# PEMILIHAN PARAMETER PENGHALUS DALAM REGRESI SPLINE LINIER

**Agustini Tripena Br.Sb.**

Fakultas Sains dan Teknik, Universitas Jenderal Soedirman
Purwokerto, Indonesia

**ABSTRAK.** Pada paper ini dibahas pemilihan parameter penghalus untuk estimasi regresi spline linier pada data beda potensial listrik dalam limbah cair. Metode yang digunakan adalah *mean square error* (MSE) dan *generalized cross validation* (GSV). Hasil penelitian menunjukkan bahwa dalam pemilihan metode *mean square error* (MSE) memberikan nilai parameter penghalus lebih kecil dari pada metode *generalized cross validation* (GCV). Ini berarti bahwa untuk kasus data beda potensial listrik dalam limbah cair metode *mean square error* (MSE) merupakan metode yang terbaik untuk mengestimasi parameter penghalus dari regresi spline linier.

**Kata Kunci: regresi spline linier, metode *mean square errorr, metode generalized cross validation*.**

***ABSTRACT.*** This paper discusses aselection of smoothing *parameters* for the *linier* spline regression estimation *o*n the data of electrical voltage differences in the wastewater. The selection methods *are based on the* mean square errorr *(MSE)* and generalized cross validation (GCV). *The results show that in selection of smooting paranceus the* mean square error (MSE) method *gives smaller value , than that of the* generalized cross validatio (GCV) method*. It means that for our data case the* errorr mean square (MSE) is the best *selection* method *of smoothing parameter for the linear spline regression estimation.*

***Keywords: linear spline regression,* mean square errorr method, generalized cross validation method**

## 1. Pendahuluan

Analisa regresi merupakan metode yang banyak digunakan untuk mengetahui hubungan antara sepasang variabel atau lebih. Misalkan $y$ adalah variabel respon dan x adalah variabel prediktor, maka hubungan variabel $x$ dan $y$ dapat dinyatakan sebagai

$$y = f(x_i) + \varepsilon_i , \qquad i = 1, 2, ..., n . \ (1) \tag{1}$$

dengan $\varepsilon_i$ adalah error random yang diasumsikan independen dengan mean nol dan variansi $\sigma^2$ sedangkan $f(x_i)$ merupakan fungsi kurva regresi. Untuk mengestimasi $f(x_i)$ ada dua estimasi yang dapat digunakan yaitu estimasi regresi parametrik dan regresi nonparametrik (Hardle, 1990). Estimasi regresi parametrik digunakan bila bentuk fungsi $f(x_i)$ diketahui dari informasi sebelumnya berdasarkan teori ataupun pengalaman masa lalu. Jadi dalam hal ini, estimasi untuk $f(x_i)$ eqivalen dengan estimasi parameter. Sementara itu, pada estimasi regresi nonparametrik tidak diberikan asumsi terhadap bentuk kurva regresi sehingga estimasi regresi nonparametrik memiliki fleksibelitas yang tinggi untuk mengestimasi kurva regresi $f(x_i)$. Dalam hal ini fungsi regresi $f(x_i)$ hanya diasumsikan termuat dalam suatu ruang fungsi tertertentu, dan pemilihan ruang fungsi tersebut biasanya dimotivasi oleh sifat kemulusan *(smoothness)* yang dimiliki oleh fungsi $f(x_i)$ .

Beberapa penulis seperti Hardle (1990), Wahba (1990), Budiantara dan Subanar (1997) menyarankan penggunaan regresi nonparametrik sebagai estimasi untuk model data, agar mempunyai fleksibelitas yang baik. Beberapa model pendekatan dalam regresi nonparametrik, yang cukup populer untuk mengestimasi fungsi $f(x_i)$ antara lain adalah regresi spline (Craven dan Wahba, 1979), kernel (Rosenblatt, 1971), dan deret Fourier dan lain-lain.. Bentuk estimator spline sangat dipengaruhi oleh nilai parameter penghalus (λ) (Budihantara, 2000). Oleh karena itu, pemilihan nilai parameter penghalus (λ) optimal mutlak diperlukan untuk memperoleh estimator spline yang sesuai dengan data.

Bentuk estimator spline juga dipengaruhi oleh lokasi dan banyaknya titik-titik knot. Nilai parameter penghalus yang sangat besar akan menghasilkan bentuk kurva regresi yang sangat halus; sebaliknya nilai parameter penghalus yang kecil memberikan bentuk kurva regresi yang sangat kasar (Wahba, 1990; Eubank, 1988; Budiantara, 1998). Pada paper ini, dibahas mengenai pemilihan parameter penghalus (λ) untuk estimasi spline linier pada data pengaruh penambahan potensial listrik dalam limbah cair.

## 2. Regresi Spline

Menurut Eubank (1988), estimasi terhadap $f(x)$ adalah $f_\lambda(x)$ yakni estimator yang mulus. Bentuk umum regresi spline orde ke-$m$ sebagai berikut:

$$y = \beta_0 + \sum_{j=1}^{m} \beta_j x^j + \sum_{k=1}^{N} \beta_{j+k} (x - K_k)_+^m + \varepsilon \qquad (2)$$

Dengan menggunakan data amatan sebanyak $n$, maka bentuk matriks dari persamaan (2) adalah

$$\mathbf{y} = \mathbf{X_1}\boldsymbol{\delta_1} + (\mathbf{X} - \mathbf{K})\boldsymbol{\delta_2} + \boldsymbol{\varepsilon} \qquad (3)$$

dengan

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}; \boldsymbol{\delta_1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}; \mathbf{X_1} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & \square_n^m \end{bmatrix}; \boldsymbol{\delta_1} = \begin{bmatrix} \beta_{m+1} \\ \beta_{m+2} \\ \beta_{m+3} \\ \vdots \\ \beta_{m+N} \end{bmatrix}$$

$$(\mathbf{X} - \mathbf{K}) = \begin{bmatrix} (x_1 - k_1)^m & (x_1 - k_2)^m & \cdots & (x_1 - k_N)^m \\ (x_2 - k_1)^m & (x_2 - k_2)^m & \cdots & (x_2 - k_N)^m \\ \vdots & \vdots & \ddots & \vdots \\ (x_n - k_1)^m & (x_n - k_2)^m & \cdots & (x_n - k_N)^m \end{bmatrix}$$

Untuk alasan kesederhanaan, maka matriks (3) dapat ditulis kembali menjadi

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (4)$$

dengan $\mathbf{X} = [\mathbf{X_1}\ (\mathbf{X} - \mathbf{K})]$ dan $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\delta_1} \\ \\ \boldsymbol{\delta_2} \end{bmatrix}$

Dalam hubungannya dengan estimasi kurva mulus $f(x)$, yang mempunyai nilai parameter penghalus ($\lambda$ ) optimal, maka untuk memilih estimator $f(x)$ yang terbaik diantara kelas estimator $C(\Lambda) = \{f_\lambda : \lambda \in \Lambda, \Lambda =$ himpunan indeks$\}$. Himpunan indeks merupakan himpunan yang berisi indeks-indeks. Dengan menggunakan model regresi spline sebagai estimasi kurva mulus $f_\lambda$, dilakukan penyesuaian persamaan menjadi

$$\mathbf{b}_\lambda = \widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X'}_\lambda \mathbf{X}_\lambda)^{-1} \mathbf{X'}_\lambda \, \mathbf{y} \tag{5}$$

Dengan $\mathbf{X}_\lambda$ adalah matriks disain dari model yang membentuk model estimasi $f_\lambda$ dengan $\lambda$ yang optimal. Dalam hal ini,

$$\begin{aligned} \boldsymbol{f}_\lambda &= \mathbf{X}_\lambda \mathbf{b}_\lambda \\ &= \mathbf{X}_\lambda (\mathbf{X'}_\lambda \mathbf{X}_\lambda)^{-1} \mathbf{X'}_\lambda \, \mathbf{y} \\ &= \mathbf{H}_\lambda \, \mathbf{y} \quad , \lambda \in \Lambda \end{aligned} \tag{6}$$

dengan $\mathbf{H}_\lambda = \mathbf{X}_\lambda (\mathbf{X'}_\lambda \mathbf{X}_\lambda)^{-1} \mathbf{X'}_\lambda$. Perlu dicatat $\mathbf{H}_\lambda$ bersifat simetris, definit positif, dan idempoten. Untuk mendapatkan kurva mulus yang mempunyai $\lambda$ optimal menggunakan data amatan sebanyak $n$, diperlukan ukuran kinerja atas estimator yang dapat diterima secara universal. Eubank (1988) menyebutkan, ukuran kinerja atas estimator tersebut adalah:

a. *Mean Squared Error* ($MSE$)

Ukuran kinerja atas estimator yang sederhana adalah kuadrat dari sisaan yang dirata-rata. Rata-rata kuadrat sisaan diberikan oleh

$$MSE(\lambda) = n^{-1}(y - f_\lambda)'(y - f_\lambda)$$

atau

$$MSE(\lambda) = n^{-1} \sum_{i=1}^{n} (y_i - f_\lambda(x_i))^2 \tag{7}$$

b. *Generalized Cross-Validation* ($GCV$)

Menurut Budihantara (2005), GCV merupakan modifikasi dari *Cross-Validation* (CV). *Cross-Validation* (CV) merupakan suatu metode untuk memilih

model berdasarkan pada kemampuan prediksi dari model tersebut. CV adalah metode untuk memilih $\lambda$ yang meminimumkan

$$CV(\lambda) = n^{-1}\sum_{i=1}^{n}\left(\frac{y_i - f_\lambda(x_i)}{1-h_{ii\cdot\lambda}}\right)^2 \quad (8)$$

dengan $h_{ii\cdot\lambda}$ adalah elemen diagonal ke-i dari matriks $\mathbf{H_\lambda}$. $GCV$ diperoleh dengan mengganti $h_{ii\cdot\lambda}$ pada persamaan (8) dengan $n^{-1}\sum_{i=1}^{n} h_{ii\cdot\lambda} = n^{-1}Tr(\mathbf{H_\lambda})$. Fungsi $GCV$ didefinisikan sebagai:

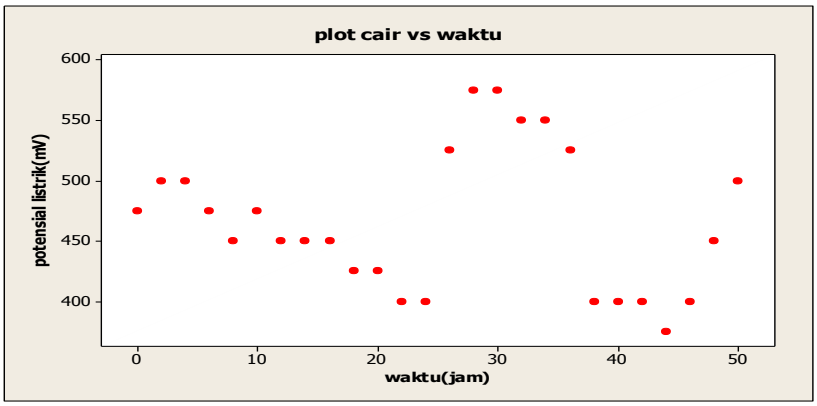
$$GCV(\lambda) = n^{-1}\frac{\sum_{i=1}^{n}(y_i - f_\lambda(x_i))^2}{(1-n^{-1}Tr(\mathbf{H_\lambda}))^2} = \frac{MSE(\lambda)}{\{n^{-1}Tr(I-\mathbf{H_\lambda})\}^2} \quad (9)$$

dengan $Tr(\mathbf{H_\lambda}) < n$. Kedua kriteria tersebut, baik $MSE(\lambda)$ ataupun $GCV(\lambda)$ diharapkan memiliki nilai yang minimum sehingga model regresi spline dapat dikatakan memiliki nilai $\lambda$ yang optimal.

## 3. Pemilihan Model Regresi Spine dengan $\lambda$ yang optimal.

### 3.1. Pembentukan Model Regresi Spline

Plot data pengaruh penambahan beda potensial listrik dalam limbah cair disajikan pada Gambar 1.




**Gambar 1.** Plot data pengaruh penambahan beda potensial listrik dalam limbah cair.

Gambar 1 plot menunjukkan bahwa ada indikasi perubahan pola perilaku dari variabel bebas pada sub-sub interval tertentu. Selanjutnya, pola data akan didekati dengan pendekatan regresi nonparametrik spline linier. Terdapat 24 titik

knot yang dapat digunakan untuk membentuk model spline. Banyaknya kombinasi titik knot yang bisa digunakan untuk membentuk model spline dengan empat titik knot adalah sebanyak 10.630 kombinasi. Persamaan regresi spline yang digunakan pada data ini adalah model spline dengan *intersep* $(\beta_0)$ karena pada awal pengukuran sudah diperoleh besarnya beda potensial listrik.

### 3.2. Estimasi Regresi Spline Linier

Model umum dari regresi spline linier adalah

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{N} \beta_{1+k}(x_i - K_k)_+ + \varepsilon_i \,; \quad \text{dengan konstanta}$$

$$y_i = \beta_1 x_i + \sum_{k=1}^{N} \beta_{1+k}(x_i - K_k)_+ + \varepsilon_i \quad ; \quad \text{tanpa konstanta}$$

Fungsi spline linier merupakan fungsi spline dengan satu orde. Bentuk fungsi spline linier dengan satu titik knot

$$f_1(x) = \beta_0 + \beta_1 x + \beta_2 (x - K)_+^1 \tag{10}$$

Persamaan (10) dapat disajikan menjadi (Tripena, 2005)

$$f_1(x) = \begin{cases} \beta_0 + \beta_1 x & , x < K \\ \beta_0 + \beta_1 x + \beta_2 (x - K), & x \geq K \end{cases} \tag{11}$$

Estimasi regresi spline linier dengan menggunakan tiga titik knot (K) dari data yang digunakan mempunyai model sebagai berikut:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - K_1)_+ + \beta_3 (x_i - K_2)_+ + \beta_4 (x_i - K_3)_+ + \varepsilon_i \tag{12}$$

Pemilihan titik knot yang optimal terletak pada nilai MSE dan GCV yang minimum. Model regresi spline linier dengan empat titik knot adalah

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - K_1)_+ + \beta_3 (x_i - K_2)_+ + \beta_4 (x_i - K_3)_+ + \beta_5 (x_i - K_4)_+ \ \varepsilon_i \quad (13)$$

Pemilihan titik knot dengan metode MSE dan GCV minimum untuk model regresi spline linier dengan empat titik knot dapat dilihat pada Tabel 1.

**Tabel 1.** Nilai $MSE$ dan $GCV$ model regresi spline linier dengan empat titik knot

| No | Titik knot | Nilai $MSE$ | Nilai $GCV$ |
|---|---|---|---|
| 1 | 8,12,18,26 | 263,8931 | 439,3273 |
| 2 | 10,12,20,24 | 173,423 | 254,5532 |

Titik knot yang optimal berada pada titik $K_1$= 10, $K_2$= 12, $K_3$= 20, dan $K_4$= 24 dengan nilai MSE minimum sebesar 173,423 dan nilai GCV minimum sebesar 254,5532. Estimasi model regresi spline linier empat titik knot diberikan pada Tabel 2 berikut.

**Tabel 2.** Estimasi model regresi spline linier dengan empat titik knot

| Parameter | Estimasi |
|---|---|
| $\beta_0$ | 536,031775 |
| $\beta_1$ | -1,317260 |
| $\beta_2$ | -65,101832 |
| $\beta_3$ | 65,275544 |
| $\beta_4$ | 27,716613 |
| $\beta_5$ | 30,384043 |

Estimasi model regresi spline linier dengan empat titik knot $K_1$=10, $K_2$=12, $K_3$=20, dan $K_4$= 4 adalah

$$\widehat{y_i} = 536{,}031775 - 1{,}317260\, x_{\mathrm{i}} - 65{,}101832(x_i - 10)_+ + 65{,}275544(x_i - 12)_+ + 27{,}716613(x_i - 20)_+ - 30{,}384043(x_i - 24)_+$$

### 3.3. Pemilihan Model Regresi Spline Terbaik

Titik knot $(K)$ yang paling optimal dengan nilai MSE dan GCV minimum adalah penggunaan empat titik knot pada regresi spline linier. Nilai MSE dan GCV model regresi spline dengan empat titik knot ditunjukkan pada Tabel 3

**Tabel 3.** Nilai MSE dan GCV beberapa model regresi spline dengan beberapa titik knot

| Orde | Model | Jumlah Knot $(K)$ | Letak Titik Knot $(K)$ | | | | Nilai $MSE\ (\lambda)$ optimal | Nilai $GCV\ (\lambda)$ optimal |
|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | | |
| 1 | Linier | 4 | 10 | 12 | 20 | 24 | 173,423 | 254,5532 |

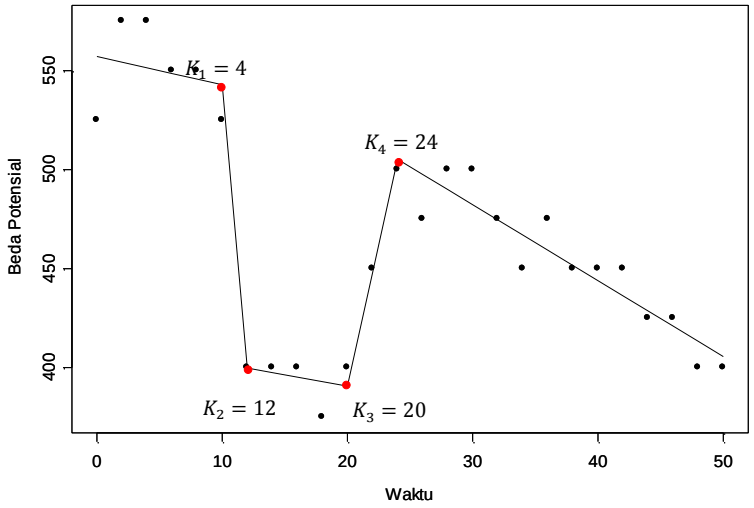
Berdasarkan Tabel 3 dapat disimpulkan bahwa model terbaik untuk data pengaruh penambahan beda potensial listrik dalam limbah cair adalah model regresi spline linier dengan empat titik knot $K_1$ = 10, $K_2$ = 12, $K_3$ = 20, dan $K_4$ = 24 yakni

$$\widehat{y_i} = 536{,}031775 - 1{,}317260\, x_{\mathrm{i}} - 65{,}101832(x_i - 10)_+ + 65{,}275544(x_i - 12)_+ + 27{,}716613(x_i - 20)_+ - 30{,}384043(x_i - 24)_+$$

Estimasi model regresi spline linier dengan empat titik knot dapat disajikan pula dalam bentuk fungsi terpotong (*truncated*) diberikan oleh

$$\hat{y}_i = \begin{cases} 536{,}031775 - 1{,}317260\, x_i,\ x_i < 10 \\ 1147{,}063257 - 66{,}418892\, x_i\ , 10 \le x_i < 12 \\ 412{,}447691 - 1{,}317060\, x_i\ , 12\ \le x_i < 20 \\ -171{,}527551 + 26{,}409553\, x_i\ , 20 \le x_i < 24 \\ 578{,}156877 - 1{,}75814672\, x_i\ , x_i \ge 24 \end{cases}$$

Semantara plot estimasi model regresi spline linier dengan empat titik knot yang merupakan model regresi spline terbaik berdasarkan kriteria nilai MSE dan GCV minimum diberikan pada Gambar 2.




**Gambar 2.** Kurva estimasi regresi spline linier dengan empat titik knot yang merupakan kurva regresi spline terbaik

Nilai koefisien determinasi $(R^2)$ sebesar 0,9344868 berarti bahwa variabel pemberian beda potensial tambahan mampu menerangkan sebesar 93,44868% terhadap potensial listrik yang dihasilkan dalam limbah cair

### 3.4. Pengujian Model Regresi Spline Terbaik

Uji hipotesis untuk pemeriksanaan model, dilakukan dengan hipotesis

$H_0$: Model tidak sesuai dengan data atau $\beta_0 = \beta_1 = \cdots = \beta_k = 0, i = 0{,}1, .., k$

$H_1$: Model sesuai dengan data atau minimal terdapat satu $\beta_i \neq 0, i = 0,1, \ldots, k$

untuk tingkat signifikansi 5%, diperoleh analisis variansi pada Tabel 4 berikut ini

**Tabel 4.** Analisis variansi untuk model regresi spline terbaik

| *Source of Variance* | *Degree of freedom (df)* | *Sum Square (SS)* | *Mean Square (MS)* | *F* |
|---|---|---|---|---|
| *Regression* | 5 | $SS_R$ = 72.732,06 | $MS_R$ = 4.546,412 | 24,96604 |
| *Error* | 20 | $SS_E$ = 3.642,077 | $MS_E$ = 182,10385 | |
| *Total* | 25 | $SS_T$ = 77.374,137 | | |

Dengan menggunakan $F_{tabel}$, diperoleh $F_{\alpha/2,p,(n-(p+1))} = F_{0.025,5,20} = 3{,}28906$, sehingga diperoleh $F_{Hitung} = 24{,}96604 \geq F_{0.025,5,20} = 3{,}28906$. Hal ini mengidentifikasi bahwa $H_0$ ditolak, artinya model berpengaruh terhadap data. Jadi dapat disimpulkan bahwa model regresi spline linier dengan titik-titik knot 10, titik 12, titik 20,dan titik 24 cukup memadai sebagai model estimasi untuk data pengaruh penambahan beda potensial listrik dalam limbah cair pada waktu tertentu.

## 4. Kesimpulan

a) Titik knot yang optimal diperoleh menggunkan empat titik knot yaitu $K_1$ = 10, $K_2$ = 12, $K_3$ = 20, dan $K_4$ = 24.

b) Pemilihan model regresi spline terbaik dengan menggunakan metode *mean square error* memberikan parameter penghalus = 173,423, dengan menggunakan metode *generalized cross validation* memberikan parameter penghalus = 254,5532, karena nilai *mean square error* paling minimum maka metode yang terbaik adalah metode *mean square error*

c) Nilai koefisien determinasi $(R^2)$ sebesar 0,9344868, berarti bahwa pemberian beda potensial tambahan pada waktu tertentu mengakibatkan perubahan sebesar 93,44868% pada beda potensial listrik yang dihasilkan dalam limbah cair.

## 5. DAFTAR PUSTAKA

Budiantara, I. N, 2005. *Penentuan Titik-Titik Knots dalam Regresi Spline* , Jurnal Jurusan Statistika FMIPA-ITS, Surabaya.

Budiantara, I. N, Subanar. 1997. *Pemilihan Parameter Penghalus dalam Regresi Spline Terbobot.* Jurnal Jurusan Statistika FMIPA-ITS, Surabaya.

Eubank, R. 1988. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

Hardle, W. 1990. *Applied Nonparametric Regression.* Cambridge University Press, New York.

Tripena, A. 2005. *Pendekatan Model Regresi Spline Linier* . Jurusan MIPA, Fakultas Sains dan Teknik, UNSOED.

Wahba, G. 1990. *Spline Models For Observasion Data*. SIAM Pensylvania.