International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 6
ISSN: 2088-5334

Nonparametric Regression Mixed Estimators of Truncated Spline and Gaussian Kernel based on Cross-Validation (CV), Generalized Cross-Validation (GCV), and Unbiased Risk (UBR) Methods

Vita Ratnasari [a,*], I Nyoman Budiantara [a], Andrea Tri Rian Dani [a]

[a] *Department of Statistics, Faculty of Science and Data Analytics, Sepuluh Nopember Institute of Technology, Surabaya, East Java, Indonesia*
Corresponding author: [*]vita.statistikaits@gmail.com

***Abstract*— Nowadays, most nonparametric regression research involves more than one predictor variable and generally uses the same type of estimator for all predictors. In the real case, each predictor variable likely has a different form of regression curve so that if it is forced, it can produce an estimation form that does not match the data pattern. Thus, it is necessary to develop a regression curve estimation model under the data pattern, namely the mixed estimator. The focus of this study is an additive nonparametric regression model, a mix of the Truncated Spline and Gaussian Kernel. There is a knot point in the Truncated Spline, while in the Gaussian Kernel, there is bandwidth. To choose the optimal knot point and bandwidth in a mixed estimator model, various methods can be used, including Cross-Validation (CV), Generalized Cross-Validation (GCV), and Unbiased Risk (UBR). This research proposes the optimal knot point and bandwidth estimation on the mixed estimator Truncated Spline and Gaussian Kernel model. Furthermore, the comparison between CV, GCV, and UBR is used to validate the proposed method. The simulation study was carried out by generating the Truncated Spline function and the Gaussian Kernel on a combination of sample size variations and variances. The simulation results show that the GCV method provides a higher coefficient of determination (R^2) value and better accuracy for each combination of sample sizes and variance variations.**

***Keywords*— Cross-validation; generalized cross-validation; mixed estimators; unbiased risk.**

Manuscript received 13 Feb. 2021; revised 21 May 2021; accepted 1 Jul. 2021. Date of publication 31 Dec. 2021.

I. INTRODUCTION

Regression analysis is one of the statistical methods used to determine the pattern of relationships between one or more variables in the functional form [1]. The relationship formed can be expressed in an equation that states the functional relationship between the response and predictor variables [2]. Early identification of the relationship pattern can be made by looking at the scatter plot [3]. If the form of the relationship pattern is known, then the parametric regression approach is used. However, not all data patterns can be clearly identified as the relationship pattern, so using nonparametric regression was proposed [4].

Along with the development of computing technology, nonparametric regression models, which generally require fairly difficult computational complexity, are being popular. The approach with the nonparametric regression model has advantages, such as being easy to use for data patterns with unknown patterns [2]. This approach has good flexibility so that the data is expected to adjust the form of regression curve estimation by itself without being influenced by the researcher's subjectivity [5]. The purpose of modeling using regression analysis is to find the appropriate form of regression curve estimation [6]. Many nonparametric regression curve estimators have been developed by researchers, including Spline [2], [7]-[11], Kernel [12]-[15], and Fourier series [16]-[20].

Models with the nonparametric regression approach developed by previous researchers assume the pattern of each predictor is considered to have the same regression curve so that only one estimator form is used for each predictor variable. However, in the real case, each predictor variable likely has a different form of regression curve. Thus, if it is enforced, it can produce an estimation form that does not match the data pattern [3], [21]. So, it is necessary to develop a mixed estimator of nonparametric regression curves, where each data pattern in the model is approximated by the appropriate curve estimator [1], [3], [4], [20]-[24]. The mixed

estimator nonparametric regression model used in this study combines the Truncated Spline and the Gaussian Kernel.

In nonparametric Gaussian Kernel regression, determining the right bandwidth is very important, while in the Truncated Spline, the important thing is determining the optimal knot point. The knot point and bandwidth in the future referred to as smoothing parameters can greatly affect the formed regression curve. The smoothing parameter that is too small can produce a very rough curve and tend to fluctuate. On the other hand, if it is too large, it can produce a curve that is too smooth which is not matched with the data pattern [13]. Thus, it becomes an interesting problem to determine the right and optimal smoothing parameter [1]. The optimal smoothing parameter can be determined using several methods, such as Cross-Validation (CV) [25], Generalized Cross-Validation (GCV) [7], and Unbiased Risk (UBR) [26].

In this research, the simulation of relationship pattern form between the response and predictor variables that follow the Truncated Spline and Kernel pattern characteristics is proposed. This proposed method is validated using many combinations of sample size variation and variance. Furthermore, the relationship pattern from the simulation data results was modeled using a mixed estimator of the Truncated Spline and Gaussian Kernel. Cross-Validation (CV), Generalized Cross-Validation (GCV), and Unbiased Risk (UBR) are used to determine the optimal smoothing parameter. This research aims to compare the performance of the CV, GCV, and UBR methods in estimating the optimal knot point and bandwidth in the mixed estimator model of nonparametric regression. Moreover, the coefficient of determination (R^2) was used as the criteria for goodness.

The structure of this paper is organized as follows: Brief explanation of material, such as the Truncated Spline, Gaussian Kernel, Mixed Estimator Nonparametric Regression, Cross-Validation (CV), Generalized Cross-Validation (GCV), Unbiased Risk (UBR), and Research Methodology in Section II. Simulation results of several case studies are given in Section III, and Conclusions are given in Section IV.

II. Material and Methods

A. Truncated Spline

The spline is a segmented polynomial model. The Truncated Spline function is a function that still maintains the properties of the polynomial function [6]. In general, a Truncated Spline nonparametric regression model can be written as follows:

$$y_i = f(x_i) + \varepsilon_i \tag{1}$$

where $f(x_i)$ is the Truncated Spline function with degrees m and ϕ is the knot point:

$$f(x_i) = \sum_{j=0}^{m} \beta_j x_i^j + \sum_{k=1}^{r} \beta_{m+k} (x_i - \phi_k)_+^m \tag{2}$$

Suppose given a paired data x_i and y_i, where $i=1,2,...,n$ follows a Truncated Spline nonparametric regression model:

$$y_i = \sum_{j=0}^{m} \beta_j x_i^j + \sum_{k=1}^{r} \beta_{m+k} (x_i - \phi_k)_+^m + \varepsilon_i \tag{3}$$

with a Truncated Spline function:

$$(x_i - \phi_k)_+^m = \begin{cases} (x_i - \phi_k)^m & x \geq \phi_k \\ 0 & x < \phi_k \end{cases} \tag{4}$$

the regression model in Equation 4 can be written in matrix form as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1\ x_1 \cdots x_1^m \vdots (x_1 - \phi_1)_+^m \cdots (x_1 - \phi_r)_+^m \\ 1\ x_2 \cdots x_2^m \vdots (x_2 - \phi_1)_+^m \cdots (x_2 - \phi_r)_+^m \\ \vdots\ \vdots\ \ddots\ \vdots\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ 1\ x_n \cdots x_n^m \vdots (x_n - \phi_1)_+^m \cdots (x_n - \phi_r)_+^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{m+r} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

so,

$$\mathbf{y} = \mathbf{X}(\phi)\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5}$$

Where $\mathbf{y}$ is the vector of the response variable of size (n x 1), $\mathbf{X}(\phi)$ is a matrix of size n x ($m + r + 1$), $\boldsymbol{\beta}$ is the vector of the regression coefficient parameter to be estimated and size ($m + r + 1$) x 1, and $\boldsymbol{\varepsilon}$ is a random error vector of size (n x 1).

B. Gaussian Kernel

Suppose given a paired data t_i and y_i, where $i=1,2,...,n$. So, a Gaussian Kernel nonparametric regression model can be written:

$$y_i = h(t_i) + \varepsilon_i \tag{6}$$

The Kernel estimator has advantages, such as flexible, easy mathematical form, and faster convergence [13].

The regression curve of $h(t_i)$ it to be approximated by the Kernel Function, the regression curve estimation can be presented in Equation 7.

$$\begin{aligned} \hat{h}_\alpha(t) &= n^{-1} \sum_{i=1}^{n} \left[\frac{K_\alpha(t - t_i)}{n^{-1} \sum_{i=1}^{n} K_\alpha(t - t_i)} \right] y_i \\ &= n^{-1} \sum_{i=1}^{n} W_{\alpha i}(t) y_i \end{aligned} \tag{7}$$

where:

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

and,

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

The kernel function used is the Gaussian Kernel:

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}\left(-t^2\right) \right); I_{[-\infty,\infty]}(t) \tag{8}$$

Based on the kernel function in Equation 7 that applies to each $t=t_1$, $t=t_2$, ..., $t=t_n$, then it can be written in matrix form as follows:

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

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

$$= \mathbf{G}(\alpha)\mathbf{y} \tag{9}$$

Where $\mathbf{y}$ is the vector of the response variable of size (n x 1), G(α) is a matrix of size (n x n).

C. Mixed Estimator Nonparametric Regression

The nonparametric regression mixed estimator is a multipredictor nonparametric regression model whose regression curve is additive, where the regression curve was approximated by two or more types of estimators [3], [27].

For example, given paired data (x_i ,t_i ,y_i) where the relationship between the predictor variables (x_i , t_i) and the response variable (y_i) follows a nonparametric regression model.

$$y_i = \mu(x_i, t_i) + \varepsilon_i \tag{10}$$

And then, the regression curve $\mu(x_i, t_i)$ is assumed to be additive such that $\mu(x_i, t_i)$ can be written into the form:

$$\mu(x_i, t_i) = f(x_i) + h(t_i) \tag{11}$$

The mixed estimator model used in this study is a combination of the Truncated Spline and the Gaussian Kernel. Furthermore, Equation 10 can be written in matrix based on Equation 5 and 9 form as follows:

$$\mathbf{y} = \mathbf{X}(\phi)\boldsymbol{\beta} + \mathbf{G}(\alpha)\mathbf{y} + \boldsymbol{\varepsilon} \tag{12}$$

Error can be written as follows:

$$\begin{aligned} \boldsymbol{\varepsilon} &= \mathbf{y} - \left[\mathbf{X}(\phi)\boldsymbol{\beta} + \mathbf{G}(\alpha)\mathbf{y}\right] \\ &= \left[\mathbf{I} - \mathbf{G}(\alpha)\right]\mathbf{y} - \mathbf{X}(\phi)\boldsymbol{\beta} \end{aligned} \tag{13}$$

The estimation of $\boldsymbol{\beta}$ can be obtained through LS optimization as follows:

$$\min_{\boldsymbol{\beta}} \left\{ \left(\left[\mathbf{I} - \mathbf{G}(\alpha)\right]\mathbf{y} - \mathbf{X}(\phi)\boldsymbol{\beta}\right)^T \left(\left[\mathbf{I} - \mathbf{G}(\alpha)\right]\mathbf{y} - \mathbf{X}(\phi)\boldsymbol{\beta}\right) \right. \tag{14}$$

Based on Equation 14, the sum squared of error can be written:

$$Q(\boldsymbol{\beta}) = \begin{pmatrix} \left\| \left[\mathbf{I} - \mathbf{G}(\alpha)\right]\mathbf{y} \right\|^2 - 2\boldsymbol{\beta}^T \mathbf{X}(\phi)^T \left[\mathbf{I} - \mathbf{G}(\alpha)\right]\mathbf{y} + \\ \boldsymbol{\beta}^T \mathbf{X}(\phi)^T \mathbf{X}(\phi)\boldsymbol{\beta} \end{pmatrix} \tag{15}$$

To get an estimator of $\boldsymbol{\beta}$, obtained by using a partial derivative of $Q(\boldsymbol{\beta})$ to $\boldsymbol{\beta}$ as follows:

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}(\phi)^T \left[\mathbf{I} - \mathbf{G}(\alpha)\right]\mathbf{y} + 2\mathbf{X}(\phi)^T \mathbf{X}(\phi)\boldsymbol{\beta} \tag{16}$$

And then, Equation 16 equal to zero. Estimate results from $\hat{\boldsymbol{\beta}}$ is:

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}(\phi)^T \mathbf{X}(\phi)\right]^{-1} \mathbf{X}(\phi)^T \left[\mathbf{I} - \mathbf{G}(\alpha)\right]\mathbf{y} \tag{17}$$

Equation 17 can be written as:

$$\hat{\boldsymbol{\beta}} = \mathbf{V}(\phi, \alpha)\mathbf{y} \tag{18}$$

with $\mathbf{V} = \left[\mathbf{X}(\phi)^T \mathbf{X}(\phi)\right]^{-1} \mathbf{X}(\phi)^T \left[\mathbf{I} - \mathbf{G}(\alpha)\right]$.

Thus, the Truncated Spline component estimator can be written as:

$$\hat{f}(x_i) = \mathbf{X}(\phi)\hat{\boldsymbol{\beta}} \tag{19}$$

so $\mathbf{X}\left[\mathbf{X}(\phi)^T \mathbf{X}(\phi)\right]^{-1} \mathbf{X}(\phi)^T \left[\mathbf{I} - \mathbf{G}(\alpha)\right]\mathbf{y}$.

Equation 19 can be written as $\hat{f}(x_i) = \mathbf{S}(\phi, \alpha)\mathbf{y}$.

As a result, the Gaussian Kernel estimator can be written as $\hat{h}_\alpha(t_i) = \mathbf{G}(\alpha)\mathbf{y}$.

Based on Equation 17 and the estimator form of each component, a mixed estimator of the nonparametric regression Truncated Spline and Gaussian Kernel were as follows:

$$\begin{aligned} \hat{\mu}_{\phi,\alpha}(x_i, t_i) &= \hat{f}_\phi(x_i) + \hat{h}_\alpha(t_i) \\ &= \left[\mathbf{S}(\phi, \alpha) + \mathbf{G}(\alpha)\right]\mathbf{y} \\ &= \mathbf{B}(\phi, \alpha)\mathbf{y} \end{aligned} \tag{20}$$

Matrix $\mathbf{B}(\phi, \alpha)$ is highly dependent with smoothing parameter (knot point and bandwidth).

D. Cross-Validation (CV)

Cross-Validation (CV) is a method developed by Craven and Wahba [25]. The formula developed is still limited to a single estimator form. Furthermore, the CV method can also be generalized to the mixed estimator form. The modified CV method formula for the mixed estimator form can be written:

$$CV(\phi_{opt}, \alpha_{opt}) = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{(y_i - \hat{y}_i)}{\left(1 - \left[\mathbf{B}_{ii}(\phi, \alpha)\right]\right)} \right]^2 \tag{21}$$

The CV method does not require variance information σ^2. The matrix $\mathbf{B}(\phi, \alpha)$ can be searched based on Equation 20. The CV method gives different weights to each observation according to its contribution [28].

E. Generalized Cross-Validation (GCV)

Generalized Cross-Validation (GCV) is a generalization of the CV method developed by Wahba [7]. The GCV method formula developed by Wahba is still limited to a single estimator form. The GCV method can be used in the form of a mixed estimator, where the modified GCV method formula can be written:

$$GCV(\phi_{opt}, \alpha_{opt}) = \left[\frac{MSE(\phi, \alpha)}{\left(n^{-1} trace\left\{\mathbf{I} - \left[\mathbf{B}(\phi, \alpha)\right]\right\}\right)^2} \right] \tag{22}$$

with,

$MSE(\phi,\alpha)$: $n^{-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

$\mathbf{I}$: Identity matrix

GCV method is same as the CV method that does not require variance information σ^2 [29]-[30].

F. Unbiased Risk (UBR)

Unbiased Risk (UBR) method was introduced by Wang [26] which can be used to determine the optimal smoothing parameter when information about σ^2 or σ^2 is known. The same thing as the CV and GCV methods, the UBR method can be generalized into a mixed estimator form so that it can be written:

$$UBR(\phi_{opt},\alpha_{opt}) = n^{-1}\left\{\begin{array}{l}\left\|\left(\mathbf{I}-\mathbf{B}(\phi,\alpha)\right)\mathbf{y}\right\|^2 + \\ \frac{\hat{\sigma}^2}{n}trace\left[\mathbf{I}-\mathbf{B}(\phi,\alpha)\right]^2 + \\ \frac{\hat{\sigma}^2}{n}trace\left[\mathbf{B}(\phi,\alpha)^2\right]^2\end{array}\right\} \quad (23)$$

with,

$$\hat{\sigma}^2 = \frac{\left\|\left(\mathbf{I}-\left[\mathbf{B}(\phi,\alpha)\right]\right)\mathbf{y}\right\|^2}{trace\left[\left(\mathbf{I}-\left[\mathbf{B}(\phi,\alpha)\right]\right)\mathbf{y}\right]} \quad (24)$$

G. Research Methodology

In this section, the step of the proposed method is presented as follows:

1. Defining a nonparametric regression model mixed estimator of the Truncated Spline and Gaussian Kernel
$$y_i = \mu(x_i,t_i) + \varepsilon_i \quad (25)$$
with $\mu(x_i,t_i) = f(x_i) + h(t_i)$.
Where $f(x_i)$ and $h(t_i)$ are smooth functions.
2. $f(x_i)$ is a smooth function defined as a Truncated Spline component with $f(x_i) = \frac{\sin(\pi x_i^2)^4}{\sin(x_i^2)}$.
Variable x_i for the Spline, generated independently of the Uniform Distribution $x_1 \sim U(0,1)$.
3. $h(t_i)$ is a smooth function defined as a Gaussian Kernel component with $h(t_i) = t_i^2 \sin t_i$.
Variable t_i for the Kernel, generated independently of the Uniform Distribution $t_1 \sim U(0,1)$.
4. An error ε_i is generated follows the Normal Distribution $\varepsilon_i \sim N(0,\sigma^2)$, where $i = 1,2,...,n$.
5. The response variable from the mixed estimator model was as follows:
$$y_i = f(x_i) + h(t_i) + \varepsilon_i = \left[\frac{\sin(\pi x_i^2)^4}{\sin(x_i^2)}\right] + \left[t_i^2 \sin t_i\right] + \varepsilon_i \quad (26)$$
6. Making a sample variation size with *n* that was tested on 25, 50, 100, and 200. For the error variance σ^2 were tested on 0.05, 0.5 and 1.
7. Replicating each generated data 20 times.
8. Making a scatter plot between response variable and each predictor variables.
9. Modeling the generated data in steps (1-8) with a nonparametric regression model mixed estimator Truncated Spline and Gaussian Kernel.
Equation 10 can be written in matrix form as:
$$\mathbf{y} = \mathbf{X}(\phi)\boldsymbol{\beta} + \mathbf{G}(\alpha)\mathbf{y} + \boldsymbol{\varepsilon} \quad (27)$$
and error is:
$$\boldsymbol{\varepsilon} = \left[\mathbf{I}-\mathbf{G}(\alpha)\right]\mathbf{y} - \mathbf{X}(\phi)\boldsymbol{\beta} \quad (28)$$
Estimation of parameter $\boldsymbol{\beta}$ can be obtained using the Ordinary Least Squares (OLS) method, where estimate results from $\boldsymbol{\beta}$ is in Equation 17.
10. The number of knot points tested is only 1 knot, for a variable set as a Truncated Spline.
11. Determining the optimal smoothing parameter using CV, GCV, and UBR. Moreover, CV, GCV, and UBR formula have been modified based on the use of a nonparametric regression model mixed estimator.
12. Calculate the Coefficient of Determination (R^2) for each modeling process carried out.
$$R^2 = 1 - \frac{SSE}{SST} \quad (29)$$
with,
$$SSE = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$
$$SST = \sum_{i=1}^{n}(y_i-\bar{y})^2$$
.
13. Calculate the average from CV, GCV, UBR, and Coefficient of Determination (R^2) for each combination of sample size variation and error variance.

III. Results and Discussion

This section describes a simulation study of a nonparametric regression model using a mixed estimator. The proposed method is a combination of the Truncated Spline and Gaussian Kernel.

$$y_i = f(x_i) + h(t_i) + \varepsilon_i \quad (30)$$

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

And $f(x_1)$ is a smooth function defined as a Truncated Spline component, $h(t_i)$ is a smooth function defined as a Gaussian Kernel.

Simulations are carried out under various regulated conditions. In this study, the sample size variations with *n* to be tested is 25, 50, 100, and 200. The combination of the error variance σ^2 to be tested is 0.05, 0.5, 1, and replication for each generated data is 20 times.

For example, a scatter plot between the response variable and each predictor variable with a sample size of $n = 50$ and $\sigma^2 = 0.05$ shown in Fig. 1.

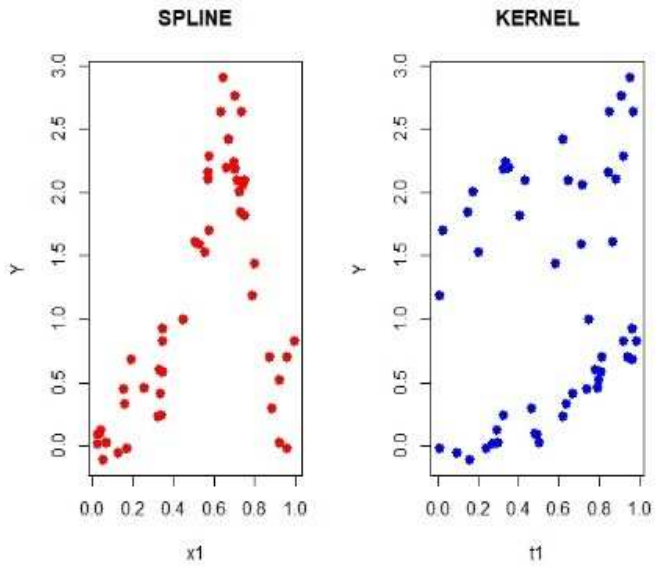


Fig. 1 Scatter plot between predictor and response variable with *n=50*

Furthermore, Fig. 2 shows the scatter plot with a sample size of $n = 200$ and $\sigma^2 = 0.05$.

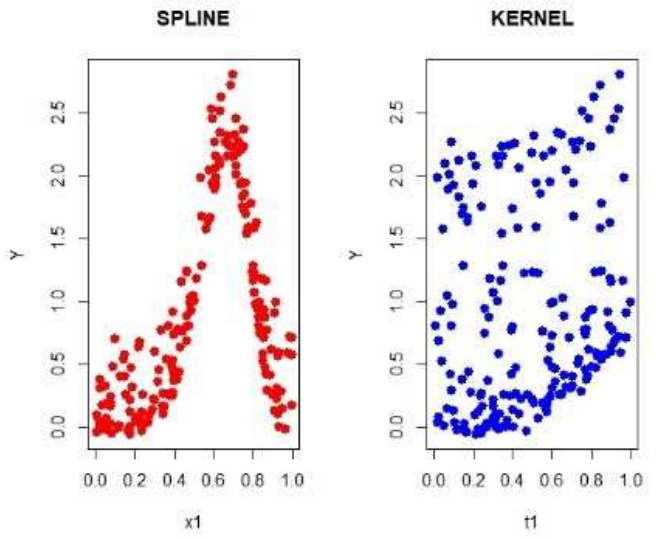


Fig. 2 Scatter plot between predictor and response variable with *n=200*

Fig. 1 and Fig. 2 show that each predictor variable has a different form of regression curve. Variable x_1 shows the characteristics of the Truncated Spline estimator, which has a changing data pattern at certain sub-intervals. In comparison, variable t_1 shows a data pattern that does not have a certain pattern, so that it was modeled with the Kernel estimator. Furthermore, based on the scatter plot in Fig. 1 and Fig. 2, a nonparametric regression model was applied using a mixed estimator of the Truncated Spline and Gaussian Kernel.

The number of knot points to be tested is only one-knot point for variables defined as a Truncated Spline component. The simulation results in the form of the average CV, GCV, UBR, and coefficient of determination (R^2) are presented in Tables 1, 2, and 3.

TABLE I
SIMULATION RESULTS WITH THE CV METHOD

Variance	Average	Number of Samples			
		n=25	*n=50*	*n=100*	*n=200*
$\sigma^2 = 0.05$	CV	0.136	0.120	0.110	0.109
	R^2	85.65%	84.75%	84.33%	84.45%
$\sigma^2 = 0.5$	CV	0.393	0.374	0.371	0.364
	R^2	66.38%	65.12%	61.84%	61.02%
$\sigma^2 = 1$	CV	1.128	1.133	1.133	1.087
	R^2	40.86%	37.42%	38.05%	35.41%

TABLE II
SIMULATION RESULTS WITH THE GCV METHOD

Variance	Average	Number of Samples			
		n=25	*n=50*	*n=100*	*n=200*
$\sigma^2 = 0.05$	GCV	1.473	2.733	5.010	8.403
	R^2	86.48%	85.30%	85.26%	85.34%
$\sigma^2 = 0.5$	GCV	1.475	3.197	4.436	10.153
	R^2	71.26%	67.25%	63.38%	61.89%
$\sigma^2 = 1$	GCV	2.168	4.033	8.096	15.192
	R^2	56.70%	44.86%	41.18%	37.15%

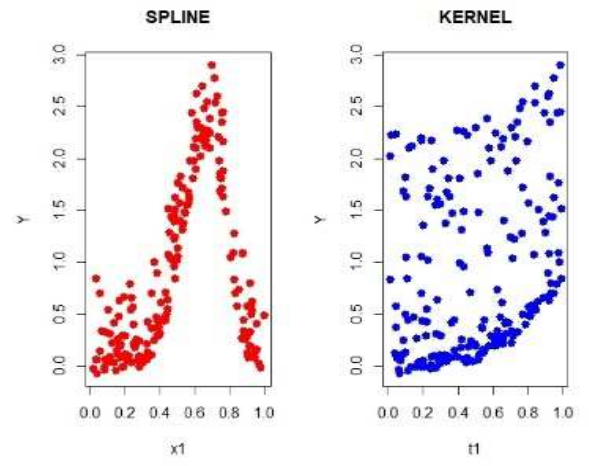
TABLE III
SIMULATION RESULTS WITH THE UBR METHOD

Variance	Average	Number of Samples			
		n=25	*n=50*	*n=100*	*n=200*
$\sigma^2 = 0.05$	UBR	0.009	0.008	0.004	0.004
	R^2	83.71%	82.74%	82.72%	82.59%
$\sigma^2 = 0.5$	UBR	0.014	0.007	0.005	0.004
	R^2	64.23%	62.38%	59.42%	59.78%
$\sigma^2 = 1$	UBR	0.016	0.009	0.006	0.004
	R^2	38.28%	35.75%	36.19%	34.04%

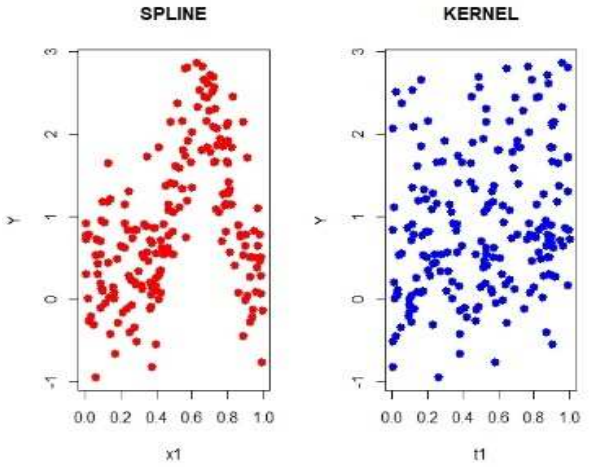
For the various sample sizes *n*, such as 25, 50, 100, and 200, with all variations of the variance tested, the GCV method provides better knot point and bandwidth estimation results compared to the CV and UBR methods. This is indicated by the value of the coefficient of determination (R^2) obtained from each experiment with GCV, which is higher than the other two methods. Furthermore, the residuals of each modeling results for each combination of sample size variation and variance follow a normal distribution.

For example, the number of samples *n=25* and the error variance is , using the GCV method in selecting the optimal knot point and bandwidth, the average GCV value is 1.473 with an R2 value of 86.48%. Meanwhile, using the CV method and the same conditions obtained an average CV value is 0.136 with R^2 value is 85.65%. Using the UBR method, it was obtained an average UBR value of 0.009 and R^2 value is 83.71%.

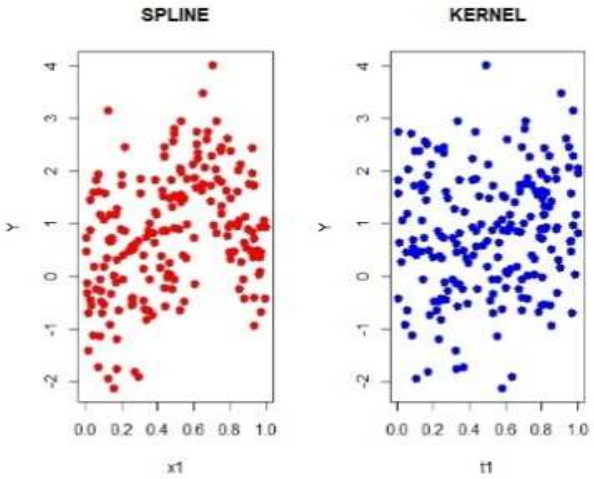
The impact of the variation variance measures σ^2 in this study has an effect on the simulation results. It can be seen that the increase of the variance tested, the value of R^2 for all methods used both CV, GCV, and UBR tend to decrease. The variance shows the deviation of the data from the average, so that the higher the variance value that is tried, then there was a tendency for the data spread far from the average value. The illustration of generated data with *n=200* and various variance conditions are shown in Fig. 3.



(a)



(b)



(c)

Fig. 3 Illustration of impact by variance (a) $\sigma^2 = 0.05$; (b) $\sigma^2 = 0.5$; (c) $\sigma^2 = 1$

Based on Fig. 3, Thus, it can be concluded that the size of the sample tested and the variance size is important. Moreover, it can be seen that in the variance $\sigma^2 = 0.05$, the Truncated Spline component has clearly shown a changing pattern at certain sub-intervals. While the Gaussian Kernel component does not appear to have a certain pattern. The increasing of the variance value, for example $\sigma^2 = 1$, the data pattern for Truncated Spline component implicitly still has shown a changing pattern in certain sub-intervals, but there is a tendency for the pattern to spread. While the Gaussian Kernel component looks more spread out and doesn't have a pattern. Based on the impact of the variance size and sample size, it can be seen that the GCV method still gives the correct estimation of knot point and bandwidth, so it can provide better coefficient of determination (R^2) value compared to the other two methods for each condition.

Based on the simulation results, the knot point and bandwidth estimation results from the CV, GCV, and UBR methods are quite good. However, the GCV method provides better performance and accuracy for each combination of sample sizes and variance variations tried. The GCV method produces optimal knot point and bandwidth to obtain the largest coefficient of determination (R^2) for each combination. As a result, the GCV method is more suitable for estimating the knot point and bandwidth in the nonparametric regression model mixed estimator of the Truncated Spline and Gaussian Kernel.

IV. Conclusion

Simulation studies on the nonparametric regression model mixed estimator of the Truncated Spline and Gaussian Kernel to compare the performance of the Cross-Validation (CV), Generalized Cross-Validation (GCV), and Unbiased Risk (UBR) methods in estimating the optimal smoothing parameter (knot point and bandwidth) have been successfully carried out. Based on the simulation results, with an error following the Normal distribution and in a combination of sample size variation and error variance. The GCV method provides better result performance and accuracy than the CV and UBR methods. The GCV method produces optimal smoothing parameters so that the largest coefficient of determination (R^2) is obtained for each combination. The results obtained in this study have the potential to contribute to the development of statistics, especially in the field of nonparametric regression.

Acknowledgment

The authors gratefully acknowledge for the financial support from the Institut Teknologi Sepuluh Nopember for this work, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI).

References

[1] A. T. R. Dani, V. Ratnasari, and I. N. Budiantara, "Optimal Knots Point and Bandwidth Selection in Modeling Mixed Estimator Nonparametric Regression," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1115, no. 1, p. 012020, 2021, doi: 10.1088/1757-899x/1115/1/012020.

[2] R. L. Eubank, *Nonparametric Regression and Spline Smoothing*, New York: Marcel Dekker, 1999.

[3] I. N. Budiantara, V. Ratnasari, M. Ratna, and I. Zain, "The Combination of Spline and Kernel Estimator for Nonparametric Regression and its Properties," *Appl. Math. Sci.*, vol. 9, no. 122, pp. 6083–6094, 2015, doi: 10.12988/ams.2015.58517.

[4] N. Y. Adrianingsih, I. N. Budiantara, and J. D. T. Purnomo, "Modeling with Mixed Kernel, Spline Truncated and Fourier Series on Human Development Index in East Java," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1115, no. 1, p. 012024, 2021, doi: 10.1088/1757-899x/1115/1/012024.

[5] D. R. Sari Saputro, K. R. Demu, and P. Widyaningsih, "Nonparametric truncated spline regression model on the data of human development index (HDI) in indonesia," *J. Phys. Conf. Ser.*, vol. 1028, no. 1, pp. 6–10, 2018, doi: 10.1088/1742-6596/1028/1/012219.

[6] N. Chamidah, B. Lestari, A. Massaid, and T. Saifudin, "Estimating mean arterial pressure affected by stress scores using Spline Nonparametric Regression model approach," *Commun. Math. Biol. Neurosci.*, vol. 2020, pp. 1–12, 2020.

[7] G. Wahba, *Spline Models for Observational Data*, Pennsylvania: SIAM, 1990.

[8] N. P. A. M. Mariati, N. Budiantara, and V. Ratnasari, "Truncated Spline Estimation of Percentage Poverty Modeling in Papua Province," *ICSA - Int. Conf. Stat. Anal. 2019*, vol. 1, pp. 69–82, 2021, doi: 10.29244/icsa.2019.pp69-82.

[9] B. Fatmawati, Budiantara, I N; Lestari, "Comparison of Smoothing and Truncated Spline Estimators in Estimating Blood Pressure Models," *Int. J. Innov. Creat. Chang.*, vol. 5, no. 3, pp. 685–707, 2019.

[10] N. P. A. M. Mariati, I. N. Budiantara, and V. Ratnasari, "Smoothing Spline Estimator in Nonparametric Regression (Application: Poverty in Papua Province)," *Proc. 7th Int. Conf. Res. Implementation, Educ. Math. Sci. (ICRIEMS 2020)*, vol. 528, no. Icriems 2020, pp. 309–314, 2021, doi: 10.2991/assehr.k.210305.044.

[11] B. Lestari, Fatmawati, and I. N. Budiantara, "Spline estimator and its asymptotic properties in multiresponse nonparametric regression model," *Songklanakarin J. Sci. Technol.*, vol. 42, no. 3, pp. 533–548, 2020, doi: 10.14456/sjst-psu.2020.68.

[12] L. R. Cheruiyot, "Local linear regression estimator on the boundary correction in nonparametric regression estimation," *J. Stat. Theory Appl.*, vol. 19, no. 3, pp. 460–471, 2020, doi: 10.2991/jsta.d.201016.001.

[13] R. Hidayat, I. N. Budiantara, B. W. Otok, and V. Ratnasari, "An extended model of penalized spline with the addition of Kernel Functions in nonparametric regression model," *Appl. Math. Inf. Sci.*, vol. 13, no. 3, pp. 453–460, 2019, doi: 10.18576/amis/130318.

[14] F. Yan, Q. S. Xu, M. L. Tang, and Z. Chen, "Kernel density-based likelihood ratio tests for linear regression models," *Stat. Med.*, vol. 40, no. 1, pp. 119–132, 2021, doi: 10.1002/sim.8765.

[15] N. Chamidah and T. Saifudin, "Estimation of children growth curve based on kernel smoothing in multi-response nonparametric regression," *Appl. Math. Sci.*, vol. 7, no. 37–40, pp. 1839–1847, 2013, doi: 10.12988/ams.2013.13168.

[16] I. Wayan Sudiarsa, "Simulations Study Combined Estimator Fourier Series and Spline Truncated in Multivariable Nonparametric Regression," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052074.

[17] D. R. S. Saputro, A. Sukmayanti, and P. Widyaningsih, "The nonparametric regression model using Fourier series approximation and penalized least squares (PLS) (case on data proverty in East Java)," *J. Phys. Conf. Ser.*, vol. 1188, no. 1, 2019, doi: 10.1088/1742-6596/1188/1/012019.

[18] A. Prahutama, Suparti, and T. W. Utami, "Modelling fourier regression for time series data - A case study: Modelling inflation in foods sector in Indonesia," *J. Phys. Conf. Ser.*, vol. 974, no. 1, pp. 0–9, 2018, doi: 10.1088/1742-6596/974/1/012067.

[19] M. F. F. Mardianto, S. M. Ulyah, and E. Tjahjono, "Prediction of national strategic commodities production based on multi-Response nonparametric regression with fourier series estimator," *Int. J. Innov. Creat. Chang.*, vol. 5, no. 3, pp. 1151–1176, 2019.

[20] I. Nyoman Budiantara *et al.*, "Modeling percentage of poor people in Indonesia using kernel and Fourier series mixed estimator in nonparametric regression," *Investig. Operacional*, vol. 40, no. 4, pp. 538–550, 2019, doi: 10.5281/zenodo.3721293.

[21] R. Hidayat, I. N. Budiantara, B. W. Otok, and V. Ratnasari, "A reproducing kernel hilbert space approach and smoothing parameters selection in spline-kernel regression," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 2, pp. 465–475, 2019.

[22] R. Hidayat, I. N. Budiantara, B. W. Otok, and V. Ratnasari, "Kernel-Spline Estimation of Additive Nonparametric Regression Model," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052028.

[23] P. Dewanti, I. Nyoman Budiantara, and A. T. Rumiati, "Modelling of SDG's Achievement in East Java Using Bi-responses Nonparametric Regression with Mixed Estimator Spline Truncated and Kernel," *J. Phys. Conf. Ser.*, vol. 1562, no. 1, 2020, doi: 10.1088/1742-6596/1562/1/012016.

[24] D. P. Rahmawati, I. N. Budiantara, D. D. Prastyo, and M. A. D. Octavanny, "Mixed Spline Smoothing and Kernel Estimator in Biresponse Nonparametric Regression," *Int. J. Math. Math. Sci.*, vol. 2021, 2021, doi: 10.1155/2021/6611084.

[25] P. Craven and G. Wahba, "Smoothing noisy data with spline functions - Estimating the correct degree of smoothing by the method of generalized cross-validation," *Numer. Math.*, vol. 31, no. 4, pp. 377–403, 1978, doi: 10.1007/BF01404567.

[26] Y. Wang, "Smoothing spline models with correlated random errors," *J. Am. Stat. Assoc.*, vol. 93, no. 441, pp. 341–348, 1998, doi: 10.1080/01621459.1998.10474115.

[27] H. Nurcahayani, I. N. Budiantara, and I. Zain, "The Curve Estimation of Combined Truncated Spline and Fourier Series Estimators for Multiresponse Nonparametric Regression," *Mathematics*, vol. 9, no. 10, p. 1141, 2021.

[28] A. R. Devi, I. N. Budiantara, and V. Ratnasari, "Unbiased risk and cross-validation method for selecting optimal knots in multivariable nonparametric regression spline truncated (case study: Unemployment rate in Central Java, Indonesia, 2015)," *AIP Conf. Proc.*, vol. 2021, no. December, 2018, doi: 10.1063/1.5062767.

[29] T. W. Utami, M. A. Haris, A. Prahutama, and E. A. Purnomo, "Optimal knot selection in spline regression using unbiased risk and generalized cross validation methods," *J. Phys. Conf. Ser.*, vol. 1446, no. 1, 2020, doi: 10.1088/1742-6596/1446/1/012049.

[30] B. A. M. Al-Talib and A. A. Hammodat, "Using Some Wavelet Shrinkage Techniques and Robust Methods to Estimate the Generalized Additive Model Parameters in Non-Linear Models," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 10, no. 6, p. 2344, 2020, doi: 10.18517/ijaseit.10.6.12767.