ALGORITMA GENETIKA DALAM PENCARIAN TITIK KNOT OPTIMAL PADA REGRESI NONPARAMETRIK SPLINE DENGAN KRITERIA GENERALIZED CROSS VALIDATION, CROSS VALIDATION, DAN UNBIASED RISK

Ade Matao Sukran[1]

[1]Universitas Mataram, adematao00@gmail.com

Abstract. The spline regression method is a highly flexible nonparametric regression analysis method that estimates regression curves effectively by using polynomials with segmented properties, allowing for effective adaptation to local data characteristics. The formation of a spline regression model must meet three criteria, namely determining the order for the model, the number of knots, and the location of the knot placement. Several Criteria are used to select the best spline model in nonparametric spline regression, including Cross Validation (CV), Unbiased Risk (UBR), Generalized Cross Validation (GCV), and Generalized Maximum Likelihood (GML). This research aims to determine the process and results of finding the optimal knot points in truncated spline regression using the best GCV, CV, and UBR methods. Optimization is performed using a genetic algorithm with various iterations such as model, standard deviation, data type, knot number, order number, and data amount. Out of a total of 300 experiments conducted, the program test results indicate that 50% of them demonstrated that the minimum MSE value was obtained when using GCV, as compared to the other criteria.

Keywords: *Cross Validation (CV), Genetic Algorithm (GA), Generalized Cross Validation (GCV), Mean Squared Error (MSE), smoothing parameter, spline regression, Unbiased Risk (UBR).*

Abstrak. Metode regresi spline merupakan salah satu metode dalam analisis regresi nonparametrik yang memiliki fleksibilitas tinggi dalam mengestimasi kurva regresi. Metode ini menggunakan polinomial dengan sifat tersegmen yang memungkinkan untuk menyesuaikan diri secara efektif terhadap karakteristik lokal data. Pembentukan model regresi spline harus memenuhi tiga kriteria yaitu menentukan orde untuk model, banyak knot, dan lokasi penempatan knot. Terdapat beberapa ukuran yang digunakan untuk memilih model spline yang terbaik dalam regresi nonparametrik spline, antara lain metode *Cross Validation* (CV), *Unbiased Risk* (UBR), *Generalized Cross Validation* (GCV), dan *Generalized Maximum Likelihood* (GML). Penelitian ini bertujuan untuk proses dan hasil pencarian titik knot optimal pada regresi *spline truncated* dengan menggunakan metode GCV, CV, dan UBR terbaik. Dilakukan optimasi menggunakan algoritma genetika dengan berbagai perulangan seperti model, banyak standar deviasi, jenis data, banyak knot, banyak orde, dan banyak data. Dari total 300 percobaan, hasil program menunjukkan bahwa lebih dari 50% percobaan nilai minimum MSE menggunakan GCV, dibandingkan dengan kriteria lainnya.

Kata kunci: *Algoritma Genetika (AG), Cross Validation (CV), Generalized Cross Validation (GCV), Mean Squared Error (MSE), Regresi spline, Titik Knot, Unbiased Risk (UBR).*

1 Pendahuluan

Statistika, sebagai salah satu cabang ilmu yang beragam, mencakup berbagai metode dan pendekatan untuk menganalisis data. Analisis regresi, salah satu bagian penting dalam statistika, digunakan untuk memahami hubungan antara variabel-variabel yang ada. Dalam analisis regresi, ada tiga jenis pendekatan umum yang digunakan: regresi parametrik, regresi semiparametrik, dan regresi nonparametrik. Regresi parametrik digunakan ketika bentuk hubungan antara variabel diketahui, sedangkan regresi nonparametrik digunakan ketika bentuk hubungan tidak diketahui.

Dari berbagai metode regresi nonparametrik yang tersedia, metode spline adalah salah satu yang paling fleksibel dan sering digunakan. Spline adalah polinomial tersegmentasi yang dapat menyesuaikan diri dengan baik terhadap karakteristik lokal data. Pembentukan model regresi spline memerlukan penentuan orde, banyaknya knot, dan lokasi knot. Knot merupakan titik perubahan perilaku dalam data.

Dalam pemilihan model spline yang optimal, berbagai metode evaluasi dapat digunakan, termasuk Generalized Cross Validation (GCV), Cross Validation (CV), Unbiased Risk (UBR), dan Generalized Maximum Likelihood (GML). Salah satu metode yang populer adalah GCV, yang memiliki keunggulan dalam sederhana dan efisiensi perhitungannya. Metode CV, metode UBR, dan GCV adalah metode regresi terboboti dengan karakteristik yang berbeda. Penelitian ini bertujuan untuk membandingkan kinerja model regresi nonparametrik spline dengan menggunakan metode GCV dan metode UBR.

Dengan membandingkan metode GCV dan UBR dalam konteks regresi nonparametrik spline dan menggunakan perangkat lunak "R" untuk menerapkannya, penelitian ini diharapkan dapat memberikan wawasan yang berharga dalam memilih metode yang paling cocok untuk situasi tertentu dalam analisis regresi.

2 Metode Penelitian

2.1 Studi Literatur

Penelitian ini membandingkan nilai MSE menggunakan metode GCV, CV, dan UBR dalam pemilihan titik knot yang optimal. Selanjutnya dioptimasi pemilihan titik knot dari metode GCV, CV, dan UBR dan dibandingkan juga nilai MSE yang tidak dioptimasi dan yang dioptimasi algoritma genetika.

2.2 Membuat Package Regresi Nonparametrik Spline Truncted Metode GCV, CV, dan UBR dan Package Optimasi Algoritma Genetika

Proses perancangan program untuk melakukan pencarian titik knot optimal dirancang menggunakan metode GCV, CV, dan UBR. Dirancang juga package untuk masalah optimasi menggunakan algoritma genetika.

2.3 Membuat Data Acak

Data yang digunakan adalah acak distribusi normal atau non-normal dengan standar deviasi 1, 3, 5, 7, dan 9 serta jumlah data mulai dari 50, 200, dan 500.

2.4 Membandingkan Metode GCV, CV, dan UBR untuk Memilih Parameter yang Optimal pada Estimator Spline Berdasarkan Nilai MSE.

Dilakukan perbandingan GCV, CV, dan UBR untuk memilih parameter yang optimal pada estimator spline berdasarkan nilai MSE. Dilakukan optimisasi titik knot agar lebih optimal dan dibandingkan nilai MSE dengan sebelumnya.

2.5 Penarikan Kesimpulan

Tahapan terakhir dari metode penelitian ini adalah dengan melakukan penarikan kesimpulan dari hasil dan pembahasan yang di tunjukkan pada penelitian ini.

3 Hasil dan Pembahasan

Dalam penelitian ini dilakukan perbandingan nilai MSE yang dilihat pengaruhnya terhadap parameter yang telah ditentukan terhadap titik knot optimal yang dihasilkan.

3.1 Proses Pembangkitan Data

Dalam penelitian ini dilakukan pembangkitan dengan beberapa parameter dengan model data, distribusi data, jumlah data, dan standar deviasi yang berbeda. Data dibangkitkan terdiri dari 50, 200, dan 500 dengan standar deviasi 1,3,5,7, dan 9 serta distribusi variabel independen x distribusi data normal atau non-normal, selain itu dibangkitkan data variabel dependen y dengan beberapa fungsi untuk melihat pengaruh pola distribusi data yang tertera pada Tabel. 3.1 sebagai berikut:

Tabel 3. 1 Model Pembangkitan Data

No.	Distribusi data	Model
1	Normal	$y = x + \varepsilon$
2	Normal	$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} + \varepsilon$
3	*Non-normal*	$y = e^x + \varepsilon$
4	*Non-normal*	$y = x^2 + \varepsilon$
5	*Non-normal*	$y = \frac{1}{x} + \varepsilon$

3.2 Proses Optimasi Titik Knot Menggunakan Algoritma Genetika

Proses optimasi dalam penelitian ini menguji 1 titik knot dan 2 titik knot, serta fungsi truncated spline orde 1 dan 2. Titik knot dipilih secara acak untuk merepresentasikan populasi. Pemilihan titik knot optimal dilakukan dengan meminimalkan nilai GCV, CV, dan UBR, menggunakan algoritma genetika. Algoritma genetika melibatkan beberapa tahapan:

3.2.1. Inisialisasi Populasi

Populasi awal dibuat secara acak dengan jumlah individu dan knot yang diinginkan. Titik knot awal juga ditentukan secara acak. Nilai GCV, CV, dan UBR dihitung untuk kombinasi orde fungsi truncated spline orde 1 dan 2.

3.2.2. Perhitungan Fitness

Nilai fitness dihitung untuk setiap individu dengan harapan mendapatkan nilai GCV, CV, dan UBR minimum. Fungsi fitness dirumuskan sesuai dengan Persamaan.

3.2.3. Seleksi

Metode seleksi Roulette Wheel digunakan, memungkinkan individu dengan fitness tinggi lebih mungkin terseleksi.

3.2.4. Crossover

Proses ini menghasilkan individu baru dari individu sebelumnya menggunakan teknik order crossover, yang mengambil angka dari parent dan angka dari data secara acak.

3.2.5. Mutasi

Mutasi terjadi pada setiap individu dalam populasi dengan probabilitas yang mengatur jumlah gen baru yang diuji. Probabilitas mutasi disesuaikan untuk menghindari gangguan acak yang berlebihan.

Setelah melalui proses tersebut, hasilnya dievaluasi untuk menentukan apakah sesuai dengan maksimum iterasi yang diinginkan. Jika iya, proses algoritma selesai; jika tidak, proses diulang sampai iterasi yang ditentukan tercapai. Selanjutnya, dihitung nilai GCV, CV, dan UBR untuk setiap individu, dan individu dengan nilai terkecil menunjukkan titik knot optimal yang telah ditemukan.

3.3 Uji Program

Berikut adalah hasil terbaik yang diperoleh dengan jumlah individu awal sebanyak 10, jumlah iterasi sebanyak 10, jumlah titik knot maksimum sebesar 2, dan orde maksimum sebesar 2, diperoleh hasil berikut:

Tabel 3. 2 Hasil Minimum Program

No.	Kriteria	Nilai Sebelum AG	Nilai Sesudah AG	MSE
1.	GCV	0.799314554	0.739338964	0.681374789
2.	CV	10.58	8	0.658400937
3.	UBR	1.35×10^{-59}	9.21×10^{-64}	0.681374789

Diantara perbandingan MSE pada tabel 5.2, dapat terlihat metode CV lebih kecil diantara kedua metode lainnya. Tiga data tersebut berada pada standar deviasi sebesar 1, model ke-2, data berdistribusi normal, jumlah titik knot 1, orde fungsi *truncated spline* 1, serta jumlah data sebanyak 50 yang sama. Hal ini membuktikan bahwa setiap data yang dibangkitkan mempengaruhi besar kecilnya nilai MSE pada model *truncated spline*.

Selanjutnya diberikan grafik dari model ke-1, berdistribusi normal, jumlah data sebanyak 50, jumlah titik knot 1, dan orde fungsi *truncated spline* 2.

Berikut diberikan intensitas terjadinya nilai MSE terkecil distribusi normal dan *non-normal*.

Tabel 3. 3 Intensitas terjadinya nilai MSE terkecil distribusi normal

Model	Kriteria						
	GCV	CV	UBR	GCV & CV	GCV & UBR	CV & UBR	GCV, CV, & UBR
$y = x + \varepsilon$	32	20	6	0	2	0	0
$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} + \varepsilon$	42	16	2	0	0	0	0
Total (120 percobaan)	74	36	8	0	2	0	0

Tabel 3. 4 Intensitas terjadinya nilai MSE terkecil distribusi *non-normal*

Model	Kriteria						
	GCV	CV	UBR	GCV & CV	GCV & UBR	CV & UBR	GCV, CV, & UBR
$y = e^x + \varepsilon$	34	2	0	8	4	0	12
$y = x^2 + \varepsilon$	36	6	4	0	0	2	12
$y = \frac{1}{x} + \varepsilon$	28	0	0	20	0	0	12
Total (240 percobaan)	98	8	4	28	4	2	36

Tabel. 3.3 dan Tabel 3.4 merupakan perbandingan nilai MSE terkecil yang dibandingkan dengan model data, distribusi data, jumlah data, orde fungsi *spline truncated,* jumlah titik knot, dan standar deviasi. Terdapat juga dua atau lebih nama parameter di tabel yang berarti terdapat dua atau lebih nilai yang minimum yang sama.

Perbedaannya merupakan Tabel 3.3 berdistribusi normal sedangkan Tabel 3.4 berdistribusi *non-normal*. Pada Tabel 3.3 metode CV pada model ke-1 dan ke-2 tidak berbeda jauh dengan metode GCV. Dan Tabel 3.4 metode CV pada model ke-5 lebih unggul dibandingkan metode GCV dan UBR. Terlihat pada Tabel 3.4 setiap model pada kriteria GCV, CV, dan UBR memiliki nilai yang sama. Nilai tersebut berasal dari standar deviasi 1. Ini menunjukkan bahwa jarak standar deviasi 1 terlalu kecil sehingga semua metode konvergen lebih cepat yang menghasilkan nilai MSE yang sama.

Berikut diberikan ringkasan intensitas terjadinya nilai MSE terkecil.

Tabel 3. 5 Ringkasan intensitas terjadinya nilai MSE terkecil

Model	Kriteria						
	GCV	CV	UBR	GCV & CV	GCV & UBR	CV & UBR	GCV, CV, & UBR
$y = x + \varepsilon$	32	20	6	0	2	0	0
$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} + \varepsilon$	42	16	2	0	0	0	0
$y = e^x + \varepsilon$	34	2	0	8	4	0	12
$y = x^2 + \varepsilon$	36	6	4	0	0	2	12
$y = \frac{1}{x} + \varepsilon$	28	0	0	20	0	0	12
Total (300 percobaan)	172	44	12	28	6	2	36

Tabel. 3.5 merupakan perbandingan nilai MSE terkecil yang dibandingkan dengan model data, distribusi data, jumlah data, orde fungsi *spline truncated,* jumlah titik knot, dan standar deviasi. Dapat dilihat bahwa lebih dari 50% kombinasi keadaan yang diujikan menunjukkan bahwa, GCV menghasilkan lebih minimum dibandingkan metode lainnya.

4 Kesimpulan

Berdasarkan penelitian dilakukan untuk mendapatkan titik knot optimal regresi nonparametrik spline truncated, algoritma genetika dapat digunakan untuk meminimumkan GCV, CV, UBR. Selanjutnya diperoleh hasil pencarian dimana titik knot optimal itu akan memperoleh MSE minimum yang dipengaruhi orde fungsi spline truncated (1 dan 2) dan standar deviasi nya, sedangkan jumlah knot (1 knot dan 2 knot) tidak mempengaruhi secara signifikan. Berdasarkan hasil uji program, 50% dari hasil menunjukkan bahwa GCV memberikan nilai lebih minimum.

5 Daftar Pustaka

Andrews, D. W. 1991. Asymptotic optimality of generalized CL, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*. 47(2-3), 359-377.

Akram, A., dkk. 2021. Implementasi Algoritma Genetika dalam Menentukan Titik Knot Optimal pada Regresi *Spline Truncated*. *Skripsi. Tidak Diterbitkan.* Fakultas Matematika dan Ilmu Pengetahuan Alam. Universitas Mataram: Mataram.

Basuki, Huda dan Santoso. 2004. *Modeling dan Simulasi*. Jakarta Selatan: IPTAQ Mulia Media.

Budiaji, W. 2019. Penerapan Reproducible Research pada RStudio dengan Bahasa R dan Paket Knitr. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, *5*(1), 1-5.

Budiantara, I. N. 2005. Model Keluarga *Spline* Polinomial *Truncated* dalam Regresi Semiparametrik. *BIMIPA*, *15*(3), 55-61.

Budiantara, I. N. 2009. *Spline* Dalam Regresi Nonparametrik dan Semiparametrik: Sebuah Pemodelan Statistika Masa Kini dan Masa Mendatang, *Pidato Pengukuhan Guru Besar Pada Jurusan Statistika*. FMIPA-ITS: Surabaya.

Budiantara, I.N., Suryadi, F., Otok, B.W., dan Guritno, S., 2006. Pemodelan B*Spline* dan MARS Pada Nilai Ujian Masuk Terhadap IPK Mahasiswa Jurusan Desain Komunikasi Visual UK. PETRA Surabaya. *Jurnal Teknik Industri*. 8(1): 1-13.

Coshall, J., & Hardle, W. 1990. Applied Nonparametric Regression. *The Statistician*, *42*(2), 195. https://doi.org/10.2307/2348990

Craven, P. dan Wahba, G. 1979. "Smoothing Noisy Data with *Spline* Functions," *Numerische Mathematik*, 31, 377-403.

Devi, A. R., Budiantara, I. N., dan Ratnasari, V. 2018. Metode Unbiased Risk (UBR) dan Cross Validation (CV) Untuk Pemilihan Titik Knot Optimal dalam Regresi Nonparametrik Multivariabel Spline Truncated (Studi Kasus: Data Tingkat Pengangguran Terbuka di Provinsi Jawa Tengah Tahun 2015). *Tesis*. *Tidak Diterbitkan.* Fakultas Matematika dan Ilmu Pengetahuan Alam. Institut Teknologi Sepuluh Nopember: Surabaya.

Dorigo, M., Birattari, M., Stutzle, T. 2006. *Ant colony optimization. IEEE Computational Intelligence Magazine. 1(4):28-39.*

Draper, N. R., dan Smith, H. 1998. *Applied regression analysis*. New York : John Wiley.

Eubank, R. L. 1999. *Nonparametric regression and spline smoothing*. New York: Marcel Dekker.

Fachrurrazi, S., 2013. Penerapan Algoritma Genetika dalam Optimasi Pendistribusian Pupuk di PT Pupuk Iskandar Muda Aceh Utara. *Jurnal Fakultas Teknik Informatika*. 2(1), 47-66.

Fitriyani, N., dan Budiantara, I. N. 2014. Metode *Cross Validation* dan *Generalized Cross Valid*ation dalam Regresi Nonparametrik *Spline* (Studi Kasus Data Fertilitas di Jawa Timur). *Prosiding Seminar Nasional Pendidikan Sains. ISBN* (pp. 978-602).

Gen, M., dan Cheng, R., 1999. *Genetic algorithms and engineering optimization* (Vol. 7). New York: John Wiley and Sons.

Green, P.J dan Silverman, B. W. 1994. *Nonparametric Regression and Generalized Linear Models; A Roughness Penalty Approach*. London: Chapman and Hall.

Hidayat, R., Yuliani., dan Sam, M., 2017. Model Regresi Nonparametrik dengan Pendekatan *Spline Truncated*. *Prosiding Seminar Nasional Matematika*. 03(1). pp. 2443-1109.

Inayati. 2010. Analisa Perbandingan Metode *Roulette Wheel Selection*, *Rank Selection* dan *Tournament Selection* pada Algoritma Genetika. *Skripsi*. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas islam Negeri Sultan Syarif Kasim Riau Pekanbaru. Pekanbaru.

Kennedy, J.; Eberhart, R. 1995. *Particle Swarm Optimization*. *Proceedings of IEEE International Conference on Neural Networks*. IV. pp. 1942–1948. doi:10.1109/ICNN.1995.488968.

Mahmudy, W. F. 2013. Algoritma Evolusi. *Program Teknologi Informasi dan Ilmu Komputer, Universitas Brawijaya, Malang*. 1-101.

Montoya, E. L, Ulloa, N, dan Miller, V. 2014. *A Simulation Study Comparing Knot Selection Methods with Equally Spaced Knots in a Penalized Regression Spline*. *International Journal of Statistics and Probability*. 3(3). 96-110.

Paputungan, I. V., 2004. Perbandingan Metode-metode dalam Algoritma genetika untuk *Travelling Salesman Problem, Seminar Nasional Aplikasi Teknologi Informasi*. J-65-72.

Pincus, Martin. 1970. *A Monte-Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems*. *Journal of the Operations Research Society of America*. 18(6): 967–1235.

Russell, S. J., dan Norvig, P. 1995. *Artificial intelligence: A modern approach.* Englewood Cliffs, N.J*:* Prentice Hall.

Sari, Sulistya U. R., Budiantara, I N., dan Wibowo, W. 2016. Perbandingan Model Regresi Nonparametrik *Spline* Multivariabel dengan Menggunakan Metode *Generalized Cross Validation* (GCV) dan *Unbiased Risk* (UBR) dalam Pemilihan Titik Knot Optimal. *Tesis*. *Tidak Diterbitkan.* Fakultas Matematika dan Ilmu Pengetahuan Alam. Institut Teknologi Sepuluh Nopember: Surabaya.

Sarvina, Y. 2017. Pemanfaatan Software Open Source "R" untuk Penelitian Agroklimat. *Jurnal Informatika Pertanian*. 26(1). Juni 2017: 23 – 30.

Sivanandam, S. N. dan Deepa, S. N. 2008. *Introduction to Genetic Algorithms.* Verlag Berlin Heidelberg: Springer.

Taha, H. A. 2002. *Operations Research-An Introduction 6th ed*. Upper Saddle River NJ 07458: Prentice Hall.

Team, R. C. 2015. *R: A language and environment for statistical computing*. Vienna, Austria; 2014.

Tripena, A., Prabowo, A., Lianawati, Y., dan Bon, A. T. 2021. *Estimated spline in nonparametric regression with a generalized cross validation and unbiased risk approach*. *Proceedings of the International Conference on Industrial Engineering and Operations Management*. 3788–3798.

Vikhar, P. A. 2016. "*Evolutionary algorithms: A critical review and its future prospects*". *Proceedings of the 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*. Jalgaon: 261–265.

Wahba, G. 1990. *Spline models for observational data*. Pennsylvania: Society for industrial and applied mathematics.

Wang, Dongshu, Tan, Dapei, Liu, Lei. 2017. *Particle swarm optimization algorithm: an overview. Soft Comput* 22, 387–408.

Wang, Y., (2011), *Smoothing Splines Methods and Aplications,* CRC Press, University of California, California.

Whitley, Darrell. 1994. *A Genetic Algorithm Tutorial. Statistics and Computing*. 4 (2): 65–85.

Wu, H., & Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*. New Jersey: John Wiley & Sons.