STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–20.
Published online in International Academic Press (www.IAPress.org)

Model Selection Criteria CV, UBR, and GCV for a Mixed Truncated Spline—Gaussian Kernel Estimator in Health Modeling

I Nyoman Budiantara [1, *], Nur Chamidah [2, 3], Andrea Tri Rian Dani [4, 5], Muhammad Anshari [6], Muhammad Fikry Al Farizi [1]

[1] *Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia*
[2] *Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya 60115, Indonesia*
[3] *Research Group of Statistical Modeling in Life Science, Faculty of Science and Technology, Universitas Airlangga, Surabaya 60115, Indonesia*
[4] *Doctoral Study Program of Mathematics and Natural Sciences, Faculty of Science and Technology, Universitas Airlangga, Surabaya 60115, Indonesia*
[5] *Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Mulawarman, Samarinda 75119, Indonesia*
[6] *Business Information Systems, Universiti Brunei Darussalam School of Business and Economics (UB-DSBE), Bandar Seri Begawan BE1410, Brunei Darussalam*

Abstract Nonparametric regression modeling has commonly applied a single estimator to all predictor variables. Although this approach is straightforward, it can be overly restrictive because predictors often exhibit heterogeneous relationships with the response variable, including linear trends, smooth nonlinear patterns, abrupt changes, or localized variations. Using a uniform estimator may therefore limit model flexibility and reduce predictive accuracy. To overcome this limitation, this study investigates a Mixed Truncated Spline—Gaussian Kernel Estimator, which allows each predictor to be modeled using the estimation technique most appropriate to its underlying data structure. The main objective of this research is to compare the performance of three smoothing parameter selection criteria, namely Cross-Validation (CV), Unbiased Risk (UBR), and Generalized Cross-Validation (GCV). These criteria are employed to determine optimal smoothing parameters, including the number and locations of knots in the truncated spline component and the bandwidth in the Gaussian kernel component. The empirical analysis is conducted using health-related data on heart disease risk factors, a domain characterized by complex and potentially nonlinear relationships. The results indicate that models incorporating three knot points consistently outperform alternative specifications. This superior performance is reflected in lower values of Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE), as well as higher coefficients of determination (R^2). Among the selection criteria examined, GCV yields the most accurate and stable model, outperforming both CV and UBR. From a methodological perspective, this study contributes to nonparametric regression by providing a systematic evaluation of smoothing parameter selection within a mixed estimator framework. From an applied standpoint, the proposed approach enhances the modeling of heart disease risk factors by offering greater flexibility and precision. Furthermore, the findings support Sustainable Development Goal (SDG) 3: Good Health and Well-Being by promoting robust, data-driven methods for evidence-based health policy formulation.

Keywords Health Modeling, Cross-Validation, Unbiased Risk, Generalized Cross-Validation, Mixed Estimators

AMS 2010 subject classifications 62G08, 62J05, 62P10

DOI: 10.19139/soic-2310-5070-3472

*Correspondence to: I Nyoman Budiantara (Email: nyomanbudiantara65@gmail.com). Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember. Teknik Mesin Street No. 175, Keputih, Sukolilo, Surabaya 60111, Indonesia.

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

1. Introduction

To date, most nonparametric regression modeling approaches have relied on a single type of estimator for all predictor variables. This approach may lead to substantial limitations, as each predictor can exhibit distinct data pattern characteristics, such as linear relationships, smooth nonlinear trends, abrupt structural changes, or localized fluctuations. Imposing a single estimator across all predictors may reduce model flexibility and hinder its ability to optimally represent complex relationships between the response variable and its predictors.

To address these limitations, Budiantara et al. (2015) introduced the concept of a mixed spline–kernel estimator for a single predictor [1], which was subsequently extended to the multivariate setting by Ratnasari et al. (2016) [2]. Further developments by Dani et al. (2021) demonstrated, using simulation data, that mixed spline–kernel estimators provide superior performance and greater flexibility compared to single-estimator approaches [3]. In general, mixed estimators in nonparametric regression combine multiple types of estimators, allowing each predictor to be modeled according to its inherent data pattern characteristics [4, 5].

Several previous studies have examined mixed estimators based on two components such as spline–Kernel [1, 6, 7], spline–Fourier series [8–10], and kernel–Fourier series [11–13] and have reported improvements in modeling performance. Nevertheless, research on smoothing parameter selection for mixed estimator models, particularly in real-world applications, remains relatively limited. This is a critical issue, as the selection of smoothing parameters such as knot locations in spline components and bandwidths in kernel components plays a decisive role in determining estimation quality.

In the context of smoothing parameter selection, Generalized Cross-Validation (GCV), as proposed by Wahba, has been widely adopted and has demonstrated strong performance across various mixed estimator models [14]. In addition to GCV, Cross-Validation (CV) and Unbiased Risk (UBR) are also commonly used, each possessing distinct characteristics, advantages, and limitations [15–17]. While previous studies generally report the superiority of GCV in simulation settings, comparative analyses of CV, UBR, and GCV using real-world application data remain scarce.

Therefore, this study conducts an in-depth investigation of a mixed truncated spline–Gaussian kernel estimator applied to health data, with a specific focus on heart disease as one of the leading non-communicable diseases. The novelty of this research lies in the comprehensive comparison of three model selection criteria CV, UBR, and GCV for the simultaneous determination of optimal smoothing parameters. This approach is expected to yield models that are more accurate, stable, and capable of effectively capturing complex relationships among health risk factors.

Beyond its methodological contributions, this study also supports the achievement of the Sustainable Development Goals (SDGs), particularly SDG 3: Good Health and Well-Being, by providing a more flexible and accurate statistical modeling framework for analyzing factors influencing heart disease. The findings are expected to serve as valuable evidence for more effective health policy formulation aimed at the prevention and control of non-communicable diseases.

2. General Methods

2.1. Nonparametric Regression Mixed Estimators

A mixed estimator in nonparametric regression is a multipredictor nonparametric regression model with an additive regression structure, in which the regression curve is approximated by two or more types of nonparametric estimators. This approach allows each predictor component to be modeled using the estimator that is most appropriate for the underlying data pattern characteristics. Budiantara et al. (2015) introduced an additive nonparametric regression model with two predictor components, where the first predictor component is approximated using truncated spline regression, while the second predictor component is approximated using kernel regression [1].

Suppose that a set of paired observations $\{(x_{1i},\ x_{2i},\ y_i),\ i = 1,\ 2,\ \ldots,\ n)\}$, is given, where the relationship between the response variable y_i and the predictor variables follows the nonparametric regression model specified

in Equation (1).

$$y_i = f(x_{1i},\ x_{2i}) + \varepsilon_i \tag{1}$$

Here $f(.)$ denotes an unknown regression function that is assumed to be smooth, in the sense that it is continuous and differentiable. The random error term ε_i is assumed to follow a normal distribution with zero mean and variance σ^2, that is $\varepsilon_i \sim \mathcal{N}(0,\ \sigma^2)$. Furthermore, the regression function $f(.)$ is assumed to have an additive structure and can therefore be expressed as shown in Equation (2) [18].

$$f(x_{1i},\ x_{2i}) = f_1(x_{1i}) + f_2(x_{2i}) \tag{2}$$

Here $f_1(.)$ and $f_2(.)$ are smooth functions, each approximated using a different nonparametric estimator according to the characteristics of the data. The primary challenge in mixed nonparametric regression lies in determining the most appropriate combination of estimators for each predictor component, as well as in selecting optimal smoothing parameters that control the degree of smoothness of the regression curves, such as knot locations in the spline component and bandwidths in the kernel component. Inappropriate choices of estimator combinations and smoothing parameters may result in overly smooth models (oversmoothing) or excessively fluctuating curves (undersmoothing), thereby reducing estimation accuracy and predictive performance.

2.2. Mixed Truncated Spline—Gaussian Kernel Estimator

The initial step is to define the Mixed Truncated Spline—Gaussian Kernel Estimator model. Suppose that paired data are observed with p predictor variables associated with the spline component and q predictor variables associated with the kernel component, such that the data can be represented as $(z_{1i},\ \ldots,\ z_{pi},\ v_{1i},\ \ldots,\ v_{qi},\ y_i)$ for $i = 1,\ 2,\ \ldots,\ n)$. It is assumed that the relationship between the predictor variables $(z_{1i},\ \ldots,\ z_{pi},\ v_{1i},\ \ldots,\ v_{qi})$ and the response variable (y_i) follows a mixed nonparametric regression model based on a spline—kernel estimator [7, 19].

$$y_i = f(z_{1i},\ z_{2i},\ \ldots,\ z_{pi}) + h(v_{1i},\ v_{2i},\ \ldots,\ v_{qi}) + \varepsilon_i \tag{3}$$

Based on Equation (3) and the functional forms of the spline and kernel components, Equation (3) can be simplified and expressed as Equation (4).

$$y_i = \mu(z_i,\ v_i) + \varepsilon_i \tag{4}$$

where:

$$\mu(z_i,\ v_i) = \sum_{j=1}^{p} f_j(z_{ji}) + \sum_{s=1}^{q} h_{\alpha s}(v_{si})$$

Equation (4) can be represented in matrix form as shown in Equation (5).

$$\boldsymbol{y} = \boldsymbol{Z}(K)\boldsymbol{\psi} \ +\ \boldsymbol{D}(\boldsymbol{\alpha})\boldsymbol{y} + \boldsymbol{\varepsilon} \tag{5}$$

where:

$$\boldsymbol{Z}(K)\boldsymbol{\psi} = \boldsymbol{f}(\boldsymbol{z_1},\ \boldsymbol{z_2},\ \ldots,\ \boldsymbol{z_p}) = \sum_{j=1}^{p} \boldsymbol{f_j}(\boldsymbol{z_j}),$$

$$\sum_{j=1}^{p} \boldsymbol{f_j}(\boldsymbol{z_j}) = \begin{bmatrix} \sum_{j=1}^{p} f_j(z_{j1}) \\ \sum_{j=1}^{p} f_j(z_{j2}) \\ \vdots \\ \sum_{j=1}^{p} f_j(z_{jn}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z_0} & \boldsymbol{Z_1}(K_1) & \boldsymbol{Z_2}(K_2) & \cdots & \boldsymbol{Z_p}(K_p) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\delta_1} \\ \boldsymbol{\delta_2} \\ \vdots \\ \boldsymbol{\delta_p} \end{bmatrix},$$

$$\begin{bmatrix} \sum_{j=1}^{p} f_j(z_{j1}) \\ \sum_{j=1}^{p} f_j(z_{j2}) \\ \vdots \\ \sum_{j=1}^{p} f_j(z_{jn}) \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & z_{21} & \cdots & z_{p1} \\ 1 & z_{12} & z_{22} & \cdots & z_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{1n} & z_{2n} & \cdots & z_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_{11} \\ \vdots \\ \beta_{1p} \end{bmatrix}$$

$$+ \begin{bmatrix} (z_{11}-K_{11})_+ & (z_{11}-K_{21})_+ & \cdots & (z_{11}-K_{r1})_+ \\ (z_{12}-K_{11})_+ & (z_{12}-K_{21})_+ & \cdots & (z_{12}-K_{r1})_+ \\ \vdots & \vdots & \ddots & \vdots \\ (z_{1n}-K_{11})_+ & (z_{1n}-K_{21})_+ & \cdots & (z_{1n}-K_{r1})_+ \end{bmatrix} \begin{bmatrix} \delta_{11} \\ \delta_{21} \\ \delta_{31} \\ \vdots \\ \delta_{r1} \end{bmatrix}$$

$$+ \begin{bmatrix} (z_{21}-K_{12})_+ & (z_{21}-K_{22})_+ & \cdots & (z_{21}-K_{r2})_+ \\ (z_{22}-K_{12})_+ & (z_{22}-K_{22})_+ & \cdots & (z_{22}-K_{r2})_+ \\ \vdots & \vdots & \ddots & \vdots \\ (z_{2n}-K_{12})_+ & (z_{2n}-K_{22})_+ & \cdots & (z_{2n}-K_{r2})_+ \end{bmatrix} \begin{bmatrix} \delta_{12} \\ \delta_{22} \\ \delta_{32} \\ \vdots \\ \delta_{r2} \end{bmatrix}$$

$$+ \cdots + \begin{bmatrix} (z_{p1}-K_{1p})_+ & (z_{p1}-K_{2p})_+ & \cdots & (z_{p1}-K_{rp})_+ \\ (z_{p2}-K_{1p})_+ & (z_{p2}-K_{2p})_+ & \cdots & (z_{p2}-K_{rp})_+ \\ \vdots & \vdots & \ddots & \vdots \\ (z_{pn}-K_{1p})_+ & (z_{pn}-K_{2p})_+ & \cdots & (z_{pn}-K_{rp})_+ \end{bmatrix} \begin{bmatrix} \delta_{1p} \\ \delta_{2p} \\ \delta_{3p} \\ \vdots \\ \delta_{rp} \end{bmatrix},$$

and

$$\boldsymbol{D}(\boldsymbol{\alpha})\boldsymbol{y} = \sum_{s=1}^{q} \hat{\boldsymbol{h}}_{\boldsymbol{\alpha s}}(\boldsymbol{v_s}), \tag{6}$$

$$\begin{aligned} \sum_{s=1}^{q} \hat{\boldsymbol{h}}_{\boldsymbol{\alpha s}}(\boldsymbol{v_s}) &= \begin{bmatrix} \sum_{s=1}^{q} \hat{h}_{\alpha s}(v_{s1}) \\ \sum_{s=1}^{q} \hat{h}_{\alpha s}(v_{s2}) \\ \vdots \\ \sum_{s=1}^{q} \hat{h}_{\alpha s}(v_{sn}) \end{bmatrix} = \begin{bmatrix} \hat{h}_{\alpha\cdot}(v_{\cdot 1}) \\ \hat{h}_{\alpha\cdot}(v_{\cdot 2}) \\ \vdots \\ \hat{h}_{\alpha\cdot}(v_{\cdot n}) \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n}\sum_{s=1}^{q} W_{\alpha_{s1}}(v_{s1}) & \frac{1}{n}\sum_{s=1}^{q} W_{\alpha_{s2}}(v_{s1}) & \cdots & \frac{1}{n}\sum_{s=1}^{q} W_{\alpha_{sn}}(v_{s1}) \\ \frac{1}{n}\sum_{s=1}^{q} W_{\alpha_{s1}}(v_{s2}) & \frac{1}{n}\sum_{s=1}^{q} W_{\alpha_{s2}}(v_{s2}) & \cdots & \frac{1}{n}\sum_{s=1}^{q} W_{\alpha_{sn}}(v_{s2}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{s=1}^{q} W_{\alpha_{s1}}(v_{sn}) & \frac{1}{n}\sum_{s=1}^{q} W_{\alpha_{s2}}(v_{sn}) & \cdots & \frac{1}{n}\sum_{s=1}^{q} W_{\alpha_{sn}}(v_{sn}) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \end{aligned} \tag{7}$$

The kernel function $W_{\alpha_{si}}$ represents a weighting function, and its functional form indicates the use of a Gaussian kernel in Equation (8).

$$W_{\alpha_{si}}(v_{si}) = \frac{K_\alpha(v_s - v_{si})}{\frac{1}{n}\sum_{i=1}^{n} K_\alpha(v_s - v_{si})}, \tag{8}$$

where:

$$K_\alpha(v_s - v_{si}) = \frac{1}{\alpha} K\left(\frac{v_s - v_{si}}{\alpha}\right).$$

In general, the elements of the matrix $\boldsymbol{D}(\boldsymbol{\alpha})\boldsymbol{y}$ in Equation (7) can be expressed in Equation (9) as follows.

$$\hat{h}_{\alpha s}(v_{si}) = \frac{1}{n}\sum_{i=1}^{n} W_{\alpha_{si}}(v_{si}) y_i. \tag{9}$$

For $s = 1$, the kernel estimator in Equation (9) can be written as:

$$\hat{h}_{\alpha 1}(v_{1i}) = \frac{1}{n}\sum_{i=1}^{n} W_{\alpha_{1i}}(v_{1i}) y_i.$$

This expression can be represented in matrix form as:

$$\hat{\boldsymbol{h}}_{\boldsymbol{\alpha 1}}(\boldsymbol{v_1}) = \boldsymbol{D}(\alpha_1)\boldsymbol{y},$$

$$\hat{\boldsymbol{h}}_{\boldsymbol{\alpha 1}}(\boldsymbol{v_1}) = \begin{bmatrix} \hat{h}_{\alpha 1}(v_{11}) \\ \hat{h}_{\alpha 1}(v_{12}) \\ \vdots \\ \hat{h}_{\alpha 1}(v_{1n}) \end{bmatrix},$$

$$\boldsymbol{D}(\alpha_1) = \begin{bmatrix} \frac{1}{n}W_{\alpha_{11}}(v_{11}) & \frac{1}{n}W_{\alpha_{12}}(v_{11}) & \cdots & \frac{1}{n}W_{\alpha_{1n}}(v_{11}) \\ \frac{1}{n}W_{\alpha_{11}}(v_{12}) & \frac{1}{n}W_{\alpha_{12}}(v_{12}) & \cdots & \frac{1}{n}W_{\alpha_{1n}}(v_{12}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}W_{\alpha_{11}}(v_{1n}) & \frac{1}{n}W_{\alpha_{12}}(v_{1n}) & \cdots & \frac{1}{n}W_{\alpha_{1n}}(v_{1n}) \end{bmatrix}.$$

Using the same procedure for $s = 2,\ 3,\ 4,\ \ldots,\ q$, the kernel estimators for the remaining predictor components can be written as:

$$\hat{\boldsymbol{h}}_{\boldsymbol{\alpha 2}}(\boldsymbol{v_2}) = \boldsymbol{D}(\alpha_2)\boldsymbol{y},\ \hat{\boldsymbol{h}}_{\boldsymbol{\alpha 3}}(\boldsymbol{v_3}) = \boldsymbol{D}(\alpha_3)\boldsymbol{y},\ \hat{\boldsymbol{h}}_{\boldsymbol{\alpha 4}}(\boldsymbol{v_4}) = \boldsymbol{D}(\alpha_4)\boldsymbol{y},\ \ldots,\ \hat{\boldsymbol{h}}_{\boldsymbol{\alpha q}}(\boldsymbol{v_q}) = \boldsymbol{D}(\alpha_q)\boldsymbol{y}. \tag{10}$$

Equation (6) can be obtained as the sum of the matrices $\boldsymbol{D}(\boldsymbol{\alpha}_s)\boldsymbol{y}$, for $s = 1,\ 2,\ 3,\ \ldots,\ q$, as expressed in Equation (11).

$$\boldsymbol{D}(\boldsymbol{\alpha})\boldsymbol{y} = \sum_{s=1}^{q} \boldsymbol{D}(\alpha_s)\boldsymbol{y} = \left[\boldsymbol{D}(\alpha_1) + \boldsymbol{D}(\alpha_2) + \ldots + \boldsymbol{D}(\alpha_q)\right]\boldsymbol{y}. \tag{11}$$

The parameter estimates of $\boldsymbol{\psi}$ can be obtained using the Least Squares (LS) method based on Equation (5). The LS approach seeks to obtain estimators by minimizing the sum of squared errors, such that the error term can be expressed as shown in Equation (12).

$$\boldsymbol{\varepsilon} = (\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\alpha}))\,\boldsymbol{y} - \boldsymbol{Z}(K)\boldsymbol{\psi} \tag{12}$$

The sum of squared errors is expressed in Equation (13).

$$Q(\boldsymbol{\psi} \mid K,\, \alpha) = \|(\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\alpha}))\,\boldsymbol{y}\|^2 - 2\boldsymbol{\psi}^T\boldsymbol{Z}(K)^T\,(\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\alpha}))\,\boldsymbol{y} + \boldsymbol{\psi}^T\boldsymbol{Z}(K)^T\boldsymbol{Z}(K)\boldsymbol{\psi} \tag{13}$$

To obtain the estimate of $\boldsymbol{\psi}$, the partial derivative of Equation (13) with respect to $\boldsymbol{\psi}$ is taken as follows in Equation (14):

$$\frac{\partial Q(\boldsymbol{\psi} \mid K,\, \alpha)}{\partial \boldsymbol{\psi}} = -2\boldsymbol{Z}(K)^T\,(\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\alpha}))\,\boldsymbol{y} + 2\boldsymbol{Z}(K)^T\boldsymbol{Z}(K)\boldsymbol{\psi}. \tag{14}$$

The resulting partial derivative is then set equal to zero, yielding the estimator of $\boldsymbol{\psi}$ as presented in Equation (15).

$$\hat{\boldsymbol{\psi}} = (\boldsymbol{Z}(K)^T\boldsymbol{Z}(K))^{-1}\,(\boldsymbol{Z}(K))^T\,(\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\alpha}))\,\boldsymbol{y} \tag{15}$$

Equation (15) can be simplified and expressed as Equation (16).

$$\hat{\boldsymbol{\psi}} = \boldsymbol{A}(K,\ \alpha)\boldsymbol{y} \tag{16}$$

where:

$$\boldsymbol{A}(K,\ \alpha) = \left((\boldsymbol{Z}(K)^T\boldsymbol{Z}(K))^{-1}\,(\boldsymbol{Z}(K))^T\,(\boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\alpha}))\right).$$

The truncated spline estimator has been previously defined as $\sum_{j=1}^{p} \boldsymbol{f_j}(\boldsymbol{z_j}) = \boldsymbol{Z}(K)\boldsymbol{\psi}$, therefore, by obtaining the estimator of $\boldsymbol{\psi}$ from the Mixed Truncated Spline—Gaussian Kernel Estimator model, it follows that:

$$\sum_{j=1}^{p} \boldsymbol{f_j}(\boldsymbol{z_j}) = \boldsymbol{Z}(K)\hat{\boldsymbol{\psi}} = \boldsymbol{R}(K,\ \alpha)\boldsymbol{y}.$$

Let the matrix:

$$\boldsymbol{R}(K,\ \alpha) = \boldsymbol{Z}(K)\,(\boldsymbol{A}(K,\ \alpha)).$$

By assuming that both estimators are additive, the resulting estimator can be expressed as shown in Equation (17).

$$\begin{aligned}\hat{\mu}(z_i,\ v_i) &= \sum_{j=1}^{p} \hat{f}_j(z_{ji}) + \sum_{s=1}^{q} \hat{h}_{\alpha s}(v_{si}) \\ &= \boldsymbol{B}(K,\ \alpha)\boldsymbol{y}\end{aligned} \tag{17}$$

where

$$\boldsymbol{B}(K,\ \alpha) = \boldsymbol{R}(K,\ \alpha) + \boldsymbol{D}(\boldsymbol{\alpha}).$$

The matrix $\boldsymbol{B}(K,\ \alpha)$ depends critically on $\boldsymbol{R}(K,\ \alpha)$, which represents the truncated spline estimator component with knot locations $K_1,\ K_2,\ \ldots,\ K_r$,, and on $\boldsymbol{D}(\boldsymbol{\alpha})$, which corresponds to the Gaussian kernel estimator component with bandwidth parameter.

2.3. Smoothing Parameter Selection Criteria

2.3.1 Cross-Validation

Cross-Validation (CV) is a method developed by Craven and Wahba. The original CV formulation proposed by Craven and Wahba was limited to single-estimator models [20]. One of the main advantages of the CV method is that it does not require prior knowledge of the error variance σ^2. Furthermore, the CV method can also be applied to mixed estimator models [21].

Let $y = (y_1,\ y_2,\ \ldots,\ y_n)^T \in \mathbb{R}^n$ denote the response vector generated from the mixed nonparametric regression model. Let the fitted response be written in linear smoothing form in Equation (17), where $\boldsymbol{B}(K,\ \alpha) \in \mathbb{R}^{n\times n}$ is the hat matrix depending on the knot configuration K for the truncated spline component and the bandwidth parameter α for the Gaussian Kernel component.

The leave-one-out fitted value for the $i-$th observation, denoted by $\hat{y}_{(-i)}$, is obtained by excluding the $i-$th observation from the estimation process. For linear smoothers, it can be shown that:

$$\hat{y}_{(-i)} = \frac{\hat{y}_i - \boldsymbol{B}_{ii}(K,\ \alpha)y_i}{1 - \boldsymbol{B}_{ii}(K,\ \alpha)}$$

Hence, the leave-one-out residual can be expressed as:

$$y_i - \hat{y}_{(-i)} = \frac{y_i - \hat{y}_i}{1 - \boldsymbol{B}_{ii}(K,\ \alpha)}$$

Substituting $\boldsymbol{B}_{ii}(K,\ \alpha) = \boldsymbol{R}_{ii}(K,\ \alpha) + \boldsymbol{D}_{ii}(\alpha)$, the Cross-Validation (CV) criterion for the mixed estimator is defines in Equation (18).

$$\begin{aligned} CV(K,\ \alpha) &= \sum_{i=1}^{n} \left[y_i - \hat{y}_{(-i)}\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(y_i - \hat{y}_i)}{(1 - [\boldsymbol{R}_{ii}(K,\ \alpha)\ +\ \boldsymbol{D}_{ii}(\boldsymbol{\alpha})])}\right]^2 \end{aligned} \tag{18}$$

where :
y_i : the observed $i-$th response variable
$\hat{y}_i$: the estimated $i-$th response variable obtained from the mixed estimator model

The CV formula presented in Equation (18) represents a modification tailored to mixed estimator models. It can be observed that the CV method assigns different weights to each observation according to its contribution to the model. The matrix $\boldsymbol{B}(K,\ \alpha))$ can be obtained from Equation (17), where $\boldsymbol{B}(K,\ \alpha)$ depends strongly on the truncated spline and Gaussian kernel component matrices. The estimated response value $(\hat{y}_i)$ is obtained from the nonparametric regression model using the mixed estimator. The smallest CV value corresponds to the optimal selection of knot locations and bandwidth parameters.

2.3.2 Unbiased Risk (UBR)

The Unbiased Risk (UBR) method can be used to determine optimal smoothing parameters when information on the error variance σ^2 is available or when σ^2 is known [22]. This method was introduced by Wang and was originally limited to single-estimator models [23]. A modified UBR formulation suitable for mixed estimator models can be expressed as shown in Equation (19).

$$\begin{aligned} UBR(K,\ \alpha) = n^{-1}\Big\{ &\|(\boldsymbol{I}_n - (\boldsymbol{R}(K,\ \alpha) + \boldsymbol{D}\,(\boldsymbol{\alpha})))\,\boldsymbol{y}\|^2 + \frac{\sigma^2}{n}\,\mathrm{trace}[\boldsymbol{I}_n - [\boldsymbol{R}(K,\ \alpha)\ +\ \boldsymbol{D}\,(\boldsymbol{\alpha})]]^2 \\ &+ \frac{\sigma^2}{n}\,\mathrm{trace}[[\boldsymbol{R}(K,\ \alpha)\ +\ \boldsymbol{D}\,(\boldsymbol{\alpha})]]^2\Big\} \end{aligned} \tag{19}$$

An estimate of σ^2 can be obtained using the following formula:

$$\hat{\sigma}^2 = \frac{\|(\boldsymbol{I}_n - (\boldsymbol{R}(K,\ \alpha) + \boldsymbol{D}(\boldsymbol{\alpha})))\,\boldsymbol{y}\|^2}{\text{trace}\|(\boldsymbol{I}_n - (\boldsymbol{R}(K,\ \alpha) + \boldsymbol{D}(\boldsymbol{\alpha})))\,\boldsymbol{y}\|^2}$$

where :
$\boldsymbol{I}_n$: $n \times n$ Identity matrix

The minimum UBR value corresponds to the optimal selection of knot locations and bandwidth parameters.

2.3.3 Generalized Cross-Validation (GCV)

The Generalized Cross-Validation (GCV) method is a generalization of the CV approach and was developed by Wahba [14, 24]. The original GCV formulation proposed by Wahba was limited to single-estimator models. Similar to the CV method, GCV does not require prior knowledge of the error variance σ^2. The GCV method can also be applied to mixed estimator models. Defines the residual vector:

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I}_n - \boldsymbol{B}(K,\ \alpha)\boldsymbol{y}) = \boldsymbol{I}_n - [\boldsymbol{R}(K,\ \alpha) + \boldsymbol{D}(\boldsymbol{\alpha})]\,\boldsymbol{y}$$

and the residual sum of squares:

$$RSS(K,\ \alpha) = \|\boldsymbol{\varepsilon}\|^2 = \|(\boldsymbol{I}_n - \boldsymbol{B}(K,\ \alpha)\boldsymbol{y})\|^2 = \|[\boldsymbol{I}_n - [\boldsymbol{R}(K,\ \alpha) + \boldsymbol{D}(\boldsymbol{\alpha})]\,\boldsymbol{y}]\|^2$$

The mean squared error is given by:

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

A modified GCV formulation suitable for mixed estimators can be expressed as shown in Equation (20).

$$\begin{aligned} GCV(K,\ \alpha) &= \sum_{i=1}^{n} \left[\frac{MSE(K,\ \alpha)}{\left(\frac{1}{n}\,\text{trace}[\boldsymbol{I}_n - [\boldsymbol{R}(K,\ \alpha) + \boldsymbol{D}(\boldsymbol{\alpha})]]\right)^2} \right] \\ &= \left[\frac{n\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\left(\text{trace}[\boldsymbol{I}_n - [\boldsymbol{R}(K,\ \alpha) + \boldsymbol{D}(\boldsymbol{\alpha})]]\right)^2} \right] \end{aligned} \tag{20}$$

where :
$\boldsymbol{I}_n$: $n \times n$ Identity matrix

The GCV formula presented in Equation (20) represents a modification specifically adapted for mixed estimator models. The GCV method can be interpreted as a weighted version of CV in which equal weights are assigned to all observations [25]. The minimum GCV value corresponds to the optimal selection of knot locations and bandwidth parameters [26].

2.4. Model Performance Evaluation

The performance of the model in this study is evaluated using the Coefficient of Determination (R^2), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). These measures are employed to assess the model's ability to explain data variability as well as its predictive accuracy.

The Coefficient of Determination (R^2) measures the proportion of the variability in the response variable that can be explained by the model and is defined as shown in Equation (21).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}. \tag{21}$$

Here, y_i denotes the $i-$th observed value, $\hat{y}_i$ represents the estimated value obtained from the mixed estimator model, and $\bar{y}$ is the mean of all observed values [27].

The Mean Squared Error (MSE) measures the average squared prediction error and is defined in Equation (22) [27].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{22}$$

The Mean Absolute Percentage Error (MAPE) measures the relative prediction error in percentage terms and is defined in Equation (23) as:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{23}$$

Smaller MAPE values indicate a higher level of predictive accuracy [28].

2.5. Data, Data Sources, and Research Variables

This study utilizes secondary data obtained from the 2023 Indonesian Health Survey (*Survei Kesehatan Indonesia* (SKI)) published by the Ministry of Health of the Republic of Indonesia. The unit of analysis consists of the 38 provinces in Indonesia, allowing the data to represent population health conditions at the regional level.

A purposive sampling approach is employed, with key considerations including data availability, the timeliness of information, and the relevance of the data to the research objectives. The use of SKI 2023 data is deemed appropriate, as it provides comprehensive and up-to-date health indicators, particularly those related to disease prevalence and its associated determinants. All variables are measured on a ratio scale. Descriptions of the variable notation, variable names, variable types, and operational definitions for each variable are presented in Table 1 as follows.

Table 1. Research Variables and Operational Definitions

Variables	Notation	Description	Type of Variables	Operational Definitions
Response	y	Prevalence of Heart Disease	Ratio	Calculated as the proportion of the population experiencing heart disease relative to the total population in a given region and year.
Predictor	X_1	Prevalence of Hypertension	Ratio	Calculated as the proportion of individuals with hypertension relative to the total population in a given region and year.
Predictor	X_2	Prevalence of Diabetes Mellitus	Ratio	Calculated as the proportion of individuals with diabetes mellitus relative to the total population in a given region and year.
Predictor	X_3	Prevalence of Daily Smoking	Ratio	Defined as the proportion of the population who smoke cigarettes daily during the observation period.

Variables	Notation	Description	Type of Variables	Operational Definitions
Predictor	X_4	Proportion of Hypertension Follow-up/ Check-ups at Health Facilities	Ratio	Calculated as the proportion of hypertension patients who routinely attend follow-up visits compared to the total number of hypertension patients.
Predictor	X_5	Proportion of Insufficient Physical Activity among Population Aged ≥ 10 Years	Ratio	Defined as the proportion of individuals aged ≥ 10 years with low levels of physical activity within a region.
Predictor	X_6	Proportion of Habitual Consumption of Fatty/ Cholesterol/ Fried Foods ≥ 1 Time	Ratio	Calculated as the proportion of the population who consume fatty, high-cholesterol, or fried foods at least once within the observation period.

2.6. Research Methodology

The primary objective of this study is to compare the selection of smoothing parameters in the Mixed Truncated Spline Gaussian Kernel Estimator model using the CV, GCV, and UBR criteria, as applied to health data, specifically the prevalence of heart disease in Indonesia. To achieve this objective, the following steps are undertaken:

1. Construct scatter plots between the response variable and each predictor variable. Identify which predictor variables are to be modeled using truncated spline estimators and which are to be modeled using Gaussian kernel estimators.
2. Detect the presence of strong relationships or correlations among predictor variables using the Variance Inflation Factor (VIF).
3. Model the prevalence of heart disease in Indonesia using a mixed estimator approach combining truncated spline and Gaussian kernel estimators.
4. Obtain the optimal CV value based on Equation (18), thereby determining the best model corresponding to the minimum CV value. Subsequently, compute the R^2, MSE, and MAPE.
5. Obtain the optimal UBR value based on Equation (19), thereby determining the best model corresponding to the minimum UBR value. Subsequently, compute the R^2, MSE, and MAPE.
6. Obtain the optimal GCV value based on Equation (20), thereby determining the best model corresponding to the minimum GCV value. Subsequently, compute the R^2, MSE, and MAPE.
7. Compare the performance of the Mixed Truncated Spline—Gaussian Kernel Estimator models based on the R^2, MSE, and MAPE.
8. Compare the performance of the best model (mixed estimator) with models where all predictors are modeled using spline estimators (full spline) and models where all predictors are modeled using kernel estimators (full kernel).

3. Results and Discussion

3.1. Scatter Plot

Based on the scatter plot analysis between the response variable, namely the prevalence of heart disease, and each predictor variable presented in Figure 1, distinct differences in the characteristics of the relationships among variables were identified.

Predictor variables X_1 (Prevalence of Hypertension) and X_2 (Prevalence of Diabetes Mellitus) exhibit relationships that are generally linear with respect to the response variable. However, this linearity is not global, as changes in slope occur across different segments of the data. This indicates the presence of structural change points (knots), suggesting that these variables are more appropriately modeled using truncated spline estimators, which provide flexibility in capturing such pattern changes.

Furthermore, variables X_3 (prevalence of daily smoking), X_5 (Proportion of Insufficient Physical Activity among Population Aged ≥ 10 Years), and X_6 (Proportion of Habitual Consumption of Fatty/ Cholesterol/ Fried Foods ≥ 1 Time) display scatter patterns that do not follow any specific parametric functional form and tend to be nonlinear and heterogeneous. Therefore, a nonparametric approach using truncated spline estimators is adopted due to its ability to capture gradual changes in the relationship without imposing a global functional form assumption.

In contrast, variable X_4 (Proportion of Hypertension Follow-up/Check-ups at Health Facilities) shows a highly fluctuating relationship without clear changes in slope across specific intervals. This characteristic reflects a smooth and continuous local pattern, making the kernel estimator more suitable, as it effectively captures local data structures through optimal bandwidth selection.

Overall, the six predictor variables exhibit different relationship characteristics with the response variable. Therefore, the modeling approach is determined based on the identification of scatter plot patterns for each variable. Based on this assessment, variables X_1, X_2, X_3, X_5, and X_6 are proposed to be modeled using truncated spline estimators, while variable X_4 is modeled using a Gaussian kernel estimator. This specification is expected to produce a model that is more flexible and representative in capturing the overall data structure.

In the final stage, a model comparison is conducted using three approaches: a model in which all predictor variables are modeled using truncated spline estimators (full spline), a model in which all variables are modeled using Gaussian kernel estimators (full kernel), and a mixed model constructed based on the identified data patterns. This comparison aims to quantitatively demonstrate the superiority of the proposed mixed model over homogeneous approaches, thereby providing stronger empirical evidence that pattern-based modeling yields more optimal performance.

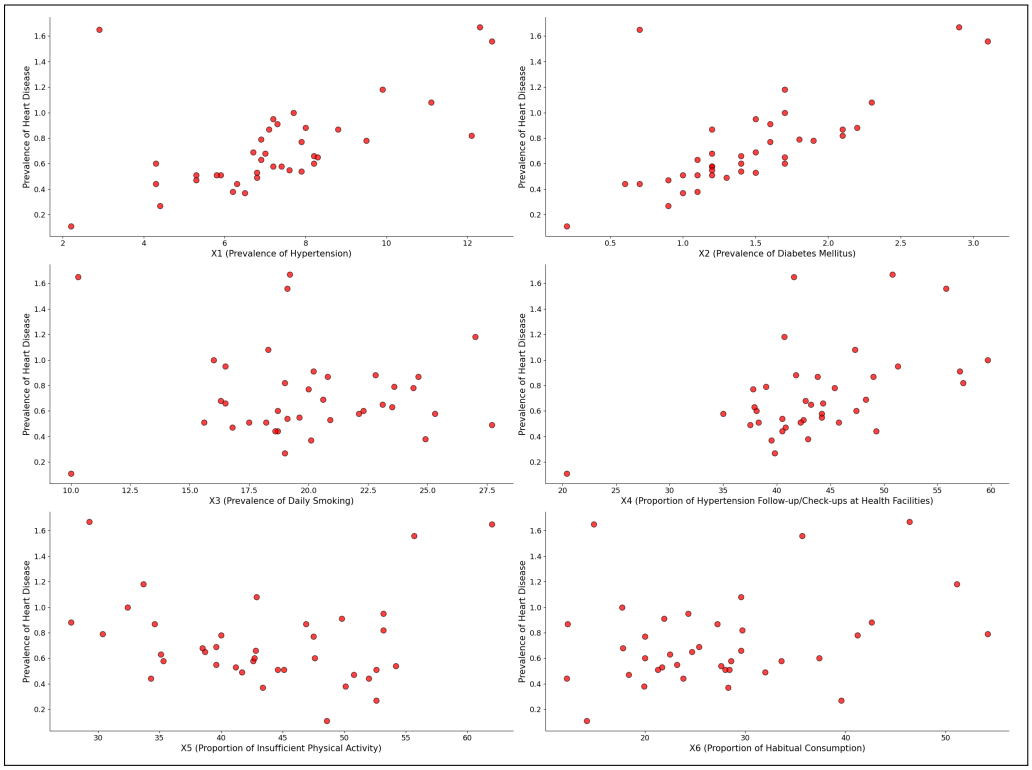

Figure 1. Scatter-plots Illustrating the Relationships between the Predictor and the Response Variable

The Pearson correlation coefficients between each predictor and the response variable are summarized in Table 2.

Table 2. Pearson Correlation Coefficients Between the Response Variable and Each Predictor Variable

Variables	X_1	X_2	X_3	X_4	X_5	X_6
Pearson (ρ) Correlation Coefficient	0.5614	0.6803	−0.0593	0.5193	−0.0507	0.2805

Accordingly, in this study, variables X_1, X_2, X_3, X_5, and X_6 are modeled using truncated spline estimators, while X_4 is modeled using a Gaussian kernel estimator. The use of distinct estimators for each predictor is grounded in both data-driven characteristics and theoretical considerations, and is expected to improve model flexibility as well as overall estimation accuracy.

3.2. Multicollinearity Assessment Using Variance Inflation Factor

Multicollinearity in this study is assessed using the Variance Inflation Factor (VIF) to identify the presence of high correlations among predictor variables. In general, VIF values below 10 indicate that multicollinearity does not pose a serious concern for the model [29].

Table 3. Variance Inflation Factor Value for Each Predictor

X_1	X_2	X_3	X_4	X_5	X_6
5.2978	5.2621	1.6549	1.7933	1.3789	1.7605

Based on the VIF calculations presented in Table 3, the obtained values are 5.2978 for X_1, 5.2621 for X_2, 1.6549 for X_3, 1.7933 for X_4, 1.3789 for X_5, and 1.7605 for X_6.

All predictor variables exhibit VIF values below the critical threshold, indicating no strong evidence of multicollinearity among the predictors. Therefore, all variables can be simultaneously included in the modeling process without causing distortion in parameter estimation.

3.3. Modeling Using the Mixed Truncated Spline and Gaussian Kernel Estimator

The modeling procedure employs a mixed truncated spline and Gaussian kernel estimator to accommodate the differing characteristics of the relationships between the response variable and each predictor. This approach provides flexibility in capturing local pattern changes through spline components while ensuring smooth functional relationships via the kernel estimator. In this study, the optimal number of knot points and kernel bandwidth are determined using three criteria: Cross Validation (CV), Unbiased Risk (UBR), and Generalized Cross Validation (GCV). Model performance is subsequently evaluated based on the Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). The modeling results based on the CV, UBR, and GCV criteria for one, two, and three knot points are summarized in Tables 4–6, respectively.

Table 4. Results of the Mixed Truncated Spline and Gaussian Kernel Estimator Modeling Based on the *Cross-Validation* (CV) Criterion

Number of Knot Points	Truncated Spline Component					Gaussian Kernel Component	CV
	Location of Knot Points					Bandwidth	
	X_1	X_2	X_3	X_5	X_6	X_4	
1	8.6551	2.00	20.9862	49.0275	21.6827	38.2689	9.327848×10^{-2}
2	3.9931; 9.3724	0.70; 2.20	13.0517; 22.2068	33.6965; 51.3862	19.4413; 41.1655	37.9448	3.001631×10^{-2}
3	**8.2965**; **9.3724**; **12.2413**	**1.90**; **2.20**; **3.00**	**20.3758**; **22.2068**; **27.0896**	**47.8482**; **51.3862**; **60.8206**	**36.8206**; **41.1655**; **52.7517**	**27.1034**	$\mathbf{2.001204 \times 10^{-2}}$

Table 5. Results of the Mixed Truncated Spline and Gaussian Kernel Estimator Modeling Based on the *Unbiased Risk* (UBR) Criterion

Number of Knot Points	Truncated Spline Component					Gaussian Kernel Component	UBR
	Location of Knot Points					Bandwidth	
	X_1	X_2	X_3	X_5	X_6	X_4	
1	3.9931	0.70	13.0517	33.6965	19.4413	37.9448	4.272226×10^{-5}
2	3.9931; 5.7862	0.70; 1.20	13.0517; 16.1034	33.6965; 39.5931	19.4413; 26.6827	31.1689	4.918123×10^{-7}
3	**3.2758**; **3.9931**; **4.7103**	**0.50**; **0.70**; **0.90**	**11.8310**; **13.0517**; **14.2724**	**31.3379**; **33.6965**; **36.0551**	**16.5448**; **19.4413**; **22.3379**	**33.8793**	$\mathbf{6.793533 \times 10^{-8}}$

Table 6. Results of the Mixed Truncated Spline and Gaussian Kernel Estimator Modeling Based on the *Generalized Cross-Validation* (GCV) Criterion

Number of Knot Points	Truncated Spline Component					Gaussian Kernel Component	GCV
	Location of Knot Points					Bandwidth	
	X_1	X_2	X_3	X_5	X_6	X_4	
1	2.5586	0.300	10.6103	28.9793	13.6482	1.3551	1.867289×10^{-2}
2	10.4482; 11.1655	2.50; 2.70	24.0379; 25.2586	54.9241; 57.2827	45.5103; 48.4069	1.3551	1.268352×10^{-2}
3	**7.9379**; **9.3724**; **11.1655**	**1.80**; **2.20**; **2.70**	**19.7655**; **22.2068**; **25.2586**	**46.6689**; **51.3862**; **57.2827**	**35.3724**; **41.1655**; **48.4068**	**1.3551**	$\mathbf{6.456789 \times 10^{-3}}$

Overall, the modeling results based on the three selection criteria CV, UBR, and GCV consistently indicate that the model with three knot points outperforms models with fewer knot points. This conclusion is supported by the

fact that the CV, UBR, and GCV values attain their minimum at the three-knot specification, suggesting an optimal balance between model goodness-of-fit and the complexity of the spline structure.

A comparison of the best-performing models under each criterion is presented in Tables 7. The CV-based model with three knot points yields an MSE of 0.0054, a MAPE of 9.871%, and an R^2 of 94.3310%. Meanwhile, the UBR-based model with three knot points produces an MSE of 0.0083, a MAPE of 9.6066%, and an R^2 of 92.9491%. The GCV-based model with three knot points demonstrates the best overall performance, achieving an MSE of 0.0052, a MAPE of 9.5559%, and the highest R^2 value of 95.4041%.

Equation (24) presents the best Mixed Truncated Spline and Gaussian Kernel Estimator model selected based on the CV criterion.

$$\begin{aligned}\hat{y}_i &= 0.3382 - 0.0105X_{1i} + 0.5068X_{2i} - 0.0125X_{3i} - 0.0202X_{5i} + 0.0026X_{6i} \\ &\quad + 0.1793\,(X_{1i} - 8.2965)_+ - 0.3066\,(X_{1i} - 9.3724)_+ - 0.3611\,(X_{1i} - 12.2414)_+ \\ &\quad - 0.1895\,(X_{2i} - 1.90)_+ + 0.8363\,(X_{2i} - 2.20)_+ - 0.1505\,(X_{2i} - 3.00)_+ \\ &\quad - 0.0862\,(X_{3i} - 20.3758)_+ + 0.1698\,(X_{3i} - 22.2069)_+ - 0.5679\,(X_{3i} - 27.0896)_+ \\ &\quad + 0.0803\,(X_{5i} - 47.8423)_+ - 0.1117\,(X_{5i} - 51.3862)_+ + 1.4073\,(X_{5i} - 60.8207)_+ \\ &\quad - 0.0813\,(X_{6i} - 36.8207)_+ + 0.1047\,(X_{6i} - 41.1655)_+ - 0.1735\,(X_{6i} - 52.7517)_+ \\ &\quad + \frac{1}{38}\sum_{i=1}^{38}\left\{\frac{\frac{1}{27.10345}K\left(\frac{X_i - X_{4i}}{27.10345}\right)}{\frac{1}{38}\sum_{i=1}^{38}\frac{1}{27.10345}K\left(\frac{X_i - X_{4i}}{27.10345}\right)}\right\}\end{aligned} \tag{24}$$

Equation (25) presents the best Mixed Truncated Spline and Gaussian Kernel Estimator model selected based on the UBR criterion.

$$\begin{aligned}\hat{y}_i &= -0.2359 + 3.7425X_{1i} - 0.4322X_{2i} - 0.9285X_{3i} + 0.0184X_{5i} + 0.0190X_{6i} \\ &\quad - 1.9488\,(X_{1i} - 3.2758)_+ - 1.5239\,(X_{1i} - 3.9931)_+ - 0.2776\,(X_{1i} - 4.7103)_+ \\ &\quad - 3.1420\,(X_{2i} - 0.50)_+ + 4.6485\,(X_{2i} - 0.70)_+ - 0.5330\,(X_{2i} - 0.90)_+ \\ &\quad - 0.4190\,(X_{3i} - 11.8310)_+ + 0.3040\,(X_{3i} - 13.0517)_+ + 1.0272\,(X_{3i} - 14.2724)_+ \\ &\quad + 0.1766\,(X_{5i} - 31.3379)_+ - 0.4214\,(X_{5i} - 33.6965)_+ + 0.2270\,(X_{5i} - 36.0551)_+ \\ &\quad + 0.0024\,(X_{6i} - 16.5448)_+ - 0.0307\,(X_{6i} - 19.4413)_+ + 0.0089\,(X_{6i} - 22.3379)_+ \\ &\quad + \frac{1}{38}\sum_{i=1}^{38}\left\{\frac{\frac{1}{33.8793}K\left(\frac{X_i - X_{4i}}{33.8793}\right)}{\frac{1}{38}\sum_{i=1}^{38}\frac{1}{33.8793}K\left(\frac{X_i - X_{4i}}{33.8793}\right)}\right\}\end{aligned} \tag{25}$$

Equation (26) presents the best Mixed Truncated Spline and Gaussian Kernel Estimator model selected based on the GCV criterion.

$$\begin{aligned}\hat{y}_i &= 0.3298 - 0.0057X_{1i} + 0.2026X_{2i} - 0.0292X_{3i} + 0.0001X_{5i} - 0.0031X_{6i} \\ &\quad - 0.0079\,(X_{1i} - 7.9379)_+ + 0.1811\,(X_{1i} - 9.3724)_+ - 0.5646\,(X_{1i} - 11.1655)_+ \\ &\quad - 0.0336\,(X_{2i} - 1.80)_+ - 0.4143\,(X_{2i} - 2.20)_+ + 2.1646\,(X_{2i} - 2.70)_+ \\ &\quad + 0.0250\,(X_{3i} - 19.7655)_+ + 0.0664\,(X_{3i} - 22.2068)_+ - 0.1307\,(X_{3i} - 25.2586)_+ \\ &\quad - 0.0488\,(X_{5i} - 46.6689)_+ + 0.0929\,(X_{5i} - 51.3862)_+ + 0.1226\,(X_{5i} - 57.2827)_+ \\ &\quad - 0.0010\,(X_{6i} - 35.3724)_+ + 0.0985\,(X_{6i} - 41.1655)_+ - 0.1826\,(X_{6i} - 48.4068)_+ \\ &\quad + \frac{1}{38}\sum_{i=1}^{38}\left\{\frac{\frac{1}{1.355172}K\left(\frac{X_i - X_{4i}}{1.355172}\right)}{\frac{1}{38}\sum_{i=1}^{38}\frac{1}{1.355172}K\left(\frac{X_i - X_{4i}}{1.355172}\right)}\right\}\end{aligned} \tag{26}$$

Table 7. Performance Comparison of the Best Models Based on CV, UBR, and GCV Criterion

Best Model	MSE	MAPE	R^2
3 Knots CV	0.0054	9.8710%	94.3310%
3 Knots UBR	0.0083	9.6066%	92.9491%
3 Knots GCV	**0.0052**	**9.5559%**	**95.4041%**

Based on the comparison in Table 7, the three-knot model selected using the GCV criterion is identified as the preferred model. The GCV-based model achieves the highest coefficient of determination and demonstrates relatively low and stable prediction errors compared to the alternative models. These results suggest that the GCV criterion provides a favorable balance between model fit and complexity, thereby reducing the risk of overfitting and yielding a model with good generalization performance. Accordingly, this model is adopted as the final model in the present study.

This finding is also consistent with previous studies [3, 22, 30], which report that the GCV method tends to outperform other selection criteria due to its ability to approximate prediction error efficiently while incorporating an implicit penalty on model complexity, resulting in more stable and reliable smoothing parameter selection.

Subsequently, a comparison was conducted between the best mixed estimator model based on the Generalized Cross Validation (GCV) criterion and the single estimator models. The comparison was carried out using three approaches: a model in which all predictor variables are modeled using truncated spline estimators (full spline), a model in which all variables are modeled using Gaussian kernel estimators (full kernel), and a mixed model constructed based on the identified data patterns.

Based on the Table 8, it can be observed that in the truncated spline model, the GCV value decreases as the number of knots increases, with the model using 3 knots yielding the smallest GCV value of 0.0132. This indicates that increasing flexibility through additional knots improves the model's ability to capture the underlying data patterns. However, the Gaussian kernel model as a single estimator produces a relatively larger GCV value of 0.0504, suggesting that the kernel-only approach is less optimal compared to the spline-based approach.

Meanwhile, the mixed truncated spline and Gaussian kernel model demonstrates the best performance among all models. This is evidenced by the lowest GCV value obtained under the 3-knot configuration, which is 0.0064. The substantial reduction in GCV for the mixed model highlights the advantage of the mixed estimator in providing greater flexibility in capturing complex data structures. Therefore, it can be concluded that the mixed truncated spline and Gaussian kernel model with 3 knots is the best model in this study, as it achieves the minimum GCV value.

Table 8. Comparison of GCV Values for Single and Mixed Model

Model	Configuration	GCV
Truncated Spline	1 Knot	0.0232
	2 Knots	0.0154
	3 Knots	0.0132
Gaussian Kernel with Optimal Bandwidth		0.0504
Mixed Truncated Spline and Gaussian Kernel Estimator	1 Knot	0.0187
	2 Knots	0.0127
	3 Knots	**0.0064**

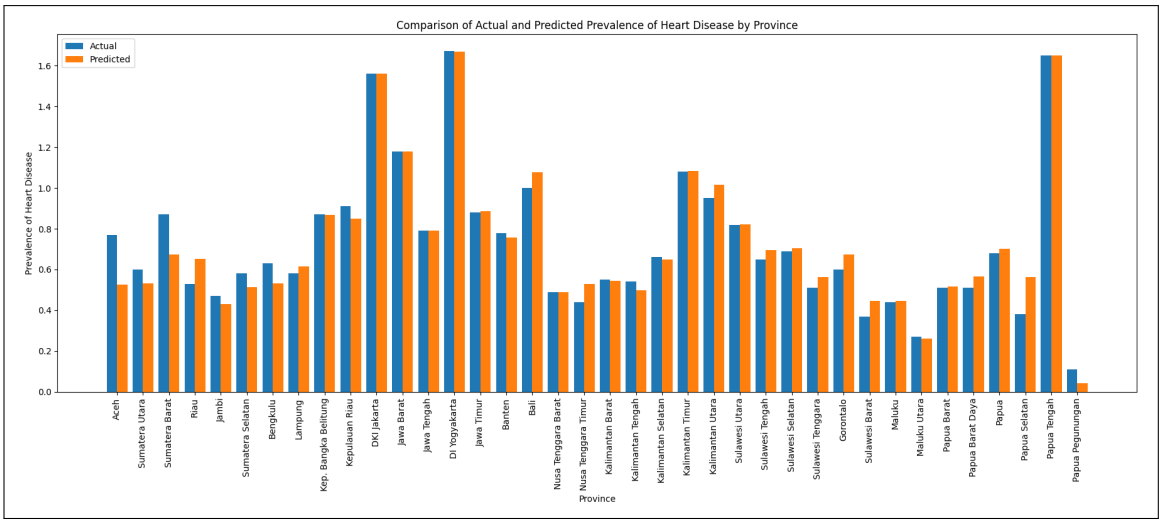

Figure 2. Comparison of Actual and Predicted Prevalence of Heart Disease Across Indonesian Provinces

Based on Figure 2, the side-by-side bar chart comparing the actual and predicted prevalence of heart disease across provinces shows that the predicted values closely follow the observed data for most regions. This indicates that the selected mixed estimator model provides a good overall fit and is able to capture the general spatial variation in heart disease prevalence. In provinces such as DKI Jakarta, West Java, Central Java, DI Yogyakarta, East Java, East Kalimantan, Papua Tengah, and Kepulauan Bangka Belitung, the predicted values are almost identical to the actual values. This suggests that the model performs particularly well in regions with moderate to high prevalence levels, where the underlying relationships between predictors and the response variable are more stable and well captured by the spline–kernel structure. Some discrepancies between actual and predicted values are observed in several provinces, including Aceh, Riau, Bali, Nusa Tenggara Timur, Papua Selatan, and Papua Pegunungan. In these regions, the model tends to slightly underpredict or overpredict the actual prevalence. Such deviations may reflect local heterogeneity, differences in healthcare access, lifestyle patterns, or unobserved regional factors that are not fully captured by the available predictors.

Notably, provinces with very low prevalence values, such as Papua Pegunungan, exhibit larger relative prediction errors, which is common in regression-based models when modeling extreme or boundary observations. However, the overall pattern remains consistent, and no systematic bias (persistent overestimation or underestimation across regions) is evident. Overall, the close alignment between actual and predicted values across most provinces supports the robustness of the final GCV-selected model. The results confirm that the mixed estimator approach effectively balances flexibility and stability, yielding reliable predictions and strong generalization performance for modeling heart disease prevalence at the provincial level.

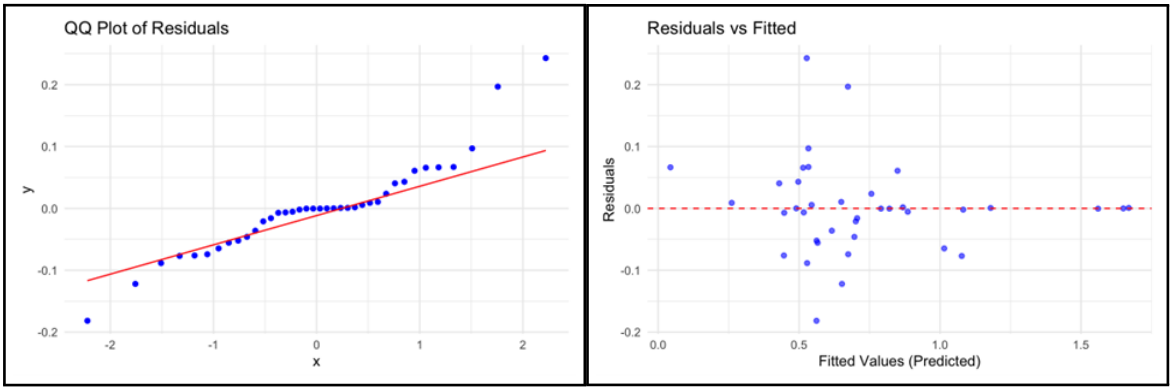

Figure 3. QQ Plot of Residuals and Plot of Residuals vs Fitted Values

Based on Figure 3, for the best model obtained, the residuals versus fitted values plot shows that the residuals are randomly scattered around the zero line without forming any specific pattern. This indicates that the model is able to adequately capture the relationship between the response variable and the predictor variables, and there is no indication of model misspecification. The relatively consistent spread of residuals across the range of fitted values further supports that the model provides a good fit to the data. Furthermore, the QQ plot of residuals shows that most of the points lie close to the diagonal line, particularly in the central part of the distribution, indicating that the normality assumption of the residuals is generally satisfied.

3.4. Discussion

The exploratory analysis using scatter plots indicates that the relationships between heart disease prevalence and the predictor variables exhibit heterogeneous characteristics. This suggests that relying on a single estimator may be insufficient to adequately capture the underlying complexity of these relationships. Certain variables, such as the prevalence of hypertension and diabetes mellitus, display approximately linear trends that are not globally consistent, while others exhibit nonlinear patterns. These findings provide strong justification for employing a nonparametric modeling approach based on a mixed truncated spline and Gaussian kernel estimator.

The modeling process was conducted using three model selection criteria: Cross-Validation (CV), Unbiased Risk (UBR), and Generalized Cross-Validation (GCV). The results show that increasing the number of knots to three consistently improves model performance, as indicated by reductions in MSE and MAPE, along with an increase in the coefficient of determination. This implies that additional knots enhance model flexibility in capturing changes in relationship patterns without substantially compromising model stability.

A comparison of the best models under each criterion reveals that the three-knot model selected using the GCV criterion provides relatively superior performance. This suggests that GCV achieves a more effective balance between model fit and spline complexity, thereby reducing the risk of overfitting and yielding a more robust model. Furthermore, a comparative analysis was conducted between the best mixed estimator model based on the GCV criterion and single-estimator models. The comparison involved three approaches: a full spline model, a full Gaussian kernel model, and a mixed model constructed based on the identified data patterns.

Based on Table 8, the truncated spline model exhibits decreasing GCV values as the number of knots increases, with the lowest value of 0.0132 obtained at three knots. This indicates that increasing model flexibility through additional knots improves the model's ability to capture underlying data patterns. In contrast, the Gaussian kernel model as a single estimator produces a relatively higher GCV value of 0.0504, suggesting that the kernel-only approach is less optimal than the spline-based approach.

Meanwhile, the mixed truncated spline and Gaussian kernel model demonstrates the best overall performance among all models. This is evidenced by the lowest GCV value of 0.0064 achieved under the three-knot configuration. The substantial reduction in GCV highlights the advantage of the mixed estimator in capturing

complex data structures more effectively. Therefore, the mixed model with three knots based on the GCV criterion can be regarded as the preferred model in this study.

Overall, the integration of truncated spline and Gaussian kernel estimators provides a flexible yet stable modeling framework capable of capturing both global structural changes and local variations. The findings confirm that the mixed estimator is effective in modeling complex relationships in health data. The selected best model, namely the three-knot GCV-based model, demonstrates good generalization performance and can serve as a reliable alternative for analyzing health-related data.

4. Conclusion

This study demonstrates that a nonparametric modeling approach using a mixed truncated spline and Gaussian kernel estimator is effective in capturing the complex relationships between health-related factors and heart disease prevalence. Among the evaluated models, the three-knot specification selected using the GCV criterion yielded the best performance, achieving a coefficient of determination (R^2) of approximately 95.4%, a minimum GCV value of 0.006456789, and relatively low prediction errors compared to the CV and UBR criteria.

In comparison, although the CV and UBR-based models produced comparable MSE values, the GCV model showed a better balance between model fit and complexity, indicating its superiority in controlling overfitting while maintaining high predictive accuracy. This result highlights that the GCV criterion is more reliable for selecting optimal smoothing parameters in the proposed mixed estimator framework.

Future research may extend this framework by incorporating spatial or spatio-temporal approaches to account for interregional dependence and temporal dynamics. Further developments in adaptive or Bayesian-based knot and bandwidth selection methods may also enhance model accuracy and robustness. The findings of this study contribute to the achievement of Sustainable Development Goal (SDG) 3: Good Health and Well-Being by providing an accurate and flexible modeling framework for identifying factors associated with heart disease prevalence.

Acknowledgement

This research is funded by the Indonesian Endowment Fund for Education (LPDP) on behalf of the Indonesian Ministry of Higher Education, Science and Technology and managed under the EQUITY Program (Contract No. 4299/B3/DT.03.08/2025 & No 3029/PKS/ITS/2025).

The Declaration of Conflict of Interest/ Common Interest

The authors declare no conflicts of interest.

REFERENCES

[1] I. N. Budiantara, V. Ratnasari, M. Ratna, and I. Zain. "The combination of spline and kernel estimator for nonparametric regression and its properties". In: *Applied Mathematical Sciences* vol. 9 (2015), pp. 6083–6094. ISSN: 13147552. DOI: 10.12988/ams.2015.58517. URL: http://www.m-hikari.com/ams/ams-2015/ams-121-124-2015/58517.html (visited on 01/12/2026).

[2] V. Ratnasari, I. N. Budiantara, M. Ratna, and I. Zain. "Estimation of nonparametric regression curve using mixed estimator of multivariable truncated spline and multivariable kernel". In: *Global Journal of Pure and Applied Mathematics* vol. 12, no. 6 (2016). Publisher: Research India Publications, pp. 5047–5057. URL: https://www.scopus.com/pages/publications/85019521760 (visited on 01/12/2026).

[3] A. T. R. Dani, V. Ratnasari, and I. N. Budiantara. "Optimal Knots Point and Bandwidth Selection in Modeling Mixed Estimator Nonparametric Regression". en. In: *IOP Conference Series: Materials Science and Engineering* vol. 1115, no. 1 (Mar. 2021). Publisher: IOP Publishing, pp. 012020. ISSN: 1757-899X. DOI: 10.1088/1757-899X/1115/1/012020. URL: https://doi.org/10.1088/1757-899X/1115/1/012020 (visited on 01/12/2026).

[4] Sifriyani, A. T. Rian Dani, M. Fauziyah, and I. N. Budiantara. "Statistical Modeling: A New Regression Curve Approximation using Mixed Estimators Truncated Spline and Epanechnikov Kernel." id. In: *Engineering Letters* vol. 31, no. 4 (Dec. 2023), pp. 1649. ISSN: 1816-093X. URL: https://openurl.ebsco.com/contentitem/gcd:173981991?sid=ebsco:plink:crawler&id=ebsco:gcd:173981991 (visited on 01/12/2026).

[5] M. A. D. Octavanny, I. N. Budiantara, H. Kuswanto, and D. P. Rahmawati. "A New Mixed Estimator in Nonparametric Regression for Longitudinal Data". en. In: *Journal of Mathematics* vol. 2021, no. 1 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/3909401, pp. 3909401. ISSN: 2314-4785. DOI: 10.1155/2021/3909401. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/3909401 (visited on 01/12/2026).

[6] R. Hidayat, I. N. Budiantara, B. W. Otok, and V. Ratnasari. "The regression curve estimation by using mixed smoothing spline and kernel (MsS-K) model". In: *Communications in Statistics - Theory and Methods* vol. 50, no. 17 (Aug. 2021). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610926.2019.1710201, pp. 3942–3953. ISSN: 0361-0926. DOI: 10.1080/03610926.2019.1710201. URL: https://doi.org/10.1080/03610926.2019.1710201 (visited on 01/12/2026).

[7] D. P. Rahmawati, I. N. Budiantara, D. D. Prastyo, and M. A. D. Octavanny. "Mixed Spline Smoothing and Kernel Estimator in Biresponse Nonparametric Regression". en. In: *International Journal of Mathematics and Mathematical Sciences* vol. 2021, no. 1 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/6611084, pp. 6611084. ISSN: 1687-0425. DOI: 10.1155/2021/6611084. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/6611084 (visited on 01/12/2026).

[8] M. A. D. Octavanny, I. N. Budiantara, H. Kuswanto, and D. P. Rahmawati. "Nonparametric Regression Model for Longitudinal Data with Mixed Truncated Spline and Fourier Series". en. In: *Abstract and Applied Analysis* vol. 2020, no. 1 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2020/4710745, pp. 4710745. ISSN: 1687-0409. DOI: 10.1155/2020/4710745. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2020/4710745 (visited on 01/12/2026).

[9] N. P. A. M. Mariati, I. N. Budiantara, V. Ratnasari, N. P. A. M. Mariati, I. N. Budiantara, and V. Ratnasari. "The Application of Mixed Smoothing Spline and Fourier Series Model in Nonparametric Regression". en. In: *Symmetry* vol. 13, no. 11 (Nov. 2021). Company: Multidisciplinary Digital Publishing Institute, Distributor: Multidisciplinary Digital Publishing Institute, Institution: Multidisciplinary Digital Publishing Institute, Label: Multidisciplinary Digital Publishing Institute. ISSN: 2073-8994. DOI: 10.3390/sym13112094. URL: https://www.mdpi.com/2073-8994/13/11/2094 (visited on 01/12/2026).

[10] M. A. D. Octavanny, I. N. Budiantara, H. Kuswanto, and D. P. Rahmawati. "The Estimation of a Regression Curve by Using Mixed Truncated Spline and Fourier Series Models for Longitudinal Data." id. In: *Engineering Letters* vol. 30, no. 1 (Mar. 2022), pp. 1. ISSN: 1816-093X. URL: https://openurl.ebsco.com/contentitem/gcd:155423506?sid=ebsco:plink:crawler&id=ebsco:gcd:155423506 (visited on 01/12/2026).

[11] N. Afifah, I. N. Budiantara, and I. N. Latra. "Mixed Estimator of Kernel and Fourier Series in Semiparametric Regression". en. In: *Journal of Physics: Conference Series* vol. 855, no. 1 (June 2017). Publisher: IOP Publishing, pp. 012002. ISSN: 1742-6596. DOI: 10.1088/1742-6596/855/1/012002. URL: https://doi.org/10.1088/1742-6596/855/1/012002 (visited on 01/12/2026).

[12] M. F. F. Mardianto, S. H. Kartiko, and H. Utami. "Forecasting Trend-Seasonal Data Using Nonparametric Regression with Kernel and Fourier Series Approach". en. In: *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*. Ed. by L.-K. Kor, A.-R. Ahmad, Z. Idrus, and K. A. Mansor. Singapore: Springer Singapore, 2019, pp. 343–349. ISBN: 978-981-13-7278-0, 978-981-13-7279-7. DOI: 10.1007/978-981-13-7279-7_42. URL: http://link.springer.com/10.1007/978-981-13-7279-7_42 (visited on 01/12/2026).

[13] I. N. Budiantara, V. Ratnasari, M. Ratna, W. Wibowo, N. Afifah, D. P. Rahmawati, and M. A. D. Octavanny. "Modeling Percentage of Poor People in Indonesia Using Kernel and Fourier Series Mixed Estimator in Nonparametric Regression." Spanish. In: *Investigación Operacional* vol. 40, no. 4 (Sept. 2019). Publisher: Editorial Universitaria de la Republica de Cuba, pp. 538–551. ISSN: 02574306. URL: https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=02574306&v=2.1&it=r&id=GALE%7CA600270106&sid=googleScholar&linkaccess=abs (visited on 01/12/2026).

[14] G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Jan. 1990. ISBN: 978-0-89871-244-5. DOI: 10.1137/1.9781611970128. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611970128 (visited on 01/12/2026).

[15] W. Härdle. *Applied Nonparametric Regression*. Econometric Society Monographs. Cambridge: Cambridge University Press, 1990. ISBN: 978-0-521-42950-4. DOI: 10.1017/CCOL0521382483. URL: https://www.cambridge.org/core/books/applied-nonparametric-regression/4C646486EADDAC0737AE339FF76A3BA7 (visited on 01/12/2026).

[16] P. Speckman. "Spline Smoothing and Optimal Rates of Convergence in Nonparametric Regression Models". In: *The Annals of Statistics* vol. 13, no. 3 (Sept. 1985). Publisher: Institute of Mathematical Statistics, pp. 970–983. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176349650. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-13/issue-3/Spline-Smoothing-and-Optimal-Rates-of-Convergence-in-Nonparametric-Regression/10.1214/aos/1176349650.full (visited on 01/12/2026).

[17] J. Rice. "Bandwidth Choice for Nonparametric Regression". In: *The Annals of Statistics* vol. 12, no. 4 (Dec. 1984). Publisher: Institute of Mathematical Statistics, pp. 1215–1230. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176346788. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-12/issue-4/Bandwidth-Choice-for-Nonparametric-Regression/10.1214/aos/1176346788.full (visited on 01/12/2026).

[18] Rismal, I. N. Budiantara, and D. D. Prastyo. "Mixture model of spline truncated and kernel in multivariable nonparametric regression". In: *AIP Conference Proceedings* vol. 1739, no. 1 (June 2016), pp. 020085. ISSN: 0094-243X. DOI: 10.1063/1.4952565. URL: https://doi.org/10.1063/1.4952565 (visited on 01/12/2026).

[19] S. Sifriyani, A. T. R. Dani, M. Fauziyah, M. N. Hayati, S. Wahyuningsih, and S. Prangga. "Spline and Kernel Mixed Estimators in Multivariable Nonparametric Regression for Dengue Hemorrhagic Fever Model". In: *Commun. Math. Biol. Neurosci.* vol. 2023, no. 0 (June 2023), pp. Article ID 11. ISSN: 2052-2541. DOI: 10.28919/cmbn/7790. URL: https://scik.org/index.php/cmbn/article/view/7790 (visited on 01/12/2026).

[20] P. Craven and G. Wahba. "Smoothing Noisy Data with Spline Functions". en. In: *Numerische Mathematik* vol. 31, no. 4 (Dec. 1978), pp. 377–403. ISSN: 0945-3245. DOI: 10.1007/BF01404567. URL: https://doi.org/10.1007/BF01404567 (visited on 01/12/2026).

[21] L. N. Berry, N. E. Helwig, L. N. Berry, and N. E. Helwig. "Cross-Validation, Information Theory, or Maximum Likelihood? A Comparison of Tuning Methods for Penalized Splines". en. In: *Stats* vol. 4, no. 3 (Sept. 2021). Company: Multidisciplinary Digital Publishing Institute, Distributor: Multidisciplinary Digital Publishing Institute, Institution: Multidisciplinary Digital Publishing Institute, Label: Multidisciplinary

Digital Publishing Institute, pp. 701–724. ISSN: 2571-905X. DOI: 10.3390/stats4030042. URL: https://www.mdpi.com/2571-905X/4/3/42 (visited on 01/12/2026).

[22] V. Ratnasari, I. N. Budiantara, and A. T. R. Dani. "Nonparametric Regression Mixed Estimators of Truncated Spline and Gaussian Kernel based on Cross-Validation (CV), Generalized Cross-Validation (GCV), and Unbiased Risk (UBR) Methods". en. In: *International Journal on Advanced Science, Engineering and Information Technology* vol. 11, no. 6 (Dec. 2021), pp. 2400–2406. ISSN: 2460-6952. DOI: 10.18517/ijaseit.11.6.14464. URL: https://ijaseit.insightsociety.org/index.php/ijaseit/article/view/14464 (visited on 01/12/2026).

[23] G. Wahba and Y. Wang. "Spline Function: Overview". In: *University of Wisconsin-Madison* (2015). URL: http://pages.stat.wisc.edu/~wahba/stat860public/bigpicture/wahba.wang.overview2015.pdf (visited on 01/12/2026).

[24] G. Wahba. "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression". In: *Journal of the Royal Statistical Society: Series B (Methodological)* vol. 40, no. 3 (July 1978), pp. 364–372. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1978.tb01050.x. URL: https://doi.org/10.1111/j.2517-6161.1978.tb01050.x (visited on 01/12/2026).

[25] G. H. Golub, M. Heath, and G. Wahba. "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter". In: *Technometrics* vol. 21, no. 2 (May 1979). Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1979.10489751, pp. 215–223. ISSN: 0040-1706. DOI: 10.1080/00401706.1979.10489751. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1979.10489751 (visited on 01/12/2026).

[26] J. D. Carew, G. Wahba, X. Xie, E. V. Nordheim, and M. E. Meyerand. "Optimal spline smoothing of fMRI time series by generalized cross-validation". In: *NeuroImage* vol. 18, no. 4 (Apr. 2003), pp. 950–961. ISSN: 1053-8119. DOI: 10.1016/S1053-8119(03)00013-2. URL: https://www.sciencedirect.com/science/article/pii/S1053811903000132 (visited on 01/12/2026).

[27] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. en. Google-Books-ID: 0yR4KUL4VDkC. John Wiley & Sons, Apr. 2012. ISBN: 978-0-470-54281-1.

[28] N. Mohamed, M. H. Ahmad, Z. Ismail, and S. Suhartono. "Short Term Load Forecasting Using Double Seasonal ARIMA Model". In: *Proc. Regional Conf. on Statist. Sci*. Vol. 10. 2010, pp. 57–73. URL: https://www.academia.edu/download/4477613/07p.pdf (visited on 01/12/2026).

[29] D. G. Kleinbaum, L. L. Kupper, A. Nizam, and E. S. Rosenberg. *Applied Regression Analysis and Other Multivariable Methods*. en. 5 ed. Google-Books-ID: v590AgAAQBAJ. Boston, MA: Cengage Learning, Aug. 2013. ISBN: 978-1-285-96375-4.

[30] 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". en. In: *Journal of Physics: Conference Series* vol. 1446, no. 1 (Jan. 2020), pp. 012049. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1446/1/012049. URL: https://doi.org/10.1088/1742-6596/1446/1/012049 (visited on 03/23/2026).