한국통계학회논문집
2010, 17권, 2호, 193-204

공간 격자데이터 분석에 대한 우위성 비교 연구 - 이상치가 존재하는 경우 -

김수정[a], 최승배[1,a], 강창완[a], 조장식[b]

[a]동의대학교 데이터정보학과, [b]경성대학교 정보통계학과

요 약

최근들어 공간적으로 분석을 필요로 하는 여러 분야에서의 연구자들은 공간통계학에 많은 관심을 가지게 되었다. 그리고 통계학 분야 역시 공간상에서 얻어진 데이터에 공간자기상관이 존재할 경우 공간적으로 분석해야 한다는 주장과 함께 많은 연구가 진행되고 있다. 공간통계학에서 다루고 있는 데이터 중에서 '공간 격자데이터 분석'은 (1) 공간이웃의 정의, (2) 공간이웃 가중치의 정의, (3) 공간모형의 적용 등의 단계를 거쳐서 행해진다. 본 연구에서는 이상치가 존재하는 공간 격자데이터를 분석할 경우 절사평균제곱오차를 이용하여 분석함으로써 예측적인 측면에서 공간통계학적 방법이 일반통계학적 방법보다 더 우수함을 보인다. 본 연구에 대한 내용의 타당성을 보이기 위해서 시뮬레이션을 통하여 공간통계학적인 방법과 일반통계학적인 방법을 비교하였다. 그리고 부산진구의 실제 범죄데이터를 이용한 적용사례를 통하여 절사평균제곱오차를 사용한 공간통계학적 방법의 유용성을 알아보았다.

주요용어: 공간이웃, 공간이웃 가중치, 공간자기상관, 절사평균제곱오차.

1. 서론

공간통계학은 공간상에서 얻어진 데이터를 다루는 학문으로서 통계학의 한 응용분야이다. 공간통계학은 "관측치들의 위치가 가까이 있을수록 상관이 더 높고, 멀리 떨어져 있을수록 관측치들의 상관은 낮아진다."라는 가정을 한다. 따라서 공간데이터를 분석할 경우, 상관관계가 존재한다면 일반통계학적 방법보다 공간통계학적인 방법을 사용하여 분석을 해야 함은 당연하다고 할 수 있다. 어떤 연구에서는 "공간데이터를 일반통계학적인 방법에 의해서 수행함으로써 오류를 범할 수 있다."라는 결론을 기술하고 있다. 대표적인 연구로서 Griffith와 Layne (1999)은 추정의 관점에서 공간데이터에 대해서 전통적인 OLS(ordinary least squares) 방법을 이용하여 모형에 대한 모수를 추정하는 연구를 수행하였다. 그들은 전통적인 OLS에 의해서 공간 특성을 반영한 모수를 추정하는 경우, OLS는 모형에 대해서 오지정(misspecification)의 문제를 갖게 되고, 추정량은 더 이상 최적선형불편 추정량(BLUE; best linear unbiased estimator)이 되지 못하며, 잘못된 통계적 추정의 결론에 도달할 위험성이 있다고 지적하였다.

공간데이터는 데이터의 형태에 따라서 다음과 같이 세 가지로 나눌 수 있다: (1) 지리통계학적 데이터(geostatistical data), (2) 공간 격자데이터(spatial lattice data) 그리고 (3) 공간 점 패턴 데이터(spatial point pattern data). 본 연구에서는 이 세 가지 데이터 종류 중에서 공간 격자데이터에 대한 분석을 다룬다. 공간 격자데이터의 격자(lattice)는 우리가 일반적으로 알고 있는 망사형태가 아니고 하나의 영역 또는 지역을 의미한다. 즉, 공간 격자데이터는 위치와 해당 위치에서 측정값의 형태로 구성된

[1] 교신저자: (614-714) 부산광역시 부산진구 가야동 산24번지, 동의대학교 데이터정보학과, 부교수.
E-mail: csb4851@deu.ac.kr

다. 여기서 위치는 하나의 점으로 표현되기는 하지만 하나의 점을 의미하는 것이 아니고 어떤 지역을 대표하는 중심점을 의미하며, 측정값은 연속형 또는 이산형이 될 수 있다.

공간 격자데이터 분석과 관련하여 연구되어 지고 있는 분야는 소지역 추정과 관련된 연구들로서 대표적인 연구는 다음과 같다. 전수영과 임성섭 (2009)은 오차항이 SAR(simultaneously autoregressive model)을 따르는 공간선형회귀모형에서 일반화 최대엔트로피 추정량에 관한 연구를 수행하였고, 김정숙 등 (2008)은 이웃정보시스템을 이용한 공간 소지역에 대한 추정량을 비교하였으며, 황희진과 신기일 (2009)은 MSPE(mean squared error of prediction)를 이용하여 임금총액에 대한 소지역을 추정하는 연구를 수행하였다. Cressie와 Chan (1989)은 지역적인 변수들에 대한 공간 모형화에 대한 연구를 수행하였고, Smirnov와 Anselin (2009)은 격자상의 공간 상호작용을 갖는 모형에 대해서 변수변환을 사용하여 로그-자코비안(Log-Jacobian)을 계산하는 병렬 방법을 제안하였다. 그리고 Besag (1974)는 격자시스템의 공간 상호작용과 공간통계학적 분석방법을 제시하고 있으며, Jhung과 Swain (1996)은 수정된 M-추정치와 마코프 확률장(Markov random field)에 근거한 베이지안 분류법을 제안하였다.

공간 격자데이터는 공간상에서 얻어지고, 공간자기상관(spatial autocorrelation)이 존재한다면 일반통계학적인 분석이 아닌 공간통계학적인 분석을 수행하는 것이 당연하다고 할 수 있다. 그리고 공간격자데이터는 이상치에 민감하다는 특성을 가지고 있다(3.1절 참조). 따라서 본 연구에서는, 이상치를 갖는 공간 격자데이터를 분석할 경우, 절사평균제곱오차(TMSE; trimed mean squared error) 통계량을 사용하여 공간통계학적 방법과 일반통계학적 방법을 예측적인 측면에서 비교한다. 그리고 시뮬레이션을 통하여 공간통계학적 분석방법이 일반통계적인 분석방법보다 예측을 수행하는데 더 우수함을 보이고 부산진구의 실제 범죄데이터를 이용한 적용사례를 통하여 절사평균제곱오차 통계량을 사용한 공간통계학적 방법의 유용성을 보인다. 본 연구에서 사용된 툴은 S-Plus의 모듈인 S+SpatialStat이다.

본 연구의 2절에서는 공간 격자데이터 분석에 대한 이해를 돕기 위해서 공간 격자데이터 분석 절차를 중심으로 간략히 소개하고, 절사평균제곱오차 통계량을 가지고 공간통계학적 방법과 일반통계학적 방법의 비교를 하기 위한 절차에 대해서 소개한다. 3절에서는 본 연구의 목적을 달성하기 위한 시뮬레이션에 대한 내용이 소개된다. 여기서 시뮬레이션을 수행하는 방법과 내용, 시뮬레이션의 수행 결과를 제시한다. 4절에서는 사례분석으로서 본 연구에서 수행하는 시뮬레이션의 순서대로 분석함으로써 본 연구의 유용성을 재확인한다. 마지막으로 본 연구에서 수행한 결과에 대한 결론 및 연구의 한계점을 제시하고, 향후 연구방향에 대해서 기술한다.

2. 분석방법

2.1. 공간 격자데이터 분석

공간 격자데이터는 위치(영역 또는 지역)와 해당위치에서 측정된 관측값으로 구성되고, 각 위치들 간의 이웃정보를 갖는 데이터이다. 공간 격자데이터 분석은 공간이웃의 정의, 공간이웃 가중치의 정의, 공간자기상관의 유무에 대한 검토, 공간모형의 적용 등의 단계를 거쳐 행해진다. 공간이웃을 정의하는 방법에는 '경계를 이루는 지역(인접하는 지역)들로 정의하는 방법'과 '두 지역의 중심점의 거리 내에 속하는 지역들을 이웃으로 정의하는 방법'이 있다. 본 연구에서는 공간이웃을 정의하는 방법을 시뮬레이션 부분과 사례분석 부분 모두에서 거리를 이용한 방법을 사용하였다. 공간이웃 가중치를 정의하는 방법은 여러 가지가 있다. 첫째, 경계하는 이웃을 공간이웃으로 정의하는 방법으로서 이웃들이 접하는 경계선의 길이에 따라서 가중치를 달리 주는 방법이다. 둘째, 행 표준화(row standardized)에 의한 가중치를 부여하는 방법으로서 각 지역의 이웃의 수에 근거하여 가중치를 부여하는 방법이다. 예를 들면, 어떤 지역에 이웃의 수가 6개 있다면 해당 지역과 6개의 지역사이의 이웃가중치는 1/6로 동일하게 부여한다. 셋째, 거리함수를 이용하는 방법으로서 가까이에 있는 이웃에 높은 가중치가 부여되도록

상관함수에 기초하여 부여하는 방법이다.

본 연구에서(시뮬레이션과 사례분석)는 공간이웃 가중치에 대한 정의 방법으로서 '거리함수'를 이용한 가중치 부여 방법으로서 다음과 같은 '거리함수'를 사용하였다 (Cressie, 1991, p.557 참조).

$$w_{ij} = \begin{cases} \left(\min\left\{d_{ij} : i = 1, 2, \ldots, n\right\}/d_{ij}\right)\left(n_j/n_i\right)^{\frac{1}{2}}, & j \in N_i, \\ 0, & \text{otherwise}, \end{cases} \tag{2.1}$$

여기서 d_{ij}는 두 이웃들 사이의 거리이고, n_i와 n_j는 각각 지역 i와 $j\,(i \neq j)$에서의 관측수이다. 그리고 N_i는 지역 i의 이웃들의 집합이다.

주어진 자료에 대해서 공간회귀모형을 적용할 지에 대해서는 공간자기상관의 존재여부에 달려 있다고 할 수 있다. 공간통계학에서 공간자기상관을 구하는 방법으로 Moran의 I통계량과 Geary의 C통계량이 대표적인 척도로서 사용되고 있다. 본 연구에서는 공간자기상관 여부를 검토하기 위해서 다음과 같이 정의되는 Moran의 I통계량을 사용하였다 (Moran, 1948).

$$I = \frac{n \sum_i \sum_j w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{W \sum_j (z_i - \bar{z})^2}, \tag{2.2}$$

여기서 n은 지역의 수, z_i와 z_j는 각각 지역 i와 j에서 얻어진 어떤 특성을 나타내는 값, $\bar{z}$는 모든 지역의 특성값들에 대한 평균을 나타낸다. 그리고 w_{ij}는 지역 i와 지역 j사이의 가중치로서 식 (2.1)에서 정의된 거리함수들의 요소들이고, W는 모든 지역들 간의 가중치들의 합을 나타낸다.

Moran의 I통계량은 피어슨의 상관계수와 같이 -1에서 1사이의 값을 갖고, 관측지역의 값이 이웃지역과 비슷한 경향을 갖는 경우에는 양의 값으로 얻어지고, 그렇지 않는 경우에는 음의 값으로 얻어진다. 피어슨의 상관계수와 마찬가지로 0에 가까울수록 공간상에 독립적으로 상관관계가 없는 것으로 해석된다.

공간자기상관의 유무에 대한 결과 공간자기상관이 존재한다는 결론이 얻어지면 공간회귀모형을 적용하게 된다. 공간회귀모형을 적합하기 위해서 다음과 같은 함수를 고려한다 (Stephen 등, 1996).

$$Z_i = \mu_i + \delta, \tag{2.3}$$

여기서 Z_i는 위치 i에서의 확률과정이고, μ_i는 위치 i에서의 평균을 의미한다. 평균은 상수일 수도 있고, 공변량을 갖는 선형모형일 수도 있다. 그리고 δ는 평균 0과 공분산 Σ를 갖는 확률변수로서 오차항이다($\delta \sim N(0, \Sigma)$). 여기서 Σ는 모든 위치에서의 확률변수들에 대한 공분산 행렬이다. 작은 규모 변동인 δ는 자기회귀 또는 이동평균 공분산 모형을 Σ에 적합시킴으로써 모형화된다. 공분산 구조 Σ의 유형에는 다음과 같이 세 가지가 있다. 조건부 공간자기회귀(CAR; conditional autoregressive model), 동시 공간자기 회귀(SAR; simultaneously autoregressive model), 이동 평균(MA; moving average). 이러한 모형들은 어떤 형태의 공분산 행렬을 갖는가에 따라서 구분되어 진다. 즉, 세 가지 공분산 행렬의 형태는 다음과 같다 (Stephen 등, 1996).

(1) CAR: $\sum = (I - \rho N)^{-1} D\sigma^2$

(2) SAR: $\sum = \left[(I - \rho N)^T D^{-1}\right]\sigma^2$

(3) MA: $\sum = (I + \rho N) D (I + \rho N)^T \sigma^2$

표 1: 예측오차(이상치인 수치들과 이상치가 아닌 수치들)

지역	1	6	14	42	96	9	16	22	50	88
실제값	3.88	2.15	3.00	2.67	2.15	1.27	0.98	1.28	1.19	1.39
예측값	1.23	1.21	−0.26	1.29	1.05	1.26	0.94	1.20	1.07	1.09
예측오차	2.65	0.94	3.26	1.38	1.10	0.01	0.04	0.08	0.12	0.30

여기에서 ρ와 σ는 공간회귀에 의해서 추정되어져야 할 모수(parameter)이고, N은 이웃행렬이고, D는 식 (2.1)에 주어진 거리함수 w_{ij}의 원소들로 구성된 공간이웃 가중치 행렬이다. 본 연구의 시뮬레이션에서는 CAR, SAR, MA 모형들을 모두 사용하였고, 사례분석에서는 CAR 모형만을 사용하였다.

2.2. 공간 격자데이터 분석의 특징

공간 격자데이터에 대해서 공간자기상관이 존재할 경우 일반통계학을 사용하는 것보다 공간통계학을 사용하는 것이 더 타당하다고 할 수 있다. 본 연구에서는 이를 보이기 위한 사전 작업에서 "공간 격자데이터 분석은 이상치에 민감하다"는 특징이 있음을 알 수 있었다. 이러한 특징은 부산지역의 각 동(100개의 동으로 데이터를 정제)에서 발생된 강도 데이터를 사용하여 CAR 공분산 모형을 적용하여 얻어진 표 1의 결과를 통하여 알 수 있다.

표 1에서 앞의 5개 수치들은 이상치라고 판단되는 수치에서의 예측오차이고, 나머지 5개의 수치들은 이상치로 볼 수 없는 대부분 몰려 있는 수치에서의 예측오차를 나타낸다. 표 1의 결과를 보면 실제값이 이상치라고 판단되는 값들에서 예측오차들의 값이 다른 값들에 비해서 큰 수치로 나타나고 있음을 볼 수 있다. 이러한 결과는 공간 격자데이터 분석이 이상치에 민감하다는 것을 암시한다고 할 수 있다. 표 1의 뒷부분에 주어져 있는 5개의 지역들은 이상치가 아닌 수치들 중에서 임의로 선택된 지역의 수치들이다.

참고로 공간자기상관 여부를 검정하기 위하여 Moran의 I통계량을 사용한 결과 p값이 0.001로 얻어져 귀무가설 "No spatial autocorrelation"이 기각되어 공간자기상관이 존재함을 알 수 있었다. 표 1에 의해서 공간 격자데이터가 이상치에 민감하다는 것을 탐색적으로 검토해 보았는데 향후 연구과제로서 보다 정확한 진단을 할 수 있는 방법론에 대한 연구를 필요로 한다.

2.3. 분석절차

공간 격자데이터가 이상치를 포함하고 있고, 공간자기상관이 존재할 경우, 적절한 공간통계학적 분석방법을 필요로 한다. 따라서 본 연구에서는 이상치가 존재한다는 전제 하에서 절사평균제곱오차 통계량을 이용한 분석절차를 다음과 같이 소개한다.

(단계 1) 데이터를 탐색하고 적절한 변환 등을 시도한다.

(단계 2) 주어진 데이터에 대한 공간이웃과 공간이웃 가중치를 정의한다.

(단계 3) 어떤 공간회귀모형(CAR, SAR, MA)을 사용할 지에 대해서 결정한다.

(단계 4) (단계 3)에서 분석할 모형에 대한 모수들을 추정한다.

(단계 5) 절사평균제곱오차를 적용하여 예측을 수행한다. 여기서 절사평균제곱오차를 사용함으로써 로버스트 공간 격자데이터 분석을 수행할 수 있다. 절사평균제곱오차는 예측을 수행할 경우 좌우 극단값을 몇 %씩 없애고 나머지 데이터를 가지고 분석을 수행한다.

참고로 (단계 5)에서 절사평균제곱오차는 모형을 적합한 후에 적용된다. 즉, 모형을 적용하여 예측하는 경우 먼저 모형을 적합시켜 모수들을 얻고 난 다음 이상치를 제거하고 제거된 개체들을 제외한 나머지 지역들에서 예측값이 얻어진다.

3. 시뮬레이션

3.1. 시뮬레이션 데이터

본 절에서의 시뮬레이션은 (1) 데이터의 특성에 따른 공간이웃과 공간이웃 가중치의 정의 그리고 (2) 위치를 생성해야 하는 문제 때문에 본 연구에서 수행되는 시뮬레이션은 일반 시뮬레이션의 방법과 달리 주어진 데이터에 의존해야 한다는 문제점을 갖는다. 본 연구의 시뮬레이션에서 사용된 데이터는 S-Plus의 공간분석을 위한 모듈인 S+SpatialStat에서 제공하고 있는 SIDS(sudden infant death syndrome)이다. SIDS 데이터는 미국의 North Carolina 주(county)에서의 100개 시(city)에서 발생된 돌발유아사망증후군 데이터이다. 이 데이터는 각 군이 위치하고 있는 좌표인 위치와 여러 가지의 변수들로 구성되어 있다.

3.2. 시뮬레이션의 내용 및 절차

2.3절에서 절사평균제곱오차를 이용한 공간 격자데이터 분석절차를 소개하였다. 본 절에서는 이러한 분석절차를 기초로 하여 이상치가 존재하는 공간 격자데이터에 대해서 공간통계학적 방법이 일반통계학적인 방법보다 우월한 가를 시뮬레이션을 통하여 검토한다. 공간통계학적 방법이 일반통계학적 방법보다 우월함을 보이기 위하여 시뮬레이션에서 사용된 모형은 공간자기상관을 고려한 공간회귀모형과 공간자기상관을 고려하지 않은 일반회귀모형이다. 그리고 우위성 판단 기준으로서 절사평균제곱오차를 사용한다. 본 연구에서 수행한 시뮬레이션의 경우는 총 9가지이다. 즉, 세 가지 공간회귀모형인 CAR, SAR, MA 모형에 따라서 데이터 셋의 수 100, 300, 500개의 경우에 대해서 시뮬레이션을 수행한다. 본 연구를 위한 시뮬레이션의 절차는 다음과 같은 단계를 거쳐 수행된다.

(단계 1) SIDS 데이터 셋에서 100개의 좌표를 그대로 이용한다.

(단계 2) SIDS 데이터 셋에 대한 공간이웃과 공간이웃 가중치를 정의한다. 여기서는 공간이웃을 거리가 30마일 이내에 존재하는 시(city)들을 이웃으로 정의하고 이웃들 간의 공간이웃 가중치는 식 (2.1)에 주어진 거리함수를 사용하여 정의하였다.

(단계 3) 고려하는 각 모형에 대해서 추정된 모수를 결정한다.

(단계 4) (단계 3)에서 결정된 모수를 가지고, 공간자기상관된 각 모형(CAR, SAR, MA)으로부터 100개의 데이터를 얻기 위해 시뮬레이션을 수행한다(모형 적합 시 종속변수 역할).

(단계 5) (단계 4)의 시뮬레이션으로 얻어진 데이터와 일정한 상관관계를 갖는 또 다른 100개의 데이터를 얻기 위해 시뮬레이션을 수행한다(모형 적합 시 설명변수 역할). 여기서 종속변수와 설명변수 사이에 상관정도가 0.7이 되도록 지정하였다.

(단계 6) (단계 4)와 (단계 5)의 과정을 데이터 셋의 수(100, 300, 500)만큼 시뮬레이션을 수행한다.

(단계 7) 얻어진 데이터에 대해서 공간통계학적인 방법으로 공간회귀모형을 고려하고 일반통계학적인 방법으로 일반회귀모형을 고려하여 주어진 데이터들에 대해서 예측을 수행한다. 이 때, 공

표 2: 각 모형별 시뮬레이션의 결과

데이터 셋의 수		모형		
		CAR	SAR	MA
100	일반통계학적 방법	0.5146787	0.5809453	0.5262878
	공간통계학적 방법	0.5137936	0.5670092	0.5236654
300	일반통계학적 방법	0.4987555	0.5192200	0.5946934
	공간통계학적 방법	0.4957263	0.5155903	0.5848662
500	일반통계학적 방법	0.5291302	0.5151110	0.5714836
	공간통계학적 방법	0.5248505	0.5131775	0.5605019

간회귀모형과 일반회귀모형의 적합을 위해서 각각 일반화 최소제곱법(GLS; generalized least squares)과 최소제곱법(LS; least squares)을 적용하였다.

(단계 8) (단계 6)의 결과를 이용하여 본 연구에서 소개한 절사평균제곱오차를 사용하여 공간통계학적 방법과 일반통계학적 방법을 비교한다.

참고로 본 연구에서 고려하는 시뮬레이션은 다음과 같은 몇 가지의 제약사항이 있다. 첫째, 데이터에 의존한다는 것이다. 즉, 공간 격자데이터에 대한 위치((위도, 경도), (동-서, 남-북) 등)는 시뮬레이션을 할 수 없기 때문에 고정시켜야 한다. 둘째, 본 연구에서 제안하는 방법은 절사평균의 개념을 사용하기 때문에 정보의 손실이 있다는 점이다. 그리고 일반회귀모형에서는 오차항에 대해서 정규성, 독립성, 등분산성을 가정하기 때문에 공간모형들을 적합하기 위하여 사용된 시뮬레이션 결과를 일반선형모형에 적합시키는데 사용하는 것은 적절하지 못하다는 단점이 있다. 그러나 (1) 두 방법의 비교를 위한 동일한 데이터가 필요하며, (2) 본 연구의 목적이 일반회귀모형의 가정이 위배되었을 경우 일반통계학적인 방법이 아닌 공간통계학적인 방법으로 접근해야 한다는 것이기 때문에 일반회귀모형을 적합할 경우 공간회귀모형에서 사용한 데이터와 같은 데이터를 사용한다.

3.3. 시뮬레이션의 결과

3.2절에서 기술한 시뮬레이션 절차를 통하여 얻어진 결과는 표 2에 주어져 있다. 표 2에 주어져 있는 수치들은 총 9가지 경우에 대해서 본 연구에서 소개한 분석절차를 이용하여 얻어진 절사평균제곱오차의 값이다. 공간통계학적 방법의 분석결과를 보면 CAR 모형이 데이터 셋의 수 100과 300에서는 SAR과 MA 모형보다 예측력이 좋게 나오기는 했지만 데이터 셋의 수가 500개인 경우는 SAR모형으로 예측한 결과가 좋게 얻어졌음을 알 수 있다. 이러한 결과는 일반통계학적인 방법에서도 마찬가지 결과로 얻어졌다. 그러나 본 연구의 목적이 이러한 패턴을 보고자 하는 것이 아니고 각 모형과 데이터 셋의 수에서 공간통계학적인 방법과 일반통계학적인 방법 중에 어느 것이 더 잘 예측하는가를 보이는 것이다.

그 결과를 보면 9가지 모든 경우에 대해서 공간통계학적인 방법에서의 수치들이 일반통계학적인 방법에서의 수치들보다 낮게 나타났기 때문에 공간통계학적인 방법이 예측력에서 우수한 결과를 보이고 있다. 이러한 결과로 볼 때, 본 연구에서 소개한 분석절차에 의해서 공간 격자데이터 분석을 수행한다면 이상치가 존재할 경우에 적절한 공간통계학적인 분석법을 적용하여 분석할 수 있다.

4. 사례분석

4.1. 적용데이터

본 연구에서 사례분석을 위해서 이용된 데이터는 2005년 1월부터 2007년 6월까지의 부산광역시의 강도에 대한 범죄 데이터로서 대검찰청에서 제공하는 정보를 바탕으로 분석 가능하도록 가공한 데

표 3: 이웃의 정의

구	기준지역	인접지역	거리
부산진구	범전동	연지동	0.745952
		부암동	1.805426
		부전동	1.106162
		양정동	1.036879
		전포동	1.537572
⋮	⋮	⋮	⋮
남구	문현동	대연동	5.920537
		범천동	5.050687
		범일동	4.294519
		전포동	6.738317
		우암동	4.283462
평균(Km)		2.913	

이터이다 (http://www.kostat.go.kr). 분석데이터의 크기는 약 13만(125,798)건이다. 이 중에서 최종적으로 사용된 분석데이터는 부산광역시 행정동 기준인 215개의 동(세부적인 동: A1동, A2동, A3동)을 105개의 동(결합한 동: A동)으로 하고, 상대적으로 많이 떨어져 있는 기장군(5개 동을 포함하고 있음)을 제외한 100개 동에서 발생한 강도 범죄(1,066건)를 대상으로 하였다. 그리고 분석의 용이성을 위하여 원래의 좌표를 TM좌표로 변환하였다.

공간 격자데이터 분석을 수행하고자 할 때, 맨 처음 수행할 작업은 공간이웃을 정의하는 것이다. 공간이웃을 정의할 때, 거리를 이용하는 방법을 적용한다면 각 지역의 중심점을 필요로 하게 된다. 본 연구에서는 공간이웃을 거리로서 정의를 하였기 때문에 중심점을 어떻게 잡는가가 중요한 문제가 될 수 있다. 본 연구에서는 각 동의 중심점을 신뢰할 수 있는 업체에 의뢰하여 TM 좌표로서 구하였다.

4.2. 공간이웃의 정의

거리를 이용하여 이웃을 정의하기 위해서 2단 집락추출법을 사용하였다. 즉, 본 연구에서는 1차 추출단위로 각 구에서 5개의 구를 표본으로서 랜덤하게 추출하였고, 추출된 구에서 2차 추출단위로서 5개의 동을 랜덤하게 추출하였다. 그리고 추출된 동과 인접된 모든 동과의 거리를 측정하였다. 이렇게 추출된 모든 동(본 연구에서는 26개 동)들에 대해서 유클리디안 거리들을 구하였고, 이 거리들에 대한 평균 거리 이내의 동들은 어떤 특정 동에 대한 이웃이라고 정의하였다. 따라서 어떤 특정한 지역에 대한 공간이웃들은 26개 동들에 대한 거리들의 평균인 2.913Km 이내의 동들로 구성된다. 이에 대한 내용들을 요약한 결과가 표 3에 주어져 있다.

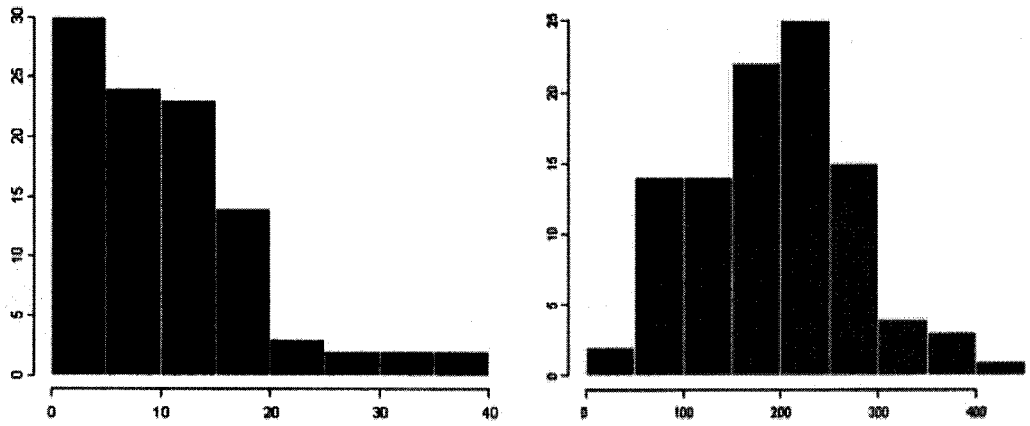
4.3. 공간이웃 가중치의 정의

본 연구에서는 이웃들의 거리들에 대한 평균인 2.913이내의 동들을 어떤 특정 동에 대한 공간이웃으로 정의하였다. 어떤 특정한 동과 이웃하는 동들에 대해서 모두 같은 가중치를 부여하는 것보다는 가까이에 있을수록 높은 가중치를 주고, 멀리 있을수록 낮은 가중치를 부여하는 것이 합리적일 것이다. 따라서 본 연구에서는 식 (2.1)에 주어진 거리함수를 이용하여 가중치를 부여하였다.

표 4는 식 (2.1)을 적용하여 얻어진 공간이웃 가중치의 결과표이다. 표 4의 'row.id'와 'col.id' 열은 이웃들의 쌍이고, 'weight'는 해당 이웃들에 대한 공간이웃 가중치를 나타낸다. 예를 들면, 지역 '1'의 이웃들은 지역들 '2', '4', '5', '6', '7' 등이고, '지역 1'과 '지역 2'에 대한 공간이웃 가중치는 0.553이다.

표 4: 길이에 따른 가중치결과

row.id	col.id	weight	matrix
1	2	0.55253183	1
1	4	0.78938383	1
1	5	0.67368451	1
1	6	0.64681814	1
1	7	0.60335200	1
⋮	⋮	⋮	⋮
95	96	0.28475932	1
95	93	0.10401318	1
96	95	0.36240424	1
98	51	0.16023798	1
100	79	0.14457693	1



(a) 원자료 (b) Freeman-Turkey변형

그림 1: *강도 범죄율의 히스토그램*

4.4. 데이터의 변환

분석에 이용되는 데이터는 정규분포라는 가정을 따라야 한다. 그러나 그림 1의 좌측 히스토그램을 통하여 알 수 있듯이 강도에 대한 범죄율은 정규성을 벗어난 분포 형태를 보이고 있다. 정규분포로 변환하기 위해서는 로그(log) 변형이나 제곱근(square root) 변환 등을 고려할 수 있지만 Freeman-Turkey의 제곱근 변형이 일반적인 제곱근 변형보다 더 안정적이다. 따라서 본 연구에서는 해당 변수에 식 (4.1)과 같이 Freeman-Turkey의 제곱근 변형을 시켰다 (Cressie, 1991).

$$Y_i = \sqrt{\frac{1000\,S_i}{n_i}} + \sqrt{\frac{1000\,(S_i+1)}{n_i}}, \tag{4.1}$$

여기서 n_i는 인구수이고, S_i는 각 동의 범죄수이다. 그리고 Y_i는 변환된 범죄율이다. Freeman-Turkey의 제곱근 변형을 이용하여 변환된 범죄율 Y_i에 대한 히스토그램이 그림 1의 우측에 주어져 있다. Freeman-Tukey 제곱근 변환한 강도 범죄율은 정규성을 보이고 있음을 알 수 있다(p = 0.157).

표 5: 강도-비거주용 건물 내 주택수: 공간회귀와 일반회귀 비교

실제값	공간예측값	공간오차 제곱합	일반예측값	일반오차 제곱합
3.87945	3.34000	0.29101	3.33000	0.30190
1.93230	1.36000	0.32753	1.38000	0.30503
1.78786	1.12000	0.44604	1.15000	0.40686
1.16262	1.82000	0.43215	1.84000	0.45885
⋮	⋮	⋮	⋮	⋮
1.21387	1.13000	0.00703	1.16000	0.00290
0.82624	1.35000	0.27432	1.37000	0.29567
1.00991	1.47000	0.21168	1.50000	0.24018
0.81375	0.97000	0.02441	1.01000	0.03851
		0.08778		0.09184

4.5. 공간자기상관 및 공간회귀 적합

본 연구에서는 공간자기상관을 보기 위해서 Moran의 I통계량을 사용하였다. 공간자기상관 분석 결과, Moran의 I값은 0.3894(p = 0.001)로 얻어져 공간자기상관이 존재하고 있음을 알 수 있다.

따라서 데이터가 공간자기상관을 갖고 있다고 판명되었기 때문에 공간회귀모형을 고려한 공간 격자데이터 분석을 수행할 수 있다. 사례분석에 대한 모형을 가정할 때, 강도에 영향을 미치는 변수들은 많이 있기 때문에 많은 설명변수를 고려하여 모형을 설정하는 것은 당연한 것이다. 그러나 본 연구에서의 목적은 강도에 영향을 미치는 영향력이 있는 설명변수들을 찾는 것이 목적이 아니기 때문에 본 연구에서는 단순공간회귀모형을 고려하였다.

본 연구에서는 분석변수들로서 종속변수는 '강도', 설명변수는 '비거주용 건물 내 주택수(House with in commercial building)'를 이용하였다 (http://www.krihs.re.kr). 그리고 공분산 행렬의 모형으로서 CAR 모형을 이용하였다. 본 연구에서 '비거주용 건물 내 주택'을 분석변수로 사용한 이유는 (1) 상업용도이기 때문에 어떤 상품이 있을 것이고, (2) 밤 또는 휴일에는 사람이 없기 때문에 범죄의 표적이 될 수 있기 때문이다. 참고로 설명변수로 사용된 '비거주용 건물 내 주택'의 정규성 검토결과 정규성을 따르지 않기 때문에 Freeman-Turkey의 제곱근 변형을 하였다. 그리고 실제 데이터를 이용하여 '강도 범죄율'과 '비거주용 건물 내 주택수'의 비율과의 상관계수가 0.68(p = 0.001)로 나타나 두 변수 간에는 높은 상관관계가 있는 것으로 나타났다.

식 (2.3)의 공간회귀모형을 적합 시킨 결과는 다음과 같으며, 회귀계수에 대한 유의성 검정결과 유의한 결과를 보였다(p = 0.001). 그리고 추정된 모수의 값은 0.3885로 얻어졌다.

$$\hat{y} = 1.1560x + 0.6283. \tag{4.2}$$

4.6. 분석결과

표 5는 본 연구에서 소개한 분석절차를 적용하여 분석한 결과이다. 표 5에 주어진 0.08778과 0.09184의 값은 다음의 절차에 의해서 얻어진 값이다. 먼저 (1) 100개 데이터에 대해서 공간회귀와 일반회귀모형으로 각각 적합시키고, (2) 적합된 모형을 이용하여 이상치라고 판단될 수 있는 양 극단값 5개씩을 제거해서 얻어진 90개의 위치에 대한 값을 예측하여 예측값와 실제값의 차이(예측오차)를 구하였다. 그리고 (3) 위의 두 값은 각각의 방법에 대해서 90개의 예측오차들을 제곱해서 평균한 값이다(즉, 10% 절사평균제곱오차). 표 5에서 '실제값'은 강도에 의한 범죄수를 Freeman-Tukey 변환을 한 값이고, '공간예측값'과 '일반예측값'은 각각 공간회귀와 일반회귀의 의한 적합값이다. 그리고 '공간오차제곱합'과 '일반오차제곱합'은 각각의 실제값과 예측값의 차를 제곱한 값이다.

표 5의 절사평균제곱오차에 대한 결과를 보면 공간통계학적 방법이 일반통계학적 방법보다 절사평균제곱오차가 작게 나왔기 때문에 일반통계학적 방법보다 공간통계학적 방법을 사용했을 때 예측력이 더 좋음을 알 수 있다.

5. 결론 및 향후 연구과제

일반적으로 공간 격자데이터는 공간상에서 얻어지고 공간자기상관이 존재함에도 불구하고 예측을 수행할 경우, 공간이웃과 공간이웃 가중치를 어떻게 적용하느냐에 따라서 공간통계학적인 방법이 우월하거나 일반통계학적인 방법이 우월하다는 일관성이 없는 결과를 보인다. 따라서 본 연구에서는 공간자기상관이 존재하는 공간데이터는 공간통계학적인 방법으로 분석해야 한다는 관점과 이상치가 존재하는 경우 보다 강건한(resistant) 분석법이 필요하다는 관점에서 보다 유용한 공간통계학적 분석방법의 필요성을 제기하였다.

이에 본 연구에서는 이상치가 존재하고, 공간자기상관이 존재하는 공간 격자데이터를 분석할 경우 절사평균제곱오차를 이용한 분석절차를 소개하였다. 본 연구에서 소개한 분석절차를 적용할 경우, 공간통계학적인 분석방법이 일반통계학적인 방법보다 우월하다는 것을 시뮬레이션을 통하여 확인하였다. 그리고 본 연구에서 소개한 절사평균제곱오차를 이용한 분석절차의 타당성을 검토하기 위하여 사례분석에서 실제 데이터에 대해서 적용하여 보았다.

본 연구의 한계점으로 공간 격자데이터에 대해서 이상치가 존재하는 경우에 한해서 적용 가능하다는 것이다. 따라서 향후 연구과제로서 이상치가 존재하지 않은 공간 격자데이터에서도 공간통계학적 방법을 적용할 수 있는 연구가 수행되어져야 할 것으로 생각한다. 이를 위하여 시뮬레이션을 통하여 어떻게 공간이웃을 정의해야지 최적인지 그리고 어떻게 공간이웃 가중치를 부여하면 최적인지를 연구할 가치가 있다. 또한 본 연구의 내용 중에서 공간 격자데이터 분석은 이상치에 민감하다는 특징을 갖는다고 주장한 바 있다. 그러나 이것은 탐색적 데이터 분석을 통한 주장이기 때문에 보다 객관적인 주장을 하기 위해서는 이에 대한 시뮬레이션을 통한 검토가 필요하다. 따라서 이에 대한 연구도 향후 연구과제로 남기고자 한다.

참고 문헌

김정숙, 황희진, 신기일 (2008). 이웃정보시스템을 이용한 공간 소지역 추정량 비교, <응용통계연구>, **21**, 855–866.

전수영, 임성섭 (2009). 오차항이 SAR(1)을 따르는 공간선형회귀모형에서 일반화 최대엔트로피 추정량에 관한 연구, <한국통계학회논문집>, **16**, 265–275.

황희진, 신기일 (2009). MSPE를 이용한 임금총액 소지역 추정, <응용통계연구>, **22**, 403–414.

Besag, J. (1974). Spatial interaction and statistical analysis of lattice systems, *Journal of Royal Statistical Society: Series B*, **36**, 192–236.

Cressie, N. and Chan, N. H. (1989). Spatial modeling of regional variables, *Journal of the American Statistical Association*, **84**, 393–401.

Cressie, N. A. C. (1991). *Statistics for Spatial Data*, John Wiley & Sons, New York.

Griffith, D. A. and Layne, L. J. (1999). *A Casebook for Spatial Statistical Data Analysis: A Compilation of Analyses of Different Thematic Data Set*, Oxford University Press, Oxford.

Jhung, Y. and Swain, P. H. (1996). Bayesian contextual classification based on modified M-estimates and Markov random fields, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **34**, 67–75.

Moran, P. (1948). The interpretation of statistical maps, *Journal of Royal Statistical Society*, **10**, 243–251.

Smirnov, O. and Anselin, L. (2009). Parallel method of computing the Log-Jacobian of the variable transformation for models with spatial interaction on a lattice, *Computational Statistics and Data Analysis*, **53**, 2980–2988.

Stephen P. K., Silvia C. V., Tamre P. C. and Alice A. S. (1996). *S+SpatialStats User's Manual*, MathSoft, Inc.: Seattle, Washington.

http://www.kostat.go.kr

http://www.krihs.re.kr

2010년 1월 접수; 2010년 3월 채택

A Comparative Study on Spatial Lattice Data Analysis - A Case Where Outlier Exists -

Sujung Kim[a], Seungbae Choi[1,a], Changwan Kang[a], Jangsik Cho[b]

[a]Department of Data Information Science, Dongeui University
[b]Department of Informational Statistics, Kyungsung University

Abstract

Recently, researchers of the various fields where the spatial analysis is needed have more interested in spatial statistics. In case of data with spatial correlation, methodologies accounting for the correlation are required and there have been developments in methods for spatial data analysis. Lattice data among spatial data is analyzed with following three procedures: (1) definition of the spatial neighborhood, (2) definition of spatial weight, and (3) the analysis using spatial models. The present paper shows a spatial statistical analysis method superior to a general statistical method in aspect estimation by using the trimmed mean squared error statistic, when we analysis the spatial lattice data that outliers are included. To show validation and usefulness of contents in this paper, we perform a small simulation study and show an empirical example with a criminal data in BusanJin-Gu, Korea.

Keywords: Spatial neighborhood, spatial neighborhood weight, spatial autocorrelation, trimed mean squared error.

[1] Corresponding author: Associate Professor, Department of Data Information Science, Dongeui University, San 24, Gaya-Dong, Busanjin-Gu, Busan 614-714, Korea. E-mail: csb4851@deu.ac.kr