Indonesian Journal of Applied Statistics ISSN 2621–086X
Volume 3 No 2 November 2020

Estimator Nadaraya-Watson dengan Pendekatan *Cross Validation* dan *Generalized Cross Validation* untuk Mengestimasi Produksi Jagung

Febriolah Lamusu[1], Tedy Machmud[2], Resmawan[3]
[1,2,3]Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Gorontalo

febriolarol@gmail.com

Abstract. Nadaraya-Watson Estimator with kernel approach depends on two-parameter, those are kernel function and bandwidth choice. However, between the two of them, bandwidth choice gave a huge impact on the result of the estimation. By minimizing the value of Mean Square Error (MSE), Cross-Validation (CV) and Generalized Cross-Validation (GCV) gave the optimal bandwidth value. In this research, corn production was considered as the dependent variable, while the planted area, harvested area, and the fertilizer as the independent variable. The result of this research showed that Nadaraya-Watson Estimator with Generalized Cross-Validation gives a better corn production estimation with optimal bandwidth value 742392,2, with $h_1 = 400, h_2 = 400,$ and $R_2 = 99{,}99\%$ with MSE 202583,9.

Keywords: kernel, estimator Nadaraya-Watson, cross validation, generalized cross validation.

1. Pendahuluan

Regresi nonparametrik merupakan metode pendugaan model yang dilakukan berdasarkan pendekatan yang tidak terikat asumsi bentuk kurva regresi tertentu, dimana kurva regresi hanya diasumsikan mulus [1]. Beberapa metode pendekatan yang paling popular digunakan dalam regresi nonparametrik adalah *spline* [2], *Fourier* [3], *wavelet* [4] dan kernel. Pada penelitian ini menggunakan kernel karena memiliki kemampuan yang baik dalam memodelkan data yang tidak mempunyai pola tertentu [5]. Selain itu regresi kernel lebih fleksibel, bentuk matematisnya mudah, dan dapat mencapai tingkat konvergen yang relatif cepat [6]. Regresi kernel memiliki beberapa fungsi, diantaranya fungsi Gaussian, fungsi *Epanechnikov*, fungsi *triangle*, fungsi *uniform*, fungsi *cosinus*, dan fungsi kuadratik [7]. Fungsi kernel yang biasa digunakan adalah fungsi Gaussian, karena fungsi Gaussian lebih umum digunakan dan lebih *smooth* dibandingkan dengan fungsi kernel yang lain [8]. Pada regresi kernel salah satu estimator yang sering digunakan yaitu estimator Nadaraya-Watson. Beberapa penelitian sebelumnya tentang estimator Nadaraya-Watson telah dilakukan oleh Nurul dkk. [9], Ida [10], dan Tiani dkk. [11].

Estimator Nadaraya-Watson dengan pendekatan kernel tergantung pada dua

parameter yaitu fungsi kernel dan pemilihan *bandwidth* yang digunakan [7]. Namun diantara keduanya pemilihan *bandwidth* memiliki pengaruh yang paling kuat terhadap estimasi yang dihasilkan [6]. Metode yang digunakan untuk mendapatkan *bandwidth* yang optimal adalah dengan menggunakan metode *Cross Validation* (CV) dan *Generalized Cross-Validation* (GCV) [12].

Provinsi Gorontalo dikenal sebagai salah satu sentra produksi jagung nasional yang mampu memenuhi kebutuhan komoditas di pasar dalam negeri, maupun untuk melakukan ekspor. Menurut dinas pertanian Provinsi Gorontalo produksi jagung di Provinsi Gorontalo terus tajam dari 692.000 ton pada tahun 2016, menjadi 1,5 juta ton pada tahun 2018. Bahkan dari total ekspor jagung 380.000 ton pada tahun 2018, sebanyak 113.000 tonnya adalah hasil produksi petani Provinsi Gorontalo. Namun, pada tahun 2019 terjadi penurunan produksi jagung di Provinsi Gorontalo [13]. Terdapat sekitar 4.405 hektar lahan pertanian padi dan jagung yang mengalami puso [14]. Oleh karena itu, perlu dilakukan upaya agar produksi jagung di Provinsi Gorontalo tetap terjaga. Salah satu upaya yang dapat dilakukan adalah dengan melakukan pemodelan untuk memprediksi dan mengetahui produksi jagung di Provinsi Gorontalo. Data produksi jagung adalah data yang berfluktuatif dan tidak membentuk suatu pola hubungan tertentu yang tidak diketahui bentuk fungsinya. Sehingga pendekatan nonparametik adalah pendekatan yang paling tepat untuk digunakan. Hasil pemodelan ini diharapkan dapat membantu pihak terkait dalam melakukan langkah-langkah strategis agar tidak mengalami kerugian yang signifikan.

2. Landasan Teori

Perbandingan metode yang digunakan untuk penelitian ini yaitu metode Cross Validation (CV) dan Generalized Cross Validation (GCV) menggunakan estimator Nadaraya-Watson.

2.1. Metode *Cross-Validation* (CV). Salah satu metode untuk menentukan nilai *bandwidth* adalah *Cross Validation* (CV). Metode *Cross Validation* atau sering disebut CV adalah metode pendugaan data untuk menunjukkan apa yang harus dilakukan jika pengulangan observasi tersedia. Pada pemilihan *bandwidth* yang optimum didasarkan nilai CV yang minimum [9].

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}(X_{\neq i})]^2$$

dengan

CV : *Cross Validation*

h : *Bandwidth*

y_i : Variabel dependen ke-*i*

$\widehat{m}_x$: Nilai estimasi

2.2. Metode *Generalized Cross Validation* (GCV). Metode *Generalized Cross Validation* (GCV) dalam regresi kernel adalah satu metode untuk memilih *bandwidth* optimal dengan meminimalkan fungsi GCV. Optimasi GCV adalah memilih *h* optimal yang meminimalkan nilai GCV [7]:

$$GCV(h) = \frac{n^{-1}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{\{1 - tr(H(h)/n\}^2}$$

dengan

GCV(*h*) : nilai GCV pada *bandwidth h*

n : GCV pada *bandwidth h*

y_i : data aktual subjek ke-*i*

$\hat{y}_i$: hasil estimasi subjek ke-*i*

tr(*H*) : jumlah dari elemen diagonal utama matriks penghalus n x n

2.3. Estimator Nadaraya-Watson. Estimator Nadaraya-Watson diperkenalkan pada tahun 1964 oleh Nadaraya dan Watson. Estimator ini untuk memperkirakan *m* sebagai rata-rata tertimbang secara lokal dengan menggunakan kernel sebagai fungsi pembobotan [15].

$$\hat{m}(X^{(1)}, X^{(2)}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\widehat{\varphi}X_i}K_{h_1}(x^{(1)} - X_i^{(1)})K_{h_2}(x^{(2)} - X_i^{(2)})$$

dalam [16] dituliskan bahwa

$$\widehat{\varphi}(X_i) = \hat{\varphi}(X_i^1)\hat{\varphi}(X_i^2) = \sum_{j=1}^{h}K_{h_1}(X_j^{(1)} - X_i^{(1)})K_{h_2}(X_j^{(2)} - X_i^{(2)})$$

dengan

K : fungsi kernel

h : nilai *bandwidth*

X_i : nilai amatan variabel prediktor ke-*i*

y_i : nilai amatan variabel respon ke-*i*

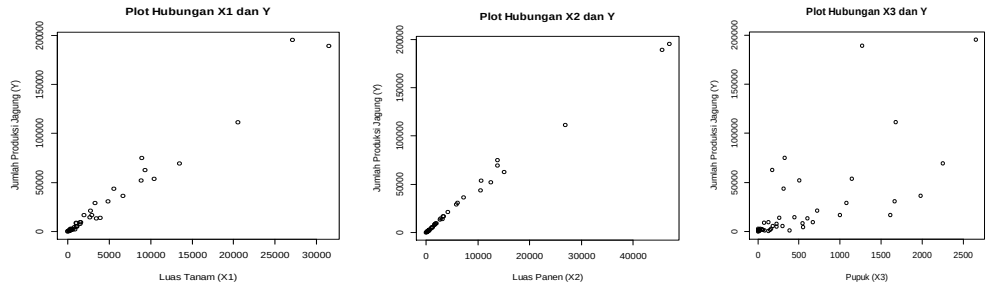
x : nilai random variabel *X* atau nilai tertentu dari variabel *X*

3. Hasil dan Pembahasan

Data yang digunakan dalam penelitian ini adalah data sekunder. Data tersebut

adalah data produksi jagung, luas panen, luas tanam dan pupuk yang didapat dari Dinas Pertanian Kabupaten Boalemo, Dinas Pertanian Kabupaten Gorontalo dan Dinas Pertanian Kabupaten Bone Bolango pada tahun 2019. Pengolahan data pada penelitian ini menggunakan software R.

3.1. *Scatter Plot*. Dalam menentukan model regresi nonparametrik kernel, langkah awal yang dilakukan adalah membuat *scatter plot* antara variabel dependen dan variabel independen.



Gambar 1. Plot Produksi Jagung dengan Luas Tanam, Luas Panen dan Pupuk

Gambar 1 menunjukkan hubungan $X1$ dan Y membentuk garis lurus. Akan tetapi data banyak terkumpul pada titik awal, sedangkan data lainnya bernilai cukup jauh berbeda sehingga terdapat beberapa data pencilan yang terlihat pada Gambar 1. Begitupun dengan hubungan $X2$ dan Y membentuk garis lurus namun data banyak terkumpul di titik awal dan data lainnya berada cukup jauh sehingga terdapat beberapa data pencilan. Pada Gambar 1 hubungan $X3$ dan Y diketahui tidak membentuk pola apapun. Data menyebar secara acak sehingga tidak jelas hubungan $X3$ dan Y. Ada juga beberapa data pencilan yang terlihat pada Gambar 1. Oleh karena itu, pola data seperti ini sangat cocok untuk didekati dengan pendekatan regresi nonparametrik.

3.2. Analisis Korelasi. Uji analisis korelasi dalam regresi nonparametrik kernel dilakukan untuk mengetahui variabel-variabel independen yang berpengaruh terhadap variabel dependen. Dibawah ini merupakan hasil uji korelasi yang disajikan dalam Tabel 1.

Tabel 1. Analisis Korelasi

Variabel	Nilai Sig	Nilai Korelasi
Luas Tanam	0,00	0,907
Luas Panen	0,00	0,947
Pupuk	0,00	0,734

Tabel 1 menunjukkan nilai signifikan dari ketiga variabel 0,00 < α=0,05 artinya variabel

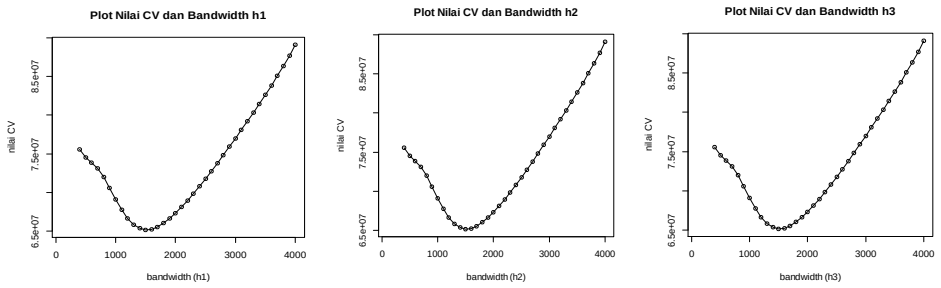
luas tanam, luas panen, dan pupuk berpengaruh signifikan terhadap variabel produksi jagung.

3.3. Pemilihan *Bandwidth* Optimum dengan *Cross-Validation* (CV). Pemilihan *bandwidth* optimum sangat penting dalam regresi kernel, karena nilai *bandwith* memiliki pengaruh yang paling kuat terhadap hasil estimasi. Berikut adalah nilai *bandwidth* optimum menggunakan metode CV dengan *software* R disajikan pada Tabel 2.

Tabel 2. Nilai *bandwidth* dengan CV

h_1	h_2	h_3	CV	MSE
1500	1500	1500	65176849	8931740

Tabel 2 menunjukkan *bandwidth* optimum berdasarkan kriteria CV adalah *bandwidth* yang memiliki nilai CV terkecil yaitu 65176849 dengan nilai MSE sebesar 8931740 dan *bandwidth* optimum adalah h_1 = 1500, h_2 = 1500, h_3 = 1500. Plot nilai *bandwidth* h_1, h_2, dan h_3 dengan CV dapat dilihat pada Gambar 2.



Gambar 2. Nilai *Bandwidth* dengan CV

Gambar 2 menunjukkan nilai *bandwidth* optimum CV h_2 berada pada titik 1500, *bandwidth* optimum CV h_2 berada pada titik 1500 dan *bandwidth* optimum CV h_3 berada pada titik 1500.

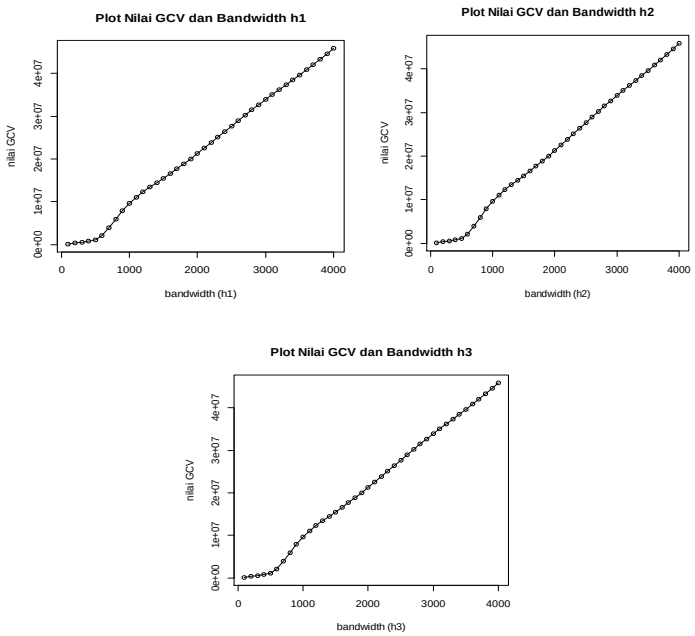
3.4. Pemilihan *Bandwidth* Optimum dengan *Generalized Cross Validation* (GCV). Nilai *bandwidth* optimum menggunakan metode GCV dapat dilihat pada Tabel 3.

Tabel 3. Nilai *Bandwidth* dengan GCV

h_1	h_2	h_3	GCV	MSE
400	400	400	742392,2	202583,9

Tabel 3 menunjukkan *bandwidth* optimum berdasarkan kriteria GCV adalah yang memiliki nilai *GCV* terkecil yaitu 742392.2 dengan *MSE* sebesar 202583.9 dan *bandwidth* optimum

adalah $h_1 = 400$, $h_2 = 400$, $h_3 = 400$. Plot nilai *bandwidth* h_1, h_2, dan h_3 dengan GCV dapat dilihat pada Gambar 3.



Gambar 3. Plot *Bandwidth* GCV

Gambar 3 menunjukkan nilai *bandwidth* optimum GCV h_1 berada pada titik 400, *bandwidth* optimum GCV h_2 berada pada titik 400 dan *bandwidth* optimum GCV h_3 berada pada titik 400.

3.5. Estimasi Nadaraya-Watson dengan Metode *Cross Validation* (CV). Berdasarkan perhitungan *bandwidth* optimum *Cross Validation* model regresi kernel Gaussian dengan estimator Nadaraya-Watson menghasilkan persamaan untuk memprediksi produksi jagung sebagai berikut:

$$\hat{m}\left(X^{(1)}, X^{(2)}, X^{(3)}\right) = -\frac{1}{(1500)(1500)(1500)} \sum_{i=1}^{44} \frac{y_i}{\hat{\varphi} X_i^{(1)} \hat{\varphi} X_i^{(2)}} K\left(\frac{x^{(1)} - X_i^{(1)}}{1500}\right) K\left(\frac{x^{(2)} - X_i^{(2)}}{1500}\right) K\left(\frac{x^{(3)} - X_i^{(3)}}{1500}\right)$$

dengan

$$\hat{\varphi}(X_i) = \hat{\varphi}(X_i^1)\hat{\varphi}(X_i^2)\hat{\varphi}(X_i^3) = \sum_{j=1}^{44} K_{1500}(X_j^{(1)} - X_i^{(1)}) K_{1500}(X_j^{(2)} - X_i^{(2)}) K_{1500}(X_j^{(3)} - X_i^{(3)})$$

3.6. Estimasi Nadaraya-Watson dengan Metode *Generalized Cross-Validation* (GCV). Berdasarkan perhitungan *bandwidth* optimum *Generalized Cross-Validation* model regresi kernel Gaussian dengan estimator Nadaraya-Watson menghasilkan

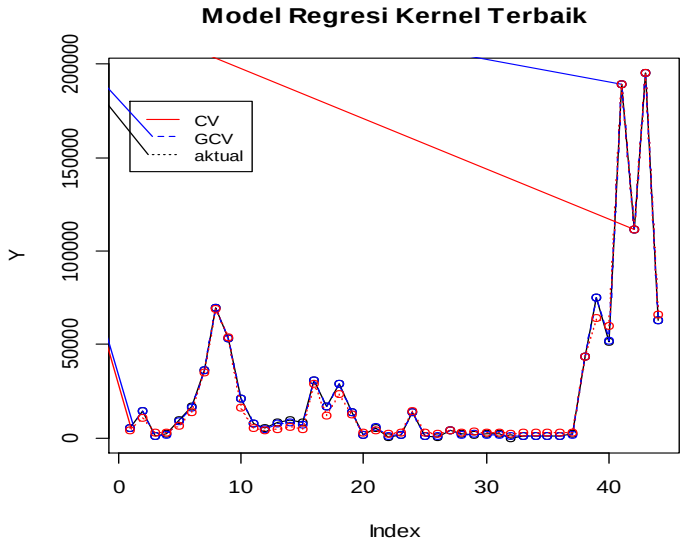
persamaan untuk memprediksi produksi jagung:

$$\hat{m}(X^{(1)}, X^{(2)}, X^{(3)}) = -\frac{1}{(400)(400)(400)} \sum_{i=1}^{44} \frac{y_i}{\hat{\varphi} X_i^{(1)} \hat{\varphi} X_i^{(2)}} K\left(\frac{x^{(1)} - X_i^{(1)}}{400}\right) K\left(\frac{x^{(2)} - X_i^{(2)}}{400}\right) K\left(\frac{x^{(3)} - X_i^{(3)}}{400}\right)$$

dengan

$$\hat{\varphi}(X_i) = \hat{\varphi}(X_i^1)\hat{\varphi}(X_i^2)\hat{\varphi}(X_i^3) = \sum_{j=1}^{44} K_{400}(X_j^{(1)} - X_i^{(1)}) K_{400}(X_j^{(2)} - X_i^{(2)}) K_{400}(X_j^{(3)} - X_i^{(3)})$$

Kurva data aktual produksi jagung dan data estimasi menggunakan metode pendekatan *Cross Validation* (CV) dan *Generalized Cross Validation* (GCV) di Provinsi Gorontalo dapat dilihat pada Gambar 4.



Gambar 4. Plot Estimasi Data Aktual dengan CV dan GCV

Gambar 4 menunjukkan bahwa kurva data aktual dan prediksi data menggunakan estimator Nadaraya-Watson dengan metode CV dan GCV memiliki pola yang sama dan saling berhimpit. Akan tetapi diantara kedua metode, GCV lebih mendekati data aktual.

3.7. Perbandingan Model. Untuk menentukan metode terbaik dalam melakukan estimasi produksi jagung di Provinsi Gorontalo dapat membandingkan nilai koefisien determinasi dan MSE yang dihasilkan. Hasil perhitungan nilai koefisien dan MSE dari metode CV dan GCV ditunjukkan pada Tabel 4.

Tabel 4. Nilai *R-Square* dan MSE

Metode	Nilai *R-Square*	MSE
Cross Validation	0,995349	8931740,0
Generalized Cross Validation	0,9998944	202583,9

Tabel 4 memperlihatkan nilai *R-Square* dari GCV sebesar 0,9998944, lebih besar daripada nilai *R-Square* CV sebesar 0,995349 sehingga metode GCV lebih baik daripada metode CV. Nilai MSE GCV sebesar 202583,9 lebih kecil daripada MSE CV sebesar 8931740,0. Oleh karena itu metode yang lebih baik untuk mengestimasi produksi jagung di Provinsi Gorontalo adalah metode GCV.

4. Kesimpulan

Berdasarkan hasil dan pembahasan dapat disimpulkan bahwa metode *Generalized Cross-Validation* (GCV) lebih baik daripada metode *Cross-Validation* (CV) untuk mengestimasi data produksi jagung di Provinsi Gorontalo. Metode GCV menghasilkan nilai *bandwidth* optimum sebesar 742392,2 dengan $h_1 = 400$, $h_2 = 400$, dan $h_3 = 400$, nilai koefisien determinasi sebesar 0,9998944, dan MSE = 202583,9.

Daftar Pustaka

[1] Eubank, R. L. *Nonparametric Regression and Spline Smoothing*. Second Edition. Marcel Dekker Inc. New York. 1999.
[2] Lestari, B., Chamidah, N., Saifudin, T. Estimasi Fungsi Regresi dalam Model Regresi Nonparametric Birespon Spline Menggunakan Estimator Smoothing Spline dan Estimator Kernel. *Jurnal Matematika, Statistika, dan Komputasi (JMSK)*, Vol. 15 No. 2. 20-24. 2019.
[3] Khairunnisa, L.R., Prahutama, A., dan Santoso, R. Pemodelan Regresi Semiparametrik dengan Pendekatan Deret Fourier. *Jurnal Gaussian*, Vol. 9 No 1. 50-63. 2020.
[4] Suparti, Rezzy, E.C., Warsito, B., dan Yasin, H. The Shift Invariant Discrete Wavelet Transform (SIDWT) With Inflation Time Series Application. *Jurnal of Mathematics*, 8 (4). 2016.
[5] Hardle, W. *Applied Nonparametric Regression*, Cambridge University Press, New York. 1994.
[6] Hardle, W. *Applied Nonparametric Regression*, Cambridge University Press, New York. 1990.
[7] Suparti, Santoso, R., Prahutama, A., dan Devi, A.R. *Regresi Nonparametrik*. Wade Group. Ponorogo. 2018.
[8] Komang, G. dan Gusti, A. Estimator Kernel dalam Model Regresi Nonparametrik. *Jurnal Matematika*, Vol. 2 No. 1. 19-30. 2012.
[9] Nurul, A., Noami, N.D., dan Shantika, M. Estimasi Model Regresi Nonparametrik Kernel Menggunakan Estimator Nadaraya-Watson. *Jurnal Matematika, Statistika dan Terapannya*, Vol. 8 No. 4, 633-638. 2019.
[10] Purwanti, I. Regresi Nonparametrik Kernel Menggunakan Estimator Nadaraya-Watson dalam Data Time Series. *Jurnal Penelitian*, Vol. 1 No. 1. 49-56. 2019.
[11] Tiani, W. U., Martyana, P., and Vega, Z. V. Kernel Nonparametric Regression for The Modelizing of The Productivity Wetland Paddy. *Jurnal International Seminar*

Education and Development of Asia. 2018.
[12] Ogden, R.T. *Essential Wavelets for Statistical Applications and Data Analysis*. Boston. Birkhauser. 1997.
[13] Badan Pusat Statistik (BPS) Provinsi Gorontalo. *Provinsi Gorontalo Dalam Angka 2019*. Gorontalo: BPS Provinsi Gorontalo. 2019.
[14] Badan Pusat Statistik (BPS) Provinsi Gorontalo. *Statistik Hortikultura Provinsi Gorontalo 2018*. Gorontalo: BPS Provinsi Gorontalo. 2018
[15] Saputra, J. A., dan Listyani, E. Pemilihan Bandwidth pada Estimator Nadaraya-Watson dengan Tipe Kernel Gaussian pada Data Time Series. *Jurnal Pendidikan Matematika dan Sains*, Vol. 1 No. 1. 1–7. 2016.
[16] Bontemps, C., Robin, J., and Van den Berg, G. Equilibrium Search with Continuous Productivity Dispersion: Theory and Non Parametric Estimation. *International Economic Review*, 41, 305-358. 2000.