Article

Modeling the Level of Drinking Water Clarity in Surabaya City Drinking Water Regional Company Using Combined Estimation of Multivariable Fourier Series and Kernel

Andi Tenri Ampa [1,2], I Nyoman Budiantara [1,*] and Ismaini Zain [1]

1   Department of Statistic, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
2   Department of Statistic, Halu Oleo University, Kendari 93132, Indonesia
*   Correspondence: i_nyoman_b@statistika.its.ac.id

Abstract: The purpose of this study is to propose an appropriate model to predict chemical composition during water purification at the Regional Water Company (PDAM) Surabaya, in order to achieve proper drinking water standards. Drinking water treatment is very expensive, so the model serves as a basis for determining the composition of chemicals used in the water purification process at PDAM Surabaya. This study examines a model of the relationship between the level of clarity of drinking water and the composition of the chemicals used. The government can obtain important benefits from the forecasting model to formulate policies for the company. One of the objectives of developing the estimation method involved in this research is to efficiently determine the exact chemical composition resulting from the water purification process, which will inform the financing and control of water quality. We used a multivariable linear approach for some parametric components, a multivariable Fourier Series approach for some nonparametric components, and a multivariable Kernel approach for semiparametric regression. Using the penalized least square (PLS) approach, a mixed estimator of the Fourier and Kernel Series was obtained with semiparametric regression. The smoothing parameters were selected using a common cross-validation technique (GCV). The performance of this technique was evaluated using the Gaussian Kernel and Fourier Series with data trends in the drinking water clarity level obtained from PDAM Surabaya. The findings showed that this technique performed well, so we recommend that the government conduct an in-depth analysis to determine correct chemical composition so that the cost of water treatment can be minimized.

Keywords: PLS; GCV; Kernel; Fourier Series; smoothing parameters

1. Introduction

In the adult human body, 70% of body weight is in the form of liquid. Therefore, drinking water is a nutritional element that is as important as carbohydrates, proteins, fats, and vitamins. Consuming good and sufficient mineral water can help the digestive process, regulate metabolism, regulate food substances in the body, and ensure body balance, provided that the quality of drinking water is assessed in terms of clarity [1]. To obtain drinking water that is fit for consumption involves costly processing. Careful planning is needed so that costs can be minimized. One approach that can be used is the application of a semiparametric regression model, as proposed in this paper.

Semiparametric regression is a regression analysis technique in addition to parametric and nonparametric regression. Both parametric and nonparametric features are integrated in semiparametric regression. The parametric features used in semiparametric regression include Fourier Series and Kernel. These three different estimators are combined in the estimate. Linear regression is the easiest and most efficient estimator compared to other nonparametric regression methods. Kernel is useful for use with un patterned data [2] and has a relatively faster convergence speed than polynomial estimators, Fourier Series, or splines [3]. Many nonparametric and semiparametric regression estimators have been

developed. For data with variable patterns independent of node location, nonparametric regression without mixed estimators, such as spline smoothing [4], Penalized Spline [5]. B-Spline [6], Weighted Partial Spline [7]. For data with changing patterns at certain sub-intervals, truncated splines are used [8,9].

Semiparametric regression estimators have been developed by several authors, including [10,11]. The authors of [12–14] used the Kernel technique and the authors of [13,14] dan [15,16] developed an approach involving Fourier Series.

The assumption used by researchers when designing nonparametric or semiparametric regression models is that each predictor in the nonparametric component will follow the same pattern. However, the actual scenario is likely to reveal different correlation patterns among the predictors. Mixed estimators for nonparametric and semiparametric regression have been developed to estimate the regression curve according to the data pattern. Studies on mixed estimators for nonparametric regression have previously been conducted on the mixed Kernel estimator and Fourier Series in nonparametric regression [17], on the mixed estimator of Fourier Series and truncated spline in nonparametric regression by [13,18,19], and on the mixed estimator Kernel and smoothing spline in nonparametric regression [20]. For mixed semiparametric regression estimators, ref. [16] developed an approach that combines spline truncated and Kernel approaches, while [12,14] constructed a model combining Kernel estimators and Fourier Series. The authors of [20] developed an approach that combines smoothing spline and Fourier Series. The smoothing parameter serves to control the smoothness between the goodness of fit and the penalty.

This paper proposes a combined estimate of semiparametric regression of the multi-variable Kernel mixed estimator and Fourier Series, in which some nonparametric components contain repeating and un patterned components. Semiparametric regression research using mixed estimates of Kernel and Series Fourier was carried out by [13] but has not been able to overcome the semiparametric data pattern of mixed multivariable Kernel and multivariable Fourier Series for parameters components; Kernel components and Fourier Series are also multivariable. This paper presents the combined estimation of the multivariable Kernel estimator and the multivariable Fourier Series, using a PLS estimation approach as the estimation method. The estimator technique uses a PLS estimate produced by combining goodness of fit and penalty.

With respect to the optimal smoothing parameter selection method using the GCV method, a small optimal smoothing parameter will produce a very rough estimate on the Fourier Series estimator, but a large smoothing parameter will produce a very fine estimate where the estimator is not able to estimate the data according to the pattern. Similarly, an optimal bandwidth is required because a very small bandwidth will result in a very coarse Kernel estimator and a very wide bandwidth will produce a Kernel estimator that is slippery and does not match the data pattern. Previous nonparametric and semiparametric regression researchers have extensively developed the GCV method. Researchers who studied the GCV technique in nonparametric regression, among others, found that the GCV technique was superior to an unbiased risk approach in this context [17]. The authors of [15] investigated semiparametric regression using the GCV method. The selection of smoothing parameters in this study was based on the development of optimal smoothing parameter selection on the balanced estimator combined multivariable Fourier Series and Kernel in semiparametric regression [13].

The model results obtained from the estimation results are anticipated to be used in modeling the level of clarity of drinking water in PDAM Surabaya. The estimation of parametric components is approximated by a multivariable parametric method, Kernel components are approximated by Gaussian Kernels and components of the Fourier Series are approximated using Fourier Series with trends. The estimation model can be used to predict the composition of the optimal use of chemicals while taking into account the threshold for drinking water. This information is anticipated to provide reference materials for planning drinking water management. As each chemical is expensive and its price

varies, it is hoped that, using this model, PDAM Surabaya can efficiently manage the costs of producing clean water.

2. Materials and Techniques

Given a data pair of n that is $(t_{1i}, \ldots, t_{pi}, x_{1i}, \ldots, x_{qi}, z_{1i}, \ldots, z_{ri}, y_i)$, with $i = 1, 2, \ldots, n$. y_i is the reaction variable. A multivariable semiparametric regression form is produced as follows:

$$y_i = \eta(t_{1i}, \ldots, t_{pi}, x_{1i}, \ldots, x_{qi}, z_{1i}, \ldots, z_{ri}) + \varepsilon_i \tag{1}$$

ε_i is a random error for which $IIDN(0, \sigma^2)$. Assuming that the regression curve μ is additive, it may be expressed as:

$$y_i = \sum_{j=1}^{p} g_j(t_{ji}) + \sum_{k=1}^{q} m_k(x_{ki}) + \sum_{l=1}^{r} h_l(z_{li}) + \varepsilon_i, \ i = 1, 2, \ldots, n \tag{2}$$

Part $\sum_{j=1}^{p} g_j(t_{ji})$ is a parametric component that can be approached using a multivariable linear function; $\sum_{k=1}^{q} m_k(x_{ki})$ is a nonparametric component that can be approached using a multivariable Kernel function, and $\sum_{l=1}^{r} h_l(z_{li})$ is a nonparametric component with h_l expected to be smooth and enclosed in continuous function space on $(0, \pi)$, so that $h_l(z_{li})$ may be approached using Fourier Series [21] $H(z_l)$ with $H(z_i) = bz_i + \frac{1}{2}a_0 + \sum_{s=1}^{S} a_s \cos sz_i$ where $b, a_{0,}, a_s, s = 1, 2, \ldots S$ are model parameters.

The following equation yields the estimator η by PLS optimization:

$$\underset{h_l \in C(0,\pi), \boldsymbol{\beta} \in R^{p+1}, \boldsymbol{\gamma} \in R^{r(S+2)}}{Min} \left\{ n^{-1} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j t_{ji} - \sum_{k=1}^{q} m_k(x_{ki}) - \sum_{l=1}^{r} \left(b_l z_{li} + \frac{1}{2}a_{0l} + \sum_{s=1}^{S} a_{sl} \cos sz_{li} \right) \right)^2 + \sum_{l=1}^{r} \breve{}_l \int_0^{\pi} \frac{2}{\pi}(h_l''(z_l))^2 dz_l \right\} \tag{3}$$

where $\breve{}_l$ are the smoothing parameters.

The function that measures goodness of fit makes up the first component of Equation (3), and the function that measures the penalty makes up the second.

3. Results and Discussion

3.1. Mixed Kernel Model and Multivariable Fourier Series in Semiparametric Regression

Several lemmas must be satisfied to generate a mixed model Kernel and multivariable Fourier Series in semiparametric regression in Equation (2). Lemma 1 presents solutions for parametric components, Lemma 2 presents solutions for Kernel components, Lemma 3 presents solutions for Fourier Series components and Lemma 4 presents solutions for goodness of fit, while Lemma 5 presents the penalty component form of Equation (3).

Lemma 1. *If the components of a linear parametric curve are multivariable, $\sum_{j=1}^{p} g_j(t_{ji})$ in Equation (2) are approximated by a multivariable linear function, then $\sum_{j=1}^{p} g_j(t_{ji})$ can be written to matrix as* $\mathbf{X}\boldsymbol{\beta}$, *where $\mathbf{X}$ is a size matrix $n \times (p+1)$, and $\boldsymbol{\beta}$ is a size vector $(p+1) \times 1$.*

Proof of Lemma 1. If function $g_j(t_{ji})$ is roughly represented by a multivariable linear function, then $g_j(t_{ji}) = \beta_{0j} + \beta_{ij} t_{ji}, \ i = 1, 2, \ldots, n; \ j = 1, 2, \ldots, q$.

For $g_j(t_{ji})$, when $i = 1, 2, \ldots, n$, then the following is obtained

$$\begin{bmatrix} g_j(t_{j1}) \\ \vdots \\ g_j(t_{jn}) \end{bmatrix} = \begin{bmatrix} \beta_{0j} + \beta_{1j}t_{j1} \\ \vdots \\ \beta_{0j} + \beta_{nj}t_{jn} \end{bmatrix}$$

so that:

$$\begin{bmatrix} \sum\limits_{j=1}^{p}(\beta_{0j} + \beta_{1j}t_{ji}) \\ \vdots \\ \sum\limits_{j=1}^{p}(\beta_{0j} + \beta_{1j}t_{ji}) \end{bmatrix} = \begin{bmatrix} \beta_{01} + \beta_{11}t_{1i} + \beta_{02} + \beta_{12}t_{2i} + \ldots + \beta_{0p} + \beta_{1p}t_{pi} \\ \vdots \\ \beta_{01} + \beta_{11}t_{1i} + \beta_{02} + \beta_{12}t_{2i} + \ldots + \beta_{0p} + \beta_{1p}t_{pi} \end{bmatrix}, \ \beta_0^* = \beta_{01} + \ldots + \beta_{0p}$$

so that the matrix can be written, such as:

$$\sum\limits_{j=1}^{p} g_j(t_{ji}) = \mathbf{X}\boldsymbol{\beta} \text{ with}$$

$$\mathbf{X} = \begin{bmatrix} 1 & t_{11} & \ldots & 1 & t_{p1} \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 1 & t_{1n} & \ldots & 1 & t_{pn} \end{bmatrix}, \ \beta_0^* = \beta_{01} + \ldots + \beta_{0p} \text{ and} \tag{4}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0^* & \beta_{11} & \ldots & \beta_{1p} \end{bmatrix}^T \text{ are linear function parameters}$$

. $\square$

Lemma 2. *If the components of the Kernel curve $\sum\limits_{k=1}^{q} m_k(x_{ki})$ in Equation (2) is approximated by a multivariable Kernel function, the Nadaraya–Watson estimator [2], then $\sum\limits_{k=1}^{q} m_k(x_{ki})$ can be written as matrix $\boldsymbol{\Omega}\mathbf{y}$, as follows:*

$$\sum\limits_{k=1}^{p} m_k(x_{ki}) = \boldsymbol{\Omega}\mathbf{y},$$

where $\boldsymbol{\Omega}$ is a size matrix size $n \times n$, $\mathbf{y}$ is a size vector $n \times 1$.

Proof of Lemma 2. If the Kernel curve $\sum\limits_{k=1}^{q} m_k(x_{ki})$ in Equation (2) is approached with a multivariable Kernel function, the Nadaraya–Watson estimator [2], then $\sum\limits_{k=1}^{q} m_k(x_{ki})$ can be written as a matrix $\boldsymbol{\Omega}\mathbf{y}$

$$m_{\varphi_k}(x_{ki}) = n^{-1}\sum\limits_{i=1}^{n} W_{\varphi_{ki}}(x_{ki})y_i$$

with:

$$W_{\varphi_k}(x_{ki}) = \frac{K_{\varphi_k}(x_k - x_{ki})}{n^{-1}\sum\limits_{i=1}^{n} K_{\varphi_k}(x_k - x_{ki})},$$

$$K_{\varphi_k}(x_k - x_{ki}) = \frac{1}{\varphi_k}K\left(\frac{x_k - x_{ki}}{\varphi_k}\right)$$

where $K_{\varphi_k}(x_k - x_{ki})$ is a Kernel function. φ_k is a bandwidth parameter.

For $i = 1, 2, \ldots, n$, obtained:

$$\hat{m}_{\varphi_{ki}}(x_{ki}) = n^{-1} \sum_{k=1}^{n} W_{\varphi_{k1}}(x_{k1}) y_1 = n^{-1} W_{\varphi_{k1}}(x_{k1}) y_1 + W_{\varphi_{k2}}(x_{k2}) y_1 + \ldots + W_{\varphi_{kn}}(x_{kn}) y_1$$

$$\vdots$$

$$\hat{m}_{\varphi_{kn}}(x_{kn}) = n^{-1} \sum_{k=1}^{n} W_{\varphi_{kn}}(x_{kn}) y_n = n^{-1} W_{\varphi_{kn}}(x_{kn}) y_n + W_{\varphi_{kn}}(x_{kn}) y_2 + \ldots + W_{\varphi_{kn}}(x_{kn}) y_n$$

as follows $k = 1, 2, \ldots, q$

$$\begin{bmatrix} \hat{m}_{\varphi_{1i}}(x_{1i}) \\ \vdots \\ \hat{m}_{\varphi_{qi}}(x_{qi}) \end{bmatrix} = \begin{bmatrix} n^{-1} \sum_{i=1}^{n} W_{\varphi_{1i}}(x_{1i}) y_i \\ \vdots \\ n^{-1} \sum_{i=1}^{n} W_{\varphi_{qi}}(x_{qi}) y_i \end{bmatrix} = \begin{bmatrix} n^{-1} W_{\varphi_{11}}(x_{1i}) y_1 + \ldots + W_{\varphi_{1n}}(x_{1i}) y_i \\ \vdots \\ n^{-1} W_{\varphi_{q1}}(x_{qi}) y_1 + \ldots + W_{\varphi_{qn}}(x_{qi}) y_i \end{bmatrix}$$

Kernel components can be written as follows:

$$\hat{g}_{\varphi_k}(t_k) = \mathbf{P}_k(\varphi_k) \mathbf{y} \tag{5}$$

$$\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \ldots & y_n \end{pmatrix}^T$$

$$\mathbf{P}_k(\varphi_k) = \begin{bmatrix} n^{-1} W_{\varphi_{11}}(x_{11}) & W_{\varphi_{12}}(x_{12}) & \ldots & W_{\varphi_{1n}}(x_{1n}) \\ & & \vdots & \\ n^{-1} W_{\varphi_{q1}}(x_{q1}) & W_{\varphi_{q2}}(x_{q2}) & \ldots & W_{\varphi_{qn}}(x_{qn}) \end{bmatrix}$$

as a result of Equation (6), the following is obtained:

$$\sum_{k=1}^{q} m_k(x_{ki}) = \mathbf{\Omega} \mathbf{y}, \blacksquare \tag{6}$$

where $\mathbf{\Omega} = \begin{bmatrix} \mathbf{P}_1(\varphi_1) & \mathbf{P}_2(\varphi_2) & \ldots & \mathbf{P}_q(\varphi_q) \end{bmatrix}^T. \square$

Lemma 3. *If the Kernel curve component* $\sum_{l=1}^{r} h_l(z_{li})$ *in Equation (2) is approximated by a Fourier Series function* $H_l(z_{li})$, *assuming regression* $h_l \in C(0, \pi)$, $l = 1, 2, \ldots, r$, *then*

$$\sum_{l=1}^{r} h_l(z_{li}) \text{ can be written as } \sum_{l=1}^{r} h_l(z_{li}) = \mathbf{D} \gamma \tag{7}$$

where $\mathbf{D}$ *is a matrix of size* $n \times r(S+2)$ *and* γ *is a vector of size* $r(S+2) \times 1$.

Proof of Lemma 3. If the components of the curve of the Fourier Series $\sum_{l=1}^{r} h_l(z_{li})$ is approximated by the Fourier Series function $H_l(z_{li})$ assuming regression $h_l(z_{li})$ $h_l \in C(0, \pi)$, $l = 1, 2, \ldots, r$ are oscillation parameters [21], then obtained follow is

$$H_l(z_{li}) = b_l z_{li} + \frac{1}{2} a_{0l} + \sum_{s=1}^{S} a_{sl} \cos sz_{li}, \; i = 1, 2, \ldots, n, \; l = 1, 2, \ldots, r \tag{8}$$

for $i = 1, 2, \ldots, n$, then the following is obtained:

$$\begin{bmatrix} H_l(z_{l1}) \\ \vdots \\ H_l(z_{ln}) \end{bmatrix} = \begin{bmatrix} b_l z_{l1} + \frac{1}{2} a_{0l} + \sum_{s=1}^{S} a_{sl} \cos sz_{l1} \\ \vdots \\ b_l z_{ln} + \frac{1}{2} a_{0l} + \sum_{s=1}^{S} a_{sl} \cos sz_{ln} \end{bmatrix} \tag{9}$$

so that $\sum\limits_{l=1}^{r} H_l(z_{li})$ in Equation (10), the above equation is presented in matrix form, as follows:

$$\sum_{l=1}^{r} h_l(z_{li}) = \mathbf{D}\boldsymbol{\gamma} \ \blacksquare \tag{10}$$

where:

$$\mathbf{D} = \begin{bmatrix} z_{11} & 1 & \cos z_{11} & \cdots & \cos Sz_{11} & \cdots & z_{r1} & 1 & \cos z_{r1} & \cdots & \cos Sz_{r1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{1n} & 1 & \cos z_{r1} & \cdots & \cos Sz_{rn} & \cdots & z_{rn} & 1 & \cos z_{rn} & \cdots & \cos Sz_{rn} \end{bmatrix}$$

and

$$\boldsymbol{\gamma} = \begin{bmatrix} b_1 & \frac{1}{2}a_{01} & a_{11} & \dots & a_{S1} & \dots & b_r & \frac{1}{2}a_{0r} & a_{1r} & \dots & a_{Sr} \end{bmatrix}^T.$$

□

Lemma 4. *If the semiparametric regression model is as in Equation (2), where the linear regression curve is given in Lemma 1, the Kernel curve is in Lemma 2 and the Fourier Series curve is in Lemma 3, then the equation's goodness of fit (3) is as follows:*

$$((I - y)\boldsymbol{\Omega} - \mathbf{X}\boldsymbol{\beta} - \mathbf{D}\boldsymbol{\gamma})^T((I - y)\boldsymbol{\Omega} - \mathbf{X}\boldsymbol{\beta} - \mathbf{D}\boldsymbol{\gamma})$$

Proof of Lemma 4. If the parametric component is approximated by a linear function so that $\sum\limits_{j=1}^{p} g_j(t_{ji}) = \mathbf{X}\boldsymbol{\beta}$ (Lemma 1), then the Kernel component is close to the Nadaraya–Watson Kernel, then $\sum\limits_{k=1}^{q} m_k(x_{ki}) = \boldsymbol{\Omega}y$, then the components of the Fourier Series are approximated by the Bilodeau Fourier Series, then $\sum\limits_{l=1}^{r} h_l(z_{li}) = \mathbf{D}\boldsymbol{\gamma}$, Equation (3) can be presented in the form of a matrix $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{y}\boldsymbol{\Omega} + \mathbf{D}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, so that the goodness of fit equation (3) is obtained $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$, as follows:

$$((I - \mathbf{y})\boldsymbol{\Omega} - \mathbf{X}\boldsymbol{\beta} - \mathbf{D}\boldsymbol{\gamma})^T((I - \mathbf{y})\boldsymbol{\Omega} - \mathbf{X}\boldsymbol{\beta} - \mathbf{D}\boldsymbol{\gamma}) \ \blacksquare$$

□

Lemma 5. *If the penalty component is given as $\sum\limits_{l=1}^{r} \breve{\ }_l \int\limits_0^{\pi} \frac{2}{\pi}(h_l''(z_l))^2 dz_l$ from Equation (2), then:*

$$\sum_{l=1}^{r} \breve{\ }_l \int_0^{\pi} \frac{2}{\pi}(h_l''(z_l))^2 dz_l = \text{fl}^T \mathbf{P}\boldsymbol{\gamma} \tag{11}$$

λ_l are smoothing parameters.

Proof of Lemma 5. Given a penalty component is given as $\sum\limits_{l=1}^{r} \breve{\ }_l \int\limits_0^{\pi} \frac{2}{\pi}(h_l''(z_l))^2 dz_l$ with $h_l(z_{li})$ expected to be smooth and enclosed in a continuous function space $C(0, \pi)$, so

that $h_l(z_{li})$ may be approached by $b_l z_{li} + \frac{1}{2} a_{0l} + \sum_{s=1}^{S} a_{sl} \cos s z_{li}$, then determine the second derivative of $h_l(z_{li})$, obtained $h_l''(z_l) = -\sum_{s=1}^{S} s^2 a_{sl} \cos s z_l$ a result

$$\int_0^\pi \frac{2}{\pi} (h''(z))^2 dz = \sum_{s=1}^{S} s^4 a_s^2 = \boldsymbol{\gamma}^T \mathbf{P} \boldsymbol{\gamma} \ \blacksquare \tag{12}$$

with $\boldsymbol{\gamma} = \begin{pmatrix} b \\ \frac{1}{2} a_0 \\ a_1 \\ \vdots \\ a_S \end{pmatrix}$ and $\mathbf{P} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1^4 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & S^4 \end{pmatrix}$. $\square$

Theorem 1 provides a detailed explanation of the mixed model of Kernel and multivariable Fourier Series, where the linear curve is provided in Equation (4), the Kernel curve in Equation (5), the Fourier Series curve in Equation (8), and the penalty in Equation (12).

Theorem 1. *Equation (1) describes a semiparametric regression model. The linear regression curve is represented by the expressions given in Equation (4), the Kernel in Equation (5), Fourier Series in Equation (8), and penalty in Equation (12). We obtain a multivariable by minimizing the PLS in Equation (3), giving us:*

$$\hat{\eta}_{(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \mathbf{s})}(\mathbf{t}, \mathbf{x}, \mathbf{z}) = \mathbf{M}^* \mathbf{y}$$

where $\mathbf{y}$ *is a vector size* $n \times 1$ *and* $\mathbf{M}^*$ *is a matrix size* $n \times n$.

Proof of Theorem 1. Based on Lemma 4, Equation (3)'s optimization can be expressed as:

$$\underset{\boldsymbol{\gamma}, \boldsymbol{\beta}}{Min} \eta(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \underset{\boldsymbol{\gamma}, \boldsymbol{\beta}}{Min} \left[n^{-1} [(\mathbf{I} - \boldsymbol{\Omega}_k)\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{D}\boldsymbol{\gamma}]^T [(\mathbf{I} - \boldsymbol{\Omega}_k)\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{D}\boldsymbol{\gamma}] + \boldsymbol{\gamma}^T \mathbf{P} \boldsymbol{\gamma} \right] \tag{13}$$

where $\boldsymbol{\gamma}^T \mathbf{P} \boldsymbol{\gamma}$ a is the penalty component.

In Equation (13), if $\mathbf{A} = (\mathbf{I} - \boldsymbol{\Omega}_k)\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, so $\zeta(\boldsymbol{\gamma}, \boldsymbol{\beta})$, It can be expressed as:

$$\frac{\partial \eta(\boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial \boldsymbol{\gamma}} = n^{-1} \frac{\partial}{\partial \boldsymbol{\gamma}} \left[\left(\mathbf{A}^T - \boldsymbol{\gamma}^T \mathbf{D}^T \right)(\mathbf{A} - \mathbf{D}\boldsymbol{\gamma}) + \boldsymbol{\gamma}^T \mathbf{P} \boldsymbol{\gamma} \right] = [2\mathbf{D}^T \mathbf{D} + 2\mathbf{P}]\boldsymbol{\gamma} - 2\mathbf{D}^T \mathbf{A} = 0$$

$$\hat{\boldsymbol{\gamma}} = [\mathbf{D}^T \mathbf{D} + \mathbf{P}]^{-1} \mathbf{D}^T \mathbf{A}$$

To get an estimator $\boldsymbol{\gamma}$, so $\eta(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is obtained by partially deriving from $\boldsymbol{\gamma}$, and the result is equal to zero.

$$\hat{\boldsymbol{\gamma}} = \mathbf{B} \left((\mathbf{I} - \boldsymbol{\Omega})\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right) \text{ where } \mathbf{B} = \left[\mathbf{D}^T \mathbf{D} + \mathbf{P} \right]^{-1} \mathbf{D}^T \tag{14}$$

Next, to get an estimation $\hat{\boldsymbol{\beta}}$, i.e., partially derivative of $\zeta(\boldsymbol{\gamma}, \boldsymbol{\beta})$ to $\boldsymbol{\beta}$, then equated to zero, as follows:

$$\frac{\partial \eta(\boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = n^{-1} \frac{\partial}{\partial \boldsymbol{\beta}} \left[\left(\mathbf{y}^T (\mathbf{I} - \boldsymbol{\Omega})^T - \boldsymbol{\beta}^T \mathbf{X}^T - \boldsymbol{\gamma}^T \mathbf{D}^T \right)((\mathbf{I} - \boldsymbol{\Omega})\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{D}\boldsymbol{\gamma}) + \boldsymbol{\gamma}^T \mathbf{P} \boldsymbol{\gamma} \right]$$

$$= n^{-1} \left[2\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} - 2\mathbf{X}^T (\mathbf{I} - \boldsymbol{\Omega})\mathbf{y} + 2\mathbf{X}^T \mathbf{D}\boldsymbol{\gamma} \right] = 0$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \left[(\mathbf{I} - \boldsymbol{\Omega})\mathbf{y} - \mathbf{D}\hat{\boldsymbol{\gamma}} \right] \tag{15}$$

Equation (14) is substitute into Equation (15), and we obtained:

$$\hat{\boldsymbol{\beta}} = \mathbf{T}^*\mathbf{y}, \text{ where } \mathbf{T}^* = \left(\mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{DBX}\right)^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\mathbf{I} - \mathbf{DB})(\mathbf{I} - \boldsymbol{\Omega}) \quad (16)$$

Next get $\hat{\boldsymbol{\gamma}}$, namely by substituting Equation (16) into the Equation (14), we obtained:

$$\hat{\boldsymbol{\gamma}} = \mathbf{B}\left((\mathbf{I} - \boldsymbol{\Omega})\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{B}(\mathbf{I} - \boldsymbol{\Omega} - \mathbf{XT}^*)\mathbf{y}$$

where $\hat{\boldsymbol{\beta}} = \mathbf{T}^*\mathbf{y}$,

The above equation can be written in matrix form, as follows

$$\hat{\boldsymbol{\gamma}} = \mathbf{C}^*\mathbf{y}, \text{where } \mathbf{C}^* = \mathbf{B}(\mathbf{I} - \boldsymbol{\Omega} - \mathbf{XT}^*)$$

$\mathbf{y}$ is a vector of size $n \times 1$ and $\mathbf{C}^*$ is a matrix of size $n \times n$,

If $\hat{\boldsymbol{\beta}}$ given in Equation (15) is substituted, we get an estimator for the parametric component:

$$\hat{\mathbf{g}}_{(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\lambda},\mathbf{s})}(\mathbf{t},\mathbf{x},\mathbf{z}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{XT}^*\mathbf{y}, \text{ with } \mathbf{T}^* \text{ is a matrix size } n \times n \quad (17)$$

where $\mathbf{T}^* = \mathbf{X}\left(\mathbf{I} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{DBX}\right)^{-1}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{DB})(\mathbf{I} - \boldsymbol{\Omega})\right)$.

After the parametric estimator is obtained, the next step is to obtain a multivariable Kernel component estimator, as follows:

$$\hat{\mathbf{m}}_{(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\lambda},\mathbf{s})}(\mathbf{t},\mathbf{x},\mathbf{z}) = \sum_{k=1}^{q} \hat{\mathbf{m}}_{(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\lambda},\mathbf{s})k}(\mathbf{t},\mathbf{x},\mathbf{z}) = \boldsymbol{\Omega}\mathbf{y} \quad (18)$$

where $\boldsymbol{\Omega} = \sum_{k=1}^{q} \mathbf{P}_k(\varphi_k) = \mathbf{P}_1(\varphi_k)\,\mathbf{P}_2(\varphi_k)\,\mathbf{P}_g(\varphi_k)$. $\boldsymbol{\Omega}$ is a matrix size $n \times n$,

The following steps are taken to obtain the estimator for the multivariable Fourier Series components:

$$\sum_{l=1}^{r} \hat{\mathbf{h}}_{(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\lambda},\mathbf{s})l}(\mathbf{t},\mathbf{x},\mathbf{z}) = \mathbf{D}\hat{\boldsymbol{\gamma}} = \mathbf{DC}^*\mathbf{y}, \text{with } \hat{\boldsymbol{\gamma}} = \mathbf{C}^*\mathbf{y} \quad (19)$$

Because $\hat{\mathbf{g}}_{(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\lambda},\mathbf{s})}(\mathbf{t},\mathbf{x},\mathbf{z})$ given in Equation (17), $\hat{\mathbf{m}}_{(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\lambda},\mathbf{s})}(\mathbf{t},\mathbf{x},\mathbf{z})$ is given in Equation (18), dan $\hat{\mathbf{h}}_{(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\lambda},\mathbf{s})}(\mathbf{t},\mathbf{x},\mathbf{z})$ is given in Equation (19), then the estimation of the Kernel mixture and the multivariable Fourier Series in the semiparametric regression, where: $\mathbf{x} = (x_1, x_2, \ldots, x_q)^T$, $\mathbf{t} = (t_1, t_2, \ldots, t_p)^T$, and $\mathbf{z} = (z_1, z_2, \ldots, z_r)^T$, so that: $\hat{\eta}_{(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\lambda},\mathbf{s})}(\mathbf{t},\mathbf{x},\mathbf{z}) = (\mathbf{XT}^* + \boldsymbol{\Omega} + \mathbf{DC}^*)\mathbf{y}$.

Furthermore, the above equation can be presented as follows:

$$\hat{\eta}_{(\boldsymbol{\beta},\boldsymbol{\varphi},\boldsymbol{\lambda},\mathbf{s})}(\mathbf{t},\mathbf{x},\mathbf{z}) = \mathbf{M}^*\mathbf{y} \ \blacksquare \quad (20)$$

where $\mathbf{M}^* = (\mathbf{XT}^* + \boldsymbol{\Omega} + \mathbf{DC}^*)$. $\mathbf{y}$ is a vector of size $n \times 1$ and $\mathbf{M}^*$ is a matrix of size $n \times n$.$\square$

3.2. Smoothing Parameter Selection

Semiparametric regression using the Kernel mixed estimator and multivariable Fourier Series is highly reliant on the selection of the best smoothing, bandwidth, and oscillation parameters. The authors of [22,23] state that the selection of the smoothing parameters

using GCV in semiparametric regression utilizing the combined estimator multivariable Fourier Series and Kernel is according to:

$$GCV(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \mathbf{s}) = \frac{n^{-1}\|(\mathbf{I} - \mathbf{M}^*)\,\mathbf{y}\|^2}{(n^{-1}trace(\mathbf{I} - \mathbf{M}^*)^2}, \text{ with } \mathbf{M}^* = (\mathbf{XT}^* + \boldsymbol{\Omega} + \mathbf{DC}^*) \tag{21}$$

The least $GCV(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \mathbf{s})$ results in the ideal smoothing parameters, oscillation parameters, and bandwidth.

4. Modeling Data

In this section, the TKAM data at PDAM Surabaya are subjected to a combined model of the multivariable Kernel and the Fourier Series in semiparametric regression. Drinking water is very important for human life and must be used wisely considering future generations [24]. Because of water pollution, purification processes are necessary which are very costly [25] and require careful planning, including the assessment of the composition of the chemical substances needed to obtain drinking water that meets required standards.

After conducting an initial study on TKAM data at PDAM Surabaya, the obtained data show that there were differences in the data patterns between each predictor variable and the reaction variable; that is, some showed a Fourier Series pattern, others showed a Kernel pattern, and some followed a linear pattern. These data were then applied to the model in Equation (19) using R software with library(pracma), library(MASS), library(lmtest) and library(gtools). The response variable y was the level of clarity of drinking water. The predictor variables thought to affect the level of water clarity included aluminum sulfate (x_1), liquid chlorine (x_2), cupric sulfate (x_3), chlorine (x_4), Dukem 108A (x_5), and the turbidity of the water after deposition (x_6), where x_2, x_3, and x_6 are parametric components, x_1 dan x_5 are Kernel components and x_4 is a Series Fourier component. The estimated result, based on the smallest GCV criterion value [23,26], from Equation (20) is 0.00209. The following estimation models were obtained:

$$\hat{y} = -0.28893 + 17.8774x_{2i} - 0.33519x_{3i} - 0.02863x_{6i} + \frac{\frac{1}{\sqrt{28}}\exp\left[-\frac{1}{2}\left(\frac{0.47351-x_{1i}}{0.33068}\right)^2\right]}{\overset{30}{\underset{i=1}{\Sigma}}\frac{1}{\sqrt{28}}\exp\left[-\frac{1}{2}\left(\frac{0.47351-x_{1i}}{0.33068}\right)^2\right]}$$

$$+ \frac{\frac{1}{\sqrt{28}}\exp\left[-\frac{1}{2}\left(\frac{0.50129-x_{5i}}{0.00177}\right)^2\right]}{\overset{30}{\underset{i=1}{\Sigma}}\frac{1}{\sqrt{28}}\exp\left[-\frac{1}{2}\left(\frac{0.50129-x_{1i}}{0.00177}\right)^2\right]} - 0.52630x_{4i} - 0.57788\cos x_{4i} - 4.51577 \times 10^{-9}\cos 2x_{4i} - 2.00623\cos 3x_{4i}$$

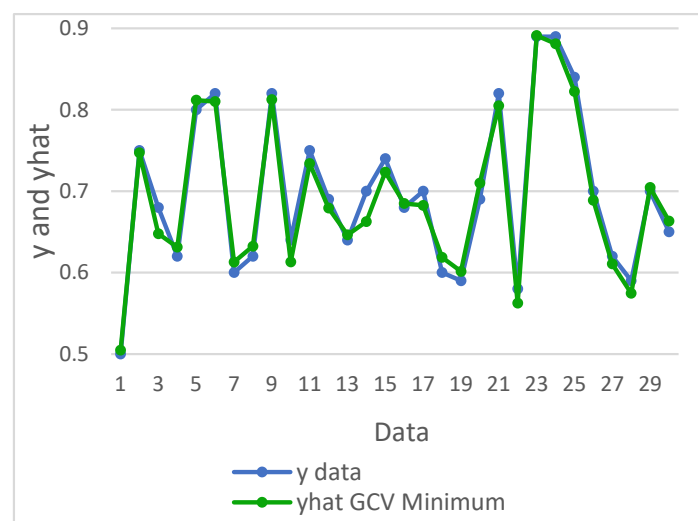An overview of the real data and the estimated results is presented in Figure 1.

Figure 1. Comparison between real data and TKAM estimation results at PDAM Surabaya.

A combination of the multivariable Kernel and multivariable Fourier Series in Semi-parametric regression has a value $R^2 = 88.2\%$ when estimating the degree of clarity of drinking water in PDAM Surabaya utilizing semiparametric regression modeling. Based on the value = 88.2% obtained, the predictor variable can explain 88.2% of the variance in the relationship between the response variables. Furthermore, this shows the suitability of this model to be used in modeling the TKAM data from PDAM Surabaya [19].

5. Conclusions

This paper presents an estimation technique for semiparametric regression using PLS. We combined the multivariable parametric estimator, the multivariable Kernel and the multivariable Fourier Series to estimate the regression curve with data having a data pattern that was partly parametric multivariable, partly multivariable Kernel and partly Fourier Series multivariable. The model was based on the smallest GCV. We considered the outcomes using various types of Kernels, while looking at the estimator features of the Fourier Series and Kernel estimator in semiparametric regression. The model obtained was more adequate compared to [19] which had a determination coefficient R of 84%. Using a mixed estimate of Kernel and multivariable Fourier Series, the coefficient of determination R was found to be 88.4%.

Author Contributions: Conceptualization, A.T.A.; data curation, A.T.A. and I.Z.; formal analysis A.T.A. and I.N.B.; funding acquisition, I.N.B.; investigation, I.N.B. and I.Z.; methodology, A.T.A. and I.N.B.; software, A.T.A. and I.Z.; supervision and validation, I.N.B.; visualization, A.T.A. and I.Z.; writing—original draft, A.T.A. and I.N.B.; writing—review and editing, A.T.A., I.N.B. and I.Z. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available to support the study's conclusions, in the Dissertation's Pane, R. Department of Statistic, Institut Teknologi Sepuluh Nopember Surabaya Indonesia, 2019.

Conflicts of Interest: The authors declare no conflict of interest.

Notations

Symbol	Meaning
y_i	Response variable to i
x_{pi}	The predictor variable to p on the parametric component x for the i-th subject.
	the ith predictor variable on the kth parametric component
	The p-th predictor variable on the parametric component for the i-th subject
t_{ki}	The predictor variable to k on the nonparametric component t for the i-th subject.
z_{ri}	The predictor variable to r on the nonparametric component z for the i-th subject.
a_0, a_S, b	Fourier Series parameters
$\boldsymbol{\beta}$	Parametric component parameter vector/regression coefficient vector
$m(x)$	Parametric function for parametric components
$g(t)$	Kernel functions for nonparametric components
$h(z)$	Fourier Series functions for nonparametric components
$\boldsymbol{\varepsilon}$	Random error vector with $\boldsymbol{\varepsilon} \sim \mathbf{N}(0, \mathbf{I}\sigma^2)$.
σ^2	Error variance
φ	Bandwidth
$\mathbf{a}$	Vector containing the parameters of the Fourier Series measuring $(S + 2) \times 1$.
$\mathbf{T}$	The matrix containing the coefficients of the multivariable Fourier Series of $n \times (S + 2)$.
λ	Smoothing parameter

S	Oscillation parameters
SSE	Sum square error
SST	Sum square total
$\mathbf{A}^{-1}$	The invers of $\mathbf{A}$ matrix
$K(t)$	Kernel function
$tr(\mathbf{A})$	The trace of $\mathbf{A}$ matrix
$\mathbf{D}^*$	Matrix D*
$E[.]$	Expectation
GCV(.)	Generalized cross validation
$\mathbf{I}$	Identity matrix
S	Oscillation of a Fourier Series function
R^2	Coefficient of determination
$\mathbf{D}$	The matrix containing the coefficients of the multivariable Fourier Series of $n \times r(S+2)$.
γ	Multivariable Fourier Series parameter matrices of size $(S+2) \times 1$
A^*	Goodness of fit component function semiparametric regression mixed with Kernel and Fourier Series
B^*	Penalty component function mixed semiparametric regression Kernel and Fourier Series
L^*	The integral of a function L
M^*	The integral of a function M
$\mathbf{P}$	Matrix containing Kernel weights of size $n \times n$.
$\mathbf{T}$	Matrix containing multivariable Fourier Series coefficients of size $n \times p(k+2)$.
Ω	Matrix containing matrices $\mathbf{P}$.
$\mathbf{X}$	Parametric component predictor variable matrix of size $n \times (r+1)$
$\mathbf{Z}$	Matrix containing univariable Fourier Series coefficients of size $n \times (k+2)$.
$\breve{}_j$	j-th smoothing parameter
β	Vector containing parameter estimates of parametric components
$\hat{\mu}$	Parameter vector containing parameter estimates of parametric and nonparametric components
R^2	Coefficient of determination
$\Re$	Real number
$\|.\|$	Norm/vector length

References

1. The Regulation of the Minister of Health of the Republic of Indonesia no. 416/Menkes/per/IX/1990. 1990. Available online: https://baristandsamarinda.kemenperin.go.id/download/PerMenKes416(1990)-Syarat&Pengawasan_Kualitas_Air.pdf (accessed on 17 May 2022).
2. Okumura, H.; Naito, K. Non-parametric Kernel Regression for Multinomial Data. *J. Multivar. Anal.* **2006**, *97*, 2009–2022. [CrossRef]
3. Nadaraya, E.A.; Watson, G.S. On Estimating Regression. *Theory Probab. Its Appl.* **1964**, *9*, 141–142. [CrossRef]
4. Eubank, R.L. *Spline Smoothing and Nonparametric Regression*; Mercel Dekker: New York, NY, USA, 1988.
5. Wood, S.N. On Confidence Intervals for Generalized Additive Models Based on Penalized Regression Spline. *Aust. N. Z. J. Stat.* **2006**, *48*, 445–464. [CrossRef]
6. Iqbal, A.; Abd Hamid, N.N.; Ismail, A.I.M.; Abbas, M. Galerkin approximation with quintic B-spline as basis and weight functions for solving second order coupled nonlinear Schrödinger equations. *Math. Comput. Simul.* **2021**, *187*, 1–16. [CrossRef]
7. Kim, T.W.; Kvasov, B. A shape-preserving approximation by weighted cubic splines. *J. Comput. Appl. Math.* **2012**, *236*, 4383–4397. [CrossRef]
8. Ruppert, D. Selecting the number of knots for penalized splines. *J. Comput. Graph. Stat.* **2002**, *11*, 735–757. [CrossRef]
9. Wahba, G. *Spline Models for Observasion Data*; SIAM: Philadelphia, PA, USA, 1990.
10. Nisa, K.; Budiantara, I.N.; Tuty, A. Multivariable Semiparametric Regression Model with Combined Estimator of Fourier Series and Kernel. *IOP Conf. Ser. Earth Environ. Sci.* **2017**, *58*, 012028. [CrossRef]
11. Bhattacharya, P.K.; Zhao, P.L. Semiparametric Inference in a Partial Linear Model. *Ann. Stat.* **1997**, *1*, 244–262. [CrossRef]
12. Smith, M.; Kohan, R.; Mathur, S.K. Bayesian Semiparametric Regression: An Exposition and Application to Print Advertising Data. *J. Bus. Res.* **2000**, *49*, 229–244. [CrossRef]
13. Cheng, M.Y.; Paige, R.L.; Sun, S.; Yan, K. Variance Reduction for Kernel Estimations in Clustered/Longitudinal Data Analysis. *J. Stat. Plan. Inference* **2010**, *140*, 1389–1397. [CrossRef]
14. Manzana, S.; Zerom, D. Kernel Estimation of a Partially Linear Additive Model. *Stat. Probab. Lett.* **2005**, *72*, 313–322. [CrossRef]
15. Amato, U.; Antoniadis, A.; De Feis, I. Fourier Series Approximation of Separable Models. *J. Comput. Appl. Math.* **2002**, *146*, 459–479. [CrossRef]

16.   Morton, J.; Silverberg, L. Fourier Series of Half-Range Functions by Smooth Extension. *Appl. Math. Model.* **2009**, *33*, 812–821. [CrossRef]
17.   Mardianto, M.F.F.; Kartiko, S.H.; Utami, H. Forecasting Trend-Seasonal Data Using Nonparametric Regression with Kernel and Fourier Series Approach. In *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (ICMS2017)*; Springer: Singapore, 2019; pp. 343–349.
18.   Nisa, K. Semiparametrik Regression Model with Combined Estimator of Truncated Spline and Fourier Series (Case Study: Life Expectancy of East Java Province). Master's Thesis, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, 2017. Available online: https://core.ac.uk/download/pdf/291461498.pdf (accessed on 27 June 2022).
19.   Pane, R.; Budiantara, I.N.; Zain, I.; Otok, B.W. Parametric and Nonparametric Estimators in Fourier Series Semiparametric Regression and Their Characteristics. *Appl. Math. Sci.* **2014**, *8*, 5053–5064. [CrossRef]
20.   Hardle, W. *Applied Nonparametric Regression*; Humboldt Universität zu Berlin: Berlin, Germany, 1994.
21.   Bilodeau, M. Fourier Smoother and Additive Models. *Can. J. Stat.* **1992**, *3*, 257–269. [CrossRef]
22.   Kayri, M.; Zirhoglu, G. Kernel Smoothing Function and Choosing Bandwidth for Nonparametric Regression Techniques. *Ozean J. Appl. Sci.* **2009**, *2*, 49–60.
23.   Lin, Y.; Zhang, H.H. Component Selection and Smoothing in Multivariate Nonparametric Regression. *Ann. Stat.* **2006**, *34*, 2272–2297. [CrossRef]
24.   Hefni, E. *Water Quality Review: For Management of Aquatic Resources and Environment*; Kanisius: Yogyakarta, Indonesia, 2003.
25.   Sutrisno, T. *Clean Water Supply Technology*; PT Rineka Cipta: Jakarta, Indonesia, 2004.
26.   Aydin, D.; Memmedli, M.; Omay, R.E. Smoothing Parameter Selection for Nonparametric Regression Using Smoothing Spline. *Eur. J. Pure Appl. Math.* **2013**, *6*, 222–238.