PROSIDING SEMINAR NASIONAL STATISTIKA
UNIVERSITAS DIPONEGORO 2011
ISBN: 978-979-097-142-4

PENENTUAN MODEL REGRESI SPLINE TERBAIK

Agustini Tripena [1]

[1)] Program Studi Matematika, Fakultas Sains dan Teknik,

Univesitas Jenderal Soedirman, Purwokerto

tripena1960@yahoo.co.id

Abstrak

Pada paper ini dibahas penentuan model regresi spline terbaik pada data polusi kadar debu (jam) dengan konsentrasi pektra cerobong asap. Metode yang digunakan adalah *mean square errorr* dan *generalized cross validation*. Untuk data yang dipunyai metode *mean square error* memberikan nilai parameter penghalus lebih kecil dari pada metode *generalized cross validation*. Dengan demikian metode *mean square errorr* merupakan metode yang terbaik untuk mengestimasi metode regresi spline untuk polusi kadar debu (jam) dengan konsentrasi pektra cerobong asap.

Hasil penelitian menunjukkan bahwa estimasi regresi spline terbaik untuk data tersebut adalah model regresi spline linier. Titik-titik knot yang optimal adalah tiga titik knot dengan nilainya masing-masing adalah K_1 = 4, K_2 = 18, dan K_3 = 24. Pemilihan model regresi spline terbaik menggunakan metode $\mathrm{MSE}(\lambda)$ dan $\mathrm{GCV}(\lambda)$ menghasilkan nilai $\mathrm{MSE}(\lambda)$ sebesar 205,243 dan nilai $\mathrm{GCV}(\lambda)$ sebesar 411,3228. $\mathrm{MSE}(\lambda)$ mempunyai nilai yang lebih kecil dibandingkan nilai $\mathrm{GCV}(\lambda)$, maka metode yang terbaik (optimal) adalah metode $\mathrm{MSE}(\lambda)$.

Kata Kunci: Nonparametrik, Regresi spline, Titik-titik knot, $\boldsymbol{MSE}$ dan $\boldsymbol{GCV}$

1. Pendahuluan

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

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

ε_i adalah error random yang diasumsikan independen dengan mean nol dan variansi σ^2 dan $f(x_i)$ merupakan kurva regresi. Untuk mengestimasi $f(x_i)$ ada dua pendekatan yang dapat digunakan yaitu pendekatan regresi parametrik dan regresi nonparametrik (Hardle, 1990).

ISBN: 978-979-097-142-4

Pendekatan regresi parametrik digunakan bila bentuk fungsi $f(x_i)$ diketahui dari informasi sebelumnya berdasarkan teori ataupun pengalaman masa lalu. Jadi dalam hal ini, mengestimasi $f(x_i)$ eqivalen dengan mengestimasi parameter. Sedangkan pendekatan regresi nonparametrik tidak memberikan asumsi terhadap bentuk kurva regresi sehingga memiliki fleksibelitas yang tinggi untuk mengestimasi kurva regresi $f(x_i)$ .Fungsi regresi $f(x_i)$ hanya diasumsikan termuat dalam suatu ruang fungsi tertertentu, dimana pemilihan ruang fungsi tersebut biasanya dimotivasi oleh sifat kemulusan *(smoothness)* yang dimiliki oleh fungsi $f(x_i)$ tersebut.

Beberapa penulis seperti Hardle (1990), Wahba (1990), Budiantara dan Subanar (1997) menyarankan penggunaan regresi nonparametrik sebagai pendekatan untuk model data, agar mempunyai fleksibelitas yang baik. Regresi spline memungkinkan untuk berbagai macam orde sehingga dapat dibentuk regresi spline linier, kuadrat, kubik maupun orde m. Spline mempunyai keunggulan dalam mengatasi pola data yang menunjukkan naik/turun yang tajam dengan bantuan titik-titik knot, serta kurva yang dihasilkan relatif mulus (Hardle, 1990). Bentuk estimator spline sangat dipengaruhi oleh nilai parameter penghalus λ (Budihantara, 2000). Oleh karena itu, pemilihan λ optimal mutlak diperlukan untuk memperoleh estimator spline yang sesuai dengan data. Bentuk estimator spline juga dipengaruhi oleh lokasi dan banyaknya titik-titik knot. Eubank (1988) menyimpulkan bahwa pemilihan λ optimal dalam regresi spline pada hakekatnya merupakan pemilihan lokasi titik knot.

Untuk nilai λ yang sangat besar akan menghasilkan bentuk kurva regresi yang sangat halus. Sebaliknya untuk nilai λ yang kecil akan memberikan bentuk kurva regresi yang sangat kasar (Wahba, 1990; Eubank, 1988; Budiantara, 1998). Akibatnya pemilihan parameter penghalus optimal merupakan hal yang sangat penting dalam regresi nonparametrik. Dalam paper ini akan dibahas penyelesaian optimal dan pemilihan parameter penghalus λ dengan menggunakan metode MSE dan GCV pada data data polusi kadar debu (jam) dengan skonsentrasi pektra cerobong asap.

2. Fungsi Spline Linier

Fungsi spline linier merupakan fungsi spline dengan satu orde. Fungsi spline linier dengan satu titik knot dapat disajikan dalam bentuk

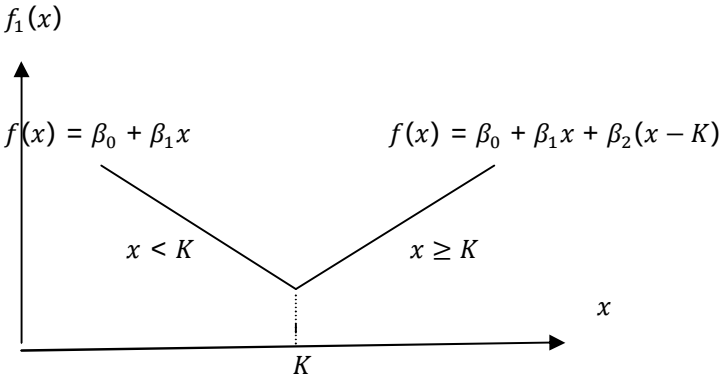
ISBN: 978-979-097-142-4

$$f_1(x) = \beta_0 + \beta_1 x + \beta_2 (x - K)_+^1 \quad (2)$$

Fungsi ini dapat pula disajikan menjadi (Tripena,2005):

$$f_1(x) = \begin{cases} \beta_0 + \beta_1 x & , x < K \\ \beta_0 + \beta_1 x + \beta_2 (x - K), & x \geq K \end{cases} \quad (3)$$

Grafik spline linier dengan satu titik knot pada $x = K$ dapat disajikan



Gambar 1. Fungsi Spline Linier dengan Satu Titik Knot pada $x = K$

3. Regresi Spline

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

$$y = \beta_0 + \sum_{j=1}^{m} \beta_j x^j + \sum_{k=1}^{N} \beta_{j+k} (x - K_k)_+^m + \varepsilon \quad (4)$$

dengan menggunakan data amatan sebanyak n, maka bentuk matriks dari persamaan (4) dapat ditulis sebagai berikut:

$$\mathbf{y} = \mathbf{X_1 \delta_1} + (\mathbf{X} - \mathbf{K})\mathbf{\delta_2} + \boldsymbol{\varepsilon} \quad (5)$$

dengan,

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

ISBN: 978-979-097-142-4

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

Untuk alasan kesederhanaan, maka matriks (5) dapat ditulis kembali menjadi:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \tag{6}$$

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

Dalam hubungannya dengan estimasi kurva mulus $f(x)$, yang mempunyai λ 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_λ, dilakukan penyesuaian persamaan menjadi:

$$\mathbf{b_\lambda} = \widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X'_\lambda X_\lambda})^{-1}\mathbf{X'_\lambda}\,\mathbf{y} \tag{7}$$

dimana $\mathbf{X_\lambda}$: matriks disain dari model yang membentuk model estimasi f_λ dengan λ yang optimal. Fungsi estimasinya adalah:

$$\boldsymbol{f_\lambda} = \mathbf{X_\lambda b_\lambda} = \mathbf{X_\lambda}(\mathbf{X'_\lambda X_\lambda})^{-1}\mathbf{X'_\lambda}\,\mathbf{y} = \mathbf{H_\lambda}\,\mathbf{y}\ \lambda \in \Lambda \tag{8}$$

dengan $\mathbf{H_\lambda} = \mathbf{X_\lambda}(\mathbf{X'_\lambda X_\lambda})^{-1}\mathbf{X'_\lambda}$, $\mathbf{H_\lambda}$ bersifat simetris, definit positif, dan idempoten.

4. Pemilihan Model Regresi Spline dengan λ yang Optimal

Pendekatan regresi nonparametrik, yakni ingin didapatkan kurva mulus yang mempunyai λ optimal menggunakan data amatan sebanyak n, maka diperlukan secara ukuran kinerja universal Eubank (1988)

i. *Mean Squared Error* (MSE)

Ukuran kinerja atas estimator yang sederhana adalah kuadrat dari sisaan yang dirata-rata. $MSE(\lambda) = n^{-1}\sum_{i=1}^{n}\left(y_i - f_\lambda(x_i)\right)^2$ (9)

ii. *Generalized Cross-Validation* (*GCV*)

Menurut Budihantara (2005), GCV merupakan modifikasi dari *Cross-Validation* (CV) adalah metode untuk memilih λ yang meminimumkan

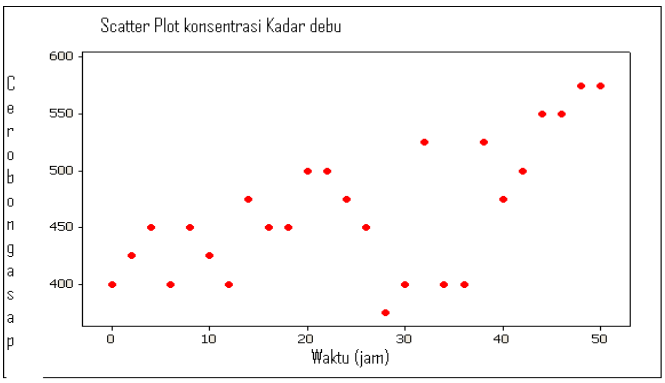
ISBN: 978-979-097-142-4

Fungsi GCV didefinisikan sebagai: $GCV(\lambda) = \frac{MSE(\lambda)}{\{n^{-1}Tr(I-\mathbf{H}_{\boldsymbol{\lambda}})\}^2}$ (10)

dengan $Tr(\mathbf{H}_{\boldsymbol{\lambda}}) < n$

4.1. Pembentukan Model Regresi Spline

Plot data polusi kadar debu (jam) dengan skonsentrasi pektra cerobong asap pada Gambar 2.



Gambar 2. Plot Data Polusi Kadar Debu (Jam) dengan Skonsentrasi Pektra Cerobong Asap

Gambar 2 Plot menunjukkan bahwa ada indikasi perubahan pola perilaku dari variabel bebas pada sub-sub interval tertentu. Terdapat 24 titik knot yang dapat digunakan untuk membentuk model spline. Banyaknya kombinasi titik knot yang bisa digunakan untuk membentuk model spline dengan tiga titik knot sebanyak 10.650 kombinasi.

4.2.Estimasi Regresi Spline Linier

Model umum dari regresi spline linier adalah

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

ISBN: 978-979-097-142-4

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

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

$$y_i = \beta_0 + \beta_1 x_i + \beta_2(x_i - K_1)_+ + \beta_3(x_i - K_2)_+ + \beta_4(x_i - K_3)_+ + \varepsilon_i \qquad (11)$$

Pemilihan titik knot yang optimal terletak pada nilai MSE dan GCV yang minimum. Nilai MSE dan GCV yang minimum dari model regresi spline linier dengan tiga titik knot disajikan pada Tabel 1 berikut ini:

Tabel 1. Nilai MSE dan GCV Model Regresi Spline Linier dengan Tiga Titik Knot

No	Titik knot	Nilai MSE	Nilai GCV
1	4,18,24	205,2419	411,3228

Pada Tabel 1. diperoleh nilai MSE minimum sebesar 205,2419 dan GCV 411,3228 yang berada pada titik knot K_1 = 4, K_2 = 18, dan K_3 = 24. Estimasi model regresi spline linier dengan tiga titik knot dapat disajikan pada Tabel 2.

Tabel 2. Estimasi Model Regresi Spline Linier dengan Tiga Titik Knot

Parameter	Estimasi
β_0	412,58156
β_1	11,12929
β_2	-20,68433
β_3	31,34783
β_4	-25,25567

ISBN: 978-979-097-142-4

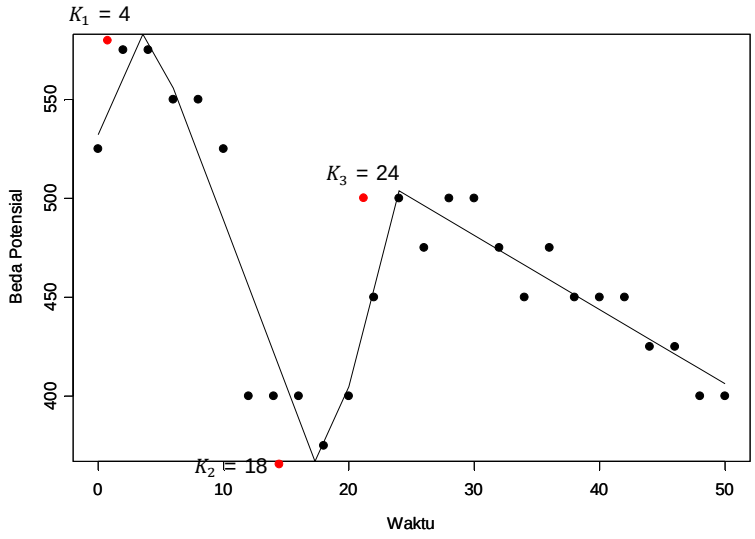
Estimasi model regresi spline linier tiga titik knot K_1 = 4, K_2 = 18, dan K_3 = 24

$$\hat{y}_i = 412{,}58156 + 11{,}12929\, x_i - 20{,}68433(x_i - 4)_+ + 31{,}34783(x_i - 18)_+ - 25{,}25567(x_i - 24)_+$$

Estimasi model regresi spline linier dengan tiga titik knot dapat disajikan pula dalam bentuk fungsi terpotong (*truncated*) sebagai berikut:

$$\hat{y}_i = \begin{cases} 412{,}58156 + 11{,}12929\, x_i, & x_i < 4 \\ 495{,}31888 - 9{,}55504\, x_i, & 4 \le x_i < 18 \\ -68{,}94206 + 21{,}79279\, x_i, & 18 \le x_i < 24 \\ 615{,}71938 - 3{,}46288\, x_i, & x_i \ge 24 \end{cases}$$

Model spline ini disajikan dalam Gambar 3:



Gambar 3. Kurva Estimasi Regresi Spline Linier dengan Tiga Titik Knot

Dari Gambar 3 terlihat bahwa kurva mempunyai *slope* baru pada titik-titik amatan awal. Kurva regresi spline dengan tiga titik knot sudah cukup mampu membentuk pola yang sesuai dengan data tingkat kemulusan kurva. Pada Gambar 3 terlihat jika terjadi perubahan pola pada K = 4, K = 18, dan K = 24. Pola data dari K = 0

ISBN: 978-979-097-142-4

sampai nilai K = 4 mempunyai kecenderungan naik secara tajam, sedangkan untuk data antara K = 4 sampai nilai K = 18 mempunyai kecenderungan turun secara tajam. Untuk data antara K = 18 sampai nilai K = 24 mempunyai kecenderungan naik secara tajam menuju nilai K = 24, sedangkan nilai diatas K = 24 mempunyai kecenderungan turun sampai waktu tertentu.

4.3. Pemilihan Model Regresi Spline Terbaik

Dengan memperhatikan hasil yang telah diperoleh, dapat disimpulkan bahwa titik knot (K) yang paling optimal dengan nilai MSE dan GCV minimum adalah penggunaan tiga titik knot pada regresi spline linier. Nilai MSE dan GCV beberapa model regresi spline dengan tiga 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 (λ) optimal	Nilai GCV (λ) optimal
			1	2	3	4		
1	Linier	3	4	18	24	-	205,243	411,3228
		3	10	22	26	-	350,2683	524,2734

Berdasarkan Tabel 3 dapat disimpulkan bahwa model terbaik untuk data polusi kadar debu (jam) dengan skonsentrasi pektra cerobong asap adalah model regresi spline linier dengan tiga titik knot K_1 = 4, K_2 = 18, K_3 = 24 yakni

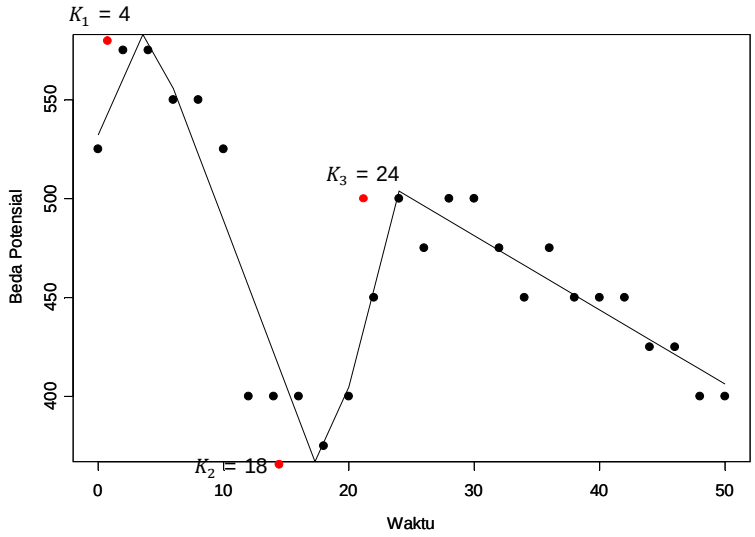
$$\hat{y}_i = 412{,}58156 + 11{,}12929\, x_i - 20{,}68433(x_i - 4)_+ + 31{,}34783(x_i - 18)_+ - 25{,}25567(x_i - 24)_+$$

ISBN: 978-979-097-142-4

Estimasi model regresi spline linier dengan tiga titik knot dapat disajikan pula dalam bentuk fungsi terpotong (*truncated*) sebagai berikut:

$$\hat{y}_i = \begin{cases} 412{,}58156 + 11{,}12929\, x_i, & x_i < 4 \\ 495{,}31888 - 9{,}55504\, x_i, & 4 \le x_i < 18 \\ -68{,}94206 + 21{,}79279\, x_i, & 18 \le x_i < 24 \\ 615{,}71938 - 3{,}46288\, x_i, & x_i \ge 24 \end{cases}$$

Sedangkan plot estimasi model regresi spline linier dengan tiga titik knot yang merupakan model regresi spline terbaik berdasarkan kriteria nilai MSE dan GCV minimum diberikan pada Gambar 4.



Gambar 4. Kurva Estimasi Regresi Spline Linier dengan Tiga Titik Knot yang Merupakan Kurva Regresi Spline Terbaik

Disamping itu, diperoleh nilai koefisien determinasi (R^2) sebesar 0,9457386. Hal ini berarti bahwa variabel polusi kadar debu (jam) mampu menerangkan sebesar 94,57386% terhadap konsentrasi pektra cerobong asap.

5. Kesimpulan

ISBN: 978-979-097-142-4

a) Estimasi regresi spline linier yang menggambarkan hubungan pengaruh polusi kadar debu waktu tertentu (jam) terhadap konsentrasi pektra

$$\hat{y}_i = 412{,}58156 + 11{,}12929\, x_i - 20{,}68433(x_i - 4)_+ 31{,}34783(x_i - 18)_+ - 25{,}25567(x_i - 24)_+$$

Model ini menghasilkan nilai koefisien determinasi $\left(R^2\right)$ sebesar 0,9045368. Hal ini berarti bahwa variabel polusi kadar debu waktu tertentu (jam) mampu menerangkan sebesar 90,45368% terhadap terhadap konsentrasi pektra .Titik knot yang optimal adalah penggunaan tiga titik knot dengan nilainya masing-masing adalah K_1 = 4, K_2 = 18, K_3 = 24.

b) Pemilihan model regresi spline terbaik dengan menggunakan metode $MSE(\lambda)$ sebesar 205,243 dan $GCV(\lambda)$411. Dilihat dari nilai kedua metode tersebut, nilai $MSE(\lambda)$ paling minimum dan metode yang terbaik .

Daftar Pustaka

Budiantara, I. N, 2002. *Aplikasi Spline Estimator Terbobot* . Jurnal Teknik Industri PETRA, Surabaya.

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.

ISBN: 978-979-097-142-4

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