PERBANDINGAN METODE REGRESI BIASA DENGAN GEOGRAPHICALLY WEIGHTED REGRESSION DALAM MEMODELKAN DATA COLUMBUS DI SOFTWARE R 2.6.1

Oleh : Alia Lestari
Dosen Prodi Matematika STAIN Palopo

Abstrak:
Metode regresi merupakan metode statistik yang paling umum digunakan. Metode regresi yaitu metode yang menghubungkan variabel respon dengan variabel bebas dengan hasil keluaran (output) utamanya adalah estimasi dari parameter yang membentuk suatu model tertentu. Metode regresi merupakan metode yang memodelkan hubungan antara variabel respon (y) dan variabel bebas (x_1, x_2, ... , x_p). Model regresi linier secara umum dinyatakan dengan $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$. *Tulisan ini akan melihat perbandingan metode regresi biasa dengan metode GWR dalam memodelkan data Columbus di Software R 2.6.1 tentang pengaruh income dan housing terhadap crime.*

Kata Kunci: *Perbandingan, metode regresi biasa, metode GWR model data Columbus*

I. Pendahuluan

Metode regresi merupakan metode statistik yang paling umum digunakan. Metode regresi yaitu metode yang menghubungkan variabel respon dengan variabel bebas dengan hasil keluaran (output) utamanya adalah estimasi dari parameter yang membentuk suatu model tertentu (Draper dan Smith, 1992). Masalah utama dari metode ini adalah jika metode ini diterapkan pada data spatial, dimana metode *Ordinary Least Square* (OLS) untuk estimasi parameter model regresi dengan asumsi *error* identik independen dan berdistribusi normal yang harus dipenuhi, maka akan diperoleh satu model taksiran untuk semua data. Hal inilah yang menyebabkan ketidaksesuaian model pada data spatial.

Sebagai contoh, model regresi klasik mengasumsikan bahwa lokasi geografis (berdasarkan *longitude* dan *latitude* bumi) lahan

pertanian tidak mempengaruhi respon model. Asumsi ini bisa menyebabkan kesimpulan yang salah dan menghasilkan *error* autokorelasi spatial sehingga dibutuhkan metode statistik yang bisa mengatasi fenomena variabilitas data spatial tersebut. Metode yang telah dikembangkan untuk analisis data pertanian dengan memperhitungkan faktor spatial yaitu tetangga terdekat (*Nearest-Neighbor*) (Cressie, 1991). Metode statistik lain yang dapat digunakan adalah *Geographically Weighted Regression* (GWR), yaitu metode yang menggunakan faktor geografis sebagai variabel bebas yang dapat mempengaruhi variabel respon (Fotheringham, Brundson, dan Charlton, 2002). Beberapa aplikasi GWR antara lain analisis GWR pada pendapatan per kapita rumah tangga di propinsi *Spanish* dengan menggunakan tiga macam pembobot yaitu fungsi Gaussian, fungsi *bisquare,* dan adaptif *bisquare* (Chasco, Garcia, dan Vicens, 2007). Penelitian lain adalah analisis GWR pada data peminjaman kredit di daerah California (Zhuang, 2006), dan aplikasi GWR terhadap tingkat kepuasan pelanggan dengan memperhitungkan faktor topografi, iklim, dan sumber daya alam (Mittal, Kamakura, dan Govind, 2004). Hasil dari penelitian-penelitian tersebut secara umum menunjukkan adanya pengaruh faktor geografis.

Tulisan ini akan melihat perbandingan metode regresi biasa dengan metode GWR dalam memodelkan data Columbus di Software R 2.6.1 tentang pengaruh *income* dan *housing* terhadap *crime*.

II. Pembahasan

a. Metode Regresi

Metode regresi merupakan metode yang memodelkan hubungan antara variabel respon (y) dan variabel bebas ($x_1, x_2, \ldots, x_p$). Model regresi linier secara umum dinyatakan dengan

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

Jika diambil sebanyak n pengamatan, maka model di atas dapat ditulis sebagai:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i \quad (1)$$

dengan $i = 1, 2, \ldots, n$; $\beta_0, \beta_1, \ldots, \beta_p$ adalah parameter model dan $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ adalah error yang diasumsikan identik, independen, dan berdistribusi Normal dengan mean nol dan varians konstan σ^2. Pada model ini, hubungan antara variabel bebas dan variabel respon dianggap konstan pada setiap lokasi geografis. Estimator dari parameter model didapat dari persamaan

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \quad (2)$$

dengan

$\boldsymbol{\beta}$: vektor dari parameter yang ditaksir berukuran n x (p+1)

$\mathbf{X}$: matrik data berukuran n x (p+1) dari variabel bebas yang elemen pada kolom pertama bernilai 1

$\mathbf{y}$: vektor observasi dari variabel respon berukuran (n x 1)

k : banyaknya variabel bebas (k = 1, 2, ... , p)

Pengujian signifikansi model secara serentak pada regresi adalah dengan analisis varians yang tertuang dalam tabel ANOVA:

Tabel 1. 1. ANOVA Model Regresi

Sumber variasi	*Sum Square*	d.b	*Mean Square*	F_{hit}
Regresi	$\hat{\boldsymbol{\beta}}^T\mathbf{X}^\mathbf{T}\mathbf{y}$	*p-1*	SSR/(*p*-1)	MSR/MSE
Error	$\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^\mathbf{T}\mathbf{y}$	*n-p*	SSE/(*n-p*)	
Total	$\mathbf{y}^T\mathbf{y}$	*n*-1		

Akan diperoleh kriteria keputusan bahwa model regresi sesuai untuk data yang digunakan adalah jika $F_{\text{hit}} > F_{\alpha;(p-1);(n-p)}$.

b. Metode GWR

Metode GWR adalah suatu teknik yang membawa kerangka dari model regresi sederhana menjadi model regresi yang terboboti (Fotheringham, et al., 2002)

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i , \qquad (3)$$

dengan

y_i : pengamatan pada lokasi ke-i (i = 1, 2, ... , n)

(u_i, v_i) : koordinat *longitude latitude* dari titik ke-i pada suatu lokasi geografis.

$\beta_k(u_i, v_i)$: realisasi dari fungsi kontinyu $\beta_k(u, v)$ pada titik ke-i

ε_i : *error* yang diasumsikan identik, independen, dan berdistribusi Normal dengan mean nol dan varians konstan σ^2

Dengan demikian setiap nilai parameter dihitung pada setiap titik lokasi georafis. Jadi setiap titik lokasi geografis mempunyai nilai parameter regresi yang berbeda-beda. Hal ini menghasilkaan variasi pada nilai parameter regresi di suatu kumpulan wilayah geografis. Jika nilai parameter regresi konstan pada tiap-tiap wilayah geografis, maka model GWR adalah model global. Artinya tiap-tiap wilayah geografis mempunyai model yang sama. Hal ini merupakan kasus spesial dari GWR.

c. Estimasi Parameter Model GWR

Pada model GWR diasumsikan bahwa data observasi yang dekat dengan titik ke-i mempunyai pengaruh yang besar pada estimasi dari $\boldsymbol{\beta}_k(u_i, v_i)$ daripada data yang berada jauh dari titik ke-i. Esensi yang bisa diambil dari hal tersebut adalah persamaan diatas mengukur hubungan model pada semua titik ke-i. Lokal parameter $\boldsymbol{\beta}_k(u_i, v_i)$ diestimasi menggunakan WLS (Leung, 2000). Pada GWR sebuah observasi diboboti dengan nilai yang berhubungan dengan titik ke-i. Bobot w_{ij}, untuk j = 1, 2, ... , n,

pada tiap lokasi (u_i, v_i) diperoleh sebagai fungsi yang kontinyu dari jarak antara titik ke-*i* dan titik data lainnya. Misal matriks berikut merupakan matriks dari lokal parameter

$$\mathbf{B} = \begin{bmatrix} \beta_0(u_1, v_1) & \beta_1(u_1, v_1) & \cdots & \beta_p(u_1, v_1) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0(u_{1n}, v_n) & \beta_1(u_{1n}, v_n) & \cdots & \beta_p(u_{1n}, v_n) \end{bmatrix} \quad (4)$$

Estimasi tiap baris adalah dengan persamaan berikut:

$$\hat{\boldsymbol{\beta}}(i) = \left(\mathbf{X}^T \mathbf{W}(i) \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}(i) \mathbf{y} \quad (5)$$

dengan

X = matrik data dari variabel bebas

Y = vektor variabel respon

W(*i*) = matriks pembobot

$$= diag\,[w_{i1}, w_{i2}, \ldots, w_{in}] \quad (6)$$

Estimasi dari (5) merupakan estimasi dari *least square* tetapi matriks pembobot tidak konstan, sehingga **W**(*i*) dihitung untuk tiap *i* dan w_{ij} mengindikasikan kedekatan atau bobot tiap titik data dengan lokasi *i*. Hal ini yang membedakan GWR dengan tradisional WLS yang mempunyai matrik bobot yang konstan.

Selain menghasilkan estimasi parameter lokal untuk tiap-tiap lokasi geografis, GWR juga menghasilkan versi lokal untuk seluruh standar regression pada seluruh lokasi geografis misalnya ukuran *goodness of fit*. Hal ini dapat memberikan informasi pada pemahaman aplikasi dari model dan untuk penelitian lebih lanjut apakah diperlukan penambahan variabel independen pada model GWR. Hal yang penting lainnya adalah titik dimana parameter lokal diestimasi dengan model GWR tidak memerlukan titik dimana data diambil. Estimasi dari parameter dapat didapat dari semua lokasi geografis. Dengan demikian, pada sistem dengan data titik lokasi geografis yang besar, estimasi model GWR dari lokal parameter.

d. Pembobotan Model GWR

Peran pembobot pada model GWR sangat penting karena nilai pembobot ini mewakili letak data observasi satu dengan lainnya. Oleh karena itu, sangat dibutuhkan ketepatan cara pembobotan (Chasco, et al., 2007). Skema pembobotan pada GWR dapat menggunakan beberapa metode yang berbeda, metode pembobotan yang biasa digunakan adalah kernel Gaussian (Bocci, et al., 2006).

Pembobot dengan Fungsi Gaussian adalah:

$$w_{ij} = \exp\left[-1/2\left(d_{ij}/b\right)^2\right]$$

dengan *b* adalah *bandwidth* atau jarak terdekat antara daerah ke-*i* dengan beberapa daerah tetangga terdekat dan d_{ij} merupakan jarak Euclidean $d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$.

Kriteria untuk penentuan nilai *M* yang tepat dapat diperoleh dengan pendekatan *least square* yaitu dengan menggunakan kriteria *cross-validation*

$$CV = \sum_{i=1}^{n}\left[y_i - \hat{y}_{i*}(b)\right]^2 \qquad (7)$$

dengan i* ≠ i dan $\hat{y}_{i*}(b)$ adalah nilai dugaan untuk y_i dengan pengamatan pada titik ke-*i* diabaikan dalam proses kalibrasinya.

e. Aplikasi Model pada Data Columbus

Pada tulisan ini aplikasi model GWR diaplikasikan pada data columbus yang disajikan pada software R.2.6.1 dengan variabel respon (*crime*) dan variabel bebas (*income* dan *housing*) dengan pembobot *gaussian.*

Tabel 4.1 Model Regresi Data Columbus

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	68,6189	4,7355	14,49	**< 2e-16**
income	-1,5973	0,3341	-4,78	**1,83E-05**
housing	-0,2739	0,1032	-2,654	**0,0109**

Tabel di atas menunjukkan bahwa semua variabel bebas signifikan terhadap variabel respon, dengan koefisien regresi variabel bebasnya bernilai negatif. Hal ini menunjukkan bahwa dengan metode regresi, income dan housing secara bersama-sama memberikan pengaruh yang negatif terhadap crime, berarti semakin tinggi variabel income dan housing, maka variabel crime akan semakin rendah. Selanjutnya akan dilakukan pemodelan GWR dengan pembobot Gauss. Hasil perhitungan diperoleh *bandwidth* global 3,217404 dengan kriteria CV minimum 11,12129 sehingga diperoleh estimasi model GWR sebagai berikut:

Tabel 4.2 Estimasi Model GWR Columbus-Gauss

Area	(Intercept)	income	housing	R2
1	50,7128	-0,4673	-0,4248	0,7872
2	63,8390	-0,3768	-0,6839	0,5947
3	72,9242	-0,2490	-0,8052	0,6324
4	80,9009	0,1011	-1,0528	0,8182
5	46,4738	-0,6860	-0,2208	0,9566
6	57,9632	-0,9844	-0,3173	0,6189
7	76,3928	-1,7949	-0,3767	0,7477
8	76,5605	-3,0268	0,0292	0,7307
9	70,8777	-1,5533	-0,1664	0,8252
10	54,2155	-1,7241	0,0320	0,9225
11	54,9113	-1,7619	0,0241	0,9587
12	54,1252	-1,6680	0,0056	0,9597
13	44,3872	-0,9391	-0,0974	0,9352
14	36,9275	-1,1777	0,2056	0,9542
15	39,8504	-2,4522	0,7946	0,9772
16	47,8158	-0,7289	-0,2153	0,9103
17	63,9024	-1,8498	-0,0658	0,9182
18	65,9502	-2,0886	-0,0153	0,6895
19	67,4941	-2,2575	0,0301	0,7351
20	69,3342	-1,6041	-0,1531	0,7223

21	68,6048	-2,6083	0,1601	0,8013
22	68,6806	-2,8321	0,2638	0,8089
23	68,7564	-3,0682	0,3729	0,7531
24	74,9821	-2,7487	0,1190	0,5901
25	66,2270	-3,1236	0,4414	0,6018
26	58,1870	-1,4991	-0,1935	0,5727
27	62,4889	-2,6581	0,2883	0,3626
28	58,4498	-2,7067	0,4926	0,1936
29	72,2619	-3,1307	0,2949	0,4607
30	73,0092	-1,2261	-0,2865	0,5987
31	71,8269	-1,9129	-0,0763	0,7054
32	64,4094	-0,3459	-0,3171	0,8020
33	70,3467	-0,9116	-0,3042	0,7340
34	70,4191	-0,4317	-0,4370	0,7569
35	64,8878	-0,0767	-0,3867	0,8127
36	65,0710	-0,2745	-0,3292	0,8649
37	64,5483	0,2016	-0,4968	0,8105
38	67,9022	-0,4444	-0,4184	0,8158
39	61,6201	1,1410	-0,6902	0,7694
40	59,0557	1,2911	-0,6317	0,8005
41	61,1625	0,6969	-0,5050	0,8275
42	64,1905	0,0443	-0,3943	0,8479
43	61,9397	-1,6154	0,0377	0,4351
44	39,2119	-0,7206	-0,1580	0,9172
45	30,6056	-0,5762	-0,0622	0,9671
46	25,7990	-0,3686	-0,0368	0,9775
47	23,9919	-0,2563	-0,0346	0,9913
48	23,2333	-0,2298	-0,0317	0,9908
49	29,7968	-0,4194	-0,0812	0,9704

Terlihat bahwa variabel *income* dan *housing* mayoritas memiliki koefisien negatif. Hal ini menunjukkan bahwa makin tinggi tingkat *income* dan *housing* maka tingkat *crime* pada masyarakat tersebut akan semakin rendah.

Pengaruh letak goegrafis atau model GWR kemudian akan diuji dengan ANOVA berikut :

Tabel 4.3 ANOVA Fotheringham Columbus-Gauss

	SSE	**d.f**	**F**	**P-value**
Model Regresi	6014,856	46,000	4,8155	0,0002152
Model GWR	1249,066	19,384		

Pada tabel ANOVA di atas terlihat bahwa P_value < 0,05 yang berarti ada pengaruh letak geografis atau model GWR pada data columbus dengan pembobot Gauss, berarti model GWR sesuai untuk digunakan pada data ini. Selanjutnya, setelah mengetahui adanya kesesuaian model GWR terhadap data dilakukan analisis untuk menguji perbedaan antara satu daerah dengan daerah lain atas suatu variabel bebas.

Tabel 4.4 Perbedaan Antar Area Pada Suatu Variabel

Variabel	**F statistic**	**Numerator d.f**	**Denumerator d.f**	**P-value**
Intercept	9,1713e-04	1,6818+01	34,97	1
Income	3,1801e-02	1,5295+02	34,97	1
Housing	2,0278e+02	6,3133+00	34,97	<2e-16

Tabel 4.4 menunjukkan bahwa pada data columbus, *housing* antara satu area dengan area lain berbeda secara signifikan.

III. Penutup

Analisis Regresi biasa dan Analisis GWR memberikan hasil yang sama pada data ini, dimana kedua variabel bebas memberikan pengaruh yang negatif terhadap variabel terikat. Namun analisis GWR dapat menjelaskan adanya perbedaan pengaruh *housing* antara satu area dengan area lain terhadap *crime*.

Daftar Pustaka

Bocci, C., Petrucci, A., dan Rocco, E. (2006), "An Application of Geographically Weighted Regression to Agricultural Data for Small Area Estimates", *Dipartimento di Statistica "G. Parenti"*, Universita degli Studi di Firenze, Italy.

Charlton, M., Fotheringham, S., dan Brunsdon, C. (2006), "Geographically Weighted Regression", *Document ESRC National Centre for Research Methods*, NCRM Methods Review Papers.

Chasco, C., Garcia, I., dan Vicens, J. (2007), "Modeling spatial variations in household disposable income with Geographically Weighted Regression", *Munich Personal RePEc Archive Paper No. 1682.*

Cressie, N.A.C. (1991), *Statistics For Spatial Data,* John Wiley & Sons, Inc. United States of America.

Draper, N.R. dan Smith, H. (1992), *Analisis Regresi Terapan,* Edisi Kedua, PT Gramedia Pustaka Utama, Jakarta.

Fotheringham, A.S., Brundson, C., dan Charlton, M. (2002) "*Geographically Weighted Regression: the analysis of spatially varying relationships*", John Wiley & Sons Ltd, England.

Leung, Y. (2000), "Statistical Tests for Spatial Non-Stationarity Based on the Geographically Weighted Regression Model", *Journal*, The Chinese University of Hong Kong, Hong Kong.

Mittal, V., Kamakura, W.A., dan Govind, R. (2004), "Geographic Patterns in Customer Service and Satisfaction: An Empirical Investigation", *Journal of Marketing,* Vol. 68, Hal. 48-62.

Yu, D. (2004), "Modeling Housing Market Modeling Housing Market Dynamics In The City Of Milwaukee: A Geographically Weighted Regression Approach", Department of Geography University of Wisconsin-Milwaukee, Wisconsin.

Zhang, L. dan Gove, J.H. (2005), "Spatial Assessment of Model Errors from Four Regression Techniques", *Forest Science,*Vol. 51, No. 4, hal. 334-346.

Zhuang, D. (2006), "Spatial Dependence and Neighborhood Effects in Mortgage Lending: A Geographically Weighted Regression Approach", *Paper,*University of Southern California, Los Angeles.