Indonesian Journal of Applied Statistics ISSN 2621 – 086 X
Volume 1 No. 2 November 2018

Penerapan *Generalized Cross Validation* dalam Model Regresi *Smoothing Spline* pada Produksi Ubi Jalar di Jawa Tengah

Trionika Dian Wahyuningsih[1], Sri Sulistijowati Handajani[2], and Diari Indriati[3]
[1,3]Program Studi Matematika FMIPA Universitas Sebelas Maret
[2]Program Studi Statistika FMIPA Universitas Sebelas Maret

rr_ssh@staff.uns.ac.id , trionikadian2@gmail.com

Abstract: Sweet Potato is a useful plant as a source carbohydrates, proteins, and is used as an animal feed and ingredient industry. Based on data from the Badan Pusat Statistik (BPS), the production fluctuations of the sweet potato in Central Java from year to year are caused by many factor. The production of sweet potato and the factors that affected it if they are described into a pattern of relationships then they do not have a specific pattern and do not follow a particular distribution, such as harvest area, the allocation of subsidized urea fertilizer, and the allocation of subsidized organic fertilizer. Therefore, the production model of sweet potato could be applied into nonparametric regression model. The approach used for nonparametric regression in this study is smoothing spline regression. The method used in regression smoothing spline is generalized cross validation (GCV). The value of the smoothing parameter (λ) is chosen from the minimum GCV value. The results of the study show that the optimum λ value for the factors of harvest area, urea fertilizer and organic fertilizer are 5.57905e-14, 2.51426e-06, and 3.227217e-13 that they result a minimum GCV i.e 2.29272e-21, 1.38391e-16, and 3.46813e-24.

Keywords: *Sweet potato, nonparametric, smoothing spline, generalized cross validation.*

1. Pendahuluan

Indonesia sebagai negara beriklim tropis, memiliki tanah subur dan hasil alam yang beraneka ragam, khususnya di bidang pertanian. Pertanian dalam arti luas terdiri atas lima sektor yaitu tanaman pangan, perkebunan, peternakan, perikanan, dan kehutanan (Soekartawi [1]). Sektor tanaman pangan yang salah satunya adalah tanaman palawija yaitu ubi kayu, ubi jalar, dan talas (Purwono dan Purnamawati [2]). Sebagai salah satu tanaman pangan, ubi jalar merupakan sumber karbohidrat dan protein.

Provinsi Jawa Tengah termasuk ke dalam lima daerah sentra produksi ubi jalar terbesar di Indonesia. Produksi ubi jalar sebagai bahan pangan pengganti merupakan suatu hal yang perlu diperhatikan agar tidak menjadi ancaman untuk ketahanan pangan Jawa Tengah dan Nasional. Produksi ubi jalar Jawa Tengah mengalami fluktuatif dari tahun ke tahun, hal ini disebabkan oleh berbagai faktor, antara lain luas panen, luas lahan, bibit, jumlah pupuk, jumlah tenaga kerja, curah hujan, dan konsumsi masyarakat.

Najwah [3] melakukan analisis efisiensi usaha tani ubi jalar (*Ipomoea batatas* L.) di Kabupaten Karanganyar dan menghasilkan kesimpulan bahwa luas lahan, bibit, pupuk phonska, pestisida, dan tenaga kerja berpengaruh nyata terhadap produksi ubi jalar. Selain itu, Defri [4] melakukan analisis pendapatan dan faktor-faktor yang memengaruhi produksi usahatani ubi jalar (studi kasus Desa Purwasari, Kecamatan Dramaga, Kabupaten Bogor). Hasil penelitian tersebut menyatakan estimasi parameter *ordinary least square* untuk fungsi produksi *Cobb-Douglass* menunjukkan bahwa variabel yang berpengaruh terhadap produksi ubi jalar adalah variabel pupuk kandang, bibit, dan tenaga kerja.

Penelitian yang dilakukan Najwah [3] dan Defri [4] belum memeroleh hasil yang memuaskan. Hal ini disebabkan produksi ubi jalar dan faktor yang memengaruhinya jika digambarkan ke suatu pola maka tidak memiliki pola tertentu. Oleh karena itu, pola tersebut tidak dapat digunakan dengan regresi parametrik sehingga digunakan regresi nonparametrik. Model regresi nonparametrik digunakan apabila tidak ada informasi sebelumnya tentang bentuk kurva regresi. Salah satu pendekatan regresi nonparametrik untuk memeroleh dugaan kurva regresi adalah *smoothing spline*. Menurut Aydin [5] *smoothing spline* memiliki hasil yang lebih baik daripada regresi kernel. Masalah utama pada saat menduga fungsi regresi *smoothing spline* adalah memilih dan menentukan parameter pemulus (Cantoni dan Hastie [6]).

Menurut Lee [7], untuk menentukan parameter pemulus pada regresi *smoothing spline* tersebut dapat digunakan metode GCV. Dalam penelitian ini digunakan metode GCV untuk mendapatkan hasil yang maksimum pada parameter pemulus (λ). Nilai λ dipilih dari nilai GCV(λ) yang minimum.

2. Model Regresi Nonparametrik

Menurut Eubank [8] model regresi nonparametrik merupakan model regresi yang digunakan untuk mengestimasi kurva regresi yang hanya tergantung pada data amatan. Model regresi nonparametrik tidak memberikan asumsi terhadap bentuk kurva regresi. Kurva tersebut hanya diasumsikan termuat dalam suatu ruang fungsi tertentu, dimana pemilihan ruang fungsi ini biasanya dimotivasi oleh sifat kemulusan (*smoothness*) yang diasumsikan dimiliki oleh fungsi regresi tersebut. Ini memberikan fleksibilitas yang lebih besar di dalam bentuk yang mungkin dari kurva regresi. Untuk mengonstruksi model regresinya dipilih ruang fungsi yang sesuai, yang mana kurva regresi diyakini termasuk didalamnya.

Diberikan n pengamatan (x_i, y_i) dengan $i = 1,2,3, \dots, n$ untuk x_i dan y_i dalam R. Variabel x_i merupakan variabel prediktor pada pengamatan ke-i, variabel y_i merupakan variabel respon pada pengamatan ke- i, dan R merupakan bilangan riil. Hubungan antara x_i dan y_i diasumsikan mengikuti model regresi berikut.

$$y_i = f(x_i) + \varepsilon_i \quad , i = 1,2,3, \dots, n, \tag{1}$$

dengan $f(x_i)$ adalah fungsi f yang tidak diketahui dan ε_i adalah sisaan yang diasumsikan berdistribusi normal independen dengan *mean* nol dan variansi σ^2 (Fox [9]).

Eubank [8] menyatakan bahwa ada beberapa teknik untuk mengestimasi kurva regresi f dalam regresi nonparametrik, salah satunya dengan *smoothing spline*.

3. Model Regresi Nonparametrik *Spline*

Salah satu model regresi dengan pendekatan nonparametrik yang dapat digunakan untuk menduga kurva regresi adalah regresi *spline*. Regresi *spline* merupakan suatu pendekatan ke arah plot data dengan tetap memperhitungkan kemulusan kurva.

Secara umum fungsi *spline* orde ke-*m* ditulis sebagai berikut :

$$f(x) = \beta_0 + \sum_{j=1}^{m} \beta_j x^j + \sum_{j=1}^{k} \theta_j \left(x - X_j\right)_+^m , \tag{2}$$

dengan fungsi terpotong sebagai berikut :

$$\left(x - X_j\right)_+^m = \begin{cases} \left(x - X_j\right)_+^m ; x \geq X_j \\ 0 \qquad\qquad ; x < X_j \end{cases},$$

dengan β_0 merupakan parameter model, β_j merupakan parameter pada variabel x^j, θ_j merupakan parameter pada variabel prediktor pemotongan knot ke-j, x^j merupakan variabel prediktor orde ke-j, X_j merupakan knot ke-j pada variabel x^j , nilai $j = 1,2, \dots, m$, m merupakan orde *spline*, dan k adalah banyak knot.

Dari bentuk matematis fungsi *spline*, ditunjukkan bahwa *spline* merupakan model polinomial yang tersegmen (*piecewise polynomial*), tetapi *spline* masih bersifat kontinu pada knot-knotnya.

4. Model Regresi Nonparametrik *Smoothing Spline*

Smoothing merupakan suatu proses yang dapat menghilangkan data kasar dengan mengikuti bentuk pola data. Berdasarkan fungsi regresi nonparametrik, f adalah fungsi pemulus dan $E(\varepsilon) = 0$, Fahrmeir dan Tuhtz [10] menduga kurva pemulus $\hat{f}(x_i)$ dapat diperoleh berdasarkan data amatan, yakni pasangan variabel prediktor dan variabel respon. Penduga fungsi pemulus merupakan penduga fungsi yang mampu memetakan

data dengan baik serta mempunyai variansi *error* yang kecil. Oleh karena itu, dengan menggunakan data amatan sebanyak *n*, $f(x_i)$ dapat diperoleh dengan meminimumkan fungsi *penalized least square* (PLS), yaitu

$$PLS = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx \quad (3)$$

dengan *n* adalah banyaknya data amatan, λ sebagai parameter pemulus dan $f''(x)$ adalah fungsi turunan tingkat dua.

5. GCV (*Generalized Cross Validation*)

Metode GCV digunakan untuk menentukan λ dalam regresi *smoothing spline*. Lee [7] mendefinisikan bentuk umum dari metode GCV sebagai berikut

$$GCV(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n}\{y_i - f_\lambda(x_i)\}^2}{\{1-n^{-1}tr(S_\lambda)\}^2}, \quad (4)$$

dengan f_λ adalah estimator dari *smoothing spline* dan $tr(S_\lambda) \leq n$. Nilai λ dipilih dari nilai GCV(λ) yang minimum.

Adapun langkah-langkah yang dilakukan untuk menentukan λ pada regresi *smoothing spline* diasumsikan sebagai berikut.

1. *Input* data $(x_{1i}, x_{2i}, x_{3i}, \dots, x_{ni}, y_i)$.
2. Menghitung matriks $\boldsymbol{T}$ dan $\boldsymbol{H}$ kemudian matriks $\boldsymbol{L}$.

$$\boldsymbol{T} = \left[\begin{pmatrix} \boldsymbol{R} + \lambda \boldsymbol{I}_n & \boldsymbol{Q}^t \\ \boldsymbol{Q} & 0 \end{pmatrix}^{-1}\right]_{(n\times n)},$$

dimana notasi $[.]_{n\times n}$ menunjukkan submatriks berukuran $n \times n$ yang dibentuk dari bagian kiri atas matriks utama

$$\left[\begin{array}{c:c} \underset{n\times n}{\boldsymbol{T}} & \underset{n\times 2}{\boldsymbol{T_1}} \\ \hdashline \underset{2\times n}{\boldsymbol{T_2}} & \underset{2\times 2}{\boldsymbol{T_3}} \end{array}\right]$$

dengan $\boldsymbol{R} = \begin{pmatrix} 0 & \frac{|x_{1i}-x_{2i}|^3}{12} & \frac{|x_{1i}-x_{3i}|^3}{12} & \dots & \frac{|x_{1i}-x_{ni}|^3}{12} \\ \frac{|x_{2i}-x_{1i}|^3}{12} & 0 & \frac{|x_{2i}-x_{3i}|^3}{12} & \dots & \frac{|x_{2i}-x_{ni}|^3}{12} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{|x_{ni}-x_{1i}|^3}{12} & \frac{|x_{ni}-x_{2i}|^3}{12} & \frac{|x_{ni}-x_{3i}|^3}{12} & \dots & 0 \end{pmatrix}$,

dan $\boldsymbol{Q} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{1i} & x_{2i} & x_{3i} & \cdots & x_{ni} \end{pmatrix}$.

Selanjutnya menentukan nilai λ yang diperoleh dari,

$$\lambda = \frac{1-q}{q}, 0 < q < 1,$$

Matriks $\boldsymbol{H}$ didefinisikan sebagai berikut $\boldsymbol{H} = (\boldsymbol{R} \quad \boldsymbol{Q^t}) \begin{pmatrix} \boldsymbol{R} + \boldsymbol{\lambda I_n} & \boldsymbol{Q^t} \\ \boldsymbol{Q} & 0 \end{pmatrix}^{-1}$ merupakan submatriks berukuran $n \times n$ yang di bentuk dari bagian kiri matriks utama,

$\left[\begin{array}{c:c} \boldsymbol{H} & \boldsymbol{H_1} \\ n \times n & n \times 2 \end{array}\right]$

dan matriks $\boldsymbol{L}$ didefinisikan sebagai $\boldsymbol{L} = \left(\boldsymbol{TH^{-1}}\right)^{\boldsymbol{t}} \boldsymbol{RTH^{-1}}$.

3. Menghitung matriks $\boldsymbol{S_\lambda}$ dengan $\boldsymbol{S_\lambda} = (\boldsymbol{I} + \boldsymbol{\lambda L})^{-1}$.
4. Menghitung f_λ untuk berbagai nilai λ dengan $f_\lambda = \boldsymbol{S_\lambda} y$.
5. Memilih f_λ yang meminimumkan GCV(λ) dengan $GCV(\lambda) = \frac{1}{n} \frac{\sum_{i=1}^{n} \{y_i - f_\lambda(x_i)\}^2}{\{1 - n^{-1} tr(\boldsymbol{S_\lambda})\}^2}$.

6. Metode Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari BPS Provinsi Jawa Tengah berupa produksi ubi jalar di Provinsi Jawa Tengah pada tahun 2015. Variabel respon yang digunakan dalam penelitian ini adalah produksi ubi jalar di 35 kabupaten/kota Jawa Tengah. Variabel prediktor yang digunakan, yaitu luas panen, alokasi pupuk urea bersubsidi, alokasi pupuk organik bersubsidi di Provinsi Jawa Tengah tahun 2015. Data alokasi pupuk urea bersubsidi dan alokasi pupuk organik bersubsidi diperoleh dari Dinas Pertanian, Tanaman Pangan dan Holtikultura Provinsi Jawa Tengah, sedangkan data luas panen dan produksi ubi jalar diperoleh dari BPS Provinsi Jawa Tengah.

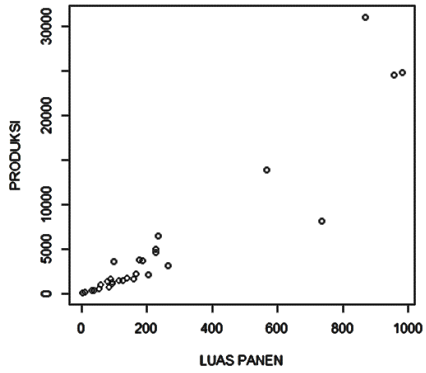
Langkah-langkah yang dilakukan untuk mencapai tujuan penelitian.

1. Membuat diagram pencar hubungan antara produksi ubi jalar terhadap masing-masing variabel prediktor.
2. Menentukan nilai λ yang digunakan pada masing-masing variabel prediktor.
3. Menentukan nilai estimator *smoothing spline* pada masing-masing variabel prediktor.
4. Menentukan λ optimum berdasarkan nilai GCV minimum pada masing-masing variabel prediktor.

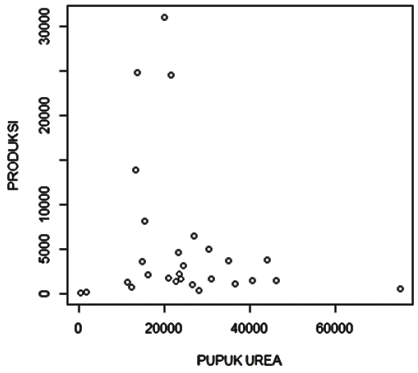
7. Hasil dan Pembahasan

7.1 Diagram Pencar Data. Diagram pencar hubungan antara produksi ubi jalar terhadap luas panen, alokasi pupuk urea bersubsidi, dan alokasi pupuk organik bersubsidi dapat dilihat pada Gambar 1.

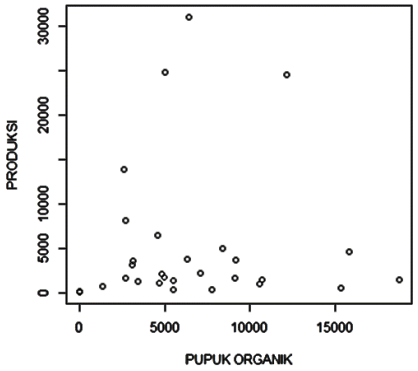
 ISSN 2621 – 086 X



(a)



(b)



(c)

Gambar 1. Diagram pencar hubungan: (a) produksi ubi jalar terhadap luas panen; (b) produksi ubi jalar terhadap alokasi pupuk urea bersubsidi; (c) produksi ubi jalar terhadap alokasi pupuk organik bersubsidi

Diagram pencar pada Gambar 1 mengindikasikan bahwa data tidak memiliki pola tertentu sehingga tidak mengikuti distribusi tertentu. Produksi ubi jalar dapat diterapkan dalam model regresi nonparametrik. Estimasi fungsi regresi nonparametrik dilakukan berdasarkan data pengamatan dengan menggunakan teknik pemulusan (*smoothing*). Pendekatan yang digunakan untuk regresi nonparametrik salah satunya adalah pendekatan dengan regresi *smoothing spline*. Adapun metode yang digunakan dalam regresi *smoothing spline* adalah metode GCV.

7.2 GCV (*Generalized Cross Validation*). Metode GCV adalah metode klasik yang digunakan untuk menentukan λ pada regresi *smoothing spline*. Nilai λ dipilih dari nilai GCV yang minimum. Langkah yang dilakukan sebelum memeroleh GCV minimum

adalah menentukan estimator *smoothing spline* dari berbagai nilai λ. Matriks estimator *smoothing spline* berdasarkan nilai λ optimum pada masing-masing faktor yang meminimumkan GCV dapat dilihat pada Tabel 4.1.

Tabel 4.1. Nilai f_λ untuk λ optimum pada masing-masing faktor

Variabel prediktor	Luas Panen	Pupuk Urea	Pupuk Organik
f_λ	4941.9999999999945	4942.000000000024	4942.
	1619.9999999999995	1619.9999999992783	1620.0000000000005
	3603.999999999982	3604.000000000146	3604.
	1681.9999999999998	1681.9999999996382	1682.0000000000002
	981.9999999999997	982.0000000000001	982.
	2136.	2136.000000000001	2136.
	8129.	8128.999999999842	8128.999999999998
	24572.999999999996	24573.000000000575	24573.
	313.9999999999995	314.00000000803686	314.
	355.00000000000074	354.9999999919489	355.
	1598.9999999999975	1599.0000000000025	1599.
	31076.	31076.000000000007	31076.
	511.0000000000009	511.00000000012307	511.
	1424.9999999999995	1424.9999999998881	1425.
	2197.0000000000014	2197.000000002583	2197.
	1403.0000000000016	1402.9999999999889	1403.
	1283.999999999979	1283.9999999999982	1284.
	1315.9999999999964	1316.000000000138	1316.
	3636.	3635.9999999999995	3636.
	24812.	24811.999999999807	24812.

4577.000000000005	4576.999999997797	4577.
6474.	6474.000000000104	6474.
13849.	13849.000000000162	13849.
738.0000000000056	737.9999999999998	738.
3075.	3075.0000000000014	3075.
1075.0000000000377	1075.000000000003	1075.
3742.999999999999	3742.9999999998713	3743.
34.00000000000449	34.00000000036755	34.
150.99999999999517	150.99999999967116	151.

Setelah ditentukan estimator *smoothing spline*, langkah selanjutnya adalah menentukan nilai GCV yang minimum. Hasil perhitungan dari berbagai nilai λ dan nilai GCV ditunjukkan pada Tabel 4.2.

Tabel 4.2. Nilai λ dan GCV pada masing-masing faktor

Luas Panen		Pupuk Urea		Pupuk Organik	
λ	GCV	λ	GCV	λ	GCV
$1.362073.10^{-17}$	$1.24793.10^{6}$	$8.932447.10^{-21}$	$8.08271.10^{7}$	$7.878948.10^{-17}$	192.012
$5.57905.10^{-14}$	$2.29272.10^{-21}$	$2.51426.10^{-6}$	$1.38391.10^{-16}$	$3.227217.10^{-13}$	$3.46813.10^{-24}$
$2.607499.10^{-5}$	0.000493612	0.001173893	$2.60289.10^{-16}$	$5.414372.10^{-6}$	$2.90679.10^{-10}$
0.0001376246	0.0137452	0.01029841	$2.32165.10^{-9}$	0.0007960934	$6.28413.10^{-6}$
0.07624897	3248.8	42.18229	0.0389435	0.08525932	0.0719865

Berdasarkan Tabel 4.2 nilai GCV minimum dan nilai λ optimum untuk masing-masing faktor adalah

1. luas panen: $\lambda = 5.57905.10^{-14}$; GCV = $2.29272.10^{-21}$
2. pupuk urea: $\lambda = 2.51426.10^{-6}$; GCV = $1.38391.10^{-16}$
3. pupuk organik: $\lambda = 3.227217.10^{-13}$; GCV = $3.46813.10^{-24}$.

8. Kesimpulan

Berdasarkan hasil dan pembahasan diperoleh kesimpulan bahwa nilai λ optimum terhadap faktor luas panen, pupuk urea, dan pupuk organik adalah $5.57905.10^{-14}$,

$2.51426.10^{-6}$, dan $3.227217.10^{-13}$ dengan GCV minimum sebesar $2.29272.10^{-21}$, $1.38391.10^{-16}$, dan $3.46813.10^{-24}$.

9. Daftar Pustaka

[1] Soekartawi, *Agribisnis Teori dan Aplikasi*, PT. Raja Grafindo Persada, Jakarta, 1999.

[2] Purwono, Purnamawati, H., *Budidaya 8 Jenis Tanaman Pangan*, Jakarta: Penebar Swadaya, 2007.

[3] Najwah, I. N., *Analisis Efisiensi Usahatani Ubi Jalar (Ipomoea batatas L.) di Kabupaten Karanganyar*, Skripsi Fakultas Pertanian, Universitas Sebelas Maret, Surakarta, 2014.

[4] Defri, K., *Analisis Pendapatan dan Faktor-faktor yang Memengaruhi Produksi Usahatani Ubi Jalar (Studi Kasus Desa Purwasari, Kecamatan Dramaga, Kabupaten Bogor)*, Skripsi Fakultas Ekonomi dan Manajemen, Institut Pertanian Bogor, Bogor, 2011.

[5] Aydin, D., *A Comparison of The Nonparametric Regression Models Using Smoothing Spline and Kernel Regression,* World Academy Science, Enginering and Technology, 2007; (36): 253-257.

[6] Cantoni, E. dan Hastie, T., *Degrees of Freedom Tests for Smoothing Splines,* Statistics Department, Stanford University, 2000.

[7] Lee, T. C. M., *Smoothing Parameter Selection for Smoothing Splines: a Simulation Study*, Computational Statistic & Data Analysis, 2003; (42): 139-148.

[8] Eubank, R., *Nonparametric Regression and Spline Smoothing. Second Edition*, New York: Marcel Dekker, 1999.

[9] Fox, J., *Nonparametric Regression* [Internet]. 2002 [cited 2009 Jan 24]. Available from: http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-nonparametric-regression.pdf, [cited 2017 Jun 02].

[10] Fahrmeir, L. dan Tuhtz, G., *Multivariate Statistical Modelling Based on Generalized Linear Models,* Springer-Verlag, New York, 1994.