M-9

OPTIMALISASI MATRIK BOBOT SPASIAL BERDASARKAN K-NEAREST NEIGHBOR DALAM SPASIAL LAG MODEL

I Gede Nyoman Mindra Jaya[1)], Bertho Tantular[2)], Zulhanif[3)]
[1,2,3)]Departemen Statistika FMIPA UNPAD
jay.komang@gmail.com, berthotantular@gmail.com, dzulhanif@yahoo.com

Abstrak

*Permasalahan dalam analisis sapatial ekonometrik yang berkaitan dengan spasial lag dependensi yaitu belum ditemukan solusi tepat dalam menentukan struktur dependensi pada data spasial. Struktur dependensi ini umumnya dinyatakan dalam matrik bobot spasial (**W**). Secara teori matrik **W** adalah fixed ditentukan berdasarkan hipotesis peneliti yang dikembangkan dari dari pemahaman terhadap fenomena yang diamati. Namun demikian, seringkali peneliti tidak memiliki informasi yang cukup untuk membangun struktur dependensi ini. Keterbatasan informasi menyebakan peneliti merujuk pada hokum Tobler yang menjelaskan bahwa semua hal saling terkait satu dengan yang lainnya namun yang lebih berdekatan lebih erat kaitannya dibandingkan yang berjauhan. Hukum ini diterjemahkan dalam berbagai cara seperti menggunkan kedekatan persinggungan antara lokasi ataupun menggunakan inverse jarak. Namun, faktanya, metode ini tidak mampu memberikan struktur **W** yang optimal menurut ukuran kebaikan model R^2 dan AIC. Penelitian ini mengusulkan satu pendekatan baru melalui metode iterasi untuk menemukan matrix **W** yang paling optimal. Metode yang digunakan adalah metode K-Nearest Neighbor (K-NN). Hasil analisis pada kasus Diare di Kota Bandung Tahun 2015 menemukan bahwa penggunaan metode K-nearest neighbor dalam memilih matrik bobot spasial yang paling optimum memberikan hasil akhir yang baik. Model dengan matrik bobot berdsarkan 3-NN memberikan nilai AIC yang paling kecil dan R^2 yang paling besar. Temuan lain dari penelitian ini adalah nilai koefisien spasial lag yang semakin tinggi tidak menjamin bahwa model spasial lag dengan ukuran AIC dan R^2 semakin baik. Dua variabel yang memiliki kontribusi terhadap angka prevalensi diare sesuai dengan fenomenanya adalah Perilaku Hidup Bersih dan Sehat (PHBS) dan Air Bersih.*

Kata Kunci: *K-NN, Optimasi, Spasial Lag.*

1. PENDAHULUAN

Analisis data spatial sangat berkembangan beberapa periode waktu terakhir. Pemanfaatan data spatial pemungkinkan peneliti untuk mempelajari karakteristik spatial secara lebih mendalam dan menggali berbagai informasi yang selama ini tidak tergali dari pendekatan non data spatial. Informasi yang dapat digali diantaranya dalah efek ketergantungan spatial dan heterogenitas spatial (Anselin, 1988 dan Jaya, dkk 2016).

Salah satu bidang ilmu yang sangat berkembang dalam kaitannya dengan data spatial adalah spatial ekonometrika. Hal ini dikarenakan adanya perkembagan riset yang sangat pesat dalam studi regional. Namun demikian, sampai saat ini masih banyak permasalahan dalam studi spatial ekonometrika yang berkaitan dengan pemodelan spasial lag dependensi yaitu belum ditemukan solusi tepat dalam menentukan struktur dependensi pada data spasial. Struktur dependensi ini umumnya dinyatakan dalam matrik bobot

spasial (**W**) (LeSage, 1999). Secara teori matrik **W** adalah fixed ditentukan berdasarkan hipotesis peneliti yang dikembangkan dari dari pemahaman terhadap fenomena yang diamati. Namun demikian, seringkali peneliti tidak memiliki informasi yang cukup untuk membangun struktur dependensi ini. Keterbatasan informasi menyebakan peneliti merujuk pada hokum Tobler yang menjelaskan bahwa semua hal saling terkait satu dengan yang lainnya namun yang lebih berdekatan lebih erat kaitannya dibandingkan yang berjauhan. Hukum ini diterjemahkan dalam berbagai cara seperti menggunkan kedekatan persinggungan antara lokasi ataupun menggunakan inverse jarak. Namun, faktanya, metode ini tidak mampu memberikan struktur **W** yang optimal menurut ukuran kebaikan model R^2 dan AIC (Perret, 2011). Penelitian ini mengusulkan satu pendekatan baru melalui metode iterasi untuk menemukan matrix **W** yang paling optimal. Metode yang digunakan adalah metode K-Nearest Neighbor (K-NN).

2. METODE PENELITIAN

a. Data

Data yang digunakan dalam penelitian ini adalah data diare di Kota Bandung tahun 2015 yang diperoleh dari Dinas Kesehatan Kota Bandung. Variabel yang diamati meliputi:

Tabel 1. Variabel Penelitian

No.	Variabel	Satuan
1.	Angka Kasus Diare	Orang
2.	Perilaku Hidup Bersih dan Sehat (PHBS)	Persentase (%)
3.	Air Bersih	Persentase (%)

b. METODE

1) Spatial Lag Depdendent

Studi spasial ekonometrika beberapa periode waktu ini sangat berkembang (LeSage, 2009). Peneliti regional memanfaatkan medel ini untuk dapat menjelaskan berbagai faktor ekonomi dan regional yang menjelaskan fenomena yang sedang ditelit. Model spasial ekonomerika merupakan sub dari model ekonometrika yang mengakomodasi adanya ketergantungan spasial dalam data. Penerapan pada model ekonometrika standar pada kasus spasial menyebabkan taksiran parameter model menjadi bias dan tidak efisien dan juga tidak konsisten (Klotz, 2004).

Model Eknometrika yang paling sering digunakan dalam penelitian adalah model spatial lag dependen (SpLag). Model SpLag dapat dituliskan dalam bentuk sebagai berikut: (Jaya dkk, 2016)

$$y_i = \rho \sum_{j=1}^{n} w_{ij} y_j + \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i, \qquad (1)$$

dengan y_i menyatakan variabel response dimana dalam penelitian ini adalah angka prevalensi diare, ρ menyatakan koefisien spasial autoregressive. Koefisien spasial lag dependen ρ menyatakan besar pengaruh dari rata-rata angka prevalensi lokasi tetangga teradap angka prevalensi lokasi yang diamati. Parmaeter model β_0 dan β_j menyatakan koefisien intersept dan slop regresi untuk variabel eskogenus ke-*k*, x_{ik} menyatakan nilai variabel eksogenus ke-*k* pada lokasi ke-*i*. Penelitian ini menggunakan empat variabel bebas yaitu PHBS dan Air Bersih. Variabel ε_i menyatakan kekeliruan acak dengan asumsi identik independen berdistribusi normal dengan rata-rata nol dan varians σ^2 (ε_i~i.i.dN(0,σ^2). Komponen w_{ij} adalah elemen dari matriks bobot spasial yang umumnya dapat ditentukan berdasarkan persinggungan lokasi ataupun jarak antar lokasi dan melalui metode optimasi. Penelitian ini menggunakan matriks bobot spasial berdasarkan metode optimasi melaluk K-NN.

2) Estimasi Parameter Model SpLag

Estimasi ML dari model SpLag melibatkan memaksimalkan fungsi kemungkinan log sehubungan dengan $\boldsymbol{\beta}$, ρ dan σ^2. Estimasi ML dari model SpLag memiliki sifat asimtotik (konsistensi, efisiensi dan normalitas asymptotic).

Fungsi Kemungkinan:

$$\ln L(\boldsymbol{\beta}, \rho, \sigma_\varepsilon^2 | \mathbf{y}, \mathbf{X}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln \sigma_\varepsilon^2 + \ln |\mathbf{I}_n - \rho \mathbf{W}_\rho| - \frac{1}{2\sigma_\varepsilon^2}\boldsymbol{\varepsilon}^{\mathbf{T}}\boldsymbol{\varepsilon} \quad (2)$$

Dengan mendefinisikan $\boldsymbol{\varepsilon}$ = $(\mathbf{I}_n$-$\rho\mathbf{W})\mathbf{y}$ - $\mathbf{X}\boldsymbol{\beta}$ maka persamaan (2) dapat ditulis

$$\ln L(\boldsymbol{\beta}, \rho, \sigma_\varepsilon^2 | \mathbf{y}, \mathbf{X}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln \sigma_\varepsilon^2 + \ln \left|\mathbf{I}_n - \rho \mathbf{W}_\rho\right| - \frac{1}{2\sigma_\varepsilon^2}\left((\mathbf{I}_n - \rho \mathbf{W}_\rho)\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\mathbf{T}}\left((\mathbf{I}_n - \rho \mathbf{W}_\rho)\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \quad (3)$$

Taksiran parameter $\boldsymbol{\beta}$ diperoleh dengan memaksimumkan persamaan (3) dan diperoleh:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \underbrace{(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{y}}_{\equiv \widehat{\boldsymbol{\beta}}_{\mathrm{O}}} - \rho \underbrace{(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{W}_\rho\mathbf{y}}_{\equiv \widehat{\boldsymbol{\beta}}_{\mathrm{L}}} = \widehat{\boldsymbol{\beta}}_{\mathrm{O}} - \rho\widehat{\boldsymbol{\beta}}_{\mathrm{L}} \quad (4)$$

Dengan parameter ρ ditaksir melalui pendekatatan numerik (Elhorst, 2012).

3) Permasalahan Pada Struktur Dependensi (W)

Struktur dependensi dapat dinyatakan dalam matrik bobot spasial (**W**). Secara teori matrik **W** adalah fixed ditentukan berdasarkan hipotesis peneliti yang dikembangkan dari dari pemahaman terhadap fenomena yang diamati. Namun demikian, seringkali peneliti tidak memiliki informasi yang cukup untuk membangun struktur dependensi ini. Keterbatasan informasi menyebakan peneliti merujuk pada hokum Tobler yang menjelaskan bahwa semua hal saling terkait satu dengan yang lainnya namun yang lebih

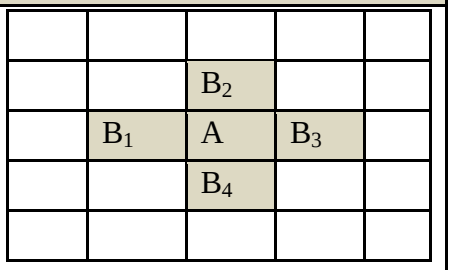
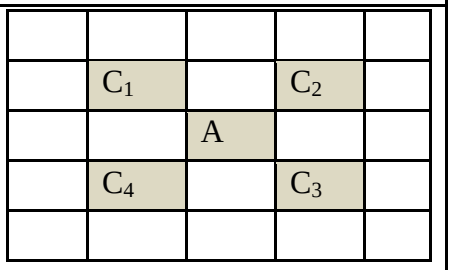
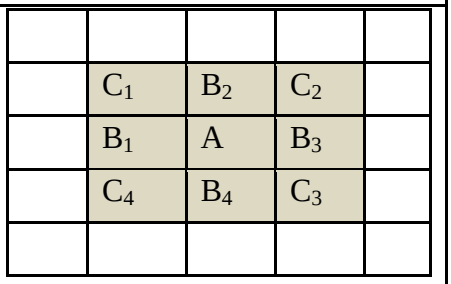
berdekatan lebih erat kaitannya dibandingkan yang berjauhan. Hukum ini diterjemahkan dalam berbagai cara seperti menggunkan kedekatan persinggungan antara lokasi ataupun menggunakan inverse jarak.

a) Bobot Spatial berdasarkan persingungan lokasi

Lokasi dikatakan saling berdekatan jika lokasi memiliki persinggungan dengan lokasi yang lain yang dapat didefinisikan sebagai berikut:

$$w_{ij} = \begin{cases} 1 \text{ jika } i \text{ dan } j \text{ bersinggungan} \\ 0 \text{ lainnya} \end{cases}$$

Terdapat tiga jenis persinggungan yang umumnya dijadikan dasar dalam penentuan matrik bobot spatial yaitu:

Tipe	Bentuk
Rook contiguity Sebuah unit spasial adalah tetangga dari unit lain jika kedua daerah berbagi tepi . Unit B1, B2, B3 dan B4 adalah tetangga unit A	
Bishop contiguity Sebuah unit spasial adalah tetangga dari unit lain jika kedua daerah berbagi sudut. Unit C1, C2, C3 dan C4 adalah tetangga unit A	
Queen contiguity: Sebuah unit spasial adalah tetangga dari unit lain jika kedua daerah berbagi sudut atau tepi . Unit B1, B2, B3, B4, C1, C2, C3, dan C4 adalah tetangga unit A	

b) Bobot Spatial berdasarkan jarak

Matrik jarak yang umumnya digunakan adalah inverse distance sebagai berikut:

$$w_{ij} = \left(\frac{1}{d_{ij}}\right)^2 \quad (5)$$

dengan d_{ij} menyatakan jarak Euclidian dari lokasi *i* ke lokaji *j*.

Namun, faktanya, metode ini tidak mampu memberikan struktur **W** yang optimal menurut ukuran kebaikan model R^2 dan AIC

c) Optimalisasi Matriks W

Optimalisasi matriks W dalam penelitian ini menggunakan pendekatan K-NN dengan tujuan menentukan bayak tetangga yang paling optimal dengan fungsi tujuannya adalah memaksimumkan nilai morans I , R^2 dan meminimumkan AIC.

K-NN dilakukan dengan tahapan:

1. Menghitung jarak Euclidiean lokasi *i* ke *j*
2. Mengurutkan jarak yang diperoleh
3. Memilih k lokasi dengan jarak terdekat sebagai nilai optimum.

Penentuan nilai nilai *k* pertama kali di dasarkan pada statistik moran I. Prosesnya dilakukan secara iterasi. Nilai k terpilih berdasarkan nilai moran I tersbesar. Selanjutnya nilai ini digunakan untuk menentukan matrik jarak yang optimum dalam pemodelan spatial Lag dependent.

Moran's *I*

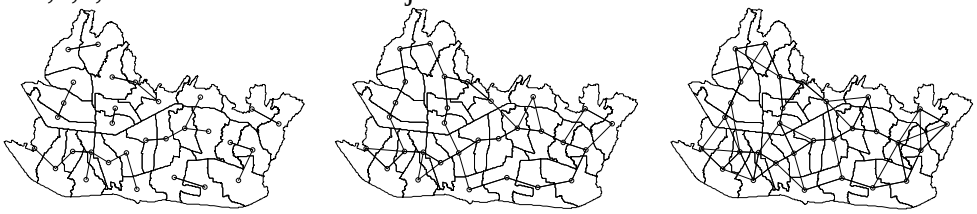
Statistik morans's dapat dihitung dengan formulasi sebagai berikut:

$$\mathbf{I} = \frac{\mathbf{e^T W e}}{\mathbf{e^T e}} \quad (6)$$

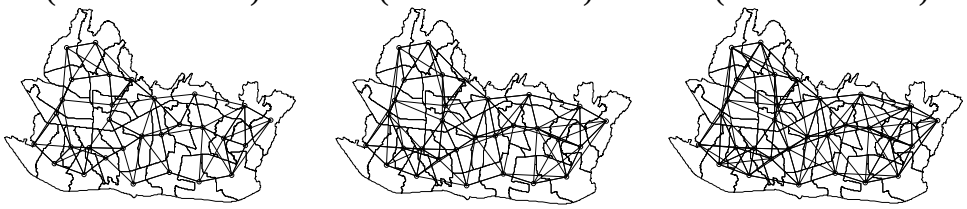
Dengan **W** matrik bobot spasial dengan **e** adalah residual yang diperole dari regresi OLS.

3. HASIL PENELITIAN DAN PEMBAHASAN

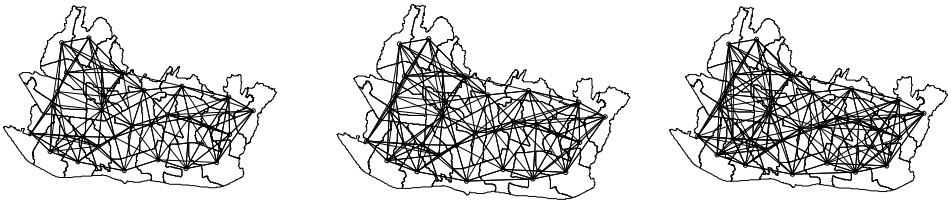
Tahap pertama dalam pemodelan adalah menentukan nilai K yang paling optimum dengan menghitung statistic moran I. Visualisasi K-NN dengan nilai k=1,2,..,9 dan nilai moran I disajikan dalam Gambar 1 berikut:



(a) 1-NN (Moran.I=0.479) (b) 2-NN (Moran.I=0.396) (c) 3-NN (Moran.I=0.417)

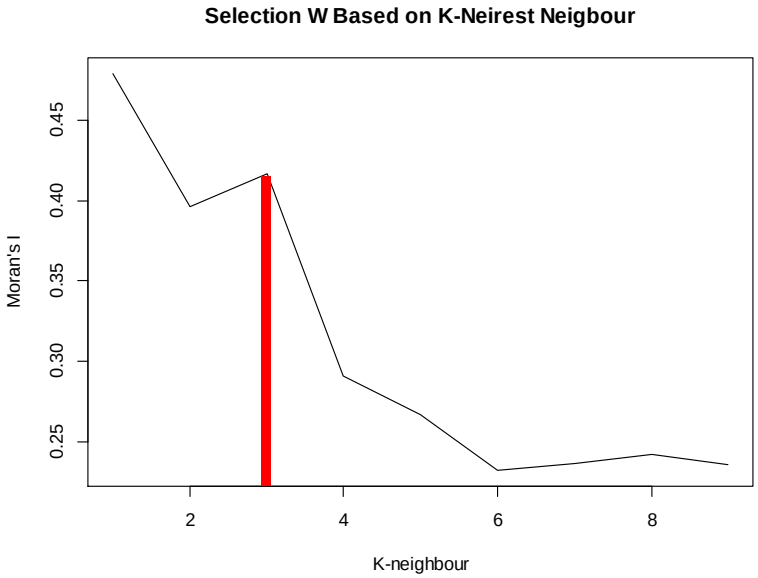


(d) 4-NN (Moran.I =0.291) (e) 5-NN (Moran.I =0.267) (f) 6-NN (Moran.I =0.232)



(g) 7-NN (Moran.I=0.237) (h) 8-NN (Moran.I =0.242) (i) 9-NN (Moran.I=0.236)

Gambar 1. Visuaslisasi K-NN dan Moran.I



Gambar 2. Nilai Optimum K

Berdasarkan hasil plot antara k-neigbour dengan Moran's I dipilih nilai k yang paling optimum adalah K=3. Selanjutnya dilakukan pengecekan pada model SpLag dengan hasil sebagai berikut:

Tabel 2. Hasil Perhitungan Model SpLag

K-NN	Rho	R^2	AIC
1	0.302	0.489	-154.20
2	0.389	0.504	-154.39
3	0.483	0.554	-156.29
4	0.412	0.444	-153.47
5	0.412	0.427	-153.26
6	0.409	0.405	-152.75
7	0.440	0.420	-153.14
8	0.485	0.440	-153.62
9	0.522	0.449	-153.87

Berdasarkan hasil perihitungan ditemukan bahwa sesuai dengan identifikasi Moran's I diperoleh k yang paling optimum adalah 3 dengan nilai R^2 paling besar

Tabel 3. Hasil Pemodelan SpLag untuk K=3

```
 Call:lagsarlm(formula = y ~ X2 + X4, data = Dataku, listw = lw)

Residuals:
        Min          1Q      Median          3Q         Max
-0.03654413 -0.00586146  0.00084374  0.00542912  0.03039668

Type: lag
Coefficients: (asymptotic standard errors)
               Estimate  Std. Error z value Pr(>|z|)
(Intercept)  6.7116e-02  4.8756e-02  1.3766   0.1686
X2          -8.1627e-05  1.5472e-04 -0.5276   0.5978
X4          -2.9571e-04  4.6712e-04 -0.6331   0.5267

Rho: 0.48254, LR test value: 6.49, p-value: 0.010848
Asymptotic standard error: 0.17022
    z-value: 2.8347, p-value: 0.0045864
Wald statistic: 8.0357, p-value: 0.0045864

Log likelihood: 83.14667 for lag model
ML residual variance (sigma squared): 0.00021292, (sigma:
0.014592)
Number of observations: 30
Number of parameters estimated: 5
AIC: -156.29, (AIC for lm: -151.8)
LM test for residual autocorrelation
test value: 0.12668, p-value: 0.7219
```

Hasil pemodelan menemukan bahw variabel PHBS dan Air bersih memberikan kontribusi sesuai dengan fenomena bahwa semakin tinggi PHBS dan Air Bersih maka akan dapat menurunkan tingkat prevalensi diare. Tingkat prevelensi diare akan turun sebesar 8.1627e-05 untuk peningkatan 1% PHBS dan menurun sebesar 2.9571e-04 untuk peningkatan 1% air bersih.

Hasil pengujian spatial lag dependen (ρ) menunjukkan hasil yang signfikan dengan nilai p.value < 0.05

4. SIMPULAN

Hasil analisis menemukan bahwa penggunaan metode K-nearest neighbor dalam memilih matrik bobot spasial yang paling optimum memberikan hasil akhir yang baik. Model dengan matrik bobot berdsarkan 3-NN memberikan nilai AIC yang paling kecil dan R^2 yang paling besar. Temuan lain dari penelitian ini adalah nilai koefisien spasial lag yang semakin tinggi tidak menjamin bahwa model spasial lag dengan ukuran AIC dan R^2 semakin baik. Dua variabel yang memiliki kontribusi terhadap angka prevalensi diare sesuai dengan fenomenanya adalah Perilaku Hidup Bersih dan Sehat (PHBS) dan Air Bersih.

5. DAFTAR PUSTAKA

Anselin, L. (1988) . *Spasial Econometrics : Methods and Models*. London: Kluwer Academic Publisher.

Ehlhorst, P. (2014), SpatialEkonometrik-From Cross-Sectional Data to Spatial Panels, Springer, Heidelberg, New York

Jaya, Mindra I. G. et al. (2016). " Bayesian Spatial Autoregressive (BSAR) Dalam Menaksir Angka Prevalensi Demam Berdarah (DB) Di Kota Bandung. Prosiding Seminar Nasional Matematika Universitas Parahyangan Bandung.

Klotz, S. (2004). Cross Sectional Dependence in Spatial Econometrics Models with an Application to German Start Up Activity Data. USA: Transaction Publisher

Lesage, J.P. 1998. *Spasial Econometrics*. Department of Economics, University of Toledo.

Perret, Jens K (2011). A Proposal for an Alternative Spatial Weight Matrix under Consideration of the Distribution of Economic Activity. Bergische Universit¨at Wuppertal SCHUMPETER DISCUSSION PAPERS. ISSN 1867-535