APLIKASI REGRESI NONPARAMETRIK SPLINE BIRESPON PADA DATA KUALITAS AIR DI DAS MAHAKAM

Oleh
Febriana Rinda Sihotang[1)], Sifriyani[2)], Surya Prangga[3)]
[1]Laboratorium Statistika Terapan, Program Studi Statistika, Jurusan Matematika, FMIPA
Universitas Mulawarman, Samarinda, Indonesia
[2,3]Program Studi Statistika, Jurusan Matematika, FMIPA, Universitas Mulawarman, Samarinda, Indonesia
e-mail: [1]rindashtg@gmail.com, [2]sifri.stat.unmul@gmail.com, [3]suryaprangge@fmipa.umul.ac.id

Abstrak

Regresi nonparametrik birespon dengan estimasi spline merupakan metode yang digunakan untuk menyelesaikan permasalahan jika bentuk kurva regresi tidak diketahui dan terdapat dua variabel respon. Spline merupakan potongan-potongan polinomial yang mempunyai sifat tersegmen kontinu. Pada penelitian ini dilakukan indentifikasi faktor yang mempengaruhi kualitas air Daerah Aliran Sungai (DAS) Mahakam, perameter kualitas air sungai yang digunakan dalam penelitian ini adalah Biochemical Oxygen Demand (BOD) dan Chemical Oxygen Demand (COD). BOD menggambarkan banyaknya oksigen yang dibutuhkan organisme untuk mengoksidasi bahan organik karbon yang terkandung sedangkan COD menggambarkan bahan organik mudah teurai maupun sukar terurai. Pengaruh BOD dan COD dalam air sungai dapat menurunkan jumlah oksigen dalam perairan karena terlalu banyak kandungan organik sehingga menyebabkan ekosistem perairan terganggu. Penelitian ini menggunakan 4 faktor yang diduga mempengaruhi BOD dan COD yaitu Total Suspended Solid (TSS), pH, suhu air, dan Dissolved Oxygen (DO). Dari hasil pembahasan dan analisis didapatkan bentuk estimasi model spline dalam regresi nonparametrik birespon terbaik dengan menggunakan kriterian nilai Generalized Cross Validation (GCV) minimum dan nilai maksimum. Model spline terbaik yang dihasilkan dalam penelitian ini adalah model spline dengan 3 titik knot dengan nilai GCV minimum sebesar 0,0612 dan nilai sebesar 94,9944%.

Kata Kunci: Birespon, BOD, COD, Regresi Nonparametrik, Spline, Titik Knot

PENDAHULUAN

Analisis regresi merupakan metode statistika yang digunakan untuk mengetahui pola hubungan antara variabel respon dengan variabel prediktor (Drapper, 1992). Secara umum terdapat tiga pendekatan dalam menganalisis regresi yaitu pendekatan parametrik, semiparametrik, dan nonparametrik. Jika pola data menunjukkan kecenderungan data mengkuti pola linear, kuadratik, atau kubik maka menggunakan pendekatan regresi parametrik, pendekatan regresi semiparametrik digunakan apabila sebagian polanya diketahui dan sebagian tidak diketahui, namun apabila *scatterplot* tidak menunjukkan kencenderungan mengikuti pola tertentu maka pendekatan yang digunakan adalah regresi nonparametrik (Budiantara 2005 dan A. Islamiyati 2017).

Dalam penelitian ini penulis menggunakan pendekatan regresi nonparametrik karena tidak semua data dilapangan diketahui bentuk pola datanya sehingga pendekatan ini mampu menyelesaikan permasalahan untuk pola data yang tidak diketahui, dan model yang dihasilkan dari

ISSN No. 1978-3787
Open Journal Systems

model regresi nonparametrik sangat fleksibel dalam mendekati pola data. Pendekatan regresi nonparametrik nantinya membentuk estimasi sendiri tanpa harus dipengaruhi oleh faktor subyektifitas peneliti, sehingga pendekatan regresi nonparametrik memiliki fleksibilitas yang tinggi. Ada beberapa estimasi dalam regresi nonparametrik, dalam penelitian ini penulis menggunakan estimasi *spline*. *Spline* merupakan potongan-potongan polinomial yang mempunyai sifat tersegmen kontinu. Sifat inilah yang membuat model menjadi lebih fleksibel dibandingkan dengan polinomial biasa, karena dalam *spline* terdapat titik knot yang merupakan titik perpaduan bersama yang menunjukkan terjadinya perubahan perilaku data. Pemilihan model *spline* terbaik dapat dilihat dari nilai knot yang paling optimum berdasarkan nilai GCV yang terkecil.

Model regresi nonparametrik *spline* birespon dalam penelitian ini diaplikasikan pada data BOD dan COD sebagai parameter kualitas air sungai Mahakam di 25 titik DAS Mahakam Provinsi Kalimantan Timur. BOD menggambarkan banyaknya oksigen yang dibutuhkan organisme untuk mengoksidasi bahan organik karbon yang terkandung, sedangkan COD menggambarkan bahan organik mudah teurai maupun sukar terurai. Secara umum parameter BOD dan COD digunakan untuk mengukur pencemaran limbah domestik (rumah tangga) karena dalam kandungan BOD dan COD ini mengandung parameter fisika dan biologi yang menentukan banyaknya jumlah bahan organik di dalam air yang dapat mengakibatkan menurunnya jumlah oksigen dalam air sehingga dapat merusak ekosistem air sungai. Pemilihan model *spline* terbaik dapat dilihat dari nilai knot yang paling optimum berdasarkan nilai GCV yang terkecil.

Berdasarkan hal tersebut penulis ingin melakukan penelitian tentang regresi nonparametrik birespon dengan estimasi *spline* untuk memodelkan BOD dan COD di DAS Mahakam dengan menggunakan empat variabel prediktor yaitu TSS, pH, Suhu Air, dan DO.

LANDASAN TEORI

Regresi Nonparametrik *Spline* Multivariabel

Dalam analisis regresi nonparametrik *spline* jika terdapat satu variabel respon dan satu variabel prediktor, maka dinamakan regresi nonparametrik *spline* univariabel. Jika dalam analisis regresi terdapat satu variabel respon dengan variabel prediktor lebih dari satu, maka regresi tersebut dinamakan regresi nonparametrik *spline* multivariabel. Secara umum, model regesi nonparametrik *spline* multivariabel dituliskan pada persamaan (1) sebagai berikut:

$$y_i = f(x_{1i,}x_{2i,}x_{3i,}...,x_{li}) + \varepsilon_i, i = 1,2,...,n \quad (1)$$

Dimana y_i sebagai variabel respon, $f(x_{1i,}x_{2i,}x_{3i,}...,x_{li})$ merupakan kurva regresi tidak diketahui bentuknya dengan $j = 1,2,...,l$, l adalah banyaknya variabel prediktor dan $i = 1,2,...,n$ yang menunjukkan banyaknya data pengamatan.

Jika dijabarkan, maka akan diperoleh model regresi nonparametrik *spline* multivariabel pada persamaan (2) sebagai berikut:

$$\begin{aligned} y_i &= \sum_{j=1}^{l}\left(\sum_{h=0}^{m}\beta_h x_{ij}^h + \sum_{k=1}^{r}\beta_{m+k}(x_{ij}-k_k)_+^m\right)+\varepsilon_i \\ &= \sum_{j=1}^{l}\sum_{h=0}^{m}\beta_{hj}x_{ij}^h + \sum_{j=1}^{l}\sum_{k=1}^{r}\beta_{m+k;j}(x_{ij}-k_{kj})_+^m + \varepsilon_i \end{aligned} \quad (2)$$

dimana:

y_i : variabel respon data pengamatan ke-i

β_{hj} : koefisien parameter regresi fungsi polinomial

x_{ij}^h : variabel prediktor data pengamatan

$\beta_{m+k;j}$: koefisien parameter regresi fungsi *truncated*

k_{kj} : knot ke-k, $k = 1,2,...,r$

ε_i : *error* ke-i

Regresi Nonparameterik Birespon *Spline* Multivariabel

Regresi birespon didefinisikan sebagai salah satu model regresi yang memiliki variabel respon lebih dari satu buah dan diantara variabel-variabel tersebut terdapat korelasi atau hubungan yang kuat, baik secara logika maupun matematis (Simila dan Tikka, 2007).

Jika regresi birespon memiliki kurva regresi yang tidak diketahui, maka pendekatan yang digunakan adalah regresi nonparametrik birespon. Fungsi yang digunakan dalam pendekatan nonparametrik adalah fungsi *spline* dengan melibatkan banyak variabel prediktor, maka model tersebut dinamakan model regresi nonparametrik *spline* birespon multivariabel. Model untuk regresi nonparametrik *spline* birespon dapat dituliskan sebagai berikut:

$$y_{1i} = \sum_{j=i}^{l} f(x_{ij}) + \varepsilon_{1i}$$
$$y_{2i} = \sum_{j=i}^{l} g(x_{ij}) + \varepsilon_{2i} \quad (3)$$

Dimana fungsi f dan g adalah kurva regresi yang tidak diketahui bentuknya dan dihampiri dengan fungsi *spline* birespon multivariabel dengan rumus sebagai berikut:

$$f(x_{ij}) = \sum_{j=1}^{l}\sum_{h=1}^{m}\theta_{hj}x_{ij}^{h} + \sum_{j=1}^{l}\sum_{k=1}^{r}\phi_{m+k;j}(x_{ij}-k_{kj})_{+}^{m} \quad \text{dan}$$
$$g(x_{ij}) = \sum_{j=1}^{l}\sum_{h=1}^{m}\psi_{hj}x_{ij}^{h} + \sum_{j=1}^{l}\sum_{k=1}^{r}\tau_{m+k;j}(x_{ij}-\lambda_{kj})_{+}^{m} \quad (4)$$

dengan:

$f(x_{ij})$: fungsi regresi respon pertama
$g(x_{ij})$: fungsi regresi respon kedua
θ_{hj} : koefisien parameter regresi polinomial respon pertama
$\phi_{(m+k)j}$: koefisien parameter regresi *spline truncated* respon pertama
ψ_{hj} : koefisien parameter regresi polinomial respon kedua
$\tau_{(m+k)j}$: koefsien parameter regresi *spline truncated* respon kedua
k_{kj} : nilai titik knot optimum respon pertama
λ_{kj} : nilai titik knot optimum respon kedua

Estimator Model Regresi Nonparametrik Spline Birespon

Untuk mendapatkan estimasi model regresi nonparametrik *spline* multivariabel dapat menggunakan metode WLS (*Weighted Least square*) dapat dituliskan dalam bentuk matriks sebagai berikut:

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

$$\mathbf{y} = \begin{pmatrix} y_1 \\ --- \\ y_2 \end{pmatrix} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n} \\ --- \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n} \end{pmatrix} \quad \text{dan} \quad (6)$$

$$\mathbf{X} = \begin{pmatrix} C & \vdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \vdots & D \end{pmatrix}$$

ISSN No. 1978-3787
Open Journal Systems

Dengan

$$\mathbf{C}=\begin{pmatrix} x_{11}^1 & \cdots & x_{11}^m & (x_{11}-k_1^1)_+^m & \cdots & (x_{11}-k_r^1)_+^m & \cdots & x_{l1}^1 & \cdots & x_{l1}^m & (x_{l1}-k_1^1)_+^m & \cdots & (x_{l1}-k_r^1)_+^m \\ x_{12}^1 & \cdots & x_{12}^m & (x_{12}-k_1^1)_+^m & \cdots & (x_{12}-k_r^1)_+^m & \cdots & x_{l2}^1 & \cdots & x_{l2}^m & (x_{l2}-k_1^1)_+^m & \cdots & (x_{l2}-k_r^1)_+^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1n}^1 & \cdots & x_{1n}^m & (x_{1n}-k_1^1)_+^m & \cdots & (x_{1n}-k_r^1)_+^m & \cdots & x_{ln}^1 & \cdots & x_{ln} & (x_{lj}-k_1^1)_+^m & \cdots & (x_{ln}-k_r^1)_+^m \end{pmatrix}$$

$$\mathbf{D}=\begin{pmatrix} x_{11}^1 & \cdots & x_{11}^m & (x_{11}-\lambda_1^1)_+^m & \cdots & (x_{11}-\lambda_r^1)_+^m & \cdots & x_{l1}^1 & \cdots & x_{l1}^m & (x_{l1}-\lambda_1^1)_+^m & \cdots & (x_{l1}-\lambda_r^1)_+^m \\ x_{12}^1 & \cdots & x_{12}^m & (x_{12}-\lambda_1^1)_+^m & \cdots & (x_{12}-\lambda_r^1)_+^m & \cdots & x_{l2} & \cdots & x_{l2}^h & (x_{l2}-\lambda_1^1)_+^m & \cdots & (x_{l1}-\lambda_r^1)_+^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1n}^1 & \cdots & x_{ln}^m & (x_{1n}-\lambda_1^1)_+^m & \cdots & (x_{1n}-\lambda_r^1)_+^m & \cdots & x_{ln}^1 & \cdots & x_{ln}^m & (x_{ln}-\lambda_1^1)_+^m & \cdots & (x_{ln}-\lambda_r^1)_+^m \end{pmatrix} \quad (7))$$

Sedangkan matriks **0** adalah matriks dengan elemen-elemen nya null yang berukuran $n\times(l(m+r))$. Untuk nilai $\boldsymbol{\beta}$ dan $\boldsymbol{\varepsilon}$ merupakan vektor parameter dan vektor *random error* yang memiliki elemen sebagai berikut :

$$\boldsymbol{\beta}=\begin{pmatrix}\beta_1\\---\\\beta_2\end{pmatrix}=\begin{pmatrix}\theta_{11}\\\theta_{12}\\\vdots\\\theta_{ml}\\\phi_{(1+1)1}\\\phi_{(1+2)1}\\\vdots\\\phi_{(m+r)l}\\---\\\psi_{11}\\\psi_{12}\\\vdots\\\psi_{ml}\\\tau_{(1+1)1}\\\tau_{(1+2)1}\\\vdots\\\tau_{(m+r)l}\end{pmatrix}, \text{dan } \boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_1\\---\\\varepsilon_2\end{pmatrix}=\begin{pmatrix}\varepsilon_{11}\\\varepsilon_{12}\\\vdots\\\varepsilon_{1n}\\---\\\varepsilon_{21}\\\varepsilon_{22}\\\vdots\\\varepsilon_{2n}\end{pmatrix} \quad (8)$$

Untuk menghitung estimasi parameter pada regresi nonparametrik *spline* birespon dapat dilakukan dengan melakukan optimasi *Weighted Least Square* (WLS). Penentuan matriks pembobot **W** dalam kasus ini yaitu dengan perhitungan nilai varian kovarian dari respon pertama dan respon kedua kemudian mendefinisikan matriks varian kovarian sebagai pembobot **W** (Oktaviana, 2011) Matriks pembobot **W** dapat juga dituliskan sebagai berikut:

$$\mathbf{W}=\left|\begin{array}{cccc|cccc}\sigma_1^2 & 0 & \cdots & 0 & \sigma_{11} & 0 & \cdots & 0\\ 0 & \sigma_1^2 & & \vdots & 0 & \sigma_{22} & & \vdots\\ \vdots & & \ddots & 0 & \vdots & & \ddots & 0\\ 0 & \cdots & 0 & \sigma_1^2 & 0 & \cdots & 0 & \sigma_{nn}\\ \hline \sigma_{11} & 0 & \cdots & 0 & \sigma_2^2 & 0 & \cdots & 0\\ 0 & \sigma_{22} & & \vdots & 0 & \sigma_2^2 & & \vdots\\ \vdots & & \ddots & 0 & \vdots & & \ddots & 0\\ 0 & \cdots & 0 & \sigma_{nn} & 0 & \cdots & 0 & \sigma_2^2\end{array}\right| \quad (9)$$

Maka untuk memperoleh estimator pada persamaan (5) dilakukan penyelesaian optimasi parameter dengan WLS sebagai berikut:

$$\min_{\beta}\left\{\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{W}\boldsymbol{\varepsilon}\right\}=\min_{\beta}\left\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{W}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \quad (10)$$

Untuk menyelesaikan persamaan (10), dilakukan penurunan secara parsial dengan memisalkan fungsi $\boldsymbol{\varphi}(\boldsymbol{\beta})=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{W}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, maka diperoleh:

$$\begin{aligned}\boldsymbol{\varphi}(\boldsymbol{\beta}) &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{W}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &= (\mathbf{y}^{\mathrm{T}}-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^{\mathrm{T}})(\mathbf{W}\mathbf{y}-\mathbf{W}\mathbf{X}\boldsymbol{\beta})\\ &= \mathbf{y}^{\mathrm{T}}\mathbf{W}\mathbf{y}-\mathbf{y}^{\mathrm{T}}\mathbf{W}\mathbf{X}\boldsymbol{\beta}-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}\mathbf{y}+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}\mathbf{X}\boldsymbol{\beta} \quad (11)\\ &= \mathbf{y}^{\mathrm{T}}\mathbf{W}\mathbf{y}-2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Selanjutnya persamaan yang diperoleh diturunkan terhadap $\boldsymbol{\beta}$ diperoleh hasil sebagai berikut :

$$\begin{aligned}\frac{\partial\boldsymbol{\varphi}(\boldsymbol{\beta})}{\partial(\boldsymbol{\beta})} &= \frac{\partial(\mathbf{y}^{\mathrm{T}}\mathbf{W}\mathbf{y}-2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \quad (12)\\ &= -2\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}+2\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Setelah dilakukan penurunan terhadap $\boldsymbol{\beta}$, hasil penurunan disamakan dengan nol dan

didapatkan hasil estimasi parameter sebagai berikut :

$$-2\mathbf{X}^T\mathbf{W}\mathbf{y} + 2\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$
$$\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{W}\mathbf{y} \quad (13)$$
$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

Sehingga bentuk estimasi model *spline* dalam regresi nonparametrik birespon menjadi sebagai berikut:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} \quad (14)$$
$$\mathbf{y} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

Korelasi Antar Variabel-Variabel

Berdasarkan definisi regresi birespon yaitu regresi dengan variabel respon sebanyak dua dan diantara variabel-variabel respon harus memiliki korelasi antara satu dengan yang lainnya. Sebelum melakukan pemodelan, terlebih dahulu harus diketahui seberapa besar hubungan atau korelasi antar variabel-variabel tersebut. Untuk mengetahui nilai korelasinya dapat digunakan koefisien korelasi Pearson yang secara umum memiliki persamaan sebagai berikut:

$$r(y_1, y_2) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_{1i} - \bar{y}_1)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_{2i} - \bar{y}_2)^2\right)} \quad (15)$$

. Nilai koefisien korelasi yang dihasilkan berkisar antara -1 sampai dengan 1, jika nilai koefisien korelasi mendekati -1 atau 1 maka hubungan atau korelasi antara variable-variabel respon semakin kuat, sedangkan jika nilai koefisiesn korelasi mendekati 0 maka hubungan atau korelasi antara variabel-variabel respon semakin lemah (Draper dan Smith, 1992).

Untuk menguji hipotesis ada hubungan yang signifikan antara variabel respon pertama dengan variabel respon kedua, maka digunakan pengujian signifikansi korelasi $t_{(hitung)}$ dengan rumus:

$$t_{(hitung)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (16)$$

Dalam menyimpulkan apakah variabel respon pertama dan kedua memiliki hubungan yang signifikan, maka harus membandingkan nilai $t_{(hitung)}$ dengan nilai $t_{(tabel)}$. Jika $t_{(hitung)} \geq t_{(tabel)}$ maka kriteria pengujiannya adalah H_0 ditolak (signifikan), dan jika $t_{(hitung)} \leq t_{(tabel)}$ maka H_1 ditolak (tidak signifikan).

Pemilihan Titik Knot Optimal

Hal penting dalam estimator *spline* adalah mencari estimator yang paling sesuai untuk sekumpulan data. Estimator Spline secara umum sangat tergantung dengan parameter penghalus, Wahba (1990) mengatakan bahwa jika nilai parameter penghalus sangat kecil atau mendekati nilai nol maka akan memberikan estimator *spline* yang sangat kasar. Sebaliknya, jika nilai parameter penghalus sangat besar atau mendekati nilai yang tak berhingga maka akan menghasilkan estimator *s*pline yang sangat mulus. Untuk itu perlu dipilih parameter penghalus yang optimal agar diperoleh estimator *spline* yang paling sesuai dengan data. Wahba (1990) memberikan suatu metode yang sangat baik untuk memilih parameter penghalus yang optimal yaitu dengan metode *Generalized Cross Validation* (GCV). Secara teoritis metode GCV mempunyai sifat optimal asimtotik yang diperlihatkan oleh Wahba (1990) yang tidak dimiliki oleh metode lainnya. Metode GCV secara umum didefinikan sebagai berikut:

$$GCV = \frac{MSE(k)}{[n^{-1}trace(\mathbf{I} - \mathbf{A}(k))]^2} \quad (17)$$

dengan:

$\mathbf{I}$ = matriks identitas
n = jumlah pengamatan
$MSE(k) = \mathbf{n^{-1}(y - X\hat{\beta})(y - X\hat{\beta})}$
$\mathbf{A}(k) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

Kriteria Pemilihan Model Terbaik

Beberapa kriteria yang digunakan untuk menentukan model regresi terbaik adalah dengan menggunakan *Mean Square Error*

ISSN No. 1978-3787
Open Journal Systems

(MSE), koefisien determinasi (R^2), *Akaike's Information Criterion* (AIC), *Schwatz's Bayesian Criterion* (SBC), Untuk penelitian ini membatasi menggunakan kriteria pemilihan model terbaik dengan menggunakan nilai R^2 maksimum. Koefisien determinasi adalah nilai dari proporsi keragaman total disekitar nilai tengah $\bar{y}$ yang dijelaskan dari model regresi (Draper and Smith, 1992). Nilai R^2 didapat diperoleh dari persamaan:

$$R^2 = \left(1 - \frac{\sum_{i-1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i-1}^{n}(Y_i - \bar{Y})^2}\right) \times 100\% \qquad (18)$$

METODE PENELITIAN

Data yang digunakan dalam penelitian ini menggunakan data skunder dari Dinas Lingkungan Hidup (DLH) Provinsi Kalimantan Timur. Berdasarkan data tersebut, diketahui bahwa terdapat 25 titik DAS Mahakam pada tahun 2018. Variabel respon yang digunakan adalah BOD (y_1), COD (y_2), TSS (x_1), pH (x_2), suhu air (x_3), dan DO (x_4).

Berikut langkah-langkah untuk menganalisis regres nonparametrik spline birespon dalam penelitian ini.

1. Analisis statistika deskriptif dari variabel respon dan variabel prediktor
2. Identifikasi pola data dengan *scatterplot*
3. Mengestimasi model *spline* menggunakan satu titik knot, dua titik knot, dan tiga titik knot.
4. Menghitung estimator parameter.
5. Memilih titik knot optimal berdasarkan GCV minimum dan nilai maksimum dari R^2
6. Interpretasi model dan menarik kesimpulan

HASIL DAN PEMBAHASAN

Analisis Statistika Deskriptif

Pada Tabel 1 berikut ini diberikan statistika deskriptif dari data penelitian yang digunakan. Statistika deskriptif yang akan ditampilkan adalah nilai minimum, nilai maksimum, dan ukuran pemusatan.

Tabel 1. Statistika Deskriptif Variabel Respon dan Prediktor

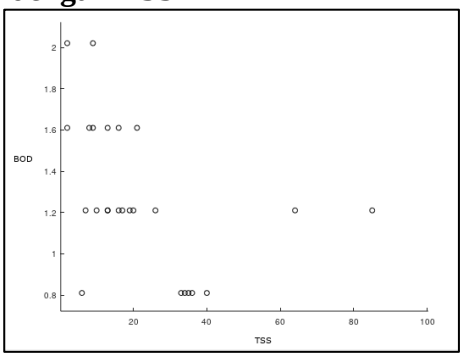
Variabel	Rata-rata	Min	Max
BOD	1,2748	0,8100	2,0200
COD	3,3852	1,8600	6,6800
TSS	22,1600	2	85
pH	6,9744	6,1300	7,9100
Suhu Air	27,496	24	32
DO	5,7884	2.8800	9,4700

Berdasarkan Tabel 1 dapat diketahui bahwa rata-rata nilai BOD di sungai Mahakam sebesar 1,2748mg/l dan nilai kandungan COD di dalam air sungai Mahakam rata-rata bernilai 3,3852mg/l. Sedangkan TSS rata-ratanya sebesar 22,1600mg/l, rata-rata pH dalam air sungai sebesar 6,9744, nilai rata-rata suhu air sungai mahakam sebesar 27,4960 dan rata-rata kadar DO di dalam air sungai mahakam sebesar 5,7884mg/l.

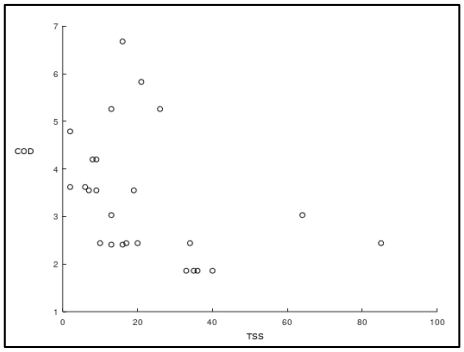
Identifikasi pola data dengan *scatterplot*

Scatterplot dibuat untuk mengetahui pola data antara dua variabel. Dalam analisis ini ingin diketahui pola data antara variabel respon pertama dan kedua dengan variabel prediktor. *Scatterplot* ini dilakukan secara satu persatu antara variabel respon dengan masing-masing variabel prediktor sehingga *scatter plot* yang terbentuk sebanyak 8 buah dan dapat dilihat dalam gambar berikut.

Gambar 1. *Scatterplot* BOD dan COD dengan TSS



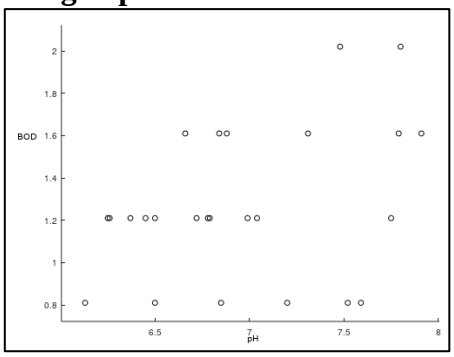
(a)



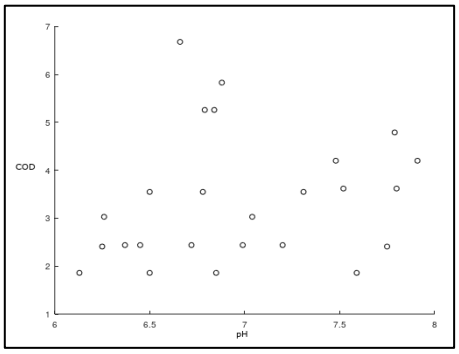
(b)

Scatterplot antara variabel BOD dan COD dengan TSS masing-masing disajikan pada gambar 1(a) dan 1(b). Dari gamber tersebut menunjukkan bahwa pola data yang terbentuk tidak mengikuti suatu pola tertentu.

Gambar 2. *Scatterplot* BOD dan COD dengan pH



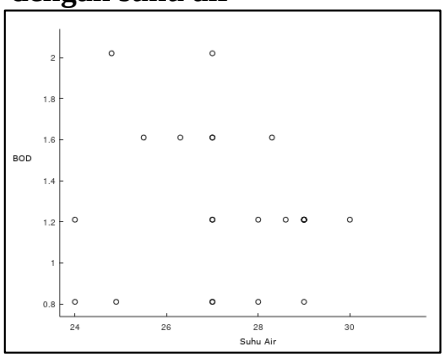
(a)



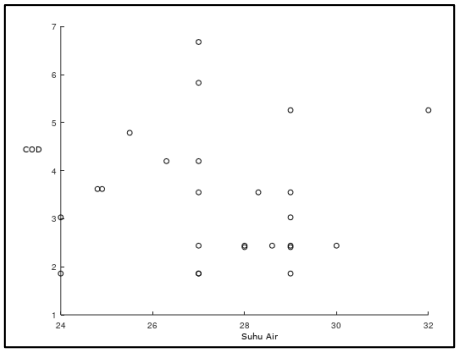
(b)

Berdasarkan gambar 2(a) dan 2(b) terlihat bahwa bentuk pola data yang terbentuk dari BOD dan COD dengan pH tidak menunjukkan suatu pola tertentu.

Gambar 3. *Scatterplot* BOD dan COD dengan suhu air



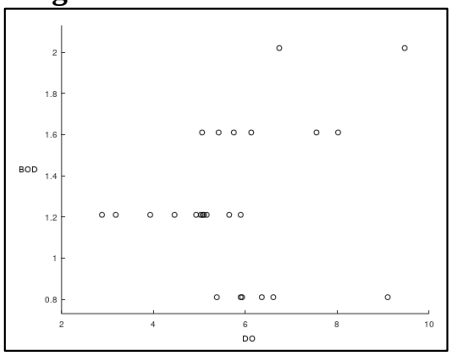
(a)



(b)

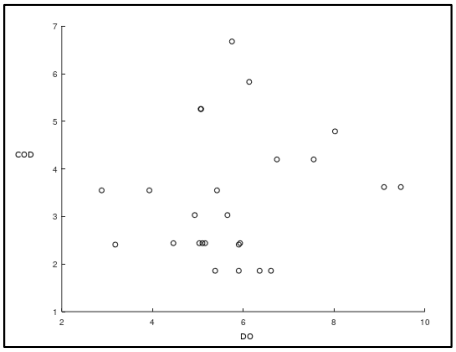
Dari gambar 3(a) dan 3(b) menunjukkan bahwa *Scatter plot* dari BOD dan COD dengan suhu air menunjukkan bahwa pola data yang terbentuk tidak mengikuti pola.

Gambar 4. *Scatterplot* BOD dan COD dengan DO



(a)

ISSN No. 1978-3787
Open Journal Systems



(b)

Dari gambar 4(a) dan 4(b) menunjukkan bahwa pola data dari BOD dengan DO tidak mengikuti pola tertentu.

Karena bentuk pola data tidak diketahui atau tidak mengikuti pola tertentu, maka indikasi untuk menyelesaikan permasalahan yaitu dengan pendekatan regresi nonparametrik.

Uji Korelasi Variabel Respon

Hipotesis

$H_0 : r \neq 0$

(Tidak terdapat hubungan yang signifikan antara BOD dan COD)

$H_1 : r = 0$

(Terdapat hubungan yang signifikan antara BOD dan COD)

Nilai koefisien korelasi Spearman dapat dihitung menggunakan rumus pada persamaan (15)

$$r(y_1, y_2) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_{1i} - \bar{y}_1)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_{2i} - \bar{y}_2)^2\right)} = 0,6510$$

Statistik uji yang digunakan untuk menguji signifikan korelasi adalah

$$t_{(hitung)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{0,6510\sqrt{25-2}}{\sqrt{1-0,4238}} = \frac{3,1220}{0,7590} = 4,1134$$

H_0 ditolak jika $t_{(hitung)} \geq t_{(tabel)(\alpha;n-1)}$

sehingga diputskan H_0 ditolak karena $t_{(hitung)} = 4,1129 \geq t_{(0,05:(25-1))} = 1,7109$ dan dapat disimpulkan Terdapat hubungan yang signifikan antara BOD dengan COD.

Pemilihan Titik Knot Optimal

Pemilihan model *spline* birespon terbaik diperoleh berdasarkan nilai GCV yang minimum dan nilai R^2 yang maksimum. Berikut ditampilkan nilai GCV dan untuk 1, 2, dan 3 titik knot yang disajikan pada tabel 2.

Tabel 2. Nilai GCV Minimum Masing-masing Titik Knot

Titik Knot	GCV Minimum	R^2
1 Titik Knot	0,1256	58,8830
2 Titik Knot	0,0717	86,7887
3 Titik Knot	0,0612	94,9944

Dari tiga nilai GCV diperoleh nilai GCV yang paling minimum dan nilai R^2 paling maksimum adalah pada 3 titik knot.

Model Regresi *Spline* Birespon Optimal

Berdasarkan titik knot optimal yang terpilih yaitu 3 titik knot, sehingga dapat dibentuk model regresi nonparametrik birespon *spline truncated* untuk variabel BOD dan COD sebagai berikut:

Estimasi model untuk respon pertama yakni BOD (*Biochemical Oxygen Demand*)

$$\hat{y}_1 = -0,4278x_1 + 0,4181(x_1 - 4,8621)_+ - 0,0503(x_1 - 24,8966)_+ + 0,2539(x_1 - 70,6897)_+ + 39,8999x_2 - 40,1220(x_2 - 6,1914)_+ + 0,4676x_2 - 6,6210)_+ - 2,5407(x_2 - 7,6031)_+ - 23,3673x_3 + 25,6321(x_3 - 24,2759)_+ - 2,3171(x_3 - 26,2069)_+ + 0,3487(x_3 - 30,6207)_+ - 0,2584x_4 + 0,3079(x_4 - 3,1072)_+ + 0,1926(x_4 - 4,6979)_+ + 1,4674(x_4 - 8,3338)_+$$

Estimasi model untuk respon kedua yakni COD (*Chemical Oxygen Demand*)

$$\hat{y}_2 = 0,2193x_1 + 1,4673(x_1 - 19,1724)_+ - 1,5695(x_1 - 22,0345)_+ + 0,3234(x_1 - 36,3448)_+ + 3,9751x_2 + 49,6206(x_2 - 6,4983)_+ -63,8785(x_2 - 6,5597)_+ - 8,6098(x_2 - 6,8666)_+ - 0,0051x_3 + 5,3864(x_3 - 25,6552)_+ - 7,6511(x_3 - 25,9310)_+ + 2,7184(x_3 - 27,3103)_+ - 1,7571x_4 + 1,4915(x_4 - 4,2435)_+ + 3,1491(x_4 - 4,4707)_+ - 3,6990(x_4 - 5,6069)_+$$

Interpretasi Model

Interpretasi untuk model regresi nonparametrik *spline* terbaik untuk variabel prediktor pertama TSS dari kedua variabel respon sebagai berikut

1. Jika variabel dari $x_2, x_3,$ dan x_4 konstan maka pemodelan dari faktor TSS (x_1) terhadap BOD adalah sebagai berikut:

$$\hat{y}_1 = 39,8999x_2 - 40,1220(x_2 - 6,1914)_+ + 0,4676(x_2 - 6,6210)_+ - 2,5407(x_2 - 7,6031)_+$$

$$= \begin{cases} 39,8999x_2 & ; x_2 < 6,1914 \\ -0,2221x_2 + 248,4113 & ; 6,1914 \le x_2 < 6,6210 \\ 0,2455x_2 + 245,3153 & ; 6,6210 \le x_2 < 7,6031 \\ -2,2954x_2 + 264,7283 & ; x_2 \ge 7,6031 \end{cases}$$

Apabila nilai TSS di suatu titik DAS kurang dari 4,8621 mg/l maka jika nilai kadar TSS akan turun sebesar 1 mg/l maka nilai BOD di sungai Mahakam akan turun sebesar 0,4278 mg/l. Apabila nilai TSS di suatu titik DAS diantara nilai 4,8621 sampai 24,8966 mg/l maka jika nilai TSS naik sebesar 1 mg/l maka nilai BOD akan turun sebesar 0,0097 mg/l. jika nilai TSS diantara 24,8966 dan 70,6897 mg/l, apabila nilai TSS naik sebesar 1 mg/l maka kadar BOD di sungai akan turun sebesar 0,0600 mg/l. Jika nilai TSS di sungai Mahakam lebih dari 70,6897 mg/l, apabila nilai TSS naik sebesar 1 mg/l maka nilai BOD akan naik sebesar 0,1939 mg/l.

2. Jika variabel dari $x_2, x_3,$ dan x_4 konstan maka pemodelan dari faktor TSS (x_1) terhadap COD adalah sebagai berikut:

$$\hat{y}_2 = 3,9751x_2 + 49,6206(x_2 - 6,4983)_+ - 63,8785(x_2 - 6,5597)_+ - 8,6098(x_2 - 6,8666)_+$$

$$= \begin{cases} 3,9751x_2 & ; x_2 < 6,4983 \\ 53,5957x_2 - 322,4495 & ; 6,4983 \le x_2 < 6,5597 \\ -10,2828x_2 + 96,5743 & ; 6,5597 \le x_2 < 6,8666 \\ -18,8926x_2 + 155,6943 & ; x_2 \ge 6,8666 \end{cases}$$

Apabila nilai TSS di suatu titik DAS kurang dari 19,1724 mg/l maka jika nilai kadar TSS akan naik sebesar 1 mg/l maka nilai COD di sungai Mahakam akan turun sebesar 0,4278 mg/l. Apabila nilai TSS di suatu titik DAS diantara nilai 19,1724 sampai 22,0345 mg/l maka jika nilai TSS naik sebesar 1 mg/l maka nilai COD akan turun sebesar 1,6866 mg/l. jika nilai TSS diantara 22,0345 dan 36,3448 mg/l, apabila nilai TSS naik sebesar 1 mg/l maka kadar COD di sungai akan turun sebesar 0,1171 mg/l. Jika nilai TSS di sungai Mahakam lebih dari 36,3448 mg/l, apabila nilai TSS naik sebesar 1 mg/l maka nilai COD akan naik sebesar 0,4405 mg/l.

3. Persamaan model *spline* terbaik dengan *truncated* untuk respon pertama (y_1) dimana jika variabel $x_1, x_3,$ dan x_4 konstan maka pemodelan dari faktor pH didalam air sungai (x_2) terhadap BOD adalah sebagai berikut:

$$\hat{y}_1 = 39,8999x_2 - 40,1220(x_2 - 6,1914)_+ + 0,4676(x_2 - 6,6210)_+ - 2,5407(x_2 - 7,6031)_+$$

$$= \begin{cases} 39,8999x_2 & ; x_2 < 6,1914 \\ -0,2221x_2 + 248,4113 & ; 6,1914 \le x_2 < 6,6210 \\ 0,2455x_2 + 245,3153 & ; 6,6210 \le x_2 < 7,6031 \\ -2,2954x_2 + 264,7283 & ; x_2 \ge 7,6031 \end{cases}$$

Apabila nilai pH kurang dari 6,1914 maka nilai kadar BOD di air sungai cenderung akan naik. Jika nilai pH berada diantara 6,1914 sampai 6,6210 maka nilai kadar BOD dalam air sungai cenderung akan turun. Sedangkan jika nilai pH berada diantara 6,6210 sampai 7,6031 maka nilai kadar BOD cenderung akan naik, dan apabila nilai pH lebih dari 7,6031 maka nilai kadar BOD cenderung akan turun.

4. Persamaan model *spline* terbaik dengan *truncated* untuk respon pertama (y_2) dimana jika variabel $x_1, x_3,$ dan x_4 konstan maka pemodelan dari faktor pH didalam air sungai (x_2) terhadap COD adalah sebagai berikut:

ISSN No. 1978-3787
Open Journal Systems

$$\hat{y}_2 = 3,9751x_2 + 49,6206(x_2 - 6,4983)_+ - 63,8785(x_2 - 6,5597)_+ - 8,6098(x_2 - 6,8666)_+$$

$$= \begin{cases} 3,9751x_2 & ; x_2 < 6,4983 \\ 53,5957x_2 - 322,4495 & ; 6,4983 \le x_2 < 6,5597 \\ -10,2828x_2 + 96,5743 & ; 6,5597 \le x_2 < 6,8666 \\ -18,8926x_2 + 155,6943 & ; x_2 \ge 6,8666 \end{cases}$$

Apabila nilai pH kurang dari 6,4983 maka nilai kadar COD di air sungai cenderung akan naik. Jika nilai pH berada diantara 6,4983 sampai 6,5597 maka nilai kadar COD dalam air sungai cendeung akan naik. Sedangkan jika nilai pH berada diantara 6,5597 sampai 6,8666 maka nilai kadar COD cenderung akan turun, dan apabila nilai pH lebih dari 6,8666 maka nilai kadar COD cenderung akan naik.

5. Persamaan model *spline* terbaik dengan *truncated* untuk respon pertama (y_1) dimana jika variabel x_1, x_2, dan x_4 konstan maka pemodelan dari faktor suhu air sungai (x_3) terhadap BOD adalah sebagai berikut:

$$\hat{y}_1 = -23,3673x_3 + 25,6321(x_3 - 24,2759)_+ - 2,3171(x_3 - 26,2069)_+ + 0,3487(x_3 - 30,6207)_+$$

$$= \begin{cases} -23,3673x_3 & ; x_3 < 24,2759 \\ 2,2648x_3 - 622,2423 & ; 24,2759 \le x_3 < 26,2069 \\ -0,0523x_3 - 561,5183 & ; 26,2069 \le x_3 < 30,6270 \\ 0,2964x_3 - 572,1957 & ; x_3 \ge 30,6270 \end{cases}$$

Apabila suhu air kurang dari $24,2759\,^{O}C$ maka nilai kadar BOD di air sungai cenderung akan turun. Jika suhu air berada diantara 24,2759 sampai $26,2069\,^{O}C$ maka nilai kadar BOD dalam air sungai cenderung akan naik. Sedangkan jika suhu air berada diantara 26,2060 sampai $30,6270\,^{O}C$ maka nilai kadar BOD cenderung akan turun, dan apabila suhu air lebih dari $30,6270\,^{O}C$ maka nilai kadar BOD cenderung akan naik.

6. Persamaan model *spline* terbaik dengan *truncated* untuk respon pertama (y_2) dimana jika variabel x_1, x_2, dan x_4 konstan maka pemodelan dari faktor suhu air sungai (x_3) terhadap COD adalah sebagai berikut:

$$\hat{y}_2 = -0,0051x_3 + 5,3864(x_3 - 25,6552)_+ - 7,6511(x_3 - 25,9310)_+ + 2,7184(x_3 - 27,3103)_+$$

$$= \begin{cases} -0,0051x_3 & ; x_3 < 25,6552 \\ 5,3823x_3 - 138,1892 & ; 25,6552 \le x_3 < 25,9310 \\ -2,2698x_3 + 60,2115 & ; 25,9310 \le x_3 < 27,3103 \\ 0,4486x_3 - 14,0289 & ; x_3 \ge 27,31035 \end{cases}$$

Apabila suhu air kurang dari $25,6552\,^{O}C$ maka nilai kadar COD di air sungai cenderung akan turun. Jika suhu air berada diantara 25,6552 sampai $25,9310\,^{O}C$ maka nilai kadar COD dalam air sungai cenderung akan naik. Sedangkan jika suhu air berada diantara 25,9310 sampai $27,3103\,^{O}C$ maka nilai kadar COD cenderung akan turun, dan apabila suhu air lebih dari $27,31035\,^{O}C$ maka nilai kadar COD cenderung akan naik.

7. Persamaan model *spline* terbaik dengan *truncated* untuk respon pertama (y_1) dimana jika variabel x_1, x_2, dan x_3 adalah konstan maka pemodelan dari faktor DO (x_4) terhadap BOD

$$\hat{y}_1 = -0,2584x_4 + 0,3079(x_4 - 3,1072)_+ + 0,1926(x_4 - 4,6979)_+ + 1,4674(x_4 - 8,3338)_+$$

$$= \begin{cases} -0,2584x_4 & ; x_4 < 3,1072 \\ 0,0495x_4 - 0,9567 & ; 3,1072 \le x_4 < 4,6979 \\ 0,2421x_4 - 1,8578 & ; 4,6979 \le x_4 < 8,3338 \\ 1,7095x_4 - 14,0868 & ; x_4 \ge 8,3338 \end{cases}$$

Apabila nilai kadar DO kurang dari 3,1072 mg/l maka nilai kadar BOD di air sungai cenderung akan turun. Jika nilai kadar DO berada diantara 3,1072 sampai 4,6979 mg/l maka nilai kadar BOD dalam air sungai cenderung akan naik. Sedangkan jika nilai kadar DO berada diantara 4,6979 sampai 8,3338 mg/l maka nilai kadar BOD cenderung akan naik, dan apabila nilai kadar DO lebih dari 8,3338 mg/l maka nilai kadar BOD cenderung akan naik.

8. Persamaan model *spline* terbaik dengan *truncated* untuk respon pertama (y_2) dimana jika variabel x_1, x_2, dan x_3 adalah konstan maka pemodelan dari faktor DO (x_4) terhadap COD

$$\hat{y}_2 = -1,7571x_4 + 1,4915(x_4 - 4,24345)_+ + 3,1491(x_4 - 4,4707)_+ - 3,6990(x_4 - 5,6069)_+$$

$$= \begin{cases} -1,7571x_4 & ; x_4 < 4,24345 \\ -0,2657x_4 - 6,3289 & ; 4,2435 \le x_4 < 4,4707 \\ 2,8835x_4 - 20,4076 & ; 4,4707 \le x_4 < 5,6069 \\ -0,8156x_4 + 0,3323 & ; x_4 \ge 5,6069 \end{cases}$$

Apabila nilai kadar DO kurang dari 4,24345 mg/l maka nilai kadar COD di air sungai cenderung akan turun. Jika nilai kadar DO berada diantara 4,2435 sampai 4,4707 mg/l maka nilai kadar COD dalam air sungai cenderung akan turun. Sedangkan jika nilai kadar DO berada diantara 4,4707 sampai 5,6069 mg/l maka nilai kadar COD cenderung akan naik, dan apabila nilai kadar DO lebih dari 5,6069 mg/l maka nilai kadar COD cenderung akan naik.

PENUTUP

Kesimpulan

Berdasarkan hasil dan pembahasan diperoleh kesimpulan yaitu Model *spline* terbaik dengan 3 titik knot dengan nilai GCV minimum sebesar 0,0612 dan nilai maksimum dari koefisien determinasi sebesar 94,9944% yang diartikan bahwa TSS, pH, Suhu Air, dan DO berpengaruh sebesar 94,9944% terhadap BOD dan COD dan sisanya sebesar 5,0056% disebabkan oleh faktor-faktor lain yang tidak diketahui.

Saran

Adapun saran yang diberikan untuk penelitian selanjutnya adalah menggunakan model spline truncated dengan tambahan 4 titik knot dan 5 titik knot.

DAFTAR PUSTAKA

[1] Draper, N. Smith, H. (1992). *Applied Regression Analysis*. Second Edition. New York: John Wiley & Sons.

[2] Budiantara, I.N. (2005). Model Spline Multivariabel Dalam Regresi Nonparametrik, *Makalah Seminar Nasional Matematika*, *Jurusan Matematika FMIPA*. ITS. Surabaya.

[3] Islamiyati, Anna. 2017. *Spline Polynomial Truncated* dalam Regresi Nonparametrik. Jurnal Matematika, Statistika dan Komputasi. Vol. 14, No.1, 54-60, Juli 2017. E-ISSN: 2614-8811. Jurnal Teknologi Pertanian Andalas. Vol. 23, No.1, Maret 2019 ISSN 1410-1920.

[4] Similia,T. Dan Tikka, J. 2007. *Input Selection and Shrinkage in Multiresponse Linear Regression* : *Preprint Submitted to Elsevier*

[5] Oktaviana, Dhina. 2011. Regresi *Spline* Birespon Untuk Memodelkan Kadar Gula Darah Penderita Diabetes Melitus. ITS-paper- 13021120000202.

[6] Wahba, G. 1990. *Spline Models for Observation Data*. SIAM. Philadelphia. CBMSNSF Regional Conference Series in Applied Mathematics. Vol. 59.

ISSN No. 1978-3787
Open Journal Systems

HALAMAN INI SENGAJA DIKOSONGKAN