted Value and the Actual Value of the Percentage of Criminal Incidents

Based on Table 3, it can be seen that the predicted value was closer enough to the actual ones. Then based on the calculation of error using Mean Absolute Error (MAE), the error value was 0.02. This value indicates that the use of semiparametric regression method with Fourier estimator can produce prediction, which is good enough. The use of MAE in this calculation was based on advantages it has than others. Those are MAE is the most natural measure of average error magnitude, and it is an unambiguous measure of average error magnitude [11].

CONCLUSIONS

Semiparametric regression estimator based on Fourier series can be determined using OLS optimization with result an estimator for the parameter vector and an estimator for curve regression. This result can be applied in several fields. One of them is the estimation of the percentage of criminal incidents based on the affected factors. Based on the analysis, the estimator result had satisfied goodness of criteria, such as minimum GCV value, minimum MSE value and high determination coefficient value for parsimony model. The analysis shows that the predicted value of the percentage of criminal incidents was getting closer value because it had the MAE value of 0,02. This indicates that the model produced using semiparametric approach based on Fourier series estimator is suitable to predict the percentage of criminal incidents in Indonesia. For further research, the data can be applied based on nonparametric regression or the other estimator in semiparametric regression.

ACKNOWLEDGMENTS

The authors give high appreciation for Central Bureau of Statistics in Indonesia that has provided data, and Universitas Airlangga that has supported this publication.

REFERENCES

1. L. J. Asrini, and I. N. Budiantara, ARPN Journal of Engineering and Applied Sciences **9**, pp. 1501 – 1506 (2014).
2. M. T. Sheykhi, Sociology and Criminology-Open Access **4,** pp. 1 – 2 (2016).
3. M. Bilodeau, The Canadian Journal of Statistics **3**, pp. 257–259 (1992).
4. M. F. F. Mardianto, E. Tjahjono and M. Rifada, ARPN Journal of Engineering and Applied Sciences **14**, pp. 2763 – 2770 (2019).
5. S. Biedermann, H. Dette and P. Hoffmann, Ann Inst Stat Math Journal **61**, pp. 143–157 (2009).

6. H. Dette, V.B. Melas and P. Shpilev, Journal of Computational Statistics and Data Analysis **26**, pp. 1–11 (2016).
7. R. Pane, I. N. Budiantara, I. Zain, and B. W. Otok, Applied Mathematical Sciences, **8**, pp. 5053-5064 (2014).
8. The Jakarta Post, Crime in Indonesia surges in late May: Police, retrived from https://www.thejakartapost.com/news/2020/06/04/crime-in-indonesia-surges-in-late-may-police.html (2020).
9. M. F. F. Mardianto, E. Tjahjono and M. Rifada, Journal of Physics: Conference Series **1227**, pp. 1–10 (2019).
10. G. Wahba, *Spline Model for Observational Data* (SIAM XII, Philadelphia, 1990), pp. 23–25.
11. C. J. Willmot and K. Matsuura, Climate Research **30**, pp. 79–82 (2005).

Fourier series estimator in semiparametric regression to predict criminal rate in Indonesia

ORIGINALITY REPORT

15%	11%	12%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

4%

★ M. Fariz Fadillah Mardianto, Nurul Afifah, Siti Amelia Dewi Safitri, Idrus Syahzaqi, Sediono. "Estimated price of shallots commodities national based on parametric and nonparametric approaches", AIP Publishing, 2021

Publication

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		