Journal of Theoretical and Applied Information Technology
31st January 2019. Vol.97. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A REPRODUCING KERNEL HILBERT SPACE APPROACH AND SMOOTHING PARAMETERS SELECTION IN SPLINE-KERNEL REGRESSION

[1]RAHMAT HIDAYAT, *[2]I NYOMAN BUDIANTARA, [3]BAMBANG W. OTOK, [4]VITA RATNASARI

[1,2,3,4]Department of Statistic, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

[1]Department of Mathematics, Cokroaminoto Palopo University, Palopo, Indonesia

E-mail: [1]dayatmath@gmail.com, *[2]i_nyoman_b@statistika.its.ac.id, [3]dr.otok.bw@gmail.com, [4]vitaratna70@gmail.com
*Corresponding author: I Nyoman Budiantara

ABSTRACT

Regression analysis studies the form of the relationship between one or more predictor variables with one response variable. The relationship of the response variable with several predictor variables in nonparametric regression does not always using one type of approach such as Spline, Kernel, or Fourier series. This fact is found in many nonparametric regression, between one predictor variable and another predictor variable that has a different pattern with the response variable. This study proposes a model that has ability to handle the different patterns in the nonparametric regression. This model was developed by adding Kernel functions to the goodness of fit component in completion of the smoothing Spline. Empirical analysis is carried out on fuel consumption data in Indonesia. The performance of the proposed model is evaluated by looking at the GCV value and comparing its coefficient of determination with the parametric regression. The result of the study shows that the proposed model is better than the compared model. In addition, this model has a highly accuracy in making predictions or forecasting.

Keywords: *Nonparametric, Regression, Spline, Kernel, GCV, Fuel Consumption*

1. INTRODUCTION

Regression analysis is one of the statistical tool that mostly used to determine the relationship between a pair or more of variables. Suppose the given data (t_i, y_i), $i = 1, 2, ..., n$ and the relationship between t_i and y_i assumed to follow the following regression model

$$y_i = f(t_i) + \varepsilon_i \; ; \; i = 1, 2, ..., n \qquad (1)$$

where f is a regression curve and ε_i is a random error that assumed to be independent and identical normally distributed with zero mean and variance σ^2.

There are two methods that can be used to estimate the function *f*, which called the parametric regression method and the nonparametric regression method. The consideration for choosing which method to use is related to the assumption of the function *f*. The parametric regression method will be appropriate if the form of function *f* is known or there are other sources that can be used to determine the form of the function *f* [1]. However, if the function *f* is unknown, then the nonparametric regression method is more suitable than the parametric one [2]. In this case, the function *f* simply assumed to be contained in a particular function space, where the selection of function space is usually motivated by the properties of smoothness of the specific function *f*.

There are several estimation techniques in nonparametric regression, such as smoothing Spline, Kernel, Wavelet, and Fourier Series [1]. The smoothing Spline was first introduced by Whitaker in 1923 as a data pattern approach. It was based on an optimization problem, developed by the Reinsc in 1967 [3]. The smoothing Spline estimator is obtained from a penalized least square optimization

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(PLS) [4], [5], [6]. The smoothing Spline has very special and very good statistical and visual interpretation [7]. Besides, the smoothing Spline also able to handle data characteristic/ function that is smooth. The smoothing Spline also has an excellent ability to handle data whose behavior changes at certain sub-intervals [5], [6], [7].

[3] is use the smoothing Spline function to approach the univariable nonparametric regression curve, then [8] developed the quantile Spline to handle the outlier data, [7] developed the M-type Spline as well as for handling the outlier data. Furthermore, [9] provides a weighted Spline estimator to handle the inequality of variance in the nonparametric regression. [10] developed a Spline estimator to estimate a robust regression curve and [11] used Bayesian in completing the Spline in the multiple nonparametric regression.

The other estimator beside the smoothing Spline which is often used in the nonparametric regression, named Kernel estimator. The first group of researchers which examine the Kernel was initiated by [12] and [13]. The Kernel estimator is the development of a histogram estimator. This estimator is a linear estimator similar to other nonparametric regression estimators, the only difference is because the Kernel estimator is the estimator which apply the use of the bandwidth method. The advantage of the Kernel estimator is having good ability in modeling data that has no specific pattern [14]. In addition, the Kernel estimator is more flexible, has simple mathematical form, and relatively fast to reach the convergence level [15]. In terms of computation, the Kernel method is easier to be done and to be implemented [16].

[17] state that, since the Spline function has its own character as well as the Kernel, then it is good to check the character of each of the predictors before the analysis started, to obtain a good estimate.

Research that related to nonparametric regression involving many predictors is limited to the use of the same type of estimator for each predictor [6]. Thus, it can be said that there are fundamental assumptions in the model, i.e., first, the pattern of each predictor in the multi-predictors nonparametric regression model is considered to have the same pattern. Second, researchers only use one form of the model estimator for each predictor. These two assumptions used in multi-predictors nonparametric regression models are rarely found, and in application, there are often cases where different patterns of each predictor variable occur, including the case of open unemployment [18]. In addition, using only one form of estimator in estimating multi-predictors nonparametric regression curves, resulting the estimator obtained will not match the data pattern. This will affect the correctness in estimation of the regression model obtained and tends to produce a large error.

Based on the results of the above research, and based on preliminary exploration of the fuel consumption data, it was found that there were predictors that were in accordance with the characteristics of Spline which changed at certain intervals and there were predictors that were in accordance with the Kernel data pattern and to obtain an estimate of the regression curve model that is match the data pattern, then in this study not only use a single regression curve estimator model, but more than one estimator model. The aim of this study is to develop new method in estimating the regression curve in nonparametric regression, which is by modifying a settlement function of Penalized smoothing Spline, by adding a Kernel function on the completion of its goodness of fit. This model is expected to be able to handle different data patterns between each predictor in nonparametric multi-predictors regression

2. MATERIALS AND METHODS

2.1 Spline Function

Spline function is a piecewise polynomial, that is a polynomial which has segmented properties. This property provides more flexibility than ordinary polynomials, making it possible to adjust effectively to the local characteristics of the function or data.

From the form of (1) if the function $f \in W_2^m[0,1] = \left\{ f : f^{(k)}, k = 0,...,m-1, f^{(m)} \in L_2[0,1] \right\}$ where $L_2[0,1]$ expresses the set of function of the integral square at the interval $[0,1]$ with the conformity of the curve to the data is $n^{-1}\sum_{i=1}^{n}\left(y_i - f(t_i)\right)^2$ and the roughness of the curve is $\int_0^1 \left(f^{"}(t)\right)^2 dt$, then the estimate of f can be obtained by minimizing the Penalized Least Square

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$n^{-1}\sum_{i=1}^{n}\left(y_i - f\left(t_i\right)\right)^2 + \Lambda\int_0^1\left(f^{"}\left(t\right)\right)^2 dt\,. \qquad (2)$$

To solve the optimization problem in (2), Reproducing Kernel of Hilbert Space can be used [18], the Spline optimization can be transformed into a problem of projection in a Hilbert space [19]. The very important characteristic of a Reproducing Kernel is that we can determine the representation of a linear functional, so that the regression curve $f \in W_2^m\left[0,1\right]$, which is the optimal solution of the equation (2).

2.2 Kernel Function

The kernel estimator has the advantage of being flexible, its mathematical form is easy and relatively fast of reaching the level of convergence. If regression curve $g(t_i)$ approached by Kernel function, the estimation of the regression curve can be presented in the form of:

$$\hat{g}_\tau\left(t\right) = n^{-1}\sum_{i=1}^{n}\left[\frac{K_\tau\left(t-t_i\right)}{n^{-1}\sum_{j=1}^{n}K_\tau\left(t-t_j\right)}\right]y_i$$

$$= n^{-1}\sum_{i=1}^{n}W_{\tau i}\left(t\right)y_i$$

where:

$$W_{\tau i}\left(t\right) = \frac{K_\tau\left(t-t_i\right)}{n^{-1}\sum_{j=1}^{n}K_\tau\left(t-t_j\right)},$$

$$K_\tau\left(t-t_i\right) = \frac{1}{\tau}K\left(\frac{t-t_i}{\tau}\right)$$

with K is the Kernel function. According to [13], the form of Kernel function K can be:

- Gaussian Kernel: $K(t) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{t^2}{2})I_{[-\infty,\infty]}(z)$
- Uniform Kernel : $K(t) = 0,5I_{[-1,1]}(t)$
- Epanechnikov Kernel: $K(t) = 0,75(1-t^2)I_{[-1,1]}(t)$
- Quadratic Kernel: $K(t) = \frac{15}{16}(1-t^2)^2 I_{[-1,1]}(t)$

and τ is the bandwidth. The Kernel approach depends on the bandwidth τ, that can use to control the smoothness of the estimation curve. The selection of the proper bandwidth is very important in Kernel regression [20], [21]. If the bandwidth that is too large then it will produce a very smooth estimation curve and it will approach the average of the response variable, whereas if the bandwidth is too small it will produce a less smooth estimation curve and it will approach the average of the data instead.

3. RESULT AND DISCUSSION

Definition:

Reproducing Kernel Hilbert Space H is a Hilbert space of real function in the interval $\left[0,1\right]$ with the properties that for every $t \in \left[0,1\right]$ there is exist the function $L_t f = f\left(t\right)$ which is defined as limited and linear functions, means exist M such that $\left|L_t f\right| = \left|f\left(t\right)\right| \le M\left\|f\right\|$.

Definition:

Reproducing Kernel of H is R function which defined in the $\left[0,1\right]\times\left[0,1\right]$ such that for every fixed point $t \in \left[0,1\right]$ imply $R_t \in H$ with $R_t\left(s\right) = R\left(s,t\right)$ and $L_t f = \left\langle R_t, f\right\rangle = f\left(t\right)$.

H is a Hilbert space and exist single reproducing Kernel R_t for point $t \in \left[0,1\right]$ in H and L_t is limited and linear function in H which maps the function f in the H space to real numbers $L_i : f \to f\left(t_i\right)$.

Suppose that the given data $\left(t_i, y_i\right)$, $i = 1,2,...,n$ and the relationship between t_i and y_i assumed to follow the regression model

$$y_i = L_i f + \varepsilon_i \qquad (3)$$

Thus, if it is assumed that the regression model is additive, then equation (3) can be modified by adding the kernel function as follows:

$$y_i = L_i f + g + \varepsilon_i \qquad (4)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Estimating function f and g are to find the function f and g which exist in the Hilbert space and minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - L_i f - g\right)^2 + \Lambda\left\|P_1 f\right\|_R^2 \qquad (5)$$

for $L_i f = f(t_i)$ and $\left\|P_1 f\right\|_R^2 = \int_0^1 \left(f^{(m)}(t)\right)^2 dt$.

3.1 Solution of Spline Function

Theorem

If $H_R = H_0 \oplus H_1$ and $\phi_1,...,\phi_m$ are both defined in the H_0 space and $T_{n\times m}$ is the $n\times m$ order of full matrix that given by: $T_{n\times m} = \{L_i\phi_v\}, i = 1,2,...,n$ and $v = 1,2,...,m$. Then f which minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \langle\eta_i, f\rangle\right)^2 + \Lambda\left\|P_1 f\right\|_R^2$$

is $\hat{f} = \sum_{v=1}^{m}\alpha_v\phi_v + \sum_{i=1}^{n}\beta_i\xi_i$

$$= T\alpha + \Sigma\beta$$

Proof:

$\hat{f}$ can be written in the form of

$$\hat{f} = \sum_{v=1}^{m}\alpha_v\phi_v + \sum_{i=1}^{n}\beta_i\xi_i$$
$$= \phi'\alpha + \xi'\beta + \psi,\ \psi \in H_R$$

which perpendicular with $\phi_1,...,\phi_m,\xi_1,...,\xi_n,$ then it should be $\psi = 0$ such that

$$\hat{f} = \sum_{v=1}^{m}\alpha_v\phi_v + \sum_{i=1}^{n}\beta_i\xi_i$$
$$= \phi'\alpha + \xi'\beta$$

for $\langle\eta_i, f\rangle$ from Riesz representation

$$\{(L_i f)\} = \{\langle\eta_i, \hat{f}\rangle\}; \eta_i \in H_R = H_0 + H_1$$

$$(L_i f) = \{\langle\eta_i, \phi'\alpha + \xi'\beta\rangle\}$$
$$= \{\langle\eta_i, \phi'\alpha\rangle\} + \{\langle\eta_i, \xi'\beta\rangle\},\ i = 1,2,...,n$$
$$\phi'\alpha = \phi_1\alpha_1 + \phi_1\alpha_2 + ... + \phi_1\alpha_m$$
$$\xi'\beta = \xi_1\beta_1 + \xi_1\beta_2 + ... + \xi_1\beta_n$$

$$= \langle\eta_i, (\phi_1\alpha_1 + \phi_1\alpha_2 + ... + \phi_1\alpha_m) + (\xi_1\beta_1 + \xi_1\beta_2 + ... + \xi_1\beta_n)\rangle$$
$$= \langle\eta_i, \phi_1\alpha_1\rangle + ... + \langle\eta_i, \phi_1\alpha_m\rangle + \langle\eta_i, \xi_1\beta_1\rangle + ... + \langle\eta_i, \xi_1\beta_n\rangle$$

$$\eta_i \in H_R = H_0 + H_1$$
$$\eta_i = \eta_{0i} + \eta_{1i},\ \eta_{0i} \in H_0,\ \eta_{1i} \in H_1$$

$$= \langle\eta_{oi} + \eta_{1i}, \phi_1\alpha_1\rangle + ... + \langle\eta_{oi} + \eta_{1i}, \phi_1\alpha_m\rangle + \langle\eta_{oi} + \eta_{1i}, \xi_1\beta_1\rangle + ... + \langle\eta_{oi} + \eta_{1i}, \xi_1\beta_n\rangle$$

$\eta_{0i} \in H_0, \eta_{1i} \in H_1$, span from H_0, ξ_i span H_1

$$= \langle\eta_{0i}, \phi_1\rangle\alpha_1 + ... + \langle\eta_{0i}, \phi_1\rangle\alpha_m + \langle\eta_{1i}, \xi_1\rangle\beta_1 + ... + \langle\eta_{1i}, \xi_1\rangle\beta_n$$
$$\eta_{oi} \in H_0, H_R = H_0 + H_1 \qquad (6)$$

According to the definition of RKHS, it is guaranteed that the linear and limited function in H_R which is $\{\langle\eta_{0i}, \phi_v\rangle\} = \{L_i\phi_v\}$ and with the Riesz representation then $\{\langle\eta_{1i}, \xi_1\rangle\} = \{\langle\xi_i, \xi_j\rangle\}$ so that the equation (6) can be written in the form of:

$$\{L_i\phi_1\}\alpha_1 + \{L_i\phi_1\}\alpha_2 + ... + \{L_i\phi_1\}\alpha_m + \langle\xi_i, \xi_1\rangle\beta_1 + \langle\xi_i, \xi_1\rangle\beta_2 + ... + \langle\xi_i, \xi_1\rangle\beta_n$$

for $i = 1,2,...,n$.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\hat{f} = \begin{pmatrix} L_1\phi_1 & L_1\phi_2 & \cdots & L_1\phi_m \\ L_2\phi_1 & L_2\phi_2 & \cdots & L_2\phi_m \\ \vdots & \vdots & \ddots & \vdots \\ L_n\phi_1 & L_n\phi_2 & \cdots & L_n\phi_m \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} + \begin{pmatrix} \langle \xi_1\xi_1 \rangle & \langle \xi_1\xi_2 \rangle & \cdots & \langle \xi_1\xi_n \rangle \\ \langle \xi_2\xi_1 \rangle & \langle \xi_2\xi_2 \rangle & \cdots & \langle \xi_2\xi_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \xi_n\xi_1 \rangle & \langle \xi_n\xi_2 \rangle & \cdots & \langle \xi_n\xi_n \rangle \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

$$\hat{f} = T\alpha + \Sigma\beta \tag{7}$$

for $\|Pf\|_R^2$

$P : H_R \to H_1$ furthermore $\|Pf\|_R^2$ can be written as

$$\begin{aligned} \|Pf\|_R^2 &= \langle Pf, Pf \rangle \\ &= \langle P(\phi'\alpha + \xi'\beta), P(\phi'\alpha + \xi'\beta) \rangle \\ &= \langle P\phi'\alpha + P\xi'\beta, P\phi'\alpha + P\xi'\beta \rangle, \\ &= \langle 0 + P\xi'\beta, 0 + P\xi'\beta \rangle = \langle P\xi'\beta, P\xi'\beta \rangle \\ &= \langle \xi'\beta, \xi'\beta \rangle \\ &= (\xi'\beta)'(\xi'\beta) \\ &= \beta'\xi\xi'\beta; \ \xi\xi' = \Sigma = \{\langle \xi_i, \xi_j \rangle\} \\ &= \beta'\Sigma\beta \end{aligned} \tag{8}$$

3.2 Solution of Kernel Function

The Nadaraya-Watson Kernel estimator is a special case of the local polynomial regression curve, that is the local polynomial regression curve which has equal order to 0 or also called the local constant regression curve. When the local polynomial regression curve has an order equal to one, then the local polynomial regression curve is also called the local linear regression curve. The local polynomial regression curve adopts the expansion of the Taylor series around *t*. If a regression curve $g(t)$ is approached by a local polynomial regression curve

$$\begin{aligned} g(t_i) &= \beta_0 + \beta_1(t_i - t) + \beta_2(t_i - t)^2 + \ldots + \beta_p(t_i - t)^p \\ &= \sum_{k=0}^{p} \beta_k (t_i - t)^k \end{aligned}$$

Related to the local polynomial regression model, the Nadayara-Watson Kernel regression model is a local polynomial regression model that only contains local constant. So, if the regression function *g* only contain the local constant, then by minimizing the function

$$L = \sum_{i=1}^{n} (y_i - \beta_0)^2 K\left(\frac{t_i - t}{\tau}\right) \tag{9}$$

will be resulting

$$\beta_0 = \sum_{i=1}^{n} \frac{K\left(\frac{t_i - t}{\tau}\right)}{\sum_{i=1}^{n} K\left(\frac{t_i - t}{\tau}\right)} y_i$$

so that

$$\begin{aligned} \hat{g}_\tau(t) &= n^{-1} \sum_{i=1}^{n} \frac{K_\tau(t_i - t)}{n^{-1}\sum_{i=1}^{n} K_\tau(t_i - t)} y_i \\ &= n^{-1} \sum_{i=1}^{n} W_{\tau i}(t) y_i \end{aligned} \tag{10}$$

The $W_{\tau i}(t)$ function is a weighted function,

$$W_{\tau i}(t) = \frac{K_\tau(t - t_i)}{n^{-1}\sum_{j=1}^{n} K_\tau(t_i - t)}$$

where $K_{\tau_j}(t_j - t_{ji})$ is the Kernel function of $K_\tau(t_i - t) = \frac{1}{\tau} K\left(\frac{t_i - t}{\tau}\right)$. The Kernel function is a function that is real, continuous, limited and symmetrical, with its integral is equal to one. The Kernel function can be a uniform Kernel, triangle Kernel, Epanechnikov Kernel, squared Kernel, tri-weight Kernel, cosine Kernel and Gaussian Kernel [22]. The Gaussian Kernel is quite often used in many studies. Gaussian Kernel function is smoother than the other kernel functions. The form of the Gaussian kernel function is

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right), \quad -\infty < t < \infty. \tag{11}$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

If the addition form in the equation (10) is explained then

$$\hat{g}_\tau(t_i) = n^{-1}W_{\tau 1}(t)\,y_1 + \ldots + n^{-1}W_{\tau n}(t)\,y_n \qquad (12)$$

From the equation of the kernel function (12), applies to every $t = t_1$ up to $t = t_n$ then .

$$\begin{bmatrix} \hat{g}_\tau(t_1) \\ \hat{g}_\tau(t_2) \\ \vdots \\ \hat{g}_\tau(t_n) \end{bmatrix} = \begin{bmatrix} n^{-1}\sum_{i=1}^{n} W_{\tau i}(t_1)y_i \\ n^{-1}\sum_{i=1}^{n} W_{\tau i}(t_2)y_i \\ \vdots \\ n^{-1}\sum_{i=1}^{n} W_{\tau i}(t_n)y_i \end{bmatrix}$$

$$= \begin{bmatrix} n^{-1}W_{\tau 1}(t_1)y_1 + \cdots + n^{-1}W_{\tau n}(t_1)y_n \\ n^{-1}W_{\tau 1}(t_2)y_1 + \cdots + n^{-1}W_{\tau n}(t_2)y_n \\ \vdots \\ n^{-1}W_{\tau 1}(t_n)y_1 + \cdots + n^{-1}W_{\tau n}(t_n)y_n \end{bmatrix}$$

$$= \begin{bmatrix} n^{-1}W_{\tau 1}(t_1) & n^{-1}W_{\tau 2}(t_1) & \cdots & n^{-1}W_{\tau n}(t_1) \\ n^{-1}W_{\tau 1}(t_2) & n^{-1}W_{\tau 2}(t_2) & \cdots & n^{-1}W_{\tau n}(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ n^{-1}W_{\tau 1}(t_n) & n^{-1}W_{\tau 2}(t_n) & \cdots & n^{-1}W_{\tau n}(t_n) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= Dy \qquad (13)$$

3.3 Penalized Spline-Kernel

Based on the equation (7), equation (8) and (13) estimation of the penalized Spline-Kernel nonparametric regression curve is presented as follows:

$$\Re(\alpha,\beta) = \frac{1}{n}\left\|(y - T\alpha - \Sigma\beta - Dy)\right\|^2 + \Lambda\beta'\Sigma\beta \qquad (14)$$

Next, will be determined α and β to minimized (14)

$$\Re(\alpha,\beta) = \left\|(y - T\alpha - \Sigma\beta - Dy)\right\|^2 + n\Lambda\beta'\Sigma\beta \qquad (15)$$

By derived partially equation (15) to β and the results is equated to zero then obtained:

$$\frac{\partial\Re}{\partial\beta} = -\Sigma'y + \Sigma'T\alpha + (\Sigma'\Sigma + n\Lambda\Sigma)\beta = 0 \qquad (16)$$

If the whole equation is completed, it is obtained

$$\beta = M^{-1}(y - T\alpha) \text{ with } M = \Sigma + n\Lambda I \qquad (17)$$

In a similar way, the equation (15) derived to α and equated to zero then obtained

$$\frac{\partial\Re}{\partial\alpha} = -T'y + T'T\alpha + T'\Sigma\beta = 0 \qquad (18)$$

If the equation (17) and (18) are solved simultaneously obtained:

$$\alpha = \left(T'M^{-1}T\right)^{-1}T'M^{-1}(I - D)\,y \qquad (19)$$

$$\beta = M^{-1}\left(I - T\left(T'M^{-1}T\right)^{-1}T'M^{-1}\right)(I - D)\,y \qquad (20)$$

From the equation (19) and (20) the estimator for the Spline smoothing component us obtained as follow:

$$\begin{aligned} \hat{f}_\Lambda &= T\alpha + \Sigma\beta \\ &= \left\{T\left(T'M^{-1}T\right)^{-1}T'M^{-1}(I - D) + \right. \\ &\quad \left. \Sigma M^{-1}\left(I - T\left(T'M^{-1}T\right)^{-1}T'M^{-1}\right)(I - D)\right\}y \\ &= \Theta_\Lambda y \end{aligned} \qquad (21)$$

while for Kernel component:

$$\hat{g}_\tau = Dy \qquad (22)$$

Based on the equation (21) and (22), the penalized Spline-Kernel estimator will be obtained as follow

$$\begin{aligned} \hat{f}^*_{\Lambda,\tau} &= T\alpha + \Sigma\beta + Dy \\ &= \left\{T\left(T'M^{-1}T\right)^{-1}T'M^{-1}(I - D) + \right. \\ &\quad \Sigma M^{-1}\left(I - T\left(T'M^{-1}T\right)^{-1}T'M^{-1}\right) \\ &\quad \left.(I - D) + D\right\}y \\ &= \Phi_{\Lambda,\tau}y \end{aligned} \qquad (23)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. SMOOTHING PARAMETER SELECTION METHODS

In producing a good estimation of regression curve, the selection of the optimal smoothing parameter of Λ and τ is an important thing. By using the Reproduction Kernel Hilbert Space to estimate the regression curve, then the optimal values of Λ and τ will be selected.

In the nonparametric regression model with one response, [2] shows that if the smoothing parameter value of Λ,τ is very small $(\Lambda,\tau \to 0)$, it will provide a very rough regression curve estimator. If the smoothing parameter value Λ,τ is very large $(\Lambda,\tau \to \infty)$, it will produce a very smooth nonparametric regression curve estimator.

As a result of that, the optimal smoothing parameter of Λ,τ will be selected to obtain the most suitable estimator for the data.

For the purpose of selecting optimal smoothing parameters Λ,τ, several methods have been developed in nonparametric and semiparametric regression, [2] providing a generalized cross validation (GCV) method.

The following method will be designed to select smoothing parameters.

Mean Square Error (MSE) of this estimator is given by:

$$\begin{aligned} MSE(\Lambda,\tau) &= \frac{1}{n}\left(y-\hat{f}^*\right)'\left(y-\hat{f}^*\right) \\ &= \frac{1}{n}\left(y-\Phi_{\Lambda,\tau}y\right)'\left(y-\Phi_{\Lambda,\tau}y\right) \\ &= \frac{1}{n}\left[\left(I-\Phi_{\Lambda,\tau}\right)y\right]'\left[\left(I-\Phi_{\Lambda,\tau}\right)y\right] \\ &= \frac{1}{n}y'\left(I-\Phi_{\Lambda,\tau}\right)'\left(I-\Phi_{\Lambda,\tau}\right)y \\ &= \frac{1}{n}\left\|\left(I-\Phi_{\Lambda,\tau}\right)y\right\|^2 \end{aligned}$$

Furthermore, the following quantity is defined as:

$$G(\Lambda,\tau) = \frac{n^{-1}\left\|\left(I-\Phi_{\Lambda,\tau}\right)y\right\|^2}{\left[n^{-1}\mathrm{tr}\left(I-\Phi_{\Lambda,\tau}\right)\right]^2} \qquad (24)$$

5. EMPIRICAL STUDY

Indonesia has officially become an oil importing country since 2004. This is due to a decrease in the level of oil production, on the other hand, the level of oil consumption continues to increase. One of the most crucial of fuel oil products is Premium. The premium is one of the three subsidized fuel products and are most demanded by the society. According to Downstream Regulatory Agency (DRA) for oil and gas data, this type of fuel consumption always increases every year. This is different from the other two types of subsidized fuel products which has a downtrend.

Although premium is the most consumed fuel in Indonesia, the demanding problems have not much resolved. This can be seen from the considerable difference between estimated of useable premium given by the government every year with the realization of its consumption. From 2007-2012, there was an average difference of 9.44%. Some researchers have conducted research related to the fuel oil, including [23], [24], [25]. Then the use of Multiple Linear Regression to predict energy needs has been done by a lot of researchers in various directions. In Italy, [26], use data historical electricity and fuel consumption, gross domestic product, gross domestic product per capita, number of cars and population as an independent variable in predicting the Italian electrical energy consumption until 2040. [27], predict the energy consumption in Turkey use four predictor variables, namely gross domestic product, population, fuel price disparity, and number of vehicles. [28] used predictors instantaneous speed and accelerate levels to estimating vehicles fuel consumption.

Whereas in New Zealand, [29] use multiple linear regression methods to predict electricity energy consumption until 2015, using three predictor variables, namely fuel price disparity, population and average selling price of electricity. The use of predictors of number of cars, population, number of exports and imports, conducted by [30] in estimating energy demand at South Korea. From the literature study conducted, it appears that the researchers most use parametric regression in knowing the relationship between fuel consumption and predictor variables that influence it.

In determining influential predictors, in addition to the literature study, this study also uses the knowledge acquisition method to experts directly related to premium consumption. In general, knowledge acquisition from experts can be done with two techniques which are the Expert Group Discussion (EGD) and Delphi Method (DM). The Expert Group Discussion is a discussion process involving experts to identify problems,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

analyze causes of problems, determine ways to solve problems, and propose various alternative solutions to problems by considering available resources. Thus, based on the research that has been done before and input from experts, then in this study two predictor variables that influence fuel consumption are used, which are the fuel price disparity and the number of vehicles. However, what distinguishes it from previous studies is the use of nonparametric regression model in this study.

To defined the penalized Spline-Kernel model using two predictors mentioned above, the historical past data for 12 years (data from 2001 to 2012), will be used in this study. The exploration of data based on complete descriptive statistics can be seen in Table 1.

Table 1: Descriptive Statistics of Data

Variable	Mean	StDev	Minimum	Maximum
Y	18.93	4.74	13.07	28.26
X1	1.913	1.474	0.490	5.080
X2	6.244	2.413	3.130	9.890

Based on the table, it is known that the average amount of fuel consumption from 2004 to 2015 is 18.93×10^6 kl. While the highest and lowest of fuel consumption are 28.26×10^6 kl and 13.07×10^6 kl, respectively. Similarly, the average of fuel price disparity is 1,913 thousand rupiahs with the highest price disparity is 5.080 and the lowest is 1.560. The disparity being 5,080 and the lowest is 1,560. The average number of cars between 2004 and 2015 was 6,244 million units with the highest number of car units was 9,890 million units and the lowest was 3,130 million units.

The next step is to test whether the pattern of the relationship between fuel price disparity, the number of cars to the fuel consumption, in the form of linear or non-linear relationships.

Test result of Ramsey Test can be shown in the following table.

Table 2: Result of Ramsey Test

Relationship	*p-value*	Remark
x_1 toward y	0.0394	non-linier
x_2 toward y	0.0445	non-linier
x_1, x_2 toward y	0.0261	non-linier

Based on the Table 2 can be concluded that the relationship between fuel price disparity, and the number of cars uses towards the fuel consumption is a non-linear relationship.

Since the pattern of the relationship fuel price disparity variable, and the number of cars uses towards the fuel consumption and the non-linear function is unknown, then it will be modeled using nonparametric regression.

For models with predictor variables fuel price disparity and the number of cars which modeled with a penalized Spline-Kernel, the optimum of smoothing and bandwidth parameters will be determined by choosing the value of the minimum GCV. The GCV's values, smoothing and different bandwidth parameters can be seen in Table 3.

Table 3: Result of GCV value

Smoothing Parameter		GCV
Λ	τ	
0.25	0.088	0.3778
0.26	0.089	0.3538
0.27	0.090	0.2692
0.28	0.091	0.2413
0.29	0.092	0.2264
0.30	**0.093**	**0.2183**
0.31	0.094	0.2359
0.32	0.095	0.3186
0.33	0.096	0.3924
0.34	0.097	0.3936
0.35	0.098	0.3940
⋮	⋮	⋮
1.25	0.187	0.3767

Based on Table 3, it is seen that the minimum GCV value is 0.2183 with smoothing and optimal bandwidth parameter values $\Lambda = 0.3$ and $\tau = 0.093$, respectively. The value R^2 obtained is 0.9314 or 93.14% of the model obtained can describe the relationship between fuel price disparity and number of cars with fuel consumption.

Journal of Theoretical and Applied Information Technology
31st January 2019. Vol.97. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

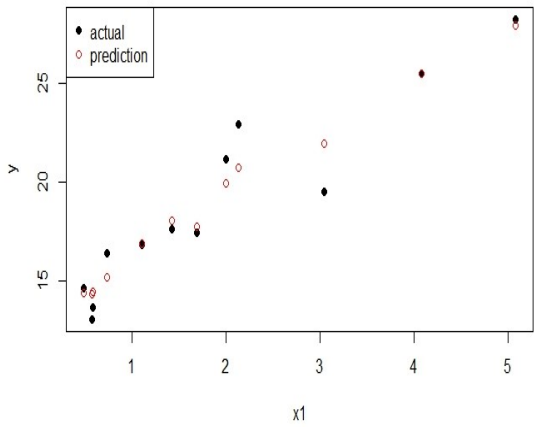

Figure 1: Plot Between Fuel Price Disparity vs Fuel Consumption

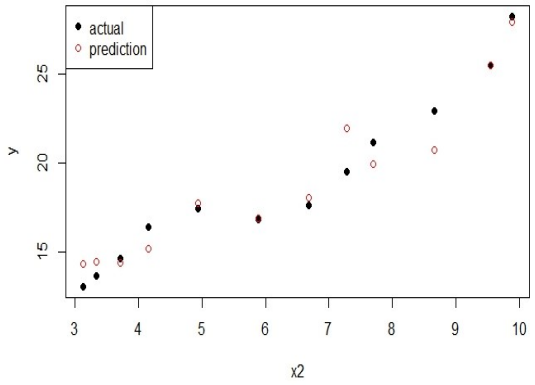

Figure 2: Plot Between Number of Cars vs Fuel Consumption

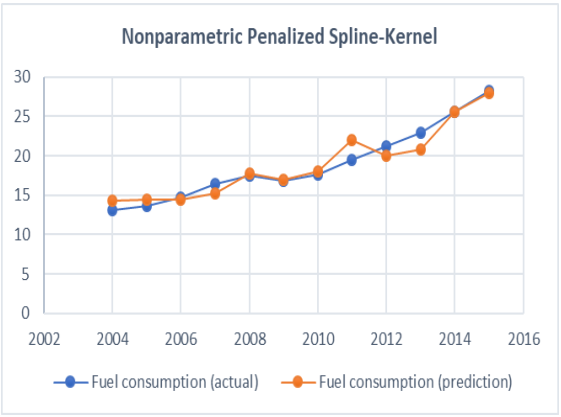

Figure 3: Plot Fuel Consumption Actual and Prediction

The fuel consumption data is also modeled using parametric regression to compare it with the result of nonparametric regression model developed in this study. The results between the actual data and predictive data using the parametric regression model are shown in the following graph.

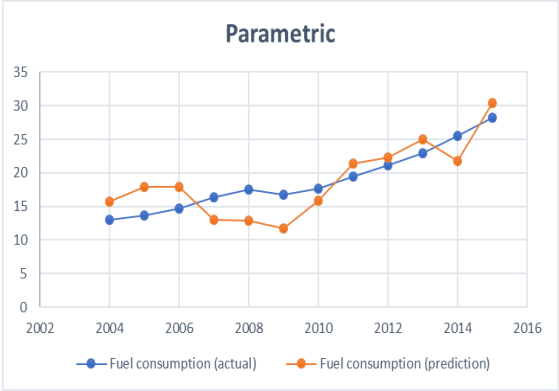

Figure 4: Plot Fuel Consumption Actual and Prediction

From the results of the analysis using parametric regression model, obtained the value of R^2 as 0.7823 or 78.23% of the model obtained can describe the relationship between fuel price and number of cars with fuel consumption. Based on the value of the determination coefficient, it can be concluded that the regression model developed in this study is better when compared with the parametric regression model.

6. MODEL VALIDATION

The next step that needs to be done is validate the model, to see the accuracy of the model in predictions. The process carried out is a cross validation evaluation by eliminating one or two out cross validations on each subject. Brief cross validation results are summarized in Table 4.

Table 4: Result of Cross Validation

Obs	Actual	Estimator			
		Leave two out		Leave one out	
		Predic	Resid	Predic	Resid
11	25.52	25.522	-0.002		
12	28.26	27.951	0.309	28.11	0.15

The results of cross validation estimation in Table 4 by removing one last observation obtained the MSE value of 0.0225, whereas the results of cross validation by removing the last two observations obtained the MSE value of 0.0477. Thus, it can be concluded that the resulting model is valid and able to describe the real phenomena that exist.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

7. CONCLUSION

From the analysis and discussion that has been done above, some conclusion can be deduct as follows:

1. If given the data (x_i, y_i) and the relationship between x_i and y_i assumed to follow the regression model

$$y_i = L_i f + \varepsilon_i$$

with $f \in H$ and L_i is a linear functional and limited to the H and H has decomposition $H = H_0 \oplus H_1$ then the estimate of f is $f \in H$ which minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - L_i f\right)^2 + \Lambda \left\| P_1 f \right\|_R^2$$

and written in the following form

$$\hat{f}_\Lambda = \sum_{v=1}^{m} \alpha_v \phi_v + \sum_{i=1}^{n} \beta_i \xi_i = T\alpha + \Sigma\beta$$

2. The developing of the Spline model with additional Kernel function can be obtained from the following solution function

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - L_i f - g\right)^2 + \Lambda \left\| P_1 f \right\|_R^2$$

and produce the following result

$$\begin{aligned}\hat{f}^*_{\Lambda,\tau} &= T\alpha + \Sigma\beta + Dy \\ &= \left\{ T\left(T'M^{-1}T\right)^{-1} T'M^{-1}\left(I - D\right) + \right. \\ &\quad \Sigma M^{-1}\left(I - T\left(T'M^{-1}T\right)^{-1} T'M^{-1}\right) \\ &\quad \left.\left(I - D\right) + D \right\} y \\ &= \Phi_{\Lambda,\tau} y\end{aligned}$$

3. The selection of the smoothing and bandwidth parameter in developing the penalized Spline-Kernel model can be obtained by the GCV method:

$$G(\Lambda,\tau) = \frac{n^{-1}\left\| \left(I - \Phi_{\Lambda,\tau}\right) y \right\|^2}{\left[n^{-1} \mathrm{tr}\left(I - \Phi_{\Lambda,\tau}\right) \right]^2}$$

REFRENCES

[1] R. L. Eubank, "*Spline Smoothing and Nonparametric Regression*," New York: CRC Press, 1999.

[2] A. A. R. Fernandes, I. N. Budiantara, B. W. Otok, Suhartono, "Reproducing Kernel Hilbert Space and Penalized Weighted Least Square in Nonparametric Regression," Vol. 8, No. 146, 2014, pp. 7289-7300.

[3] G. Wahba, "*Spline Models for Observational Data*," Vol. 59, SIAM: Pennsylvania, 1990.

[4] S. N. Wood, "On Confidence Intervals for Generalized Additive Models Based on Penalized Regression Spline", *Aus. N. Z. J. Stat*., Vol. 48, 2006, pp. 445-464.

[5] H. Becher, G. Kauermann, P. Khomski, and B. Kouyate, "Using Penalized Splines to Model Age and Season of Birth Dependent Effects of Childhood Mortality Risk Factors in Rural Burkina Faso," *Biometrical Journal*, Vol. 51, No. 1, 2009, pp. 110-122.

[6] I. N. Budiantara, M. Ratna, I. Zain, I., and W. Wibowo, "Modeling the Percentage of Poor People in Indonesia Using Spline Nonparametric Regression Approach", *International Journal Basic & Applied Sciences*, Vol. 12, 2012, pp. 118-124.

[7] D. D. Cox, and F. O'Sullivan, "Penalized Type Estimator for Generalized Nonparametric Regression", *Journal of Multivariate Analysis*, Vol. 56, 1996, pp. 185-206.

[8] R., Ng., P. Koenker, and S. Portnoy, "Quantile Smoothing Spline", *Biometrika*, Vol. 81, 1994, pp. 673-680.

[9] A. A. R. Fernandes, I. N. Budiantara, B. W. Otok, and Suhartono, "Spline Estimator for Bi-responses Nonparametric Regression Model for Longitudinal Data," *Applied Mathematical Sciences*, Vol 8, No. 114, 2014, pp. 5653-5665.

[10] G. Qin, Z. Zhu, and W. K. Fung, "Robust Estimation of Covariance Parameters in Partial Linear Model for Longitudinal Data, *Journal of Statistical Planning and Inference*, Vol. 139, 2009, pp. 558 – 570.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[11] Y.R. Yue, D. Simpson, F. Lindgren, and H. Rue, "Bayesian Adaptive Smoothing Spline using Stochastic Differential Equation", *Bayesian Analysis*, 2014, pp. 397-424.

[12] E. A. Nadayara, "On Estimating Regression," *Theory of Probability and Its Applications*, Vol. 9, No. 1, 1964, pp. 141-142.

[13] G. S. Watson, "Smooth Regression Analysis," *Sankhya: The Indian Journal of Statistics, series A*, Vol. 26, No. 4, 1964, pp. 359-372.

[14] W. Hardle, "*Applied Nonparametric Regression*," Cambridge University Press, New York, 1990.

[15] B. Lestari, I. N. Budiantara, S. Sunaryo, and M. Mashuri, "Spline Smoothing Estimator in Multi-response Nonparametric Regression with Unequal Correlation of Errors", *Journal of Mathematics and Statistics*, Vol. 6, No. 3, 2010, pp. 327-332.

[16] J. S. KlemelA, "*Multivariate Nonparametric Regression and Visualization: with R and Applications to Finance*," New Jersey: John Wiley and Sons, 2014.

[17] X. Lin, N. Wang, A. H. Welsh, and R. J. Carrol, "Equivalent Kernels of Smoothing Spline in Nonparametric Regression for Clustered/ Longitudinal Data," *Biometrika*, Vol. 91, No. 1, 2004, pp. 177-193.

[18] Rismal, I. N. Budiantara, "Mixture Model of Spline Truncated and Kernel in Multivariable Nonparametric Regression", *AIP Conference Proceedings*, 2016, pp. 1-7.

[19] G. Campos, D. Gianola, and G. J. Rosa, "Reproducing Kernel Hilbert Spaces Regression: A General Framework for Genetic Evaluation", *Journal of Animal Science*, Vol. 6, No. 11, 2009, pp. 1-19.

[20] C. J. Nuzman and H. V. Poor, "Reproducing Kernel Hilbert Space Methods for Wide-Sense Self-Similar Processes", *The Annals of Applied Probability*, Vol. 11, No. 4, pp. 1199-1219.

[21] M. Kayri, and G. Zirhhoglu, "Kernel Smoothing Function and Choosing Bandwitdh for Nonparametric Regression Methods," *Ozean Journal of Applied Sciences*, Vol. 2, No. 1, 2009, pp. 49-54.

[22] H. Okumura, and K. Naito, "Non-Parametric Kernel Regression for Multinomial Data", *Journal of Multivariate Analysis*, Vol. 97, 2006, pp. 2009-2022.

[23] A. Sözen and E. Arcaklioglu, "Prediction of Net Energy Consumption Based on Economic Indicators (GNP and GDP) in Turkey", *Energy Policy*, Vol. 35, 2007, pp.: 4981 -4992.

[24] L. Jinke, S. Huang, and G. Dianming, "Causality Relationship between Coal Consumption and GDP: Difference of Major OECD and non-OECD Countries", *Applied Energy*, Vol. 85, 2008, pp. 421-429

[25] I. Ozturk, and A. Acaravci, "The Casual Relationship between Energy Consumption and GDP in Albania, Bulgaria, Hungary, and Romania: Evidence from ARDL Bound Testing Approach", *Applied Energy*, Vol. 87, 2008, pp. 1938-1943

[26] V. Bianco, O. Manca, and S. Nardini, "Electricity Consumption Forecasting in Italy Using Linear Regression Models", *Energy*, Vol. 34, 2009, pp. 1413-1421.

[27] M. Kankal, A. Akpinar, M. I. Komurcu, and T. S. Ozsahin, "Modeling and Forecasting of Turkey's Energy Consumption using Socio-Economic and Demographic Variables", *Applied Energy*, Vol. 88, 2011, pp. 1927-1939.

[28] K. Ahn, H. Rakha, A. Trani, and M. Van Aerde, "Estimating Vehicle Fuel Consumption and Emissions based on Instantaneous Speed and Acceleration Levels", Journal of Transportation Engineering, Vol. 128, No. 2, 2002, pp. 182–190, 2002.

[29] Z. Mohamed, and P. Bodger, "Forecasting Electricity Consumption in New Zealand using Economic and Demographic Variables", *Energy*, Vol. 30, 2005, pp.: 1833-1843.

[30] Z.W. Geem, and W.E. Roper, "Energy Demand Estimation of South Korea using Artificial Neural Network", *Energy Policy*, Vol. 37, 2009, pp. 4049-4054.