STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, April 2026, pp 3255–3281.
Published online in International Academic Press (www.IAPress.org)

A Penalized Least Squares Estimation of Fourier Series Semiparametric Regression: Theory, Simulation, and Application

Ihsan Fathoni Amri [1], Nur Chamidah [2,*], Toha Saifudin [2], Budi Lestari [3], Dursun Aydin [4], Febrian Hikmah Nur Rohim [1]

[1]*Department of Data Science, Faculty of Science and Agriculture Technology, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia*
[2]*Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*
[3]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, The University of Jember, Indonesia*
[4]*Statistics Department, Faculty of Science, Muğla Sıtkı Koçman University, Muğla 48000, Turkey*

Abstract In regression analysis, the functional relationship between the response and predictor variables may follow a semiparametric regression model composed of both parametric and nonparametric components, where the nonparametric component is a time–dependent function approximated using a Fourier Series. In this study, we develop a penalized least squares smoothing technique to estimate the Fourier Series Semiparametric Regression (FSSR) model. The penalized least squares method is particularly effective when the generalized cross validation method fails to select optimal parameters due to the neglected overfitting effect in the model.We also provide a numerical illustration through a simulation study and apply the proposed method to real data for predicting Earth's surface temperature based on relative humidity. The results show that the FSSR model produces a MAPE value of 1.068%, indicating a very high level of prediction accuracy. In addition, the low RMSE value of 0.2816 demonstrates that the model's prediction errors remain stable for both in-sample and out-of-sample data. This stability further confirms that the FSSR model is capable of mitigating potential overfitting, thereby providing consistent and reliable estimates.

Keywords Estimation, Fourier Series Semiparametric Regression Model, Penalized Least Square, Generalized Cross Validation, Mean Average Percentage Error.

DOI: 10.19139/soic-2310-5070-3139

1. Introduction

Regression analysis is one of statistical techniques to analyze the functional relationship between response and predictor variables. In the regression analysis, the functional relationship between response and predictor variables is called a regression function [1, 2, 3, 4, 5]. If the form of regression function is known then we will use the parametric regression approach. In this approach, the regression function estimation is equivalent to the estimation of the parameters within the parametric regression model [6, 7]. The nonparametric regression approach is applied when the regression function is still unrecognized, or there is incomplete knowledge regarding the form of the data. The nonparametric regression approach has high flexibility, because its regression function does not have specification in some forms. Its regression function is just assumed to be smooth, so that to estimate it we can use several smoothing techniques [2, 4, 5, 7]. Next, if we combine the parametric regression model and the nonparametric regression model, we will have a new regression model called semiparametric regression model where its regression function is constructed by two components namely parametric component and nonparametric component [8]. It implies that the estimating regression function of the semiparametric model is equivalent to

*Correspondence to: Nur Chamidah (Email: nur-c@fst.unair.ac.id). Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

estimating these components [1, 3]. There are several smoothing techniques in nonparametric regression and semiparametric regression approaches which were used in many cases, for examples, kernel that was used for estimating asset pricing model [9], and for estimating nonparametric regression models with some modified assumptions [10, 11, 12] local linear that was used for designing some locally standard growth charts of toddlers [13, 14, 15] for identifying the number of Mycobacterium tuberculosis [16] , and for estimating the HIV and AIDS model [17]; local polynomial that was used for designing children growth charts [13, 18] ; smoothing spline which was used for estimating nonparametric regression models in several cases [19, 20, 21, 22, 23], and for determining asymptotic properties of estimators [1, 2, 3, 24]; least square spline that was used for estimating model of blood pressures affected by stress scores [25] , for estimating mean arterial pressure affected by stress scores [26] , and for designing standard growth charts [23, 27]; truncated spline was used for estimating semiparametric regression model [1]; Fourier series smoothing techniques were discussed [28, 29, 30, 31]. Additionally [11], discussed spline and kernel for estimating multiresponse nonparametric regression model; [12, 32] discussed spline and kernel smoothing techniques for selecting optimal smoothing parameter; and for estimating coefficient in a rates model, respectively.

In regression modeling, for estimating regression functions we usual apply several optimization methods, for examples, least square and weighted least square that have been discussed by some previous researchers mentioned above [33]; penalized least square and penalized weighted least square that have been discussed by [2, 3, 12, 15, 28] . According to [33], penalized least square is very good to use if GCV (Generalized Cross Validation) method cannot choose really good parameters because of over-fitting effect in the model is negligible. That is why penalized least square must be chosen to avoid over-fitting effect. Meanwhile, applying Fourier Series (FS) using both sine and cosine as the estimators in developing the semiparametric regression model has not been discussed by previous researchers. In this study, the nonparametric component of the semiparametric regression model is a function of time which will be approximated by an Fourier series. In the following discussion, we will call the model for cases like this as the Fourier Series Semiparametric Regression (FSSR) model. Therefore, in this study, we develop a mathematical estimation method for the Fourier Series Semiparametric Regression (FSSR) model for time series data by using Penalized Least Square (PLS) smoothing technique.

In this study, a semiparametric approach using Penalized Least Squares (PLS) is combined with the Penalized Fourier Series estimator and validated using simulation data. The simulation results are then integrated with 60 observational data points of land surface temperature and humidity at a height of 2 meters. The response variable used is land surface temperature, while the predictor variable is humidity at a height of 2 meters. In this model, the relationship between land surface temperature and observation time is treated as a nonparametric component, whereas the relationship between land surface temperature and humidity is treated as a parametric component. The contribution of this research lies in the development of a semiparametric model based on the Fourier Series estimator for time series data analysis, achieving high accuracy in modeling meteorological variables such as land surface temperature, relative humidity, and time. By integrating parametric and nonparametric components, this model effectively captures complex relationship patterns that cannot be explained linearly. The use of Penalized Least Squares (PLS) helps prevent overfitting, while optimal parameter selection through Generalized Cross Validation (GCV) ensures stable estimation with low prediction error.

2. Research Methods

Suppose that we have a paired dataset $(y, g(t), u_1, u_2, \ldots, u_p)$ that satisfies a semiparametric regression model as follows:

$$y = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \ldots + \beta_p u_p + g(t) + \varepsilon \tag{1}$$

Hence, for every $i = 1, 2, \ldots, n,$, the paired dataset $(y_i, t_i, u_{1i}, u_{2i}, \ldots, u_{pi})$ satisfies a regression model as follows:

$$y_i = \beta_0 + \beta_1 u_{1i} + \beta_2 u_{2i} + \ldots + \beta_p u_{pi} + g(t_i) + \varepsilon_i. \tag{2}$$

Next, we can write the model in **Equation (2)** as follows:

$$y = \mathbf{U}^{\mathrm{T}}\boldsymbol{\beta} + g(t) + \varepsilon \tag{3}$$

where is a vector of responses, $\mathbf{U}$ is a matrix of predictors for parametric component, $\boldsymbol{\beta}$ is a vector of parameters for parametric component, $g(t)$ is a vector of nonparametric regression functions, and is a vector of random errors, where $\varepsilon \sim N_n(0, \sigma^2\mathbf{I})$ that is multivariate normally distributed. The function $g(t)$ can be approximated by using Fourier series estimator that has high flexibility, then really good to use in volatile data. The Fourier series function used in this study differs from the recommended form [35]. In this research, the Fourier function is generated based on the complex exponential formulation, which is mathematically expressed as follows: :

$$g(t_i) = a_0 + \sum_{j=1}^{J} \left[c_j \cos(2\pi j t_i) + d_j \sin(2\pi j t_i)\right]. \tag{4}$$

According to Wang et al [**?**]. penalized least square (PLS) is a good optimization method to avoid over-fitting effect. Therefore, here we use the PLS for estimating the semiparametric model based on Fourier series estimator. The PLS optimization in the semiparametric regression based on Fourier series estimator is given by:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1},\, g\in\mathcal{C}(0,1)}{\text{Min}} \left[n^{-1} \sum_{i=1}^{n} \left(y_i - \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\beta} - g(t_i)\right)^2 + \lambda \int_0^1 \left[g^{(2)}(t_i)\right]^2 dt_i \right]$$

where λ (Lambda)represents a smoothing parameter which controls trade-off between goodness of fit and smoothness of an estimation curve.

Also, we provide a simulation study for giving an illustration about implementation of the proposed method. In this simulation study, we use four different samples sizes, namely, $n = 100$ that represents large samples, $n = 40$ and $n = 30$ that represent moderate samples, and $n = 20$ that represents small samples with Fourier coefficients, $k = 1, 2, ..., 10$.

3. Results And Discussion

In this section we provide results and discussions of this study that consist of the theoretical estimation result of the Fourier series semiparametric regression (FSSR) model using penalized least squares (PLS) smoothing technique, and estimation result of the Fourier series semiparametric regression (FSSR) model based on a simulation study.

3.1. Estimation Result of the FSSR Model

Consider a paired dataset $(y, g(t), u_1, u_2, \ldots, u_p)$ that follows the semiparametric regression model presented in **Equation (1)** as follows:

$$y = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \ldots + \beta_p u_p + g(t) + \varepsilon$$

Hence, for every $i = 1, 2, \ldots, n$,, we have the following semiparametric regression model:

$$y_i = \beta_0 + \beta_1 u_{1i} + \beta_2 u_{2i} + \ldots + \beta_p u_{pi} + g(t_i) + \varepsilon_i.$$

We can express the semiparametric regression model in the matrix equation form as follows:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & u_{11} & u_{21} & \cdots & u_{p1} \\ 1 & u_{12} & u_{22} & \cdots & u_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_{1n} & u_{2n} & \cdots & u_{pn} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{bmatrix} g(t_1) \\ g(t_2) \\ \vdots \\ g(t_n) \end{bmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Thus, we have the semiparametric regression model as follows:

$$\mathbf{y} = \mathbf{U}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{g}(t) + \boldsymbol{\varepsilon}.$$

In the semiparametric regression model presented in **Equation (2)**, the nonparametric function, $g(t)$ is unknown and will be estimated by using penalized least square (PLS) smoothing technique based on Fourier series estimator, and then in the next discussion, the model will be called the Fourier Series Semiparametric Regression (FSSR) model. Here, we assume that $g(t_i) \in L_2[a,b]$. In other words, the function $g(t_i)$ is contained in a Hilbert space, $L_2[a,b]$, such that $g(t_i)$ can be expressed as linear combination of bases elements of $L_2[a,b]$, Furthermore, if $\{z_j\}_{j=1}^{\infty}$ is orthonormal complete system $(z_1, z_2, \ldots, z_\infty)$ of $L_2[a,b]$, then we have:

$$g(t_i) = \sum_{j=1}^{\infty} \alpha_j z_j(t_i). \tag{5}$$

where α_j is a scalar. Hence, based on **Equations (2)** and **Equation (5)**, for j = $1, 2, ..., 10$, we have:

$$y_i = \beta_0 + \beta_1 u_{1i} + \beta_2 u_{2i} + \ldots + \beta_p u_{pi} + \sum_{j=1}^{\infty} \alpha_j z_j(t_i) + \varepsilon_i. \tag{6}$$

Let $\alpha_j = \langle g, z_j \rangle$ where $\{z_j\}_{j=1}^{\infty}$ is an orthonormal complete system $(z_1, z_2, \ldots, z_\infty)$ of $L_2[a,b]$ and $g \in L_2[a,b]$, then α_j is a Fourier coefficient for g which satisfies $\sum_{j=1}^{\infty} |\alpha_j|^2 = \|g\|^2$, because $\sum_{j=1}^{\infty} |\alpha_j|^2 < \infty$ and $z_j \to 0$. While, if n is infinite, then the regression function g can be approached by the following function:

$$g(t_i) = \sum_{j=1}^{\lambda} \alpha_j z_j(t_i) \tag{7}$$

where λ is an integer. Based on **Equation (7)**, for $j = 1, 2, ..., 10$, we can write the **Equation (6)** as follows:

$$y_i = \beta_0 + \beta_1 u_{1i} + \beta_2 u_{2i} + \ldots + \beta_p u_{pi} + \sum_{j=1}^{\lambda} \alpha_j z_j(t_i) + \varepsilon_i \tag{8}$$

where $t_1, t_2, ..., t_n$ are assumed have the same distance in $[a,b]$. The unknown Fourier coefficient can be estimated using the optimal value of λ for obtaining smooth regression function.

Further, if observation is time and shows a periodic pattern then function g can be estimated by using linear model of sine and cosine functions where the complete function is constructed from sine and cosine functions with an orthonormal complete system of $L_2[a,b]$ that is defined by:

$$z_j(t_i) = e^{2\pi l j t_i}, \quad l = \sqrt{-1} \text{ and } j = 0, \pm 1, \ldots \tag{9}$$

The estimation of the nonparametric regression function, $g(t_i)$, is given by Theorem 1.

Theorem 1. *By considering **Equation (9)**, if the estimated nonparametric regression function, $\hat{g}(t_i)$, which is a sine and cosine function can be expressed into a Fourier series nonparametric based on the nonparametric regression model $\boldsymbol{y} = \boldsymbol{g}(t_i) + \varepsilon$ then $\hat{g}(t_i)$is given by:*

$$\hat{g}(t_i) = \hat{a}_0 + \sum_{j=1}^{J} \left[c_j \cos(2\pi j t_i) + d_j \sin(2\pi j t_i) \right].$$

Proof of Theorem 1. Suppose we have a paired dataset $(y_i, t_i, u_{1i}, u_{2i}, ..., u_{ni})$. Next, by considering **Equation (8)** and **Equation (9)**, we have the Fourier series semiparametric regression model as follows:

$$\boldsymbol{y} = \boldsymbol{z}_\lambda \boldsymbol{\alpha}_\lambda + \boldsymbol{\varepsilon} \tag{10}$$

where:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{z}_\lambda = \begin{bmatrix} e^{2\pi i(-\lambda)1} & e^{2\pi i(-\lambda+1)1} & \dots & e^{2\pi i\lambda 1} \\ e^{2\pi i(-\lambda)2} & e^{2\pi i(-\lambda+1)2} & \dots & e^{2\pi i\lambda 2} \\ \vdots & \vdots & \ddots & \vdots \\ e^{2\pi i(-\lambda)n} & e^{2\pi i(-\lambda+1)n} & \dots & e^{2\pi i\lambda n} \end{bmatrix}, \quad \boldsymbol{\alpha}_\lambda = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_\lambda \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Based on **Equation (10)**, the least square estimator for $\boldsymbol{\alpha}_\lambda$ which minimizes the sum square errors (SSE), $\boldsymbol{Q} = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}$. , can be obtained as follows:

$$Q = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{z}_\lambda\boldsymbol{\alpha}_\lambda)^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{z}_\lambda\boldsymbol{\alpha}_\lambda) = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2\boldsymbol{z}_\lambda^{\mathrm{T}}\boldsymbol{\alpha}_\lambda^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{z}_\lambda^{\mathrm{T}}\boldsymbol{\alpha}_\lambda^{\mathrm{T}}\boldsymbol{z}_\lambda\boldsymbol{\alpha}_\lambda.$$

Next, we determine the partially differential of **Q** with respect to $\boldsymbol{\alpha}_\lambda$ as follows:

$$\frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{\alpha}_\lambda} = 0 \iff \hat{\boldsymbol{\alpha}}_\lambda = (\boldsymbol{z}_\lambda^{\mathrm{T}}\boldsymbol{z}_\lambda)^{-1}\boldsymbol{z}_\lambda^{\mathrm{T}}\boldsymbol{y}. \tag{11}$$

Hence, the estimation of nonparametric regression function, $g(t_i)$, is given by:

$$\hat{g}(t_i) = \sum_{j=1}^{\lambda} \hat{\alpha}_j z_j(t_i) \tag{12}$$

where $z_j = \{z_\lambda\}_{j=-\lambda}^{\lambda}$. Also, we can express the estimator for the Fourier series nonparametric regression function as follows:

$$\hat{\mathbf{g}}(t_i) = \mathbf{z}_\lambda\boldsymbol{\alpha}_\lambda \tag{13}$$

and the elements of matrix $\mathbf{z}_\lambda = \left\{e^{2\pi ijx}\right\}_{j=-\lambda}^{\lambda}$ in **Equation (13)** with dimension $n \times (2\lambda + 1)$ are given as follows:

$$\mathbf{z}_\lambda = \begin{bmatrix} e^{2\pi i(-\lambda)1} & e^{2\pi i(-\lambda+1)1} & \dots & e^{2\pi i\lambda 1} \\ e^{2\pi i(-\lambda)2} & e^{2\pi i(-\lambda+1)2} & \dots & e^{2\pi i\lambda 2} \\ \vdots & \vdots & \ddots & \vdots \\ e^{2\pi i(-\lambda)n} & e^{2\pi i(-\lambda+1)n} & \dots & e^{2\pi i\lambda n} \end{bmatrix}$$

and the transpose of matrix $\mathbf{z}_\lambda$ is given by:

$$\mathbf{z}_\lambda^T = \begin{bmatrix} e^{2\pi i(-\lambda)1} & e^{2\pi i(-\lambda)2} & \dots & e^{2\pi i(-\lambda)n} \\ e^{2\pi i(-\lambda+1)1} & e^{2\pi i(-\lambda+1)2} & \dots & e^{2\pi i(-\lambda+1)n} \\ \vdots & \vdots & \ddots & \vdots \\ e^{2\pi i\lambda 1} & e^{2\pi i\lambda 2} & \dots & e^{2\pi i\lambda n} \end{bmatrix}.$$

Next, by using the definition of an orthonormal sequence, we have:

$$\int_a^b z_j(t)z_k(t)\,dt = \begin{cases} 0 & \text{for } j \neq k, \\ 1 & \text{for } j = k. \end{cases}$$

Hence, we obtain:

$$\mathbf{z}_\lambda^T\mathbf{z}_\lambda = \begin{pmatrix} e^{2\pi i(-\lambda)1} & e^{2\pi i(-\lambda)2} & \dots & e^{2\pi i(-\lambda)n} \\ e^{2\pi i(-\lambda+1)1} & e^{2\pi i(-\lambda+1)2} & \dots & e^{2\pi i(-\lambda+1)n} \\ \vdots & \vdots & \ddots & \vdots \\ e^{2\pi i\lambda 1} & e^{2\pi i\lambda 2} & \dots & e^{2\pi i\lambda n} \end{pmatrix}^T \begin{pmatrix} e^{2\pi i(-\lambda)1} & e^{2\pi i(-\lambda+1)1} & \dots & e^{2\pi i\lambda 1} \\ e^{2\pi i(-\lambda)2} & e^{2\pi i(-\lambda+1)2} & \dots & e^{2\pi i\lambda 2} \\ \vdots & \vdots & \ddots & \vdots \\ e^{2\pi i(-\lambda)n} & e^{2\pi i(-\lambda+1)n} & \dots & e^{2\pi i\lambda n} \end{pmatrix}$$

$$= \begin{pmatrix} n & 0 & \cdots & 0 \\ 0 & n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n \end{pmatrix} = n \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Therefore, we have: $(\mathbf{z}_\lambda^T \mathbf{z}_\lambda)^{-1} = \frac{1}{n}\mathbf{I}$, where $\mathbf{I}$, where is an identity matrix. Also, in this step we have:

$$\mathbf{z}_\lambda^T \mathbf{y} = \begin{bmatrix} e^{2\pi i(-\lambda)1} & e^{2\pi i(-\lambda)2} & \cdots & e^{2\pi i(-\lambda)n} \\ e^{2\pi i(-\lambda+1)1} & e^{2\pi i(-\lambda+1)2} & \cdots & e^{2\pi i(-\lambda+1)n} \\ \vdots & \vdots & \ddots & \vdots \\ e^{2\pi i\lambda 1} & e^{2\pi i\lambda 2} & \cdots & e^{2\pi i\lambda n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} e^{2\pi i(-\lambda)1}y_1 + e^{2\pi i(-\lambda)2}y_2 + \cdots + e^{2\pi i(-\lambda)n}y_n \\ e^{2\pi i(-\lambda+1)1}y_1 + e^{2\pi i(-\lambda+1)2}y_2 + \cdots + e^{2\pi i(-\lambda+1)n}y_n \\ \vdots \\ e^{2\pi i\lambda 1}y_1 + e^{2\pi i\lambda 2}y_2 + \cdots + e^{2\pi i\lambda n}y_n \end{bmatrix} = \sum_{i=1}^{n} y_i e^{-2\pi i j t_i} \quad \text{for } -\lambda \le j \le \lambda.$$

Thus, the **Equation (12)** can be written as follows:

$$\hat{a}_j = n^{-1} \sum_{i=1}^{n} y_i e^{-2\pi i j t_i}. \tag{14}$$

Based on the **Equation (10), (13)**, and **Equation (14)**, the Fourier series estimator for $g(t_i)$, namely, $\hat{g}(t_i)$, is:

$$\hat{\theta}(t_i) = \sum_{j=-\lambda}^{\lambda} \hat{a}_j e^{-2\pi i j t_i} = \sum_{j=-\lambda}^{\lambda} \left(n^{-1} \sum_{i=1}^{n} y_i^* e^{-2\pi i j t_i} \right) e^{2\pi i j t_i}. \tag{15}$$

By taking $[a, b] = [0, 1]$ and t_i is equidistance in $[0, 1]$ that is $t_i = \frac{i-1}{n}$ for $i = 1, 2, \ldots, n$ the Fourier series estimator in **Equation (15)** can be written as:

$$\hat{\theta}(t_i) = \sum_{j=-J}^{J} \hat{a}_j e^{\frac{2\pi i j(i-1)}{n}}. \tag{16}$$

where $i = 1, 2, ..., n$ and $l = \sqrt{-1}$. Here, $\hat{g}(t_i)$ has real component and imaginary component. If it is defined that $\hat{\alpha}_{-j}$ is a compound complex of $\hat{\alpha}_j$ where

$$c_j = \hat{a}_j + \hat{a}_{-j} \quad \text{and} \quad d_j = i(\hat{a}_j + \hat{a}_{-j}) \tag{17}$$

then based on **Equation (17)** the values of c_j and d_j can be determined by the following equations:

$$e^{it} = \cos(t) + i\sin(t) \quad \text{and} \quad e^{-it} = \cos(t) - i\sin(t).$$

So, c_j and d_j can be written as follows:

$$\begin{aligned}
c_j &= \hat{a}_j + \hat{a}_{-j} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i e^{2\pi i j t_i} + \sum_{i=1}^{n} y_i e^{-2\pi i j t_i}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} y_i(\cos(2\pi j t_i) + i\sin(2\pi j t_i)) + \sum_{i=1}^{n} y_i(\cos(2\pi j t_i) - i\sin(2\pi j t_i))\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}(y_i\cos(2\pi j t_i) + y_i i\sin(2\pi j t_i)) + \sum_{i=1}^{n}(y_i\cos(2\pi j t_i) - y_i i\sin(2\pi j t_i))\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\cos(2\pi j t_i) + i\sin(2\pi j t_i) + \cos(2\pi j t_i) - i\sin(2\pi j t_i))y_i \\
&= \frac{2}{n}\sum_{i=1}^{n} y_i\cos(2\pi j t_i)
\end{aligned}$$

$$c_j = \frac{2}{n}\sum_{i=1}^{n} y_i \cos\left(\frac{2\pi j(i-1)}{n}\right) \tag{18}$$

$$\begin{aligned}
d_j &= l(\hat{a}_j + \hat{\alpha}_{-j}) = l\frac{1}{n}\left(\sum_{i=1}^{n} y_i e^{-2\pi l j t_i} - \sum_{i=1}^{n} y_i e^{2\pi l j t_i}\right) \\
&= l\left(\frac{1}{n}\sum_{i=1}^{n} y_i(e^{-2\pi l j t_i} - e^{2\pi l j t_i})\right) \\
&= l\left(\frac{1}{n}\sum_{i=1}^{n} y_i(\cos(2\pi l j t_i) - l\sin(2\pi l j t_i) - \cos(2\pi l j t_i) - l\sin(2\pi l j t_i))\right) \\
&= l\left(\frac{1}{n}\sum_{i=1}^{n} y_i(-2l\cos(2\pi l j t_i))\right) = \frac{2}{n}\sum_{i=1}^{n} y_i\sin(2\pi l j t_i)
\end{aligned}$$

$$d_j = \frac{2}{n}\sum_{i=1}^{n} y_i \sin\left(\frac{2\pi j(i-1)}{n}\right). \tag{19}$$

Based on **Equation (18)** and **Equation (19)**, the Fourier series estimator $\hat{g}(t_i)$ can be written as follows:

$$\begin{aligned}
\hat{g}(t_i) &= \sum_{j=-J}^{J} \hat{a}_j e^{2\pi j t_i} \\
&= \sum_{j=-J}^{J}(\hat{a}_j e^{2\pi j t_i} + \hat{a}_{-j} e^{-2\pi j t_i}) = \hat{a}_0 + \sum_{j=1}^{J}(\hat{a}_j e^{2\pi j t_i} + \hat{a}_{-j} e^{-2\pi j t_i}) \\
&= \hat{a}_0 + \sum_{j=1}^{J}(\hat{a}_j(\cos(2\pi j t_i) + \iota\sin(2\pi j t_i)) + \hat{a}_{-j}(\cos(2\pi j t_i) - \iota\sin(2\pi j t_i))) \\
&= \hat{a}_0 + \sum_{j=1}^{J}((\hat{a}_j + \hat{a}_{-j})\cos(2\pi j t_i) - \iota(\hat{a}_j - \hat{a}_{-j})\sin(2\pi j t_i))
\end{aligned}$$

$$\hat{a}_0 + \sum_{j=1}^{J'} \left(c_j \big(\cos(2\pi j t_i) \big) + d_j \big(\sin(2\pi j t_i) \big) \right). \tag{20}$$

Thus, the Fourier series estimator given in **Equation (20)** can be written as follows:

$$\hat{g}(t_i) = \hat{a}_0 + \sum_{j=1}^{J'} \left[c_j \big(\cos(2\pi j t_i) \big) + d_j \big(\sin(2\pi j t_i) \big) \right] \tag{21}$$

where $c_j = \frac{2}{n} \sum_{i=1}^{n} y_i \cos(2\pi j t_i)$ and $d_j = \frac{2}{n} \sum_{i=1}^{n} y_i \sin(2\pi j t_i)$. Thus, **Theorem 1** is proved.

The completely estimated FSSR model is obtained by estimating the parameter $\boldsymbol{\beta}$ and the function $\boldsymbol{g}(t_i)$ by using the PLS optimization which is given in the **Theorem 2**.

Theorem 2. *If given the semiparametric regression model presented by **Equation (3)**, and the regression function $g(t_i)$is approximated by **Equation (4)**, then the estimator for parameter $\boldsymbol{\beta}$ and estimator for function $\boldsymbol{g}(t_i)$ can be determined by applying the following PLS optimization:*

$$\min_{\beta \in \mathbb{R}^{r+1},\, g \in \mathcal{C}(0,1)} \left[n^{-1} (\mathbf{y}^* - \mathbf{g})^{\mathrm{T}} (\mathbf{y}^* - \mathbf{g}) + \lambda \int_0^1 \left(g^{(2)}(t) \right)^2 dt \right]$$

where $\mathbf{y}^* = \mathbf{y} - \mathbf{U}^{\mathrm{T}} \boldsymbol{\beta}$, $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{U}^{\mathrm{T}} = \begin{bmatrix} 1 & u_{11} & u_{21} & \cdots & u_{p1} \\ 1 & u_{12} & u_{22} & \cdots & u_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_{1n} & u_{2n} & \cdots & u_{pn} \end{bmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$, *and the PLS optimization will give:* $\hat{\boldsymbol{\beta}} = [\mathbf{U}^{\mathrm{T}} \mathbf{V} \mathbf{U}]^{-1} \mathbf{U}^{\mathrm{T}} \mathbf{V} \mathbf{y}$; $\hat{g}(t_i) = \mathbf{H}(\mathbf{I} - \mathbf{H}\mathbf{U}[\mathbf{U}^{\mathrm{T}} \mathbf{V} \mathbf{U}]^{-1} \mathbf{U}^{\mathrm{T}} \mathbf{V}) \mathbf{y}$; and $\hat{y}(u,t) = \mathbf{U}\hat{\boldsymbol{\beta}} + \hat{g}(t_i)$. **Proof of Theorem 2.** Parameter $\boldsymbol{\beta}$ and the function $\boldsymbol{g}(t_i)$ in the semiparametric regression model presented by **Equation (3)** can be estimated by using the following PLS optimization:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{r+1},\, g \in \mathcal{C}(0,1)} \left[n^{-1} \sum_{i=1}^{n} \left(y_i - \mathbf{U}_i^{\mathrm{T}} \boldsymbol{\beta} - g(t_i) \right)^2 + \lambda \int_0^1 \left(g^{(2)}(t_i) \right)^2 dt_i \right]$$

Based on **Equation (5)**, we have the penalty $p(g)$ as follows:

$$P(g) = \lambda \int_0^1 \left(g^{(2)}(t_i) \right)^2 dt_i = \lambda \int_0^1 \left[\left(\hat{a}_0 + \sum_{j=1}^{J} \left(c_j \cos(2\pi j t_i) + d_j \sin(2\pi j t_i) \right) \right)^{(2)} \right]^2 dt_i$$

$$= P(g) = \lambda \int_0^1 \left[\frac{d^2}{dt^2} \left(\hat{a}_0 + \sum_{j=1}^{J} \left(c_j \cos(2\pi j t_i) + d_j \sin(2\pi j t_i) \right) \right) \right]^2 dt_i \tag{22}$$

The squared second derivative of the function $g(t_i)$ in **Equation (22)** can be obtained as follows:

$$\left(g^{(2)}(t_i) \right)^2 = \left[\frac{d^2}{dt^2} \left(\hat{a}_0 + \sum_{j=1}^{J} \left(c_j \cos(2\pi j t_i) + d_j \sin(2\pi j t_i) \right) \right) \right]^2$$

where $g^{(1)}(t_i) = \frac{d}{dt}\left[\sum_{j=1}^{J}(-2\pi j c_j \sin(2\pi j t_i)) + \sum_{j=1}^{J}(2\pi j d_j \cos(2\pi j t_i))\right]$. and $g^{(2)}(t_i) = \sum_{j=1}^{J}(-4\pi^2 j^2 c_j \cos(2\pi j t_i)) - \sum_{j=1}^{J}(4\pi^2 j^2 d_j \sin(2\pi j t_i))$. Hence, we have:

$$\begin{aligned}\left(g^{(2)}(t_i)\right)^2 &= \left(\sum_{j=1}^{J}(-4\pi^2 j^2 c_j \cos(2\pi j t_i)) - \sum_{j=1}^{J}(4\pi^2 j^2 d_j \sin(2\pi j t_i))\right)^2 \\ &= \left(\sum_{j=1}^{J}(-4\pi^2 j^2 c_j \cos(2\pi j t_i))\right)^2 + \left(\sum_{j=1}^{J}(32\pi^4 j^4 c_j \cos(2\pi j t_i)\, d_j \sin(2\pi j t_i))\right) + \\ &\qquad \left(\sum_{j=1}^{J} -4\pi^2 j^2 d_j \sin(2\pi j t_i)\right)^2. \end{aligned} \tag{23}$$

Next, the **Equation (23)** is divided into three parts, and then taking the solution part by part. Hence, by considering **Equation (22)**, the part one of **Equation (23)**, namely, $P(g_{(1)})$, is as follows:

$$\boldsymbol{P(g_{(1)}) = \lambda \int_0^1 \left(\sum_{j=1}^{J} -4\pi^2 j^2 c_j \cos(2\pi j t_i)\right)^2 dt_i.} \tag{24}$$

To take the solution to **Equation (24)**, we give an illustration for $j = 1, 2, 3$ as follows:

$$\begin{aligned}\boldsymbol{P(g_{(1)})} &= \boldsymbol{\lambda \int_0^1 \left(\sum_{j=1}^{3} -4\pi^2 j^2 c_j \cos(2\pi j t_i)\right)^2 dt_i} \\ &= \left(-4\pi^2 c_1 \cos(2\pi t_i) - 4\pi^2 4 c_2 \cos(4\pi t_i) - 4\pi^2 9 c_3 \cos(6\pi t_i)\right)^2 \\ &= \left[(-4\pi^2 c_1 \cos(2\pi t_i))^2 + (-4\pi^2 4 c_2 \cos(4\pi t_i))^2 + (-4\pi^2 9 c_3 \cos(6\pi t_i))^2\right] \\ &\quad + 2[-4\pi^2 c_1 \cos(2\pi t_i)][4\pi^2 4 c_2 \cos(4\pi t_i)] + 2[-4\pi^2 c_1 \cos(2\pi t_i)][4\pi^2 9 c_3 \cos(6\pi t_i)] \\ &\quad + 2[-4\pi^2 4 c_2 \cos(4\pi t_i)][4\pi^2 9 c_3 \cos(6\pi t_i)] \\ &= \sum_{j=1}^{3}(-4\pi^2 j^2 c_j \cos(2\pi j t_i))^2 + \sum_{j=1}^{3}(-4\pi^2 j^2 c_j \cos(2\pi j t_i))^2 \\ &\quad + 2\sum_{j=1}^{3}\sum_{i<z}(-4\pi^2 j^2 c_j \cos(2\pi j t_i))(-4\pi^2 z^2 c_z \cos(2\pi z t_i)).\end{aligned}$$

The **Equation (24)** can be rewritten completely as follows:

$$\begin{aligned}P(g_{(1)}) = \lambda \int_0^1 \Bigg[&\sum_{j=1}^{J}\left(-4\pi^2 j^2 c_j \cos(2\pi j t_i)\right)^2 + \\ &2\sum_{\substack{j<z \\ j=1}}^{J}\sum_{z=1}^{J}\left(-4\pi^2 j^2 c_j \cos(2\pi j t_i)\right)\left(-4\pi^2 z^2 c_z \cos(2\pi z t_i)\right)\Bigg] dt_i. \end{aligned} \tag{25}$$

We can write the **Equation (25)** as follows:

$$P(g_{(1)}) = \lambda[A + B] \quad \text{where} \quad A = \int_0^1 \left[\sum_{j=1}^{J} \left(-4\pi^2 j^2 c_j \cos(2\pi j t_i)\right)\right]^2 dt_i,$$

and $B = 2\int_0^1 \left[\sum_{\substack{j<z \\ j=1}}^{J} \sum_{z=1}^{J} \left(-4\pi^2 j^2 c_j \cos(2\pi j t_i)\right)\left(-4\pi^2 z^2 c_z \cos(2\pi z t_i)\right)\right] dt_i$.
In this step, we have $A = \int_0^1 \left[\sum_{j=1}^{J} \left(-4\pi^2 j^2 c_j \cos(2\pi j t_i)\right)\right]^2 dt_i = \sum_{j=1}^{J} \left[16\pi^4 j^4 c_j^2 \left(\int_0^1 \cos^2(2\pi j t_i)\, dt_i\right)\right]$.
Since $\cos 2t = 2\cos^2 t - 1$; $\cos^2 t = \frac{1+\cos 2t}{2}$; and $\cos^2(2t) = \frac{1}{2} + \frac{1}{2}\cos(4t)$, then we have:

$$A = \sum_{j=1}^{J} \left[16\pi^4 j^4 c_j^2 \left(\int_0^{\frac{1}{2}} dt_i + \frac{1}{2}\int_0^1 \cos(4\pi j t_i)\, dt_i\right)\right].$$

Furthermore, let $u = 4\pi j t_i$. It implies $du = 4\pi j\, dt_i$ or $\frac{du}{4\pi j} = dt_i$. Based on this step, we obtain:

$$\begin{aligned} A &= \sum_{j=1}^{J} \left[8\pi^4 j^4 c_j^2 \left(\int_0^1 1\, dt_i + \int_0^1 \cos(u)\, \frac{du}{4\pi j}\right)\right] \\ &= \sum_{j=1}^{J} \left[8\pi^4 j^4 c_j^2 \left(t_i + \frac{1}{4\pi j}\sin(4\pi j t_i)\right)_0^1\right] \\ &= \sum_{j=1}^{J} \left[8\pi^4 j^4 c_j^2 (1-0)\right] = \sum_{j=1}^{J} \left[8\pi^4 j^4 c_j^2\right]. \end{aligned}$$

In the similar way, we also obtain:

$$\begin{aligned} B &= 2\int_0^1 \left[\sum_{\substack{j<z \\ j=1}}^{J} \sum_{j=1}^{J} (-4\pi^2 j^2 c_j \cos(2\pi j t_i))\,(-4\pi^2 z^2 c_z \cos(2\pi z t_i))\right] dt_i \\ &= 2\sum_{\substack{j<z \\ j=1}}^{J} \sum_{j=1}^{J} \int_0^1 \left[(-4\pi^2 j^2 c_j \cos(2\pi j t_i))\,(-4\pi^2 z^2 c_z \cos(2\pi z t_i))\right] dt_i \\ &= 2\sum_{\substack{j<z \\ j=1}}^{J} \sum_{j=1}^{J} 16\pi^4 j^2 z^2 c_j c_z \int_0^1 \left[\cos(2\pi j t_i)\cos(2\pi z t_i)\right] dt_i. \end{aligned}$$

The next step, we first determine value of the following integral:

$$\begin{aligned}\int_0^1 [\cos(2\pi j t_i)\cos(2\pi z t_i)]\,dt_i &= \int_0^1 \tfrac{1}{2}[\cos(2\pi(j+z)t_i)+\cos(2\pi(j-z)t_i)]\,dt_i \\ &= \tfrac{1}{2}\int_0^1 [\cos(2\pi(j+z)t_i)]\,dt_i + \tfrac{1}{2}\int_0^1 [\cos(2\pi(j-z)t_i)]\,dt_i \\ &= \tfrac{1}{2}\left[\tfrac{1}{2\pi(j+z)}\sin(2\pi(j+z)t_i)\right]_0^1 + \tfrac{1}{2}\left[\tfrac{1}{2\pi(j-z)}\sin(2\pi(j-z)t_i)\right]_0^1 \\ &= \tfrac{1}{2}\left[\tfrac{1}{2\pi(j+z)}\sin(2\pi(j+z)t_i)-0\right] + \tfrac{1}{2}\left[\tfrac{1}{2\pi(j-z)}\sin(2\pi(j-z)t_i)-0\right] \\ &= 0.\end{aligned}$$

Hence, we obtain:

$$\begin{aligned}B &= 2\sum_{\substack{j<z\\ j=1}}^{J}\sum_{j=1}^{J} 16\pi^4 j^2 z^2 c_j c_z \int_0^1 [\cos(2\pi j t_i)\cos(2\pi z t_i)]\,dt_i \\ &= 2\sum_{\substack{j<z\\ j=1}}^{J}\sum_{j=1}^{J} 16\pi^4 j^2 z^2 c_j c_z \cdot 0 = 0.\end{aligned}$$

Since $B=0$, we obtain: $P(g_{(1)}) = \lambda\displaystyle\int_0^1 \left(\sum_{j=1}^{J} -4\pi^2 j^2 c_j \cos(2\pi j t_i)\right)^2 dt_i \quad = \lambda\sum_{j=1}^{J}[8\pi^4 j^4 c_j^2]$. Next, the part two of **Equation (23)**, namely, $P(g_{(2)})$, is as follows:

$$\begin{aligned}P(g_{(2)}) &= \lambda\int_0^1 \left(\sum_{j=1}^{J} 32\pi^4 j^4 c_j \cos(2\pi j t_i)\right) d_j \sin(2\pi j t_i)\,dt_i \\ &= \lambda\int_0^1 \left(\sum_{j=1}^{J} 16\pi^4 j^4 c_j 2(\cos(2\pi j t_i) d_j \sin(2\pi j t_i))\right) dt_i \\ &= \lambda\sum_{j=1}^{J} 16\pi^4 j^4 c_j \int_0^1 2(\cos(2\pi j t_i) d_j \sin(2\pi j t_i))\,dt_i.\end{aligned}$$

Since $\sin(A+B) = \cos B + \cos A \sin B; \sin(2t+2t) = \cos 2t \sin 2t + \cos 2t \sin 2t;$ and $\sin(4t) = 2\cos 2t \sin 2t$, then we have: $P(g_{(2)}) = \lambda\sum_{j=1}^{J} 16\pi^4 j^4 c_j \int_0^1 \sin(4\pi j t_i)\,dt_i$. Let $4\pi j t_i. = u$. It gives $t_i = \frac{1}{4\pi j t_i}\,du$, and we have the value of $P(g_{(2)})$ as follows:

$$\begin{aligned}P(g_{(2)}) &= \lambda\sum_{j=1}^{J} 16\pi^4 j^4 c_j \int_0^1 \frac{\sin u\,du}{4\pi j} \\ du &= \lambda\sum_{j=1}^{J} 16\pi^4 j^4 c_j \left(\frac{1}{4\pi j}\right)[-\cos(4\pi j t_i)]_0^1 \\ &= \lambda\sum_{j=1}^{J} 16\pi^4 j^4 c_j \left(\frac{1}{4\pi j}\right)[-1-(-1)] = 0.\end{aligned}$$

Furthermore, the part three of **Equation (23)** is $P(g_{(3)}) = \lambda \int_0^1 \left(\sum_{j=1}^{J} -4\pi^2 j^2 d_j \sin(2\pi j t_i)\right)^2 dt_i$.
Next, for $j = 1, 2, 3$ we have:

$$\begin{aligned}
\left(\sum_{j=1}^{J} -4\pi^2 j^2 d_j \sin(2\pi j t_i)\right)^2 &= \left(\sum_{j=1}^{3} -4\pi^2 j^2 d_j \sin(2\pi j t_i)\right)^2 \\
&= (-4\pi^2 d_1 \sin(2\pi t_i) - 4\pi^2 \cdot 4 d_2 \sin(4\pi t_i) - 4\pi^2 \cdot 9 d_3 \sin(6\pi t_i))^2 \\
&= ((-4\pi^2 d_1 \sin(2\pi t_i))^2 + (-4\pi^2 \cdot 4 d_2 \sin(4\pi t_i))^2 + (-4\pi^2 \cdot 9 d_3 \sin(6\pi t_i))^2) \\
&\quad + 2[-4\pi^2 d_1 \sin(2\pi t_i) \cdot 4\pi^2 \cdot 4 d_2 \sin(4\pi t_i)] \\
&\quad + 2[-4\pi^2 d_1 \sin(2\pi t_i) \cdot 4\pi^2 \cdot 9 d_3 \sin(6\pi t_i)] \\
&\quad + 2[-4\pi^2 \cdot 4 d_2 \sin(4\pi t_i) \cdot 4\pi^2 \cdot 9 d_3 \sin(6\pi t_i)] \\
&= \sum_{j=1}^{3} \left(-4\pi^2 j^2 d_j \sin(2\pi j t_i)\right)^2 + 2 \sum_{\substack{j<z \\ j=1}}^{3} \sum_{z=1}^{3} \left(-4\pi^2 j^2 d_j \sin(2\pi j t_i)\right) \left(-4\pi^2 z^2 d_z \sin(2\pi z t_i)\right).
\end{aligned}$$

From this step, we obtain the part three completely for $j = 1, 2, ..., J$ as follows:

$$\begin{aligned}
P(g_{(3)}) &= \lambda \int_0^1 \left(\sum_{j=1}^{J} -4\pi^2 j^2 d_j \sin(2\pi j t_i)\right)^2 dt_i \\
&= \lambda \int_0^1 [\sum_{j=1}^{3} \left(-4\pi^2 j^2 d_j \sin(2\pi j t_i)\right)^2 + \\
&\quad 2 \sum_{\substack{j<z \\ j=1}}^{3} \sum_{z=1}^{3} \left(-4\pi^2 j^2 d_j \sin(2\pi j t_i)\right) \left(-4\pi^2 z^2 d_z \sin(2\pi z t_i)\right)] dt_i. \qquad (26)
\end{aligned}$$

We can write **Equation (26)** as $P(g_{(3)}) = \lambda[E + F]$, where $E = \int_0^1 \left[\sum_{j=1}^{J} \left(-4\pi^2 j^2 d_j \sin(2\pi j t_i)\right)^2\right] dt_i$, and $F = 2\int_0^1 \left[\sum_{\substack{j<z \\ j=1}}^{J} \sum_{z=1}^{J} (-4\pi^2 j^2 d_j \sin(2\pi j t_i)) \cdot (-4\pi^2 z^2 d_z \sin(2\pi z t_i))\right] dt_i$. Hence, we have $E = \int_0^1 \left[\sum_{j=1}^{J} \left(-4\pi^2 j^2 d_j \sin(2\pi j t_i)\right)\right]^2 dt_i = \sum_{j=1}^{J} \left[16\pi^4 j^4 d_j^2 \left(\int_0^1 \sin^2(2\pi j t_i)\, dt_i\right)\right]$. Since $2\sin^2 a = 1 - \cos(2a)$ or $\sin^2 a = \frac{1}{2}(1 - \cos(2a))$, and $\sin^2(2t) = \frac{1}{2}(1 - \cos(4t))$, then we obtain $E = \sum_{j=1}^{J} \left[16\pi^4 j^4 d_j^2 \left(\int_0^1 \frac{1}{2}\, dt_i - \frac{1}{2}\int_0^1 \cos(4\pi j t_i)\, dt_i\right)\right]$. Let $u = 4\pi j t_i$. It gives $du = 4\pi j\, dt_i$, or $\frac{du}{4\pi j} = dt_i$. Hence, we have:

$$\begin{aligned}
E &= \sum_{j=1}^{J} \left[8\pi^4 j^4 d_j^2 \left(\int_0^1 1\, dt_i - \int_0^1 \cos(u) \frac{du}{4\pi j}\right)\right] \\
&= \sum_{j=1}^{J} \left[8\pi^4 j^4 d_j^2 \left(t_i - \frac{1}{4\pi j} \sin(4\pi j t_i)\right)_0^1\right] \\
&= \sum_{j=1}^{J} \left[8\pi^4 j^4 d_j^2 (1 - 0)\right] = \sum_{j=1}^{J} [8\pi^4 j^4 d_j^2].
\end{aligned}$$

Similarly, in this step we have:

$$F = 2\int_0^1 \left[\sum_{\substack{j<z\\ j=1}}^{J}\sum_{z=1}^{J}(-4\pi^2 j^2 d_j \sin(2\pi j t_i))(-4\pi^2 z^2 d_z \sin(2\pi z t_i))\right] dt_i$$

$$= 2\sum_{\substack{j<z\\ j=1}}^{J}\sum_{z=1}^{J}\int_0^1 (-4\pi^2 j^2 d_j \sin(2\pi j t_i))(-4\pi^2 z^2 d_z \sin(2\pi z t_i))\, dt_i$$

$$= 2\sum_{\substack{j<z\\ j=1}}^{J}\sum_{z=1}^{J} 16\pi^4 j^2 z^2 d_j d_z \int_0^1 [\sin(2\pi j t_i)\sin(2\pi z t_i)]\, dt_i.$$

Since $\sin A \sin B = \frac{1}{2}\cos(A-B) + \frac{1}{2}\cos(A+B)$, then we determine the value of the following integral as follows:

$$\begin{aligned}\int_0^1 [\sin(2\pi j t_i)\sin(2\pi z t_i)]\, dt_i &= \int_0^1 \tfrac{1}{2}[\cos(2\pi j - 2\pi z)t_i + \cos(2\pi j + 2\pi z)t_i]\, dt_i\\ &= \tfrac{1}{2}\int_0^1 [\cos(2\pi j - 2\pi z)t_i]\, dt_i + \tfrac{1}{2}\int_0^1 [\cos(2\pi j + 2\pi z)t_i]\, dt_i\\ &= \tfrac{1}{2}\left[\tfrac{1}{2\pi j - 2\pi z}\sin(2\pi j - 2\pi z)t_i\right]_0^1 - \tfrac{1}{2}\left[\tfrac{1}{2\pi j + 2\pi z}\sin(2\pi j + 2\pi z)t_i\right]_0^1\\ &= \tfrac{1}{2}\left[\tfrac{1}{2\pi j - 2\pi z}\sin(2\pi j - 2\pi z)t_i - 0\right] - \tfrac{1}{2}\left[\tfrac{1}{2\pi j + 2\pi z}\sin(2\pi j + 2\pi z)t_i - 0\right] = 0.\end{aligned}$$

Hence, we obtain:
$F = 2\sum_{\substack{j<z\\ j=1}}^{J}\sum_{z=1}^{J} 16\pi^4 j^2 z^2 d_j d_z \int_0^1 [\sin(2\pi j t_i)\sin(2\pi z t_i)]\, dt_i = 2\sum_{\substack{j<z\\ j=1}}^{J}\sum_{z=1}^{J} 16\pi^4 j^2 z^2 d_j d_z \cdot 0 = 0,$ and
$P(g_{(3)}) = \lambda\int_0^1 \left(\sum_{j=1}^{J} -4\pi^2 j^2 d_j \sin(2\pi j t_i)\right)^2 dt_i = \sum_{j=1}^{J}[8\pi^4 j^4 d_j^2]$. Finally, by combining part one, part two, and part three, we obtain the expression of **Equation (22)** as follows:

$$P(g) = \lambda\sum_{j=1}^{J}[8\pi^4 j^4 c_j^2] + \sum_{j=1}^{J}[8\pi^4 j^4 d_j^2]. \tag{27}$$

Also, we can write $p(g)$ given by **Equation (27)** in the notation of matrix as follows:
$P(g) = \lambda\int_0^1 \left(g^{(2)}(t)\right)^2 dx = \lambda\left[\sum_{j=1}^{J}(8\pi^4 j^4 c_j^2) + \sum_{j=1}^{J}(8\pi^4 j^4 d_j^2)\right] = \lambda^*\left[\sum_{j=1}^{J} j^4 c_j^2 + \sum_{j=1}^{J} j^4 d_j^2\right].$ where
$\lambda^* = 8\pi^4\lambda = \lambda[1^4 c_1^2 + 2^4 c_2^2 + \ldots + J^4 c_J^2 + 1^4 d_1^2 + 2^4 d_2^2 + \ldots + J^4 d_J^2] = \lambda[Y^T D Y];$

$$\mathbf{Y}^T = (a_0\ c_1\ c_2\ \cdots\ c_J\ d_1\ d_2\ \cdots\ d_J); \quad \text{and}$$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1^4 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 2^4 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 3^4 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & J^4 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1^4 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 2^4 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 3^4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & J^4 \end{bmatrix}.$$

Based on **Equation (21)** and **Equation (22)**, the PLS optimization can be solved as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{r+1},\, g\in\mathcal{C}(0,1)} \left[n^{-1}\sum_{i=1}^{n}\left(y_i - \mathbf{u}_i^T\boldsymbol{\beta} - g(t_i)\right)^2 + \lambda\int_0^1 \left(g^{(2)}(t_i)\right)^2 dt_i \right] = \min_{\boldsymbol{\beta}\in\mathbb{R}^{r+1},\, g\in\mathcal{C}(0,1)} \{W(\boldsymbol{\gamma})\}.$$

Since $\mathbf{y}^* = y_i - \mathbf{u}_i^T\boldsymbol{\beta},$ then $W(\boldsymbol{\gamma})$ is given by:

$$\begin{aligned} W(\boldsymbol{\gamma}) &= n^{-1}\sum_{i=1}^{n}\left[(\mathbf{y}^* - g(t_i))^T(\mathbf{y}^* - g(t_i))\right] + \lambda\int_0^1 \left(g^{(2)}(t_i)\right)^2 dt_i \\ &= n^{-1}\sum_{i=1}^{n}\Big[(\mathbf{y}^* - (a_0 + \sum_{j=1}^{J'}(c_j\cos(2\pi j t_i) + d_j\sin(2\pi j t_i))))^T \\ &\quad \times (\mathbf{y}^* - (a_0 + \sum_{j=1}^{J'}(c_j\cos(2\pi j t_i) + d_j\sin(2\pi j t_i))))\Big] + \lambda^*[\mathbf{Y}^T\mathbf{D}\mathbf{Y}]. \end{aligned}$$

It is easy to show that $\sum_{i=1}^{n}\left((\mathbf{y}^* - g(t_i))^T(\mathbf{y}^* - g(t_i))\right) = (\mathbf{y}^* - \mathbf{F}\mathbf{y})^T(\mathbf{y}^* - \mathbf{F}\mathbf{y})$ such that we have:

$$\begin{aligned} W(\mathbf{y}) &= n^{-1}\sum_{i=1}^{n}\left((\mathbf{y}^* - g(t_i))^T(\mathbf{y}^* - g(t_i))\right) + \lambda\int_0^1 g^{(2)}(t_i)\,dt_i \\ &= n^{-1}(\mathbf{y}^* - \mathbf{F}\mathbf{y})^T(\mathbf{y}^* - \mathbf{F}\mathbf{y}) + \lambda\,\mathbf{y}^T\mathbf{D}\mathbf{y} \\ &= n^{-1}(\mathbf{y}^{*T}\mathbf{y}^* - \mathbf{y}^{*T}\mathbf{F}\mathbf{y} - \mathbf{y}^T\mathbf{F}^T\mathbf{y}^* + \mathbf{y}^T\mathbf{F}^T\mathbf{F}\mathbf{y}) + \lambda\,\mathbf{y}^T\mathbf{D}\mathbf{y} \\ &= n^{-1}\mathbf{y}^{*T}\mathbf{y}^* - n^{-1}\mathbf{y}^{*T}\mathbf{F}\mathbf{y} - n^{-1}\mathbf{y}^T\mathbf{F}^T\mathbf{y}^* + n^{-1}\mathbf{y}^T\mathbf{F}^T\mathbf{F}\mathbf{y} + \lambda\,\mathbf{y}^T\mathbf{D}\mathbf{y} \\ &= n^{-1}\mathbf{y}^{*T}\mathbf{y}^* - n^{-1}\mathbf{y}^T\mathbf{F}\mathbf{y} - n^{-1}(\mathbf{y}^T\mathbf{F}\mathbf{T}\mathbf{y}^*)^T + \mathbf{y}^T(n^{-1}\mathbf{F}^T\mathbf{F} + \lambda\mathbf{D})\mathbf{y}. \end{aligned}$$

Next, we determine:

$$\begin{aligned} \frac{\partial W(\mathbf{y})}{\partial\mathbf{y}} = 0 &\iff \frac{\partial}{\partial\mathbf{y}}\left\{n^{-1}\mathbf{y}^{*T}\mathbf{y}^* - n^{-1}\mathbf{y}^{*T}\mathbf{F}\mathbf{y} - n^{-1}(\mathbf{y}^T\mathbf{F}^T\mathbf{y}^*)^T + \mathbf{y}^T(n^{-1}\mathbf{F}^T\mathbf{F} + \lambda\mathbf{D})\mathbf{y}\right\} = 0 \\ &\iff -2n^{-1}\mathbf{F}^T\mathbf{y}^* + 2(n^{-1}\mathbf{F}^T\mathbf{F} + \lambda\mathbf{D})\mathbf{y} = 0 \\ &\iff \hat{\mathbf{y}} = (n^{-1}\mathbf{F}^T\mathbf{F} + \lambda\mathbf{D})^{-1}n^{-1}\mathbf{F}^T\mathbf{y}^*. \end{aligned}$$

Hence, we obtain $\hat{\boldsymbol{g}}(t_i)$ as follows:

$$\hat{g}(t_i) = \mathbf{F}^T\hat{\mathbf{y}} = \mathbf{F}(n^{-1}\mathbf{F}^T\mathbf{F} + \lambda\mathbf{D})^{-1}n^{-1}\mathbf{F}^T\mathbf{y}^* = \mathbf{H}(\mathbf{y} - \mathbf{U}\boldsymbol{\beta}) = \mathbf{H}\mathbf{y}^*. \tag{28}$$

where $\mathbf{U} = \begin{pmatrix} 1 & u_{11} & u_{21} & \cdots & u_{p1} \\ 1 & u_{12} & u_{22} & \cdots & u_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_{1n} & u_{2n} & \cdots & u_{pn} \end{pmatrix}$, and $\mathbf{H} = \mathbf{F}(n^{-1}\mathbf{F}^T\mathbf{F} + \lambda\mathbf{D})^{-1}n^{-1}\mathbf{F}^T$.

Next, to find the estimation of parameter $\boldsymbol{\beta}$, we consider the following model:

$$\mathbf{y} = \mathbf{U}\boldsymbol{\beta} - \mathbf{F}\mathbf{y}^T + \boldsymbol{\varepsilon} = \mathbf{U}\boldsymbol{\beta} - \hat{g}(t_i) + \boldsymbol{\varepsilon}.$$

This means that $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{U}\boldsymbol{\beta} - \hat{g}(t_i)$.. Hence, we obtain:

$$\begin{aligned}
\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} &= [\mathbf{y} - \mathbf{U}\boldsymbol{\beta} - \mathbf{H}(\mathbf{y} - \mathbf{U}\boldsymbol{\beta})]^T[\mathbf{y} - \mathbf{U}\boldsymbol{\beta} - \mathbf{H}(\mathbf{y} - \mathbf{U}\boldsymbol{\beta})] \\
&= [\mathbf{y} - \mathbf{U}\boldsymbol{\beta} - \mathbf{H}\mathbf{y} + \mathbf{H}\mathbf{U}\boldsymbol{\beta}]^T[\mathbf{y} - \mathbf{U}\boldsymbol{\beta} - \mathbf{H}\mathbf{y} + \mathbf{H}\mathbf{U}\boldsymbol{\beta}] \\
&= [\mathbf{y}^T(\mathbf{I} - \mathbf{H})^T - \boldsymbol{\beta}^T\mathbf{U}^T\mathbf{H}^T][(\mathbf{I} - \mathbf{H})\mathbf{y} - (\mathbf{I} - \mathbf{H})\mathbf{U}\boldsymbol{\beta}] \\
&= [\mathbf{y}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y} - [\mathbf{y}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{U}\boldsymbol{\beta}] + \\
&\quad - [\boldsymbol{\beta}^T\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y}] + [\boldsymbol{\beta}^T\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{U}\boldsymbol{\beta}] \\
&= \mathbf{y}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y} - [\mathbf{y}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{U}\boldsymbol{\beta}]^T + \\
&\quad - [\boldsymbol{\beta}^T\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y}] + [\boldsymbol{\beta}^T\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{U}\boldsymbol{\beta}] \\
&= \mathbf{y}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y} - \boldsymbol{\beta}^T\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y} + \\
&\quad - \boldsymbol{\beta}^T\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y} + \boldsymbol{\beta}^T\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{U}\boldsymbol{\beta} \\
&= \mathbf{y}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y} + \boldsymbol{\beta}^T\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{U}\boldsymbol{\beta}.
\end{aligned}$$

Hereinafter, we obtain:

$$\frac{\partial(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon})}{\partial\boldsymbol{\beta}} = 0 \iff -2\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y} + 2\mathbf{U}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{U}\boldsymbol{\beta} = 0.$$

$$\iff \hat{\boldsymbol{\beta}}(\mathbf{U}^T\mathbf{V}\mathbf{U})^{-1}\mathbf{U}^T\mathbf{V}\mathbf{y} \quad \text{where} \quad \mathbf{V} = (\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H}). \tag{29}$$

Therefore, based on **Equation (28)** and **Equation (29)**, we obtain the estimation of regression function of the Fourier Series Semiparametric Regression (FSSR) model as follows:

$$\hat{g}(t_l) = \mathbf{H}(\mathbf{y} - \mathbf{U}\hat{\boldsymbol{\beta}}) = \mathbf{H}\big[\mathbf{I} - \mathbf{H}\mathbf{U}(\mathbf{U}^T\mathbf{V}\mathbf{U})^{-1}\mathbf{U}^T\mathbf{V}\big]\mathbf{y}. \tag{30}$$

Hence, the completely estimated Fourier series semiparametric regression (FSSR) model is:

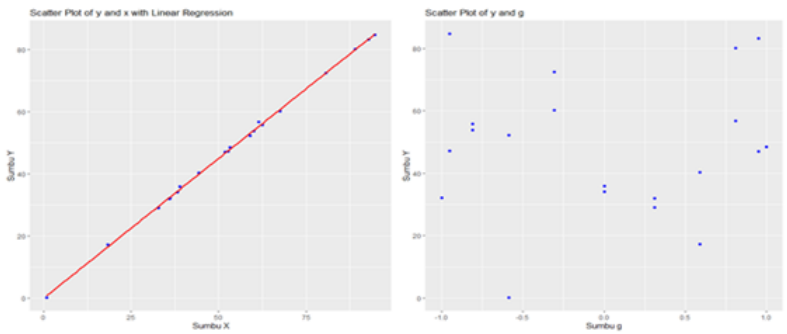
$$\hat{y}(\mathbf{u}, t) = \mathbf{U}(\mathbf{U}^\top\mathbf{V}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{V}\mathbf{y} + \mathbf{H}\big[\mathbf{I} - \mathbf{H}\mathbf{U}(\mathbf{U}^\top\mathbf{V}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{V}\big]\mathbf{y} = \mathbf{U}\hat{\boldsymbol{\beta}} + \hat{\mathbf{g}}(t_i). \tag{31}$$

Thus, the **Theorem 2** is proved.

3.2. Simulation Study

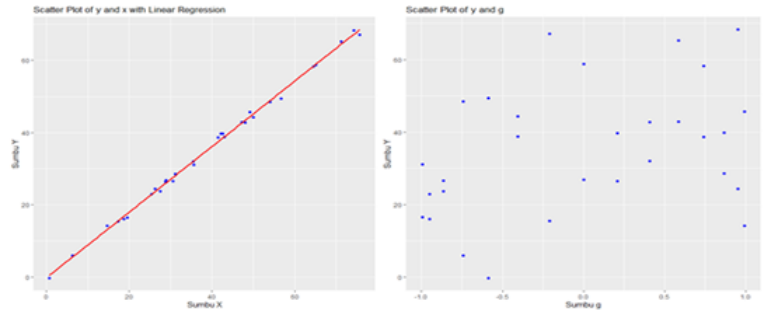
In this simulation study, we use four different samples sizes that represent small sample $(n = 20)$, moderate sample $(n = 40$ and $n = 30)$, and large sample $(n = 100)$, and with Fourier coefficients $k = 1, 2, \ldots, 10$ through the scenario as follows: (1). Given samples size, $n = 20, 30, 40, 100$; (2). Set time values (t) equals to the number of sample size (n); (3). Input time values (t) into equation: $g(t) = \sin(2\pi t)$; (4). Generate values of u from Normal distribution: $u < -rnorm(n, 50, 25)$; The generation of these values is designed such that the variable u contains n observations whose distribution follows the characteristics of the original data, namely a mean of approximately 50 and a standard deviation of 25. (5). Generate values of error (ϵ) from Normal distribution: $< -rnorm(n, 0, sqrt(0.5))$; (6). Determine values of y using equation: $y = \beta u + g(t) +$ where $\beta = 0.9$; The selection of the parameter value β is intended to ensure that, when the relationship between y and u is visualized through a plot, the resulting pattern exhibits a tendency toward linearity. (7). Plot y versus u, and plot y versus $g(t)$; (8). Input values of lower and upper bounds of lambda (λ) with ranges and increments as follows: (8a). $= [0.0001, 0.01]$ and increment equals to 0.0001; (8b). $= [0.01, 1]$ and increment equals to 0.01; (8c). $= [1, 100]$ and increment equals to 1; (9). Input Fourier coefficient values, $j = 1, 2, \ldots, 10$, which minimize value of GCV; (10). Determine MAPE (Mean Average Percentage Error) values for each Fourier coefficient.

Next, based on the scenario above, we obtain scatter plots of observations y versus u, and observations y versus values of function g for $n = 100, n = 40, n = 30$, and $n = 20$ which are given in Figures 1-4, respectively.



(a) (b)

Figure 1. Scatter Plots for $n = 20$ of y versus u (a), and y versus g (b).



(a) (b)

Figure 2. Scatter Plots for $n = 30$ of y versus u (a), and y versus g (b).

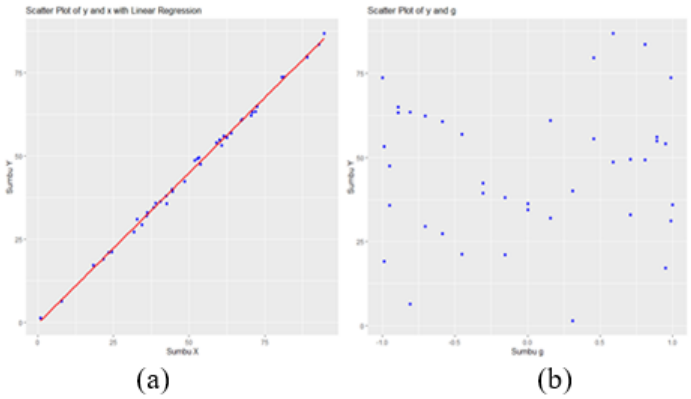


Figure 3. Scatter Plots for $n = 40$ of y versus u (a), and y versus g (b).

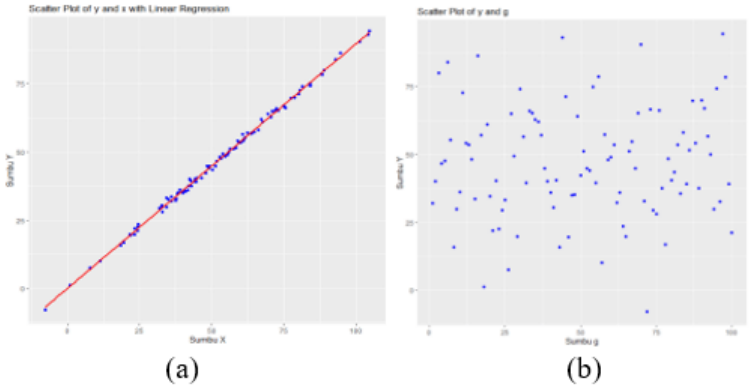
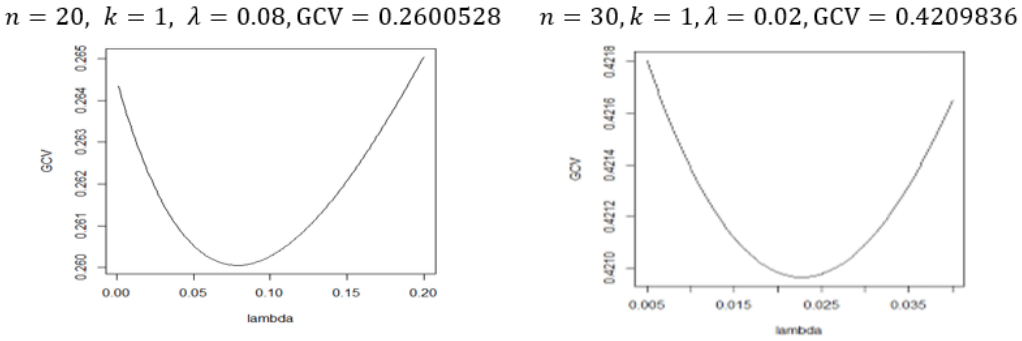


Figure 4. Scatter Plots for $n = 100$ of y versus u (a), and y versus g (b).

Figure 1–4 show that all scatter plots y versus u for $n = 20, 30, 40, 100$ provide a linear curve pattern, and all scatter plots y versus g for $n = 20, 30, 40, 100$ do not show any particular curve pattern. This means that to analyze the data we have, it is suitable to use a semiparametric regression model approach. Further, we obtain plots of GCV versus bandwidth which is given in Figure 5.



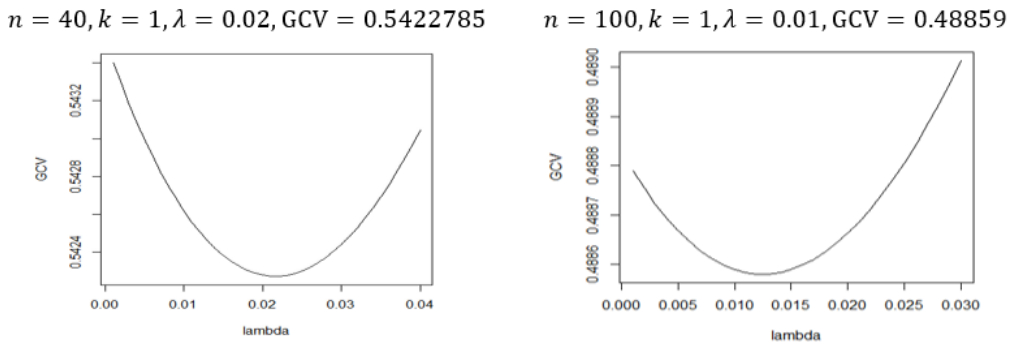


Figure 5. Plots of GCV versus Lambda for $n = 20, 30, 40, 100$.

Figure 5 shows minimum values of GCV and optimal lambda for samples sizes $n = 20, 30, 40, 100$. Next, in Table 1 we present the estimation results of the simulation study for samples sizes $n = 20, 30, 40, 100$ including the number of Fourier coefficients (k), values of optimal lambda (λ), minimum values of GCV, values of parameters $(\hat{\beta}_0, \hat{\beta}_1, \hat{\alpha}_0, \hat{c}_0, \hat{c}_1)$, and values of MAPE.

Table 1. Estimation Results of the Simulation Study for $n = 20, 30, 40, 100$.

Estimation	**Samples Size**			
	$n = 20$	$n = 30$	$n = 40$	$n = 100$
k	1	1	1	1
Optimal λ	0.08	0.02	0.02	0.01
$\text{GCV}_{\min}$	0.2600528	0.4209836	0.5422785	0.48859
$\hat{\beta}_0$	7.4187466e-18	-3.210626e-18	-3.893704e-18	-1.947083e-18
$\hat{\beta}_1$	-180.8980365	-190.8947903	-180.9021829	-170.8985376
$\hat{\alpha}_0$	0.06888385	0.3295847	-0.1163614	0.00037622
$\hat{c}_0$	-0.08215611	0.04767274	-0.02987295	-0.02677371
$\hat{c}_1$	0.1667841	0.3506139	0.3618541	0.2911945
MAPE	9.706166	7.107682	1.9989046	1.66728

Based on estimation results of the simulation study for $n = 20, 30, 40, 100$ given in Table 1, and by considering Equations (8), (21), and Equations (31), we obtain the estimated Fourier series semiparametric regression (FSSR) model using PLS smoothing technique that is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 u_{1i} + \hat{\beta}_2 u_{2i} + \ldots + \hat{\beta}_p u_{pi} + \hat{\alpha}_0 + \sum_{j=1}^{J} \left[\hat{c}_j \cos(2\pi j t_i) + \hat{d}_j \sin(2\pi j t_i) \right]. \tag{32}$$

for $i = 1, 2, \ldots, n$. Hence, for $n = 20, 30, 40, 100$, we obtain some results as follows:

(a). For $n = 20$, we have Fourier coefficient $k = 1$, minimum GCV of 0.2600528, and the optimal lambda of 0.08. Hence, by using Equation (32), we obtain:

$$\begin{aligned} \hat{y}_i = {} & 7.418746\text{e}{-18} - 180.8980365\, u_{1i} + 0.06888385 \\ & - 0.08215611 \cos(2\pi t_i) + 0.1667841 \sin(2\pi t_i), \quad i = 1, 2, \ldots, 20. \end{aligned} \tag{33}$$

(b). For $n = 30$, we have Fourier coefficient $k = 1$, minimum GCV of 0.4209836, and the optimal lambda of 0.02. Hence, by using Equation (32), we obtain:

$$\begin{aligned} \hat{y}_i = {} & -3.210626\text{e}{-18} - 190.8947903\, u_{1i} + 0.3295847 \\ & + 0.04767274 \cos(2\pi t_i) + 0.3506139 \sin(2\pi t_i), \quad i = 1, 2, \ldots, 30. \end{aligned} \tag{34}$$

(c). For $n = 40$, we have Fourier coefficient $k = 1$, minimum GCV of 0.5422785, and the optimal lambda of 0.02. Hence, by using Equation (32), we obtain:

$$\hat{y}_i = -3.893704\text{e}{-18} - 18.9021829\, u_{1i} - 0.1163614 - 0.02987295\cos(2\pi t_i) + 0.3618541\sin(2\pi t_i), \quad i = 1, 2, \ldots, 40. \tag{35}$$

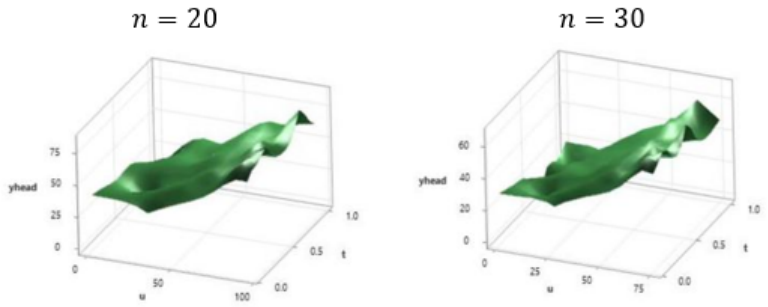
(d). For $n = 100$, we have Fourier coefficient $k = 1$, minimum GCV of 0.48859, and the optimal lambda of 0.01. Hence, by using Equation (32), we obtain:

$$\hat{y}_i = -1.947083\text{e}{-18} - 17.08985376\, u_{1i} + 0.0003762195 - 0.02677371\cos(2\pi t_i) + 0.2911945\sin(2\pi t_i), \quad i = 1, 2, \ldots, n. \tag{36}$$

Table 1 also shows the MAPE values of estimation results for the Fourier Semiparametric Regression (FSSR) models using PLS smoothing technique on simulation study with sample sizes $n = 20, 30, 40, 100$. The results show that the MAPE values for $n = 20, 30, 40, 100$ are less than 10, this means that the estimation results of the FSSR models using PLS smoothing technique has highly accuracy criteria [34]. It can also be seen that the larger the sample size, the smaller the MAPE value. This means that the estimations of the FSSR models using PLS smoothing technique have increasingly accurate criteria.

Hereinafter, based on Equations (33)–(36) that are the estimation results of the FSSR models using PLS smoothing technique for samples sizes $n = 20, 30, 40, 100$, we obtain the surface plots of estimation results of the FSSR models for samples sizes $n = 20, 30, 40, 100$ that are given in Figure 6.

From Figure 6 we can see that surface plots of all the estimated FSSR models for all samples sizes simulated $(i.e., n = 20, 30, 40, 100)$ show the existence of fluctuating patterns and trend patterns (in this case it is an upward trend) that is the value of the estimated response variable $(\hat{y})$, which in Figure 6 it is expressed as yhead, tends to increase as the value of the predictor variables u and t increases. This means that all the illustrations produced from this simulation study theoretically support the concept of the estimation method proposed for estimating FSSR model, namely, that penalized least square must be chosen to avoid over-fitting effect, and the use of the Fourier series is also appropriate, namely, this is indicated by the presence of fluctuating patterns over time which tends to have trend and seasonal patterns.



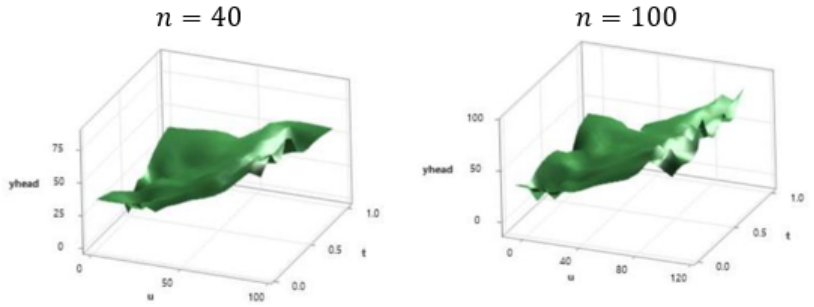


Figure 6. Surface Plots of Estimation Results of the FSSR Models Using PLS Smoothing Technique for Sample Sizes $n = 20, 30, 40, 100$.

3.3. Comparison of the FSSR Model and the ARIMAX Model Using Real Data

The data used in this study consists of 60 data points, which are divided into two groups, namely the in-sample and out-sample sets. The data were obtained from NASA POWER through the website https://power.larc.nasa.gov/data-access-viewer/, with the data collection location situated in Sragen Regency, Central Java Province, Indonesia, covering the period from January 1, 2025 to March 2, 2025. For the in-sample distribution, the data are divided into 50 data points, while the out-sample distribution consists of 10 data points. The first step in this study is to examine the relationship or correlation between earth surface temperature, denoted as EST, and relative humidity at 2 meters, denoted as RH2M, within the in-sample data. Correlation statistical analysis aims to determine the strength and direction of the relationship between these two variables, namely earth surface temperature and relative humidity at 2 meters. The Pearson correlation coefficient ranges from -1 to 1, where a positive value indicates a positive correlation, and a negative value indicates a negative correlation. The following Figure 7 and Table 2 describe the correlation between earth surface temperature and relative humidity at 2 meters.

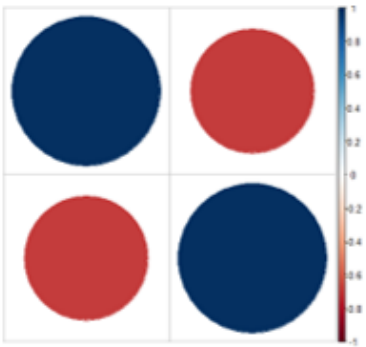


Figure 7. Correlation Plot of Earth Surface Temperature and Relative Humidity at 2 Meters.

Table 2. Correlation Matrix of Earth Surface Temperature and Relative Humidity at 2 Meters

	EST	**RH2M**
EST	1.00000	-0.69739
RH2M	-0.69739	1.00000

In Table 2, the correlation value of -0.69739 presented in the correlation matrix shows a correlation between Earth Surface Temperature (EST) and Relative Humidity at 2 Meters (RH2M). This negative correlation suggests an inverse relationship between Earth Surface Temperature and Relative Humidity, meaning that as Earth Surface Temperature increases, the Relative Humidity decreases, and vice versa.

The next step is to create a time series scatter plot for the in-sample scenario, plotting Earth Surface Temperature against Relative Humidity to examine the data distribution. The following Figure 8 is the scatter plot of Earth Surface Temperature versus Relative Humidity.

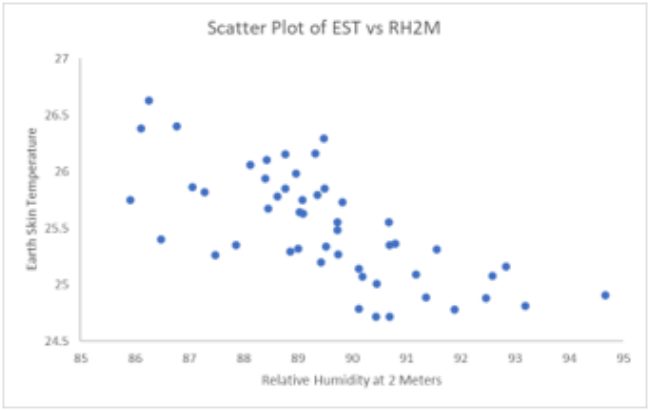


Figure 8. Scatter Plot of Earth Surface Temperature versus Relative Humidity.

Based on the Figure 8, it shows that the in-sample data follows a linear pattern with a downward trend, indicating a negative parametric relationship between Earth Surface Temperature (EST) and Relative Humidity at 2 Meters (RH2M). To reinforce this assumption, a linearity test was conducted using a linear regression model, and the obtained results are presented in Table 3.

Table 3. Significance Test of the Linear Model Parameters

	Estimate	**Std. Error**	**t value**	**Pr(¿|t|)**
(Intercept)	41.67236	2.39854	17.374	$<$ 2e-16
RH2M	-0.18057	0.02678	-6.742	1.83e-08

As shown in Table 3, the coefficient value is less than $\alpha = 0.05$, indicating that the linear model coefficient is statistically significant. Therefore, it can be concluded that Earth Surface Temperature (EST) and Relative Humidity at 2 Meters (RH2M) have a linear relationship. Next, a scatter plot of Earth Surface Temperature and time is created, as shown in the following Figure 9.

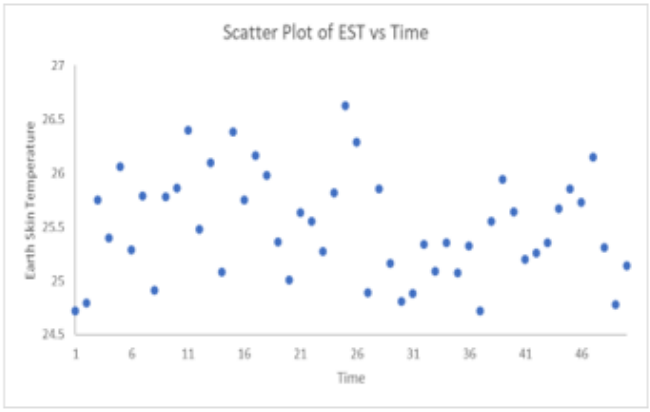


Figure 9. Scatter Plot of Earth Surface Temperature and Time.

Based on Figure 9, the scatter plot between Earth Surface Temperature and time does not show any specific pattern. This indicates that a nonparametric regression approach can be applied. Therefore, the functional relationship between Earth Surface Temperature and Relative Humidity is linear, while the functional relationship between Earth Surface Temperature and time does not form any particular pattern. Before applying semiparametric regression, this study will also employ the ARIMAX method between Earth Surface Temperature and Relative Humidity to compare the prediction results of both methods.

The first step in applying the ARIMAX method is to test whether Earth Surface Temperature requires a differencing process. The test was conducted using the Augmented Dickey–Fuller (ADF) method, which produced an ADF Statistic of –5.6914 with a p-value of 0.000. The p-value, which is far below the 0.05 significance level, indicates that the Earth Surface Temperature data are stationary in mean; therefore, no differencing is required to stabilize the data. The results of this test are presented in Table 4.

Table 4. Augmented Dickey-Fuller (ADF) Test Result

	ADF Statistic	**p-value**
Value	-5.691412771	0.000

The next step is to construct several tentative ARIMAX models using various combinations of ARIMA orders while including Relative Humidity as an exogenous variable. Each candidate model is then evaluated based on the Akaike Information Criterion (AIC), where the model with the lowest AIC value is selected as the best model because it provides the optimal balance between model complexity and goodness-of-fit. A summary of the ARIMAX model selection based on AIC values is presented in Table 5.

Table 5. AIC Comparison for ARIMAX Model Orders

Order (p, d, q)	**AIC**
(1, 0, 0)	35.297
(0, 0, 1)	36.631
(1, 0, 1)	37.608
(0, 0, 0)	41.427

Based on the model selection results in Table 5, the best ARIMAX model is obtained at the order (p, d, q) = (1, 0, 0) with an AIC value of 35.297, which is the smallest among all candidate models. Therefore, this model is selected for forecasting both in-sample and out-of-sample data in order to compare its predictive performance with that of the semiparametric regression approach.

After selecting the ARIMAX model, assumption tests are conducted to ensure that the model is appropriate for forecasting purposes. The evaluation covers three main aspects: residual autocorrelation, residual normality, and the presence of heteroskedasticity. The Ljung–Box test is used to assess whether the residuals are free from autocorrelation, the Shapiro–Wilk test is used to verify whether the residuals follow a normal distribution, and the ARCH LM test is performed to detect the presence of heteroskedasticity in the residuals. The results of these assumption tests are presented in Table 6.

Table 6. Assumption Tests for Model Diagnostics

Assumption Test	**p-value**
Ljung–Box Test	0.8389
Shapiro–Wilk Test	0.3422
ARCH LM Test	0.8196

Based on the assumption test results in Table 6, the Ljung–Box Test p-value of 0.8389 indicates that there is no autocorrelation in the residuals, as the value is greater than the 0.05 significance level. The Shapiro–Wilk Test

p-value of 0.3422 shows that the residuals follow a normal distribution, meaning that the normality assumption is satisfied. The ARCH LM Test produces a p-value of 0.8196, which indicates that there is no evidence of heteroskedasticity in the residuals.

Returning to the semiparametric regression approach, its application in this case involves determining the minimum Generalized Cross Validation (GCV) value using Fourier series estimation. In this study, the Fourier series coefficient limit is set to 10. Figure 10 presents the GCV plot for each Fourier coefficient based on the in-sample data.

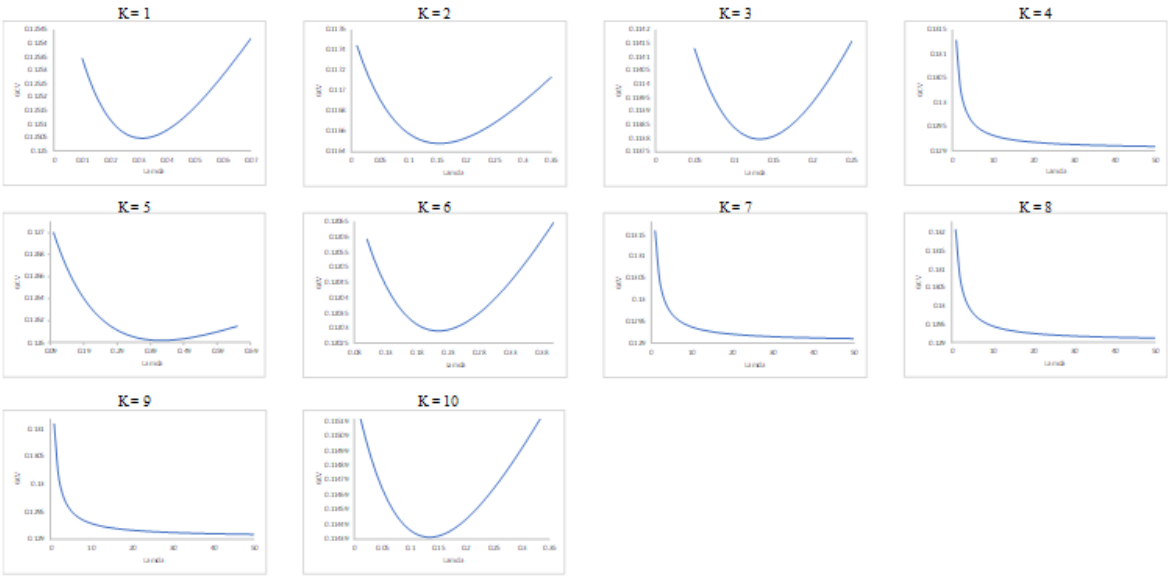


Figure 10. GCV Plots for Each Fourier Coefficient.

Based on Figure 10, it can be observed that each Fourier coefficient has a minimum value of GCV along with its corresponding lambda value. The following Table 7 presents the GCV and lambda values for each Fourier coefficient.

Table 7. GCV Value for $K = 1, 2, \ldots, 10$.

K	GCV	Lambda
1	0.1250457	0.031
2	0.1164853	0.15
3	0.113797	0.13
4	0.1290923	49
5	0.1260202	0.42
6	0.1202909	0.21
7	0.1291032	49
8	0.1291203	49
9	0.1290847	49
10	0.1144038	0.14

Based on the Figure 10 and Table 7, the GCV value for each Fourier coefficient in the in-sample data reaches its minimum at the third Fourier coefficient. It can be concluded that the best semiparametric model is obtained at the third Fourier coefficient, with a minimum GCV value of 0.113797 and a lambda of 0.13. The best FSSR model

based on the minimum GCV can be expressed as follows:

$$\begin{aligned} y_i = {} & 4.578279\text{e}{-16} + (-0.184205)\, u_{1i} + 41.97897 + 0.009824081 \cos(2\pi t_i) \\ & + (-0.003549639) \sin(2\pi t_i) + (-0.001217795) \cos(4\pi t_i) + 0.0296664 \sin(4\pi t_i) \\ & + 0.004221657 \cos(6\pi t_i) + (-0.0002334817) \sin(6\pi t_i) \end{aligned}$$

After obtaining the best semiparametric regression and ARIMAX models, the next step is to evaluate their accuracy and performance. This evaluation is carried out using the Fourier series estimator with the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). MAPE measures the percentage of prediction error, where a lower value indicates higher accuracy [34]. This analysis aims to ensure that the resulting model effectively captures the relationships among the variables. Plots of the actual and estimated values using the best Fourier coefficients and model orders are presented in Figure 11.

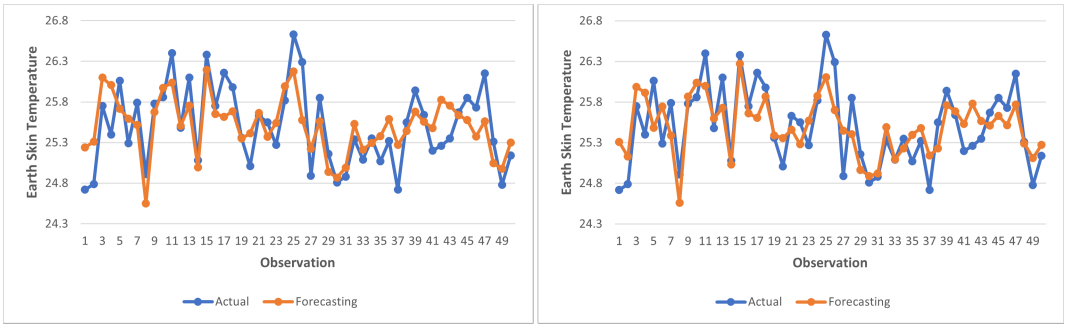


Figure 11. (a). FSSR, (b). ARIMAX, Plots of the Actual Values (blue) and Estimating Values (red) Using the Best Fourier Coefficient and orders.

Based on Figure 11, it can be observed that the model demonstrates strong performance. For the in-sample data, both the model with the best Fourier coefficients and the ARIMAX model with the optimal order are able to explain the data variability with a low error rate. Subsequently, both models are used to evaluate estimation performance by comparing the predicted data with the out-sample data. The plots comparing the estimated and actual values of the best FSSR and ARIMAX models are presented in Figure 12, and a quantitative evaluation of both models' performance is conducted by calculating the RMSE and MAPE values, which are presented in Table 8.

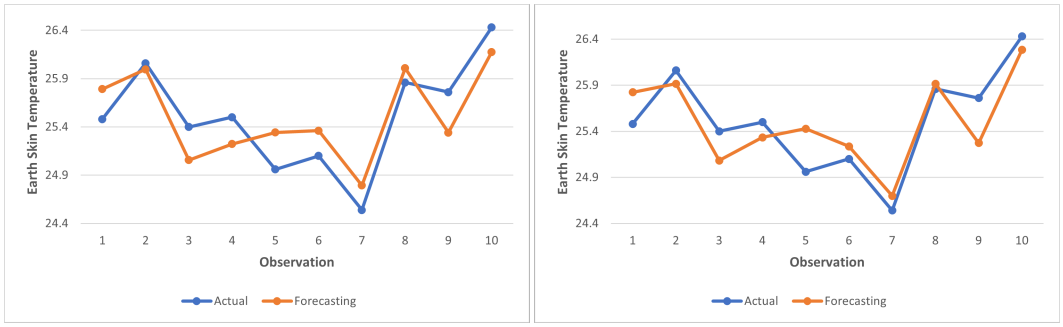


Figure 12. (a). FSSR, (b). ARIMAX, Plots of the Actual Values (blue) and Estimating Values (red) Using the Best Fourier Coefficient and orders.

Table 8. Performance Comparison Between FSSR and ARIMAX Models

	FSSR	**ARIMAX**
RMSE	0.2816	0.2889
MAPE	1.0684	0.9539

Based on Figure 12 and Table 8, it can be observed that the forecasting results of both models do not differ significantly from the actual values. The RMSE values show that the FSSR model has an error level of 0.2816, while the ARIMAX model yields a slightly higher RMSE of 0.2889. The lower RMSE of the FSSR model indicates that it produces smaller and more consistent prediction errors, suggesting that FSSR does not experience overfitting and is able to maintain its performance on new data. In terms of error percentage, the FSSR model obtains a MAPE of 1.0684%, while ARIMAX achieves a MAPE of 0.9539%. Although ARIMAX shows a slightly lower MAPE, the stability of the RMSE in the FSSR model makes it superior overall, as it provides more stable estimations, effectively mitigates potential overfitting, and offers better flexibility in capturing the relationships among Earth Surface Temperature, Relative Humidity, and time.

4. Conclusion

The estimated regression function of the Fourier Series Semiparametric Regression (FSSR) model using the Penalized Least Square (PLS) smoothing technique is a combination of the estimated parametric component and the estimated nonparametric component in the FSSR model. In addition, both the estimated parameters in the parametric component and the estimated function in the nonparametric component are linear with respect to the observed data. This implies that the estimated regression function of the FSSR model with the PLS smoothing technique is also linear in the observations, making the overall estimation process of the FSSR model linear in the data. Furthermore, the theoretical results of this study can be utilized for advancing statistical inference theory and for analyzing time series data that exhibit mixed patterns where part of the data follows linear, quadratic, or cubic trends, while the remaining portion shows fluctuating patterns over time that tend to display trend and seasonal components. In addition, the simulation study results demonstrate that the FSSR model estimated using the PLS smoothing technique satisfies high accuracy criteria. The implementation of this model shows that FSSR is capable of effectively explaining data variability in a semiparametric framework and producing accurate predictions. This is reflected in the MAPE value of 1.068%, which indicates an excellent level of predictive accuracy. Moreover, the FSSR model yields an RMSE value of 0.2816, which is lower than that of the ARIMAX model (0.2889), indicating that FSSR produces smaller prediction errors. This difference also demonstrates that FSSR is superior, as its stable RMSE values for both in-sample and out-of-sample data suggest no indication of overfitting. With consistent performance and more stable estimation results compared to ARIMAX, the FSSR model can be considered valid and reliable for prediction purposes. Therefore, this approach has strong potential for application to rainfall data or other environmental variables in future analysis and forecasting efforts.

Authors Contribution

All authors have contributed to this research article, namely Ihsan Fathoni Amri: Conceptualization, Methodology, Software, Writing–Original Draft Preparation, Validation; Nur Chamidah: Conceptualization, Methodology, Supervision, Validation, Software, Writing–Review & Editing; Toha Saifudin: Conceptualization, Methodology, Software, Supervision, Validation; Budi Lestari: Methodology, Supervision, Validation, Writing–Review & Editing; Dursun Aydin: Supervision, Validation, Writing–Review & Editing; and Febrian Hikmah Nur Rohim: Writing. All authors have read and approved the published version of the manuscript.

Acknowledgement

I would like to thank all the authors for their contributions, and I also extend my gratitude to Universitas Muhammadiyah Semarang (UNIMUS) for the support and research facilities provided.

REFERENCES

1. N. Nurhaswinda *et al.*, *Analisis regresi linier sederhana dan penerapannya*, Jurnal Cahaya Nusantara, vol. 1, no. 2, pp. 69–78, 2025.
2. B. Lestari, N. Chamidah, D. Aydin, and E. Yilmaz, *Reproducing kernel Hilbert space approach to multiresponse smoothing spline regression function*, Symmetry (Basel), vol. 14, no. 11, p. 2227, 2022.
3. B. Lestari, N. Chamidah, I. N. Budiantara, and D. Aydin, *Determining confidence interval and asymptotic distribution for parameters of multiresponse semiparametric regression model using smoothing spline estimator*, Journal of King Saud University–Science, vol. 35, no. 5, p. 102664, 2023.
4. C. Kurniawan, *Analisis Data Hubungan Antar Variabel Sebagai Metode Alternatif Penentukan Hubungan Kausalitas*, Sinteks: Jurnal Teknik, vol. 5, no. 2, 2016.
5. A. Anandari, *Analisis Regresi Deret Fourier: Aplikasi Data Curah Hujan*, CV Jejak (Jejak Publisher), 2023.
6. N. Ravishanker, Z. Chi, and D. K. Dey, *A first course in linear model theory*, Chapman and Hall/CRC, 2021.
7. R. Hidayat, Y. Yuliani, and M. Sam, *Model regresi nonparametrik dengan pendekatan spline truncated*, in *Prosiding Seminar Nasional*, 2018.
8. K. Sya'baniah, *Regresi semiparametrik kernel dengan fungsi Epanechnikov untuk memodelkan inflasi di Indonesia (Tesis)*, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Available: https://etheses.uin-malang.ac.id/64981/, 2024.
9. P. Kafi, R. Eyvazloo, and M. Asima, *Performance of semi-parametric asset pricing model in Tehran stock exchange*, Financial Research Journal, 24(3), 375–390, 2022.
10. D. Aydına, Ö. İ. Günerib, and A. Fitc, *Choice of bandwidth for nonparametric regression models using kernel smoothing: A simulation study*, International Journal of Sciences: Basic and Applied Research (IJSBAR), vol. 26, no. 1, pp. 47–61, 2016.
11. Z. Rahasia, R. Resmawan, and D. R. Isa, *Pemodelan Data Time Series dengan Pendekatan Regresi Nonparametrik B-Spline*, AKSIOMA: Jurnal Matematika dan Pendidikan Matematika, vol. 11, no. 1, pp. 9–16, 2020.
12. L. P. S. Pratiwi and I. M. P. P. Wijaya, *Pemodelan Produk Domestik Regional Bruto di Indonesia dengan Regresi Nonparametrik Menggunakan Estimator Spline*, Jurnal Statistika dan Aplikasinya, vol. 6, no. 2, pp. 223–233, 2022.
13. L. R. Cheruiyot, *Local linear regression estimator on the boundary correction in nonparametric regression estimation*, Journal of Statistical Theory and Applications, vol. 19, no. 3, pp. 460–471, 2020.
14. W. A. Anam, A. Massaid, N. A. Amesya, and N. Chamidah, *Modeling of diabetes mellitus risk based on consumption of salt, sugar, and fat factors using local linear estimator*, in *AIP Conference Proceedings*, AIP Publishing LLC, 2020, p. 030009.
15. W. S. Cleveland, E. Grosse, and W. M. Shyu, *Local regression models*, in *Statistical Models in S*, Routledge, 2017, pp. 309–376.
16. N. Chamidah, Y. S. Yonani, E. Ana, and B. Lestari, *Identification the number of Mycobacterium tuberculosis based on sputum image using local linear estimator*, Bulletin of Electrical Engineering and Informatics, vol. 9, no. 5, pp. 2109–2116, 2020.
17. A. Tohari, N. Chamidah, F. Fatmawati, and B. Lestari, *Modelling the number of HIV and AIDS cases in East Java using biresponse multipredictor negative binomial regression based on local linear estimator*, Communications in Mathematical Biology and Neuroscience (CMBN), vol. 2021, no. 73, pp. 1–17, 2021.
18. J. Fan, *Local polynomial modelling and its applications: Monographs on Statistics and Applied Probability 66*, Routledge, 2018.
19. D. I. Purnama, *A comparison between nonparametric approach: Smoothing spline and B-spline to analyze the total of train passengers in Sumatra Island*, EKSAKTA: Journal of Sciences and Data Analysis, pp. 73–80, 2020.
20. A. Araveeporn, *The estimating parameter and number of knots for nonparametric regression methods in modelling time series data*, Modelling, 5(4), 2024.
21. I. N. Fatmawati, B. L. Budiantara, and B. Lestari, *Comparison of smoothing and truncated spline estimators in estimating blood pressure models*, International Journal of Innovation, Creativity and Change, vol. 5, no. 3, pp. 1177–1199, 2019.
22. A. Iriany and A. A. R. Fernandes, *Hybrid Fourier series and smoothing spline path non-parametrics estimation model*, Frontiers in Applied Mathematics and Statistics, vol. 8, p. 1045098, 2023.
23. N. P. A. M. Mariati, I. N. Budiantara, and V. Ratnasari, *Combination estimation of smoothing spline and Fourier series in nonparametric regression*, Journal of Mathematics, vol. 2020, no. 1, p. 4712531, 2020.
24. B. Lestari and I. N. Budiantara, *Spline estimator and its asymptotic properties in multiresponse nonparametric regression model*, Songklanakarin Journal of Science and Technology, vol. 42, no. 3, pp. 533–548, 2020.
25. J. L. Kirkby, Á. Leitao, and D. Nguyen, *Spline local basis methods for nonparametric density estimation*, Statistic Surveys, vol. 17, pp. 75–118, 2023.
26. I. Sriliana, I. N. Budiantara, and V. Ratnasari, *A truncated spline and local linear mixed estimator in nonparametric regression for longitudinal data and its application*, Symmetry (Basel), vol. 14, no. 12, p. 2687, 2022.
27. D. A. Widyastuti, A. A. R. Fernandes, and H. Pramoedyo, *Spline estimation method in nonparametric regression using truncated spline approach*, in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 012027.
28. R. Pane and A. T. Ampa, *Estimation of heteroskedasticity semiparametric regression curve using Fourier series approach*, Journal of Research in Mathematics Trends and Technology, 2(1), 14–20, 2020.
29. S. Sifriyani, J. P. Sitinjak, and A. T. R. Dani, *Penerapan model regresi semiparametrik deret Fourier untuk mengidentifikasi faktor penentu angka harapan hidup dalam konteks SDGs 3*, Jurnal Gaussian, 14(2), 302–313, 2025.

30. A. Prahutama and T. W. Utami, *Modelling Fourier regression for time series data: A case study of modelling inflation in the food sector in Indonesia*, in *Journal of Physics: Conference Series*, IOP Publishing, 2018, p. 012067.
31. M. F. F. Mardianto and H. U. Gunardi, *An analysis about Fourier series estimator in nonparametric regression for longitudinal data*, Mathematics and Statistics, vol. 9, no. 4, pp. 501–510, 2021.
32. F. Osmani, E. Hajizadeh, and P. Mansouri, *Kernel and regression spline smoothing techniques to estimate coefficient in rates model and its application in psoriasis*, Medical Journal of the Islamic Republic of Iran, vol. 33, p. 90, 2019.
33. M. Maharani and D. R. S. Saputro, *Generalized cross validation (GCV) in smoothing spline nonparametric regression models*, Journal of Physics: Conference Series, Vol. 1808, No. 1, p. 012053, IOP Publishing, 2021.
34. M. F. F. Mardianto, E. Tjahjono, and M. Rifada, *Semiparametric regression based on three forms of trigonometric function in Fourier series estimator*, Journal of Physics: Conference Series, 1277(1), Article 012052, 2019.
35. W. Wibowo, S. Haryatmi, and N. Budiantara, *Penalized Least Squares for Semiparametric Regression*, International Journal of Academic Research, vol. 4, no. 6, 2012.