Journal of the Korean Data & Information Science Society
2017, **28**(3), 505–513

http://dx.doi.org/10.7465/jkdi.2017.28.3.505
한국데이터정보과학회지

공간이웃정보를 고려한 공간회귀분석

김수정[1]

[1]한국한의학연구원

접수 2017년 4월 12일, 수정 2017년 5월 17일, 게재확정 2017년 5월 18일

요 약

최근, 더욱 상세하고 정확한 추정 결과를 위해 소지역추정 (small area estimation; SAE)의 연구가 많이 진행되고 있다. 그 중 공간회귀모형 (spatial regression model)을 이용한 방법이 주를 이루고 있는데 이를 사용하기 위해서는 공간이웃 (spatial neighbor)의 정의가 필요하다. 본 연구에서는 공간이웃을 정의하는 방법으로 도로네 삼각망 (Delaunay triangulation; DT)을 소개하고 $k-$최근접 ($k-$nearest neighbor; KNN)과 비교하여 분석한다. 두 가지 공간이웃을 정의하는 방법중에서 어떤 방법으로 이웃을 정의하는 것이 효율적인지 알아보기 위해 시뮬레이션을 실시하였고, 지가 (land price)데이터를 이용하여 실 데이터를 분석하였다.

주요용어: 공간이웃, 도로네 삼각망, 소지역 추정.

1. 서론

최근 공간통계학을 접목한 읍, 면, 동 단위의 소지역 추정의 연구가 많이 진행되고 있다. 특히, 소지역 추정은 미국, 영국, 호주 등 통계선진국가를 중심으로 다양한 분야에서 현실적인 비용과 시간 등의 차원에서 널리 사용되고 있다.

특히, 전국 단위나 광역시 · 도 단위와 같은 큰 행정단위로 분석하는 경우 소지역 단위의 값을 놓치거나 왜곡 할 수 있다 (Panczak, 2016). 소지역 추정과 관련된 연구로 Lee (2000)는 경제활동자료 (economic activity data)를 이용하여 2가지의 소지역 추정방법을 비교하였고, Kim 등 (2008)은 이웃정보시스템 (neighborhood information system)을 이용하여 소지역 추정량을 여러 통계량을 이용하여 비교하였고, Hwang과 Shin (2009)은 MSPE (mean squared percentage error)를 최소화하는 추정량을 이용하여 소지역 추정법을 제안하였다. 소지역 추정은 지역 내 표본수나 발생 빈도가 크지 않아 특정 소지역의 특징이 분석결과에 영향을 미칠 수 있다. 이런 소지역 추정에서 공간이웃의 정보를 정의하는 것은 자료 분석의 결과에 큰 영향을 미치는 중요한 부분이다. 특히, 공간회귀모형을 적용할 경우 공간이웃과 공간이웃가중치를 어떻게 정의하느냐에 따라서 분석결과가 달라질 수 있다. Kim 등 (2008)은 공간이웃의 방법으로 위치와 거리를 기준으로 이웃을 정의하였고, Lee과 Shin (2008)은 지역 경계와 거리를 기준으로 이웃을 정의하고 공간이웃 정의방법에 대해 비교 분석하였다. Kim 등 (2010)은 거리를 이용하여 공간이웃을 정의하고, 공간회귀모형과 일반선형회귀모형을 비교하였다.

기존 공간이웃정의 방법은 단순 가까운 거리 또는 일정 거리를 기준으로 정의되어 졌다. 그 중 $k-$최근접은 거리 측도를 이용시 많이 쓰여진 방법 중 하나이다 (Lee 등, 2015). 최근 공간분석을 통해 상권의 최적입지, 마켓팅, 프렌차이즈 관리 등 여러 분야에서 활용되고 있는데 단순 거리로 공간상에 분

[1] (34054) 대전광역시 유성구 유성대로 1672, 미병연구단, 선임연구원. E-mail: sjkim@kiom.re.kr

포하고 있는 객체들을 측정하는데는 한계가 있다 (Jung, 2009). 공간데이터 사이의 공통적인 연관성을 찾는 공간데이터미이닝에서도 주로 기하연산에 기반을 두고 한정된 클러스터링 방법을 많이 다루었는데 (Son 등, 1998), 최근 공간 클러스터링 (spatial clustering)방법으로 임의의 분포에서 클러스터 발견, 이상치의 효율적인 처리 등 기존 방법에서 몇 가지 단점이 보안 된 도로네 삼각망을 이용한 공간 클러스터링이 많이 이용되고 있다 (Yang과 Cui, 2010). 본 연구에서는 공간이웃을 정의하는 방법으로 도로네 삼각망에 대해 소개하고 $k-$최근접과 비교 분석하였다. 또한 시뮬레이션에서는 공간 영역 (spatial area)을 달리하여 시뮬레이션을 실행하였고, 지가데이터를 이용하여 실 데이터를 분석하였다.

본 연구의 구성은 다음과 같다. 2절에서는 공간데이터분석의 이해를 돕기 위해 본 연구의 분석 절차를 중심으로 공간이웃의 정의, 공간이웃 가중치의 정의, 공간자기상관, 공간회귀모형 등에 대해 설명한다. 3절에서는 두 가지의 공간이웃의 정의방법을 달리하여 어떤 방법으로 이웃을 정의하는 것이 효율적인지 알아보기 위하여 시뮬레이션을 실시하였다. 시뮬레이션의 예측성능은 공간회귀모형인 공간자기회귀 (spatial auto-regression; SAR)모형을 이용하여 잔차제곱합 (residual sum of squares; RSS)을 기준으로 도로네 삼각망과 $k-$최근접의 두 가지 공간이웃 정의방법을 비교한다. 4절에서는 일본의 지가데이터를 이용하여 실 데이터로 공간이웃의 정의 방법을 비교하고, 5절에서는 본 연구에서 수행한 결과에 대해 결론을 내리고 연구의 한계점과 향후 연구방향을 제시한다. 본 연구는 R (ver 2.15.2)을 이용하여 분석하였다.

2. 분석방법

공간회귀분석은 공간이웃의 정의, 공간이웃 가중치의 정의, 공간자기상관의 확인, 공간회귀모형의 적용 등의 단계로 분석이 이루어진다.

2.1. 공간이웃의 정의

어떤 지역이 다른 지역의 주변지역과 이웃하고 있는 상태를 공간이웃이라고 한다. 공간이웃을 정의하는 방법은 몇 가지가 있다. 두 지역의 중심점의 유클리드 거리 (Euclidean distance)로 이웃관계를 정의하는 방법, 최근접 k지역을 이웃으로 정의하는 방법, 도로네 삼각망을 이용한 방법, 일정 반경 이내를 포함하는 지역에 대해서 공간이웃을 정의하는 방법 등이 있다. 본 연구에서는 시뮬레이션과 실 데이터 분석에서 도로네 삼각망방법과 최근접 방법을 이용하여 비교하였다.

$k-$최근접은 지역 간의 거리를 기준으로 임의의 공간객체에서 가장 가까운 k지역만을 추출하고, 그것들을 이웃지역으로 정의하는 방법이다. 따라서, k의 값에 정의에 따라서 이웃지역도 달라진다. 본 연구에서는 이웃지역 k를 2∼5개로 최근접으로 지정하여 지역의 이웃관계를 정의하였다.

도로네 삼각망은 해당점에서 가까운 점끼리를 연결하여, 삼각형을 형성하는 방법이다. 예를 들어 점들 중에서 임의로 세 점을 선택하고 그 외접원을 그린다. 이 때, 원내에 선택한 3점 이외의 점이 포함되지 않았을 때 세 점을 삼각형으로 묶는다. 모든 세 점의 조합에 따라 얻어지는 삼각형 분할을 도로네 삼각망이라고 한다 (Figure 2.1). 도로네 삼각망의 특징은 경계가 인접해 있는지에 관계없이 이웃관계가 정의되며, 지역이 밀집되어 있는 지역은 더욱 밀집되어 네트워크가 형성되는 특징을 가지고 있다. 도로네 삼각망은 보로노이 다이어그램 (Voronoi diagram)의 쌍대도형이며, 점간의 자연스러운 인접관계를 나타내는 것으로 알려져 있다. 그러나, 단순히 지역의 분포 패턴에 따라 멀리 떨어져 있는 두 점을 인접하다고 간주해 버리는 위험성이 있다. 본 연구에서는 멀리 떨어져 있는 두 지역을 이웃으로 정의하는 단점을 보안하기 위해 산과 강 등의 지형 등을 고려하여 이웃을 재정의하였다.

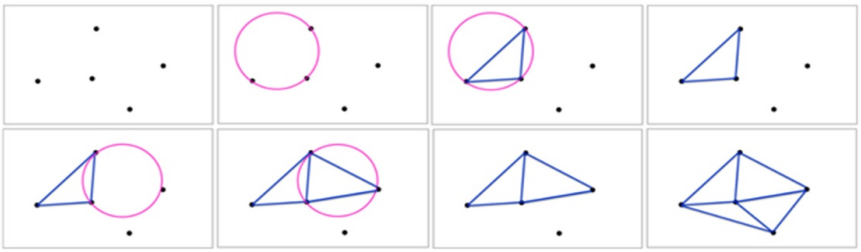

Figure 2.1 Delaunay triangulation procedure

2.2. 공간이웃 가중치

공간이웃이 정의되면, 행렬식을 이용해서 이웃관계의 가중치를 부여하여 지역간의 근접성을 표현할 수 있다. 이것을 공간가중치행렬 (spatial weight matrix)이라고 한다. 공간가중치행렬에는 ① 이웃행렬을 그대로 이용한 방법 (W_B), ② 이웃행렬의 행합으로 표준화 하는 방법 (W_W), ③ 이웃행렬의 전 요소의 합으로 표준화하는 방법 (W_C), ④ 거리행렬을 이용해 표준화하는 방법 (W_S) 등이 있다. 여기서는 행 표준화 (row-standardization) 방법이라고도 불리는 이웃행렬의 행합으로 표준화하는 방법 (W_W)을 이용해서 공간가중치행렬을 계산하였다. 이 방법은 행을 표준화 할 경우 주변지역의 값이 한 지점에 평균적으로 얼마나 영향을 미치는지 계량화 할 수 있는 장점이 있다 (Lee, 2015). 단위 지역 단위의 인접성을 기초로 하여 두 지역 단위가 인접하면 1, 인접하지 않으면 0으로 설정하여, 행의 합계를 1로 표준화하여 공간이웃가중치를 계산한다. 요소 C_{ij} $(i, j = 1, 2, \cdots, n)$에서 공간이웃행렬 C가 이웃행렬의 행합으로 표준화한 W_W의 요소 w_{ij}는 식 (2.1)과 같다.

$$w_{ij} = \frac{c_{ij}}{\sum_{i=1}^{n} c_{ij}}. \tag{2.1}$$

2.3. 공간자기상관

일반적으로 상관관계 (correlation coefficient)라고 하면 두 변수 사이의 관련성을 의미하지만, 공간자기상관 (spatial autocorrelation)은 하나의 변수와 여러 관측값들 사이에서의 관계를 의미한다. 즉, 가까이 위치하는 데이터 속성과 멀리 위치하는 속성을 비교하여 비슷한 여부에 대한 관련성을 나타내는 계수이다. 대표적으로 사용되는 척도로는 Moran's I 통계량, Geary's C통계량 등이 있다 (Han, 2016). 본 연구에서는 Moran's I로 공간자기상관을 나타내었다.

지역 i, j에서 얻어진 공간가중치행렬의 요소를 w_{ij}라고 하면 Moran's I는 식 (2.2)와 같다. 여기에서 N은 대상지역의 수, x_i와 x_j는 각각 지역 i, j에서 얻어진 어떤 특성값, $\bar{x}$는 모든 지역의 특성값들에 대한 평균을 나타내다. Moran's I는 -1과 1사이의 값을 가지고 $|I|$가 1에 가까울수록 높은 공간적 상관관계를 가진다 (Moran, 1948).

$$\text{Moran's } I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \cdot \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\,(x_i - \bar{x})\,(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{2.2}$$

2.4. 공간회귀모형

공간자기상관에 대한 결과 공간자기상관이 존재하면 공간자기상관을 통제 할 수 있는 공간회귀모형을 통해 분석한다. 본 연구에서는 공간이웃 정의방법의 비교를 위해 시뮬레이션과 실 데이터분석에서 공간회귀모형인 SAR을 이용하여 분석하였다. SAR 모형은 공간적 파급효과를 정식화한 모형이며, 일반

선형모형에서 공간자기회귀계수 ρ, 공간가중치행렬 W, 반응변수의 곱인 ρWy가 설명변수로 추가된 모형이다 (Hur, 2007). $(I-\rho W_1)^{-1}$은 공간승수효과 (spatial multiplier)를 나타내는 것으로 공간 간의 상호작용 효과를 나타내며, 각 지역에서 다른 모든 지역들이 서로 연관되어 있음을 의미한다 (Anselin, 2001). X를 독립변수벡터, y를 종속변수, ρ를 공간자기회귀계수, β를 회귀계수벡터, ϵ를 오차항벡터라고 하면, SAR 모형식은 (2.3)과 같이 나타낼 수 있다.

$$y = \rho Wy + \beta X + \epsilon, \quad \epsilon \sim N\left(0, \sigma^2 I\right). \tag{2.3}$$

식 (2.3)에서 우변의 변수 ρWy를 좌변으로 이항한 후 정리하면 식(2.4)과 같다.

$$\begin{aligned} y - \rho Wy &= \beta X + \epsilon, \\ (I - \rho W)\, y &= \beta X + \epsilon, \\ y &= (I - \rho W)^{-1} \beta X + (I - \rho W)^{-1} \epsilon. \end{aligned} \tag{2.4}$$

3. 시뮬레이션

본 연구에서는 이웃을 정의하는 두 방법이 추정에 얼마나 영향을 미치는지 시뮬레이션을 시행하였고, 예측성능은 공간회귀모형인 SAR을 실행하여 RSS의 평균을 기준으로 비교하였다. 도로네 삼각망이 지점의 수가 밀집된 지역에서 더욱 밀집되게 네트워크가 형성되는지 특징을 살펴보기 위해서 지역사이즈를 달리하여 데이터를 생성하였다. 시뮬레이션 순서는 아래와 같이 이루어졌다.

(1) 정방영역으로 크기를 지정하여 각각 x좌표, y좌표 ([0,1]×[0,1], [0,2]×[0,2], [0,3.5]×[0,3.5], [0,5]×[0,5], [0,10]×[0,10])로 Gaussian 분포를 따르는 point data를 geoR패키지의 grf()함수를 이용하여 (x, y)의 좌표와 데이터를 각각 생성함.

(2) 생성한 데이터에서 영역 크기별로 100개씩 데이터를 추출함.

(3) 각각의 데이터로 도로네 삼각망과 $k-$최근접법으로 공간이웃을 정의함.

(4) 행 표준화로 공간이웃의 가중치를 계산하고, Moran's I 통계량으로 공간자기상관을 구함.

(5) 추출한 데이터로 SAR 모형을 이용하여 영역 사이즈별 30회씩 실행하여 RSS의 평균으로 비교함.

각각 도로네 삼각망과 최근접 (k=2, 3, 4, 5)으로 이웃을 정의한 것은 Figure 3.1과 같다 (area size= [0,1] × [0,1]). 이웃의 정의 방법과 이웃수 k에 따라서 이웃의 지점과 이웃의 관계가 확연히 다른 것을 알 수 있다.

영역별로 시행한 도로네 삼각망과 $k-$최근접에 따른 시뮬레이션 결과는 Table 3.1과 같다. SAR 모형을 이용해서 RSS를 나타낸 것이다. 영역 사이즈가 1×1은 도로네 삼각망의 RSS가 7.207로 가장 작은 값을 나타내고 있다. 2×2는 도로네 삼각망 (13.500)과 최근접 k=3 (13.516)이 근소한 차이가 나는 것을 알 수 있다. 영역 사이즈가 3.5×3.5는 최근접 k=3 (21.744), 영역 사이즈가 5×5, 10×10은 최근접 k=2 (area size 5×5=28.211, area size 10×10=52.173)의 RSS가 값이 가장 작은 값을 나타내고 있으며, 영역사이즈 (5×5, 10×10)가 넓어질수록 도로네 삼각망보다 $k-$최근접의 RSS의 값이 작은 값을 나타내고 있는 것을 알 수 있다.

4. 실제데이터분석

본 연구에서 사례분석을 위해 사용된 데이터는 2009년 일본 관동지방의 데이터이다. 일본의 관동지방은 도쿄 도 (Tokyo), 이바라키 현 (Ibaraki ken), 토치기 현 (Tochigi ken), 군마 현 (Gunma ken),

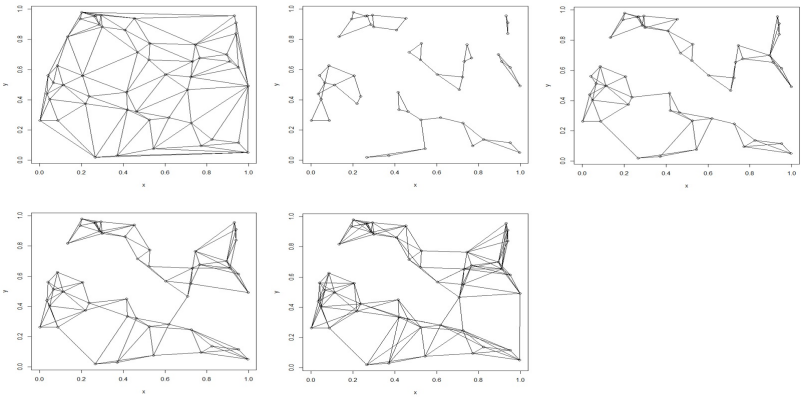

Figure 3.1 Spatial neighborhood (top left: Delaunay triangulation, clockwise from middle left: $k = 2, 3, 4, 5$ nearest neighbors)

Table 3.1 Residual sum of squares for DT and KNN

	Delaunay	$k-$nearest neighbors			
area size	triangulation	$k = 2$	$k = 3$	$k = 4$	$k = 5$
1×1	7.207	7.894	7.816	7.853	8.168
2×2	13.500	14.056	13.516	13.893	13.954
3.5×3.5	22.488	21.818	21.744	22.998	23.623
5×5	28.280	28.211	29.083	29.400	30.347
10×10	56.192	52.173	52.686	53.902	55.234

DT: Delaunay triangulation, KNN: k-nearest neighbor.

사이타마 현 (Saitama ken), 치바 현 (Chiba-ken), 카나가와 현 (Kanagawa ken)의 1도 6현을 말한다. 일본의 시군구-읍면동별 지가데이터 (주택 표준공시지가의 평균가격, 만엔/m^2)를 종속변수로 야간인구밀도 (천명/m^2), 제 3차 산업 종사자 인구밀도 (천명/m^2)등의 데이터를 독립변수로 분석하였다 (n=362). Figure 4.1은 각각 지가데이터와 야간인구밀도, 제3차 산업 종사자 인구밀도의 분포를 나타내고 있다. Figure 4.1의 지가 (left)데이터를 보면 도쿄 도의 일부와 카나가와 현의 일부를 중심으로 짙은 색의 높은 밀도를 나타내고 있어 다른 지역에 비해 지가가 높게 나타나고 있는 것을 알 수 있다. 야간인구밀도 (middle)와 제3차 산업종사자 인구밀도 (right)도 도쿄 도의 일부와 카나가와 현의 일부를 중심으로 짙은 색의 높은 밀도를 나타내 지가데이터와 비슷한 분포를 보이고 있다.

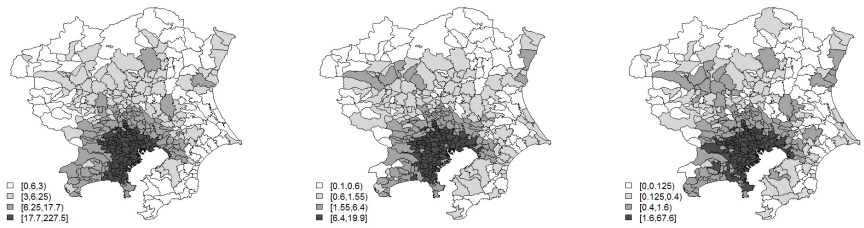

Figure 4.1 Data distribution (left: land price, middle: NPD, right: TPD)

우선, 앞에서 제시한 도로네 삼각망 (Figure 4.2)과 $k-$최근접 (Figure 4.3)으로 각각 공간이웃을 정의하였다. Figure 4.2 (left)는 도로네 삼각망으로 이웃을 정의한 것으로, 멀리 떨어져 있는 두 지역까지 이웃관계를 정의하고 있는 것을 알 수 있다. 본 연구에서는 도로네 삼각망으로 이웃으로 정의된 지역에

서 산, 강, 바다 등의 지형을 고려하여 다시 이웃을 정의하였으며, Figure 4.2 (right)에서 붉은선이 수정 후 이웃을 재정의 한 것이다. Figure 4.3은 최근접 지역을 k=2, 3, 4, 5로 지정하여 이웃관계를 정의한 것으로 이웃지역 k에 따라 이웃의 지역과 모양이 달라지는 것을 알 수 있다.

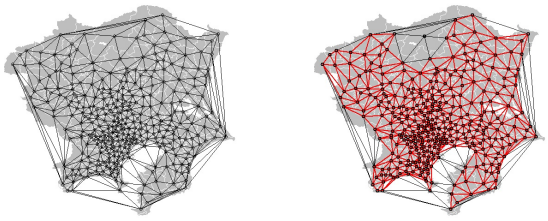

Figure 4.2 Delaunay triangulation neighborhood before revisions(left) after revisions (right)

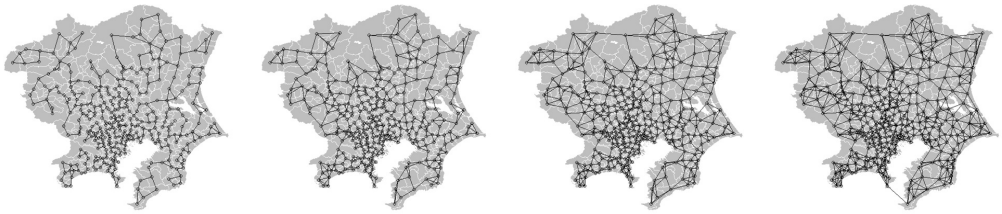

Figure 4.3 Spatial neighborhood (from left k=2, 3, 4, 5 nearest neighbors)

다음은 행 표준화로 공간이웃가중치를 주고, Moran's I의 통계량을 이용하여 공간자기상관을 구하였다 (Table 4.1). 도로네 삼각망은 $k-$최근접에 비해 멀리 있는 지역 (상대적으로 공간상관도가 낮은)까지 이웃으로 정의하기 때문에 최근접방법에 비해서 공간자기상관관계가 낮은 것으로 보인다. 최근접은 변수마다 차이가 있지만 k의 수가 늘어남에 따라 Moran's I의 통계량이 낮아지는 것을 알 수 있다.

Table 4.1 Moran's I statistic

	Delaunay	$k-$nearest neighbors			
variable	triangulation	$k=2$	$k=3$	$k=4$	$k=5$
land price	0.753^{***}	0.825^{***}	0.860^{***}	0.816^{***}	0.815^{***}
NPD	0.867^{***}	0.909^{***}	0.911^{***}	0.914^{***}	0.884^{***}
TPD	0.631^{***}	0.798^{***}	0.755^{***}	0.677^{***}	0.689^{***}

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.
NPD: night time of population density, TPD: tertiary industry practitioners of population density.

다음은 지가데이터를 종속변수로 두고, 야간인구밀도와 제 3차 산업 종사자 인구밀도 등을 독립변수로 SAR 모형을 이용하여 분석한 결과이다 (Table 4.2). RSS의 관점에서 볼 때 최근접 k=3이 19390.2로 가장 작고, 좋은 결과를 얻었다고 볼 수 있으며, AIC (Akaike information criterion)통계량도 최근접 k=3이 2493.9으로 가장 작고 적합이 잘되었다고 할 수 있다.

5. 결론

본 연구에서는 공간데이터의 분석에서 공간이웃을 정의하는 방법으로 도로네 삼각망과 $k-$최근접을 이용하여 시뮬레이션과 일본의 소지역 행정 단위인 시군구-읍면동별 지가데이터를 이용하여 공간이웃을 정의하는 방법을 비교하였다.

Table 4.2 Result of spatial auto-regression model

	Delaunay	$k-$nearest neighbors			
variable	triangulation	$k=2$	$k=3$	$k=4$	$k=5$
(Intercept)	1.795^{*}	2.071^{***}	1.830^{***}	1.795^{***}	1.858^{***}
NPD	0.832^{***}	1.066^{***}	0.828^{***}	0.705^{***}	0.813^{***}
TPD	1.673^{***}	1.808^{***}	1.546^{***}	1.585^{***}	1.656^{***}
spatial lag(ρ)	0.387^{***}	0.271^{***}	0.389^{***}	0.419^{***}	0.372^{***}
AIC	2518.2	2514.8	2493.9	2495.9	2507
RSS	21007.9	20870.7	19390.2	19559.6	20459.5

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.
NPD: night time of population density, TPD: tertiary industry practitioners of population density.

공간이웃을 정의하는 방법에 따라 공간자기상관을 나타내는 Moran's I의 통계량과 잔자제곱합과 모형의 적합도 등의 결과가 달라짐을 알 수 있었다. 시뮬레이션에서는 SAR 모형을 이용해서 구한 RSS를 기준으로 영역 사이즈가 1×1, 2×2의 경우 도로네 삼각망의 잔차가 수치적으로 작게 나왔고, 영역 사이즈가 커짐에 따라 $k-$최근접의 RSS가 작은 값을 나타내었다. 실제 지가데이터를 이용한 결과는 RSS를 기준으로 도로네 삼각망보다 최근접 (k=3)이 작은 수치를 나타내었고, AIC를 기준으로도 최근접 (k=3)이 조금 더 적합하다고 볼 수 있었다. 이는 시뮬레이션의 결과에서도 영역사이즈 (지역의 크기)가 넓어질수록 최근접이 작은 수치를 나타내었는데 추후 지역의 크기와 범위를 고려하여 분석해 보아야 하겠다. 또한, 시뮬레이션과 실 데이터 분석을 통해 공간이웃을 정의 할 때 분석데이터와 연구 목적에 따라 공간이웃의 정의방법을 고려해야 한다고 생각되며 (Wheeler, 2007), 향후 여러 실제데이터를 이용하여 데이터 형태나 종류에 따라서도 공간이웃의 정의방법을 고려하여 분석해야 하겠다. 또한, 공간데이터인 만큼 분석 주제에 따라 산과 강, 바다 등의 지형의 형태에 따라서도 달라져야 하겠다.

References

Anselin, L. (2001). *Spatial externalities, spatial multipliers and spatial econometrics*, Regional Economics Applications Laboratory, Illinois.

Han, J. H. and Lee, M. J. (2016). Cancer cluster detection using scan statistic. *Journal of the Korean Data & Information Science Society*, **27**, 1193-1201.

Hwang, H. J. and Shin, K. I. (2009). A small area estimation for monthly wage using mean squared percentage error. *Korean Journal of Applied Statistics*, **22**, 403-414.

Hur, Y. (2007). A study on the estimation of house price in regard of spatial effects. *House Studies Review*, **15**, 5-23.

Jung, D. Y. and Son, Y. G. (2009). A analysis on the spatial features of the neighborhood trade area using positive spatial autocorrelation method. *Journal of the Korean Society for GeoSpatial Information System*, **17**, 141-147.

Kim, J. S., Hwang, H. J. and Shin, K. I. (2008). Comparison of spatial small area estimators based on neighborhood information systems. *Korean Journal of Applied Statistics*, **21**, 855-866.

Kim, S. J., Choi, S. B., Kang, C. W. and Cho, J. S. (2010). A comparative study on spatial lattice data analysis. *Communications for Statistical Applications and Methods*, **17**, 193-204.

Lee, K. O. (2000). On application of small area estimation to the unemployment statistics of si-gun-gu. *The Korean Journal of Applied Statistics*, **13**, 275-286.

Lee, K. S. and Sine, K. I. (2008). Comparison of neighborhood information systems for lattice data. *The Korean Journal of Applied Statistics*, **21**, 387-397.

Lee, W. J. and Park, C. Y. (2015). Prediction of apartment prices per unit in Daegu-Gyeongbuk areas by spatial regression models. *Journal of the Korean Data & Information Science Society*, **26**, 561-568.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, **10**, 243-251.

Panczak, R., Held L., Moser, A., Jones, P. A., Ruhli, F. J. and Staub, K. (2016). Finding big shots: small-area mapping and spatial modelling of obesity among Swiss male conscripts. *BMC Obesity*, **3**,

1-12.
Son, E. J., Kang, I. S., Kim, T. W. and Li, K. J. (1998). A spatial data mining method by clustering analysis. *Korea Information Science Society*, **25**, 161-163.
Wheeler, D. C. (2007). A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003. *International Journal of Health Geographics*, **6**, 13.
Yang, X. and Cui, W. (2010). A novel spatial clustering algorithm based on Delaunay Triangulation. *Journal of Software Engineering & Applications*, **3**, 141-149.

Journal of the Korean Data & Information Science Society
2017, **28**(3), 505–513

http://dx.doi.org/10.7465/jkdi.2017.28.3.505
한국데이터정보과학회지

A study on the spatial neighborhood in spatial regression analysis

Sujung Kim[1]

[1]Mibyeong Research Center, Korea Institute of Oriental Medicine

Received 12 April 2017, revised 17 May 2017, accepted 18 May 2017

Abstract

Recently, numerous small area estimation studies have been conducted to obtain more detailed and accurate estimation results. Most of these studies have employed spatial regression models, which require a clear definition of spatial neighborhoods. In this study, we introduce the Delaunay triangulation as a method to define spatial neighborhood, and compare this method with the $k-$nearest neighbor method. A simulation was conducted to determine which of the two methods is more efficient in defining spatial neighborhood, and we demonstrate the performance of the proposed method using a land price data.

Keywords: Delaunay triangulation, small area estimation, spatial neighbor.

[1] Senior researcher, Mibyeong Research Center, Korea Institute of Oriental Medicine, Daejeon 1672, Korea. E-mail: sjkim@kiom.re.kr