

피어슨 상관계수의 공간화: 세 관련 기법 간의 비교 실험 연구

이상일* · 조대현** · 이민파***

Spatializing the Pearson's Correlation Coefficient: An Experimental Comparison of Three Relevant Techniques

Sang-Il Lee* · Daeheon Cho** · Minpa Lee***

요약 : 본 연구는 두 변수 간의 상관성을 측정하는데 지배적인 통계기법으로 사용되어 온 피어슨 상관계수를 공간화하는 방식에 대해 다루고 있다. 이변량 공간적 자기상관이 존재할 경우, 피어슨 상관계수값과 그것에 대한 유의성 검정 결과가 갖는 통계학적 의미는 훼손될 수 밖에 없다. 본 연구는 이변량 상관관계에서의 공간적 자기상관의 문제를 해결하기 위해 제시된 세 가지 연구 기법(수정 t -검정, 공간필터 상관계수, 이변량 공간적 자기상관 통계량)에 대한 상세한 리뷰를 제공하고, 다소 독립적으로 발전해 온 세 기법이 얼마나 일관성 있는 결과를 보여주는지를 실험 연구를 통해 살펴보고자 했다. 주요 결과는 다음의 두 가지로 요약된다. 첫째, 몇몇 예외를 제외한다면, 세 가지 접근법의 결과는 상당한 정도의 상호 일관성을 갖는 것으로 나타났다. 즉, L^* 에 의거해 높은 이변량 공간적 자기상관을 보여주는 패턴 쌍일수록 공간필터 상관계수와 유효표본크기(자유도)는 작은 반면, 유의 확률은 높게 나타났다. 둘째, L^* 와 가장 일관성 있는 결과를 보여준 것은 고유벡터공간필터링(ESF, eigenvector spatial filtering) 기법에 기반한 공간필터 상관계수 기법이었다. 즉, L^* 가 커질수록 공간필터 상관계수가 감소하는 거의 완벽한 경향성을 보여주었다. 본 연구의 가장 큰 의미는 피어슨 상관계수가 본질적으로 비공간적인 통계량임을 명확히 하고, 이 문제점을 해결하기 위해 제안되어 온 세 접근법이 개별적 특성에도 불구하고 일관성 있는 결과를 보여준다는 점을 실험 연구를 통해 밝혔다는 점이다.

주요어 : 피어슨 상관계수, 수정 t -검정, 공간필터 상관계수, 이변량 공간적 자기상관 통계량, 고유벡터공간필터링

Abstract : This study deals with spatializing the Pearson's correlation coefficient as a dominant statistical technique for measuring and assessing bivariate relationships. With the presence of bivariate spatial autocorrelation in a pair of variables under investigation, not only Pearson's correlation coefficients themselves but their statistical significance are deemed to be questionable. This study provides a comprehensive review on the three different approaches to the problem of spatial autocorrelation in the bivariate correlation (modified t -test, spatially filtered correlation coefficients, and bivariate spatial autocorrelation statistics), and examines

본 연구는 국토교통부 국토공간정보연구사업의 연구비지원(과제번호14NSIP-B080144-01)에 의해 수행되었습니다. 이 논문의 일부 내용은 2016년 미국 텍사스 주 덴턴(Denton, Texas)에서 개최된 미국지리학회 남서지역분과 연례학술대회(Annual Meeting of Southwest Division of the American Association of Geographers)에서 발표되었음.

* 서울대학교 지리교육과 교수(Professor, Department of Geography Education, Seoul National University, si_lee@snu.ac.kr)

** 가톨릭관동대학교 지리교육과 조교수(Assistant Professor, Department of Geography Education, Catholic Kwandong University, dhcho@gmail.com)

*** (주)망고시스템 기술연구소 연구소장(Director of R&D, Institute of Technology, Mango System Inc., minpa.lee@mango-system.com)

how compatible the results from the three different camps might be by conducting a simulation experiment. The main findings are twofold. First, with some exceptional cases, the three approaches are quite correspondent to one another in terms of experimental results; the higher the degree of bivariate spatial autocorrelation as measured by L^* , the lower the spatially filtered correlation coefficients, the smaller the effective sample size, and the higher the p -values. Second, the most compatible results are found between L^* and the spatially filtered correlation coefficients based on the eigenvector spatial filtering (ESF) approach; there is an almost perfect negative relationship between the statistics and the correlation coefficients. The major contribution of this study to spatializing the Pearson's statistic lies in reaffirming that the statistic is aspatial in nature and in clarifying in an experimental simulation that the three different approaches yield consistent results to some extent.

Key Words : Pearson's correlation coefficient, modified t -test, spatially filtered correlation coefficients, bivariate spatial autocorrelation statistics, eigenvector spatial filtering

1. 서론

공간데이터에 대한 통계학적 분석은 보통 여러 개의 변수를 동시에 고려해야 하는 다변량 상황에서 이루어진다. 다변량 통계분석이 제대로 이루어지기 위해서는 개별 변수에 대한 기술통계적 요약이나 데이터 탐색이 선행되어야 하지만, 가장 중요한 사항은 역시 변수들간의 상관성을 정식화하는 것이다. 이런 의미에서 보면 상관분석은 매우 중요한 위치를 차지한다. 상관분석은 그 자체로 가장 단순한 형태의 다변량 통계분석일 뿐만 아니라 다중회귀분석, 주성분분석, 정준상관분석, 관별분석 등과 같은 보다 복잡한 형태의 다변량 통계분석을 수행하기 전에 반드시 이루어져야 하는 통계 기법이기도 한 것이다. 그런데, 주지하는 바처럼, 다변량 통계분석 기법을 공간데이터에 적용할 때는 그렇지 않은 데이터에 적용할 때보다 훨씬 더 많은 주의를 기울여야 한다(Griffith and Amehin, 1997).

개별 변수에 공간적 의존성(spatial dependence) 혹은 공간적 자기상관(spatial autocorrelation)이 존재한다는 것은 표준 통계학이 기반하고 있는 '독립관측 가정(independent observations assumption)'을 위배한다는 것을 의미하며, 결국 유효표본크기(effective sample size)의 삭감 혹은 자유도(degree of freedom)의 하락으로 이어지게 됨으로써, 통계학적 결론의 오류 가능성이 증대되는 결과가 초래된다. 따

라서 공간데이터분석 혹은 공간통계학의 발전은 일반 통계학 기법을 다양한 방식으로 공간화하려는 시도와 맞물려 진행되어 온 것으로 이해할 수 있다. 그런데 상관분석이 일변량 분석과 다변량 분석의 상호 역할을 하는 중요한 분석 기법임에도 불구하고 상대적으로 적은 관심을 받아왔다는 점은 지적할 필요가 있다. 최근 공간데이터분석 혹은 공간통계학의 발전은 패턴 탐지에 초점을 둔 일변량 분석(이상일 등, 2015; 2016)과 다양한 공간적 회귀분석 기법의 발달(Fotheringham *et al.*, 2002; Griffith, 2003; Anselin, 2009; Anselin and Rey, 2014)로 특징지어지는 다변량 분석에 보다 초점이 맞추어져 있는 듯하다¹⁾.

따라서 본 연구는 등간/비율 척도의 변수간 상관성을 측정하는 통계량으로 가장 널리 사용되고 있는 피어슨 상관계수(Pearson's correlation coefficients)에 초점을 맞추고자 한다. 보다 구체적으로, 연구 데이터에 공간적 자기상관이 존재할 경우, 피어슨 상관계수의 측도로서의 본질적 성격과 통계적 유의성에 어떠한 문제가 발생하는지에 대해 살펴보고자 한다. 공간적 자기상관과 피어슨 상관계수간의 관련성에 대한 주목은 이미 1980년대 초반 경에 이루어졌다(Bivand, 1980; Griffith, 1980; Haining, 1980; Richardson and Hémon, 1981). 이후 이러한 문제에 해결책을 제시하기 위한 시도가 다각도로 진행되어 왔는데 이는 세 가지 정도로 정리될 수 있다(Lee, 2017). 첫째, 피어슨 상관계수의 유의성 검정법을 수정하는 것으로 공간적 자기상관의 정도를 감안할 수

있는 새로운 표준오차 계산법을 제시한다(수정 t -검정). 둘째, 공간필터링(spatial filtering) 기법을 활용하여 ‘진정한(genuine)’ 혹은 ‘순수한(pure)’ 피어슨 상관계수 값을 산출한다(공간필터 상관계수). 셋째, 개별 변수의 공간적 자기상관과 두 변수간의 비공간적 상관을 결합한 새로운 측도 혹은 통계량을 제시한다(이변량 공간적 자기상관 통계량).

위의 세 가지 기법들은 어느 정도 독립적으로 발전해 왔기 때문에, 세 기법의 상대적인 특성이나 연구 결과의 일관성이 어느 정도인지에 대한 검토는 거의 이루어지지 않았다. 이에 본 연구는 지금까지 진행되어 온 이 세 가지 방향의 연구가 공간적 자기상관과 피어슨 상관계수와의 관련성에 대한 우리의 이해를 어떻게 확장시켜 줄 수 있는지를 가상 데이터와 실험적 시뮬레이션을 통해 검토해 보고자 한다. 이를 통해 피어슨 통계량의 비공간성을 보다 명확히 재확인하고, 이 문제를 해결하기 위한 가능한 대안이 무엇인지에 대한 논의의 장을 제공하고자 한다.

2. 연구 방법 및 절차

1) 분석 방법론

(1) 수정 t -검정

피어슨 상관계수는 전형적인 교차곱 형태의 통계량으로 다음의 수식에 의해 주어진다.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{1}{n} \sum_{i=1}^n z_{x_i} z_{y_i} \quad (1)$$

여기서 z_{x_i} 와 z_{y_i} 는 각각 $(x_i - \bar{x})/\sqrt{\sum_i (x_i - \bar{x})^2/n}$ 과 $(y_i - \bar{y})/\sqrt{\sum_i (y_i - \bar{y})^2/n}$ 으로 정의한다²⁾. 이러한 변형식을 이용하면 피어슨 상관계수가 보다 간명하게 표시될 수 있는데, 두 변수의 표준점수를 서로 곱한 값들의 평균으로 정의될 수 있는 것이다. 이 때 각 데이터 포인트에서의 두 표준점수의 곱을 국지적 피어슨 통계량(local Pearson's r_i)으로 규정할 수 있으며 이변량 탐색 기법의 하나로 사용될 수 있다(Lee, 2001b;

2004a; 이상일, 2008; 이화정 등, 2013). 모집단에서의 무상관성이 귀무가설로 주어질 경우, 피어슨 상관계수의 검정통계량과 표준오차는 다음과 같이 주어진다.

$$t = \frac{r}{\hat{\sigma}_r}, \hat{\sigma}_r = \sqrt{\frac{1-r^2}{n-2}} \quad (2)$$

수식 (2)가 의미하는 바는 피어슨 상관계수의 표본분포가 자유도가 $n-2$ 인 t -분포를 보인다는 것인데, 문제의 핵심은 데이터에 공간적 자기상관이 존재할 경우 표준오차의 과소추정 문제가 발생한다는 것이다(Clifford and Richardson, 1985; Dutilleul, 1993). 공간적 의존성 혹은 공간적 자기상관은 “공간단위 간의 지리적 근접성과 공간단위가 보유한 속성값 간의 수치적 유사성, 이 둘 간의 특정한 관련성”으로 정의될 수 있는데(이상일 등, 2015), 이는 피어슨 상관계수를 비롯한 대부분의 표준적 통계기법이 전제하고 있는 ‘독립관측 가정’을 위배하는 것이다(Haining, 1991). 공간적 자기상관이 존재하게 되면 공간단위에서의 값에 대한 정보를 가지고 인접 공간단위에서의 값을 추측하는 것의 합리성이 발생하게 되는데, 이는 개별 공간단위에서의 값들이 독립적으로 표집되지 않았다는 것을 의미한다. 이처럼 특정한 양의 정보가 인접 공간단위 사이에서 중복적으로 존재한다면, 독립관측 가정에 기반한 자유도(혹은 유효표본크기)는 더 이상 신뢰하기 어렵게 된다. 양의 공간적 자기상관이 존재할 경우 자유도는 하락해야 하기 때문에 식(2)에 제시되어 있는 표준오차 추정식은 ‘진정한’ 표준오차를 과소추정하게 된다. 표준오차에서의 과소추정은 측정통계값을 부풀리게 되고, 이는 유의확률의 감소로 이어지기 때문에, 결국 귀무가설이 옳음에도 불구하고 그것을 기각하게 되는 ‘제1종 오류(Type I error)’를 범할 가능성이 높아지게 되는 것이다(Lee, 2017).

수정 t -검정은 이러한 문제를 공간통계학적으로 해결하고자 한 노력의 결과이다. 몇 가지 서로 다른 방법이 제안되어 왔지만, 여기서는 클리포드-리처드슨이 제안한 방식(Clifford-Richardson's solution)에 집중하고자 한다(Clifford and Richardson, 1985;

Clifford *et al.*, 1989; Dutilleul, 1993; Griffith and Paelinck, 2011). 이 방식은 표준오차의 추정치를 계산하는 수식을 다음과 같이 제시한다.

$$\hat{\sigma}_r = \sqrt{\frac{1-r^2}{n'-2}} \quad (3)$$

여기서 n' 는 공간적 자기상관의 정도를 감안한 새로운 유효표본크기를 의미한다. 제공된 기호 속의 분모 전체($n'-2$)를 본다면 자유도를 재계산하는 것과 동일한 것이다. 그리고 유효표본크기는 다음의 수식에 의거해 계산된다.

$$n' = 1 + n^2 [\text{tr}(\hat{\mathbf{R}}_x \hat{\mathbf{R}}_y)]^{-1} \quad (4)$$

여기서 $\hat{\mathbf{R}}_x$ 와 $\hat{\mathbf{R}}_y$ 는 두 변수 각각의 개별 공간단위에서의 공간적 자기상관의 정도를 나타내주는 공분산 매트릭스이다. 따라서 두 매트릭스의 곱을 통해 생성되는 새로운 매트릭스의 주대각 요소는 각 공간단위에서의 이변량 공간적 자기상관의 상대적 정도를 나타내 준다³⁾. 결국 $\text{tr}(\hat{\mathbf{R}}_x \hat{\mathbf{R}}_y)$ 를 통해 이변량 공간적 자기상관의 총체적인 강도가 측정되는 것이다(Lee, 2017). 만일 공간적 자기상관이 존재하지 않는다면 $\text{tr}(\hat{\mathbf{R}}_x \hat{\mathbf{R}}_y) = n$ 이 되기 때문에 $n' \cong n$ 의 관계가 성립한다(Haining, 1991). 만일 양의 이변량 공간적 자기상관이 존재한다면 $n' < n$ 의 관계가 성립할 것이고 강도가 강해질수록 유효표본크기는 점점 더 감소하게 될 것이다.

하나의 예를 들어 식(4)를 설명해 보면 다음과 같다. 표본의 크기가 50인 두 변수간의 피어슨 상관계수가 0.3이라고 하자. 식(2)에 따라 가설검정을 행하면, 검정통계값은 2.179, 자유도는 48, 유의확률은 0.0343으로 계산된다. 두 변수간의 상관관계는 통계적으로 유의한 것으로 결론지어진다. 그런데 모든 공간단위에서 평균적으로 2.0 정도의 양의 이변량 공간적 자기상관이 존재한다고 하자. 식(4)에 의거해 유효표본크기를 구하면 $26(1+50^2/100)$ 이 되며, 이를 이용해 유의성 검정을 하면 유의확률이 0.1365으로 계산된다(자유도 24의 t -검정 적용). 공간적 자기상관을 감안할 경우 두 변수간 상관관계의 통계적 유의성은 사라지게 되는 것이다. 이 접근법은 측정통계량으로

서의 피어슨 상관계수는 그대로 사용하되 검정통계량에서 표준오차 부분만 수정함으로써 다른 통계학적 결론을 이끌어내는 전략을 택하고 있다. 양의 공간적 자기상관이 존재한다면 검정통계값은 감소할 것이고 유의확률은 높아질 것이다.

(2) 공간필터 상관계수

공간적 필터링(spatial filtering)은 공간적 자기상관이 존재하는 변수를 공간적으로 독립적인, 즉 독립관측 가정을 위배하지 않는 변수로 변환하는 것을 의미하는데(Getis and Griffith, 2002; Griffith, 2003; 2010; 2017; Griffith and Chun, 2014), 이것의 요체는 특정한 공간적 필터를 이용해 변수에 내재되어 있는 공간적으로 구조화된 요소를 제거하는 것이다. 따라서 어떠한 방식으로 공간적 필터를 생성할 것인가에 따라 다양한 기법이 가능할 수 있는데(Griffith, 2010), 여기서는 공간회귀분석 기법으로 널리 사용되고 있는 SAR(simultaneous autoregressive) 모형에 기반한 방식(SAR 방식)과 고유벡터공간필터링(ESF, eigenvector spatial filtering) 기법에 기반한 방식(ESF 방식)에 대해서만 다루기로 한다.

SAR 방식은 공간오차모형(spatial error model) 중 가장 널리 사용되고 있는 SAR 모형에 기반하고 있고 다음과 같은 방식으로 정식화 된다(Baily and Gatrell, 1995). 우선 일반적인 OLS(ordinary least squares) 모형은 식(5)와 같이 주어진다. 만일 잔차에 공간적 자기상관이 존재할 경우 식(6)에서처럼 잔차(ϵ)는 공간적으로 구조화된 잔차($\rho \mathbf{V}\epsilon$)와 공간적으로 독립적인 잔차(η)로 구분된다. 식(6)을 식(5)에 대입하면 식(7)에 나타나 있는 SAR 모형의 기본 수식이 완성된다. 이 때 식(6)을 재정리 하면 식(8)과 식(9)에서 보이는 두 수식이 도출된다.

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \quad (5)$$

$$\epsilon = \rho \mathbf{V}\epsilon + \eta \quad (6)$$

$$\mathbf{y} = \mathbf{X}\beta + \rho \mathbf{V}\epsilon + \eta \quad (7)$$

$$\epsilon = (\mathbf{I} - \rho \mathbf{V})^{-1} \eta \quad (8)$$

$$\eta = (\mathbf{I} - \rho \mathbf{V})\epsilon \quad (9)$$

식(5)~(9)에서 \mathbf{y} 는 종속변수 벡터, \mathbf{X} 는 독립변수 매

트릭스, β 는 회귀계수 벡터, ϵ 는 잔차 벡터, η 는 공간적 자기상관이 제거된, 즉 공간적으로 독립적인 잔차 벡터, \mathbf{I} 는 단위행렬(unit matrix), \mathbf{V} 는 공간근접성행렬(spatial proximity matrix) 혹은 공간가중치행렬(spatial weights matrix), ρ 는 공간적 자기회귀 계수(spatial autoregressive coefficient)이다. 여기서 주목해야 할 것은 식(9)이다. 왜냐하면 공간적 자기회귀 계수가 추정되면 공간적 자기상관이 존재하는 잔차(ϵ)로부터 공간적으로 구조화된 잔차($\rho\mathbf{V}\epsilon$)를 차감함으로써 공간적 자기상관으로부터 자유로운 잔차(η)가 도출되기 때문이다. 이때, $(\mathbf{I}-\rho\mathbf{V})$ 가 바로 공간적 필터 역할을 한 것이다.

문제는 회귀분석의 잔차에 적용되는 수식(9)를 일반적인 변수의 공간적 필터링에 어떻게 적용할 것인가인데, 그 절차는 다음과 같다. 식(7)에 나타나 있는 SAR 기본 모형에 공간적 필터링을 적용하고자 하는 변수(x)를 종속변수에 두고 독립변수를 투입하지 않는 모형(절편만으로 구성된 모형)을 구성하면 식(10)과 같이 주어진다.

$$\mathbf{x} = \mu_x \mathbf{1} + \rho_x \mathbf{V} \epsilon + \eta_x \quad (10)$$

여기서 $\mathbf{1}$ 은 모든 구성요소가 1인 열벡터(column vector)이다. 이 단순 모형을 통해 공간적 자기회귀 계수의 추정값이 획득되면($\hat{\rho}_x$), 식(9)를 변형한 다음의 식(11)을 통해 공간필터 변수(spatially filtered variable)(\mathbf{x}^{SF})가 생성된다.

$$\mathbf{x}^{SF} = (\mathbf{I} - \hat{\rho}_x \mathbf{V}) \mathbf{x} \quad (11)$$

공간필터 변수를 생성하는 두 번째 방법은 ESF 방식이다. 이것은 SAR 방식과 마찬가지로 특정한 공간적 회귀분석 기법에 기반하고 있다. ESF 기법은 다니엘 그리피스(Daniel A. Griffith) 교수의 선구적인 연구(1996; 2000; 2003)에 기반하고 있는데, SAR 모형과 마찬가지로 잔차에서 공간적 자기상관을 제거하고자 하는 목적을 가졌지만 기본 개념과 추정 방식에서는 상이한 입장을 취한다. ESF 회귀분석의 기본 모형은 다음과 같이 주어진다.

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{E}\gamma + \eta \quad (12)$$

$$\epsilon = \mathbf{E}\gamma + \eta \quad (13)$$

여기서 \mathbf{E} 는 공간근접성행렬로부터 도출된 고유벡터 중 특정한 방식으로 선정된 고유벡터들의 매트릭스이고, γ 는 고유벡터 매트릭스에 대한 회귀계수 벡터이며, η 는 앞에서와 마찬가지로 공간적 자기상관이 존재하지 않는 잔차 벡터이다. 고유벡터는 특정한 형태의 공간근접성행렬(혹은 조정된 공간근접성행렬)을 분해함으로써 얻어지는데, 각 고유벡터는 특정한 수준과 특정한 형태의 공간적 자기상관 패턴을 내재하고 있다(Boots and Tiefelsdorf, 2000; Griffith, 2003; Tiefelsdorf and Griffith, 2007; 이상일 등, 2015; 2016; 2017). 따라서 ESF 회귀분석은 식(5)에 나타나 있는 OLS 회귀분석의 구조를 그대로 따르는 대신 추출된 고유벡터를 일종의 통제변수(control variables)로 투입함으로써 잔차(ϵ)를 공간적으로 구조화된 부분($\mathbf{E}\gamma$)과 그렇지 않은 부분(η)으로 분리한다(자세한 내용은 이상일 등(2017) 참조).

이러한 ESF 모형에 기반해 공간필터 변수를 추출하기 위해서는 식(10)에서 보는 바처럼, 공간적 필터링을 적용하고자 하는 변수(x)를 종속변수에 두고 선정된 고유벡터가 독립변수 역할을 하는 ESF 모형을 구성하면 된다(Griffith, 2010).

$$\mathbf{x} = \mu_x \mathbf{1} + \mathbf{E}_x \beta_x + \eta_x \quad (14)$$

여기서 $\mathbf{1}$ 은 모든 구성요소가 1인 열벡터, \mathbf{E}_x 는 변수 x 에 대해 선정된 고유벡터의 매트릭스, η_x 는 공간적으로 독립적인 잔차벡터이다. 동일한 정의를 또 다른 변수(y)에 대해서도 적용될 수 있다. 그런데 어떤 고유벡터는 두 변수 모두에 사용될 수 있고(공통 고유벡터), 어떤 고유벡터는 한쪽 변수에만 사용될 수 있다(특수 고유벡터). 이 점을 감안하면 식(14)는 다음과 같이 변형된다(Griffith, 2010).

$$\mathbf{x} = \mu_x \mathbf{1} + \mathbf{E}_C \beta_{C_x} + \mathbf{E}_{U_x} \beta_{U_x} + \eta_x \quad (15)$$

여기에서 \mathbf{E}_C 는 공통 고유벡터의 매트릭스, β_{C_x} 는 공통 고유벡터와 결부된 회귀계수 벡터, \mathbf{E}_{U_x} 는 변수 x 에만 해당하는 특수 고유벡터 매트릭스, β_{U_x} 는 특수 고유벡터와 결부된 회귀계수 벡터이다. 이러한 방식으

로 도출된 잔차(η_x) 그 자체가 바로 공간필터 변수가 된다.

$$\mathbf{x}^{SF} = \eta_x \quad (16)$$

식 (11)에 나타나 있는 SAR 방식에 의하건 식(16)에 나타나 있는 ESF 방식에 의하건 또 다른 변수(y)에 대한 공간필터 변수(\mathbf{y}^{SF})가 도출될 수 있고, 이 새로운 두 공간필터 변수 간의 피어슨 상관계수가 바로 공간필터 상관계수(spatially filtered correlation coefficient)로 정의된다. 이 기법은 측정통계량은 그대로 둔 채 투입되는 변수값을 변화시킴으로써 새로운 피어슨 상관계수 값을 산출하는 전략을 취한다. 양의 공간적 자기상관의 정도가 심할수록 공간필터 상관계수의 값은 원 피어슨 상관계수 값과 비교해 더 작아질 것이다.

(3) 이변량 공간적 자기상관 통계량

Lee(2017)는 공간데이터에 피어슨 상관계수를 사용하는 것의 문제는 “적어도 하나의 변수에 공간적 자기상관이 존재하기 때문”이라고 말하기 보다는 “두 변수 간에 이변량 공간적 자기상관(bivariate spatial autocorrelation)이 존재하기 때문”이라고 말하는 것이 보다 적절하다고 지적한 바 있다. 이변량 공간적 자기상관에 대한 개념규정은 공간적 자기상관 일반에 대한 개념규정(이상일 등, 2015)을 이변량 상황으로 확장하기만 하면 된다. 즉, 이변량 공간적 자기상관은 “관측개체들의 위치 유사성과 이변량 연관성(bivariate association)에서의 유사성간의 특정한 관련성”으로 개념규정할 수 있다(Lee, 2017). 개별 공간단위에서 이변량 연관성은 네 종류로 구분될 수 있는데, 평균 이상의 x_i 값이 평균 이상의 y_i 값과 연관되어 있을 수도 있고($H=H$)⁴⁾, 평균 이상의 x_i 값이 평균 미만의 y_i 값과($H=L$), 평균 미만의 x_i 값이 평균 이상의 y_i 값과($L=H$), 평균 미만의 x_i 값이 평균 미만의 y_i 값과($L=L$) 연관되어 있을 수도 있다. 이변량 공간적 자기상관이란 특정한 유형의 이변량 연관성을 보이는 공간단위가 공간적으로 무작위적으로 분포하지 않을 때 발생한다(Lee, 2017). 예를 들어, $H=H$ 연관성을 보이는 공간단위가 클러스터를 이루어 분포하고 있다

면 최소한 그 하위 지역에서는 양의 이변량 공간적 자기상관이 존재한다고 말할 수 있다.

이변량 공간적 자기상관 통계량 혹은 이변량 공간 연관성 통계량(bivariate spatial association statistics)은 이변량 공간적 의존성 혹은 이변량 공간적 자기상관을 측정하기 위해 고안된 것이다. 상당히 오래 전에 교차-모런(cross-Moran) 통계량(혹은 이변량 모런 통계량)이 제안된 바가 있고(Wartenberg, 1985; Griffith, 1993; Reich *et al.*, 1994), 최근 이변량 기어리(bivariate Geary) 통계량도 제안된 바 있지만(이상일, 2007; 2008), 본 연구에서는 Lee(2001a; 2001b; 2004b; 2009; 2017)가 개발한 L 통계량에 집중하고자 한다. Lee(2001b)는 이변량 공간적 자기상관 통계량이 갖추어야 할 두 가지 조건을 제시한 바 있는데, 첫 번째 조건은 피어슨 상관계수의 방향성과 규모가 가능한 보존되어야 한다는 것이고, 두 번째 조건은 두 변수 모두의 일변량 공간적 자기상관의 정도가 반영되어야 한다는 것이다. 이 두 가지 조건을 결합함으로써, 이변량 공간적 자기상관 통계량은 매 지점별 이변량 연관성(국지적 피어슨 상관계수가 측정) 뿐만 아니라 이변량 연관성의 공간적 집중도도 함께 측정해야만 한다. Lee의 통계량은 이 두 조건을 모두 만족하는 것으로 여겨지며 다음의 수식을 통해 주어진다.

$$L = \frac{n}{\sum_i (\sum_j v_{ij})^2} \frac{\sum_i \sum_j [v_{ij}(x_j - \bar{x})][v_{ij}(y_j - \bar{y})]}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (17)$$

공간근접성행렬의 주대각 요소가 0인가 그렇지 않은가의 여부가 공간적 자기상관 통계량의 성격에 지대한 영향을 끼친다는 주장(Lee, 2004b; 2009; 이상일 등, 2015; 2016)에 근거하면 식(17)에 나타나 있는 일반 L 통계량은 L^0 통계량과 L^* 통계량으로 구분될 수 있다. 이 둘 중 측도로서의 의미가 훨씬 더 큰 L^* 통계량에 집중하고자 하는데(이에 대해서는 Lee, 2017 참조), 행-표준화(row-standardized) 공간근접성행렬이 사용되면 식(17)은 좀 더 간명한 형태로 전환된다.

$$L^* = \frac{1}{n} \sum_i (w_{ij}^* z_{x_i}) (w_{ij}^* z_{y_i}) = \frac{1}{n} \sum_i \tilde{z}_{x_i} \tilde{z}_{y_i} \quad (18)$$

여기서 w_{ij} 는 주대각 요소가 0이 아닌 값을 갖는 행 표준화 공간근접성행렬의 한 요소이고, \tilde{z}_{xi} 와 \tilde{z}_{yj} 는 각 공간단위에서의 표준점수의 공간이동평균(spatial moving average) 값이다. 식(18)을 식(1)과 비교해 보면 L^* 통계량이 피어슨 상관계수와 구조적으로 동일한 교차곱 통계량이라는 사실을 쉽게 알 수 있다. 이 통계량의 가치는 다음과 같이 설명할 수 있다. 공간단위의 개수가 n 인 두 변수가 있다고 하자. 각 공간단위에서의 두 변수 값을 결합한 상태에서 공간단위의 위치를 무작위적으로 재배열 한다고 하자. 이렇게 하면 피어슨 상관계수는 동일하지만 ‘패턴 일치도’에서는 차이를 보이는 두 변수가 새로이 생성된다. 이러한 방식의 재배열은 이론적으로 $n!$ 만큼 가능한데, 피어슨 상관계수는 동일하지만 ‘공형화(共型化, co-patterning)’의 정도에서는 서로 다른 $n!$ 쌍이 생성되는 것이다(Lee, 2001a; 2001b; 2017). 이 서로다른 공형화의 정도를 측정해주는 것이 바로 L^* 통계량인 것이다.

L^* 통계량이 피어슨 상관계수를 공간화하는 문제에 가장 크게 기여하는 바는 L^* 통계량이 피어슨 상관계수에 대한 대체 통계량 혹은 적어도 보완 통계량으로 간주될 수 있다는 점이다. L^* 통계량은 두 변수 간의 피어슨 상관계수가 클수록, 그리고 개입되는 두 변수의 공간적 자기상관이 클수록 큰 값을 갖게 된다(Lee, 2001a; 2001b; 2004b). 따라서 특정한 변수 쌍에 대한 L^* 통계값은 그 통계량의 기댓값(피어슨 상관계수에 비례)과 두 변수 간의 피어슨 상관계수 값 사이의 어느 지점에 위치하게 되는 데, 피어슨 상관계수 값에 가까울수록 공간적 자기상관의 정도가 강한 것을 의미한다. 그러므로 L^* 통계량은 피어슨 상관계수에 기반하고 있음과 동시에 이변량 공간적 자기상관도 반영하고 있으므로 후자에 대한 대체 통계량 혹은 보완 통계량으로서 간주될 수 있는 것이다.

2) 분석 절차

앞에서 설명한 바처럼 피어슨 상관계수를 공간화하려는 세 가지 접근법은 사실상 서로 독립적으로 발전해 온 것이기 때문에 서로 다른 접근법이 얼마나 일

관성 있는 분석 결과를 보여줄지에 대한 기존 연구는 존재하지 않는다. 본 연구는 이를 위해 다음과 같은 실험 디자인을 마련했다.

첫째, 피어슨 상관계수는 동일하지만 서로 다른 이변량 공간적 자기상관을 보여주는 가상의 데이터를 생성한다. 이를 위해 그림 1에 나타나 있는 두 패턴을 이용하고자 한다(이상일, 2001a; 2001b). 가상의 연구 대상 지역은 37개의 육각형으로 구성되어 있다. 두 패턴 모두 3의 값을 갖는 7개의 육각형(검은색), 2의 값을 갖는 17개의 육각형(회색), 1의 값을 갖는 13개의 육각형(흰색)으로 구성되어 있다. 따라서 두 변수의 평균과 분산은 동일하다. 그러나 일변량 공간적 자기상관의 정도라는 측면에서는 두 패턴은 완전히 다른 것이다. 패턴 A의 모런 통계값(Moran's I)은 0.681로 극단적으로 높은 양의 공간적 자기상관을 보여주지만, 패턴 B의 모런 통계값은 -0.186으로 음의 공간적 자기상관을 보여주고 있다. 두 패턴 간의 피어슨 상관계수는 0.422인데, 식(2)에 나타나 있는 표준적 가설검정에 따르면 유의확률은 0.009로 높은 통계적 유의성을 보여준다. 이 데이터로부터 특정한 수준의 이변량 공간적 자기상관을 갖는 패턴의 쌍을 생성하기 위해 다음과 같은 시뮬레이션을 수행한다. 우선 모든 공간단위에서 두 값을 결합한다. 앞에서 사용한 용어로 말하자면 모든 공간단위에서 국지적 피어슨 상관계수는 그대로 유지한다. 그리고 나서 공간단위의 순서를 무작위적으로 재배열한다. 이 재배열의 결과 변하지 않는 것은 피어슨 상관계수이고, 변하는 것은 이변량 공간적 자기상관의 정도(혹은 두 변수의 일변량 공간적 자기상관)이다. 본 가상 데이터의 경우 이렇게 무작위 재배치를 통해 도출 가능한 변수 쌍은 모두 9.07×10^{26} 개이다⁵⁾. 이 들 중 L^* 에 의거해 특정한 수준의 이변량 공간적 자기상관 통계량을 보이는 패턴의 쌍을 도출한다. 본 연구에서는 특정한 L^* 를 먼저 정한 후 무작위 재배치를 계속 반복하면서 근접한 값을 보이는 패턴 쌍을 발견하는 방식을 채택했다.

둘째, 서로 다른 이변량 공간적 자기상관을 보이는 패턴 쌍에 대해 수정 t -검정과 공간필터 회귀계수 기법을 적용한다. 만일 세 가지 접근법의 결과가 완벽히 상호 일관성을 갖는 것이라면, L^* 에 의거해 보다 높은

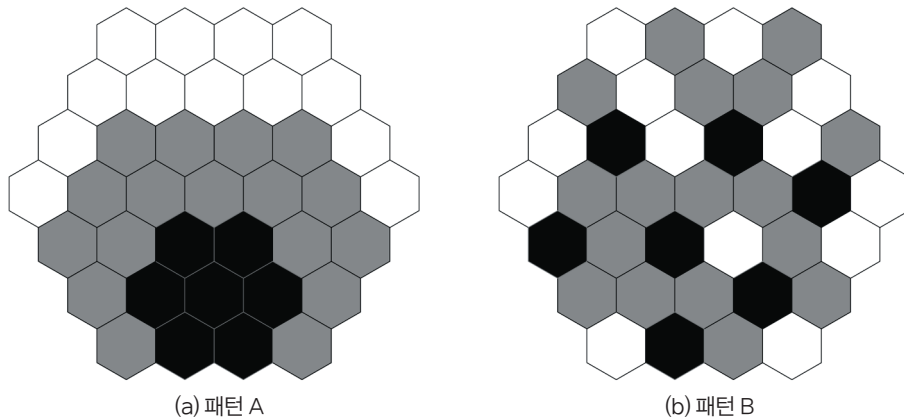


그림 1. 실험 연구를 위한 가상의 두 패턴

이변량 공간적 자기상관을 보이는 패턴 쌍일수록 수정 t -검정에서 유의하지 않을 가능성이 높아져야 하고(낮은 자유도와 높은 유의확률), 공간필터 회귀계수는 보다 낮은 값을 보여야 한다.

셋째, ESF 공간필터링을 행하기 위해서는 고유벡터를 선정하는 원칙을 결정할 필요가 있다. 본 연구에서는 가장 널리 사용되고 있는 방식을 사용하고자 하는데, 최대 공간적 자기상관의 최소한 25% 수준을 보여주는 고유벡터를 선정한 후, 단계적 회귀분석(stepwise regression)을 실행하여 최종적인 고유벡터를 추출하는 방식이다(Griffith, 2010; Chun and Griffith, 2013; Chun *et al.*, 2016; Griffith, 2017).

분석 및 시각화는 R과 국토교통부 국토공간정보연구사업의 공간정보 SW활용을 위한 오픈소스 가공기술개발 과제를 통해 개발된 분석도구(국토교통과학기술진흥원, 2018)를 활용하였다. 이 가운데, 수정 t -검정을 위해서는 SpatialPack이라고 하는 R 패키지(Vallejos *et al.*, 2013; 2019; Osorio *et al.*, 2018)를 활용하였다.

3. 연구 결과

표 1은 이변량 공간적 자기상관의 수준을 달리하는 가상의 8쌍의 패턴을 보여주고 있다. 모든 쌍은 앞







에서 설명한 바처럼 모두 5% 유의수준에서 유의미한 0.422의 피어슨 상관계수 값을 가진다. 그러나 L^* 에 의거해 측정된 이변량 공간적 자기상관의 수준은 0.000, 0.050, 0.100, 0.150, 0.200, 0.250, 0.302, 0.371로 서로 다르다. PAIR1과 PAIR2는 L^* 통계량의 기댓값(0.066)보다 낮은 값으로 음의 이변량 공간적 자기상관을 보여준다. 일변량 공간적 자기상관의 경우도 모런 통계량을 기준으로 할 때 기댓값(-0.028)보다 낮은 값을 보인 PAIR1-X, PAIR1-Y, PAIR2-X는 음의 공간적 자기상관을 보여주고 있고, 기어리 통계량(Geary's c)과 S^* 통계량(이상일 등, 2015; 2016; 2017)도 유사한 경향성을 보여주고 있다. 그러나 통계적으로 유의미한 L^* 값을 보여주고 있는 쌍은 PAIR5~PAIR8이다. 이 중 PAIR5, PAIR6, PAIR7은 두 변수 중 한 변수의 패턴만 통계적으로 유의미한 공간적 자기상관을 보여주고 있는데 반해 PAIR8은 두 변수 모두에서 매우 높은 수준의 일변량 공간적 자기상관을 보여주고 있다. 유의미한 이변량 공간적 자기상관이 한 변수만의 높은 일변량 공간적 자기상관에 기인한 경우와 두 변수 모두에서의 높은 일변량 공간적 자기상관에 기인한 경우가 결과에 어떠한 영향을 미칠지를 살펴보는 것은 흥미로운 사항이 될 것이다.




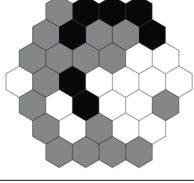
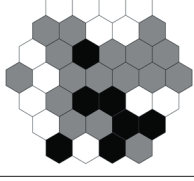
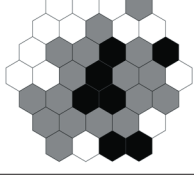
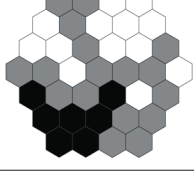
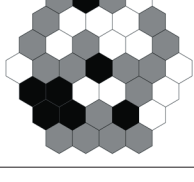
표 2는 표 1에 나타나 있는 가상의 8쌍에 대해 공간필터 상관계수 기법과 수정 t -검정 기법을 적용한 결과이다. 반복적으로 얘기하지만 8쌍은 동일한 피어슨 상관계수를 공유하지만 서로 다른 L^* 값을 가진



다. 우선 SAR 방식의 공간필터 상관계수를 살펴보면, 0.218~0.500 사이의 값을 나타내고 있는데, 전체적으로는 L^* 값이 증가할수록 공간필터 상관계수가 감

소하는 경향을 보여주고 있다. 가장 큰 값은 PAIR1의 0.439이며 가장 작은 값은 PAIR8의 0.218이다. 이는 이변량 공간적 자기상관이 높을수록 거기에 비례해

표 1. 이변량 공간적 자기상관의 정도가 서로 다른 8개의 가상의 패턴 쌍

패턴 쌍	구성 변수		이변량 공간적 자기상관 L^*	일변량 공간적 자기상관		
	변수명	패턴		모런 통계량	기거리 통계량	S^* 통계량
PAIR1	PAIR1-X		0.000 (0.126)	-0.065 (0.705)	0.999 (0.989)	0.135 (0.706)
	PAIR1-Y			-0.102 (0.451)	1.062 (0.555)	0.162 (0.912)
PAIR2	PAIR2-X		0.050 (0.720)	-0.039 (0.906)	1.009 (0.931)	0.118 (0.497)
	PAIR2-Y			-0.003 (0.803)	0.946 (0.605)	0.120 (0.513)
PAIR3	PAIR3-X		0.100 (0.423)	0.086 (0.249)	0.936 (0.537)	0.227 (0.200)
	PAIR3-Y			0.137 (0.094)	0.788* (0.043)	0.258 (0.066)

PAIR4	PAIR4-X		0.150 (0.050)	0.060 (0.370)	0.988 (0.909)	0.177 (0.701)
	PAIR4-Y			0.136 (0.095)	0.830 (0.104)	0.298* (0.010)
PAIR5	PAIR5-X		0.200** (0.002)	0.019 (0.634)	0.967 (0.752)	0.183 (0.623)
	PAIR5-Y			0.253** (0.004)	0.746* (0.015)	0.369** (0.000)
PAIR6	PAIR6-X		0.250** (0.000)	0.143 (0.082)	0.809 (0.068)	0.290* (0.016)
	PAIR6-Y			0.240** (0.006)	0.694** (0.003)	0.327** (0.002)
PAIR7	PAIR7-X		0.302* (0.000)	0.461** (0.000)	0.515** (0.000)	0.513** (0.000)
	PAIR7-Y			0.079 (0.277)	0.936 (0.537)	0.228 (0.192)

PAIR8	PAIR8-X		0.371** (0.000)	0.407** (0.000)	0.652** (0.001)	0.426** (0.000)
	PAIR8-Y			0.359** (0.000)	0.662** (0.001)	0.427** (0.000)

(괄호 속의 값은 유의확률임. *와 **는 각각 0.05와 0.01 수준에서 유의함을 나타냄. 모런과 기어리 통계량에는 주대각 요소가 모두 0이고 연접한 공간단위에 대해서는 1의 값을 갖는 C^0 공간근접성행렬을, Lee의 통계량에는 주대각 요소와 연접한 공간단위에 대해서 1의 값을 갖는 C^1 공간근접성행렬의 행-표준화 버전인 W 를 적용함.)

피어슨 상관계수가 부풀려졌다는 것을 의미한다. 다시 말해 높은 피어슨 상관계수는 인접한 공간단위에서 유사한 이변량 연관성이 발생하는 자기상관 매커니즘 때문에 발생한 것이지 변수간의 본질적인 상관성을 나타낸 것은 아니라는 점을 함축하고 있는 것이다. 따라서 공간적 필터링을 통해 그러한 공간적 효과를 배제해버리면 실질적인 피어슨 상관계수는 높지 않을 수도 있다는 것을 보여주고 있다. 그런데 이러한 경향성에 반하는 사례가 PAIR5이다. 0.500은 PAIR1의 0.439보다 큰 값일 뿐만 아니라 피어슨 상관계수 값인 0.422보다 훨씬 더 큰 값이다. 이에 대해서는 적절한 설명을 제공하기가 어렵다.

ESF 방식의 결과를 설명하기 전에 ESF 방식이 어떻게 작동하는 지를 보다 명확하게 보여주기 위해 PAIR5를 사례로 도해하고자 한다. 식(15)는 각 변수가 네 개의 하위 부분으로 구분될 수 있음을 보여주고 있는데 표 3은 PAIR5의 두 변수가 어떻게 네 부분으로 나뉘어지는지를 시각적으로 보여주고 있다. 우선 두 변수 모두 동일한 평균값(1.838)을 공유한다. 공통 요소는 8번 고유벡터⁶⁾와 관련되어 있는데, 두 변수 각각을 8번 고유벡터로 회귀분석 했을 때의 예측값(모런 통계량 0.324)이다. 변수 x 의 특수 요소는 5번 고유벡터와 관련되어 있는데, 마찬가지로 5번 고유벡터로 변수 x 를 회귀분석 했을 때의 예측값(모런 통계량 0.678)이다. 변수 y 의 고유 요소는 두 개의 고

유벡터(2번 고유벡터와 4번 고유벡터)와 관련되어 있는데 변수 y 를 두 개의 고유벡터로 회귀분석 했을 때의 예측값(모런 통계량 0.823)이다. 마지막의 잔차는 앞의 세 요소를 합하고 그것을 원변수에서 차감했을 때 남는 것이다. 잔차의 모런 통계값은 각각 -0.134와 -0.127로 공간적 자기상관이 존재하지 않는다. 이 두 잔차 간의 피어슨 상관계수가 바로 ESF 기법에 의거한 공간필터 상관계수이다(0.190).

다시 표 2로 돌아가면, ESF 방식이 SAR 방식보다 훨씬 더 일관성 있는 결과를 보여주고 있음을 알 수 있다. 가장 큰 값은 PAIR1의 0.422이며, 가장 작은 값은 PAIR8의 0.011이다. L^* 값이 커질수록 공간필터 상관계수가 감소하는 거의 완벽한 경향성을 보여주고 있다. 앞의 SAR 방식의 결과에서 문제가 되었던 PAIR5도 경향성과 일치하는 결과를 나타내고 있다. PAIR6와 PAIR7 사이에서 다소 경향성에 반하는 결과가 나왔는데, 이는 앞에서 언급한 개별 패턴의 일변량 공간적 자기상관과 관련되어 있는 것으로 보인다. PAIR6는 유의미한 공간적 자기상관을 보이는 y ($p=0.006$)와 통계적 유의성에 근접한 x ($p=0.082$)로 구성되어 있는 반면⁷⁾, PAIR7은 매우 유의미한 x 와 유의미하지 않은 y 로 구성되어 있는 차이가 있다. 아마도 ESF 방식은 두 패턴의 개별 일변량 공간적 자기상관에 보다 민감하고, L^* 통계량은 전체적인 이변량 공간적 자기상관에 보다 민감하기 때문에 이러한 결과가

표 2. 공간필터 상관계수와 수정 *t*-검정 결과

쌍	이변량 통계량		공간필터 상관계수		수정 <i>t</i> -검정	
	<i>r</i>	<i>L</i> *	SAR	ESF	<i>df</i>	<i>p</i> -값
PAIR1	0.422	0.000	0.439	0.422	29.8	0.017*
PAIR2		0.050	0.411	0.422	43.4	0.004**
PAIR3		0.100	0.384	0.357	33.1	0.012*
PAIR4		0.150	0.384	0.303	35.1	0.009**
PAIR5		0.200	0.500	0.190	31.8	0.013*
PAIR6		0.250	0.337	0.120	25.6	0.026*
PAIR7		0.302	0.279	0.128	20.7	0.047*
PAIR8		0.371	0.218	0.011	13.1	0.115

(*와 **는 각각 0.05와 0.01 수준에서 유의함을 나타냄.)

표 3. ESF 방식의 도해(PAIR5의 경우)

PAIR5	원변수	분해			
		평균	공통 요소	특수 요소	잔차
PAIR5-X					
PAIR5-Y					

(음영의 차이는 각 패턴 내에서의 상대적인 값의 높낮이만을 나타냄.)

나타난 것으로 판단된다.

ESF 방식이 SAR 방식에 비해 갖는 단점 역시 지적 될 필요가 있는데, 상관계수 값이 SAR 방식에 비해 너무 낮다는 점에 주목할 필요가 있다. 특히 PAIR5 이후 상관계수 값이 극도로 줄어드는데, 특히 PAIR8의 상관계수 값은 0.011에 불과하다. 이는 ESF가 일종의 과필터링(over-filtering)을 하고 있을 수도 있다는 의심의 근거가 된다. 공간적 의존성은 1차 효과(first-order effect)와 2차 효과(second-order effect) 모두에 의해 발생하는데(Bailey and Gatrell, 1995), 엄밀히 말하면 공간적 자기상관은 2차 효과와만 관련되어 있다. 다시 말하면 공간적 자기상관 없이도 공간 변수는 본원적으로 특정한 패턴을 보유했을 수

있다. 이런 관점에서 보면 ESF 방식은 1차 효과와 2차 효과의 구분 없이 과도하게 공간적 패턴을 제거하는 방법론일 수 있다. 이와 관련해서는 보다 면밀한 후속 연구가 뒤따라야 할 것이다.

마지막으로 수정 *t*-검정 결과를 살펴보면, PAIR1~PAIR4에서 다소 일관성이 없는 결과가 도출되었지만, PAIR4~PAIR8에서는 매우 체계적인 결과가 도출되었다. 즉, 이변량 공간적 자기상관이 높아질수록 자유도는 하락하고, 유의확률은 높아진다. 특히 PAIR4는 표준적인 피어슨 상관계수의 통계적인 결과(자유도와 유의확률)와 거의 동일한 결과를 보여주고 있다. 유의확률은 계속 높아져 PAIR8에 이르러서는 마침내 5% 수준에서 유의하지 않다는 결과가 나타난다.

따라서 수정 t -검정의 결과는 L^* 와 공간필터 상관계수 기법이 보여준 결과에 상당히 조응하는 것으로 판단된다. 그러나 PAIR1~PAIR4의 결과에 대해서는 좀 더 심도 깊은 분석이 필요할 것으로 보인다. 자유도가 가장 높게 측정된 것은 PAIR2인데, L^* 의 기준으로 보면 기댓값에 가장 근접한, 즉 공간적 무작위성이 가장 두드러진 패턴 쌍이다. 이는 수정 t -검정이 공간적 자기상관이 없는 패턴에 매우 민감하다는 점을 함축하고 있다. 이는 PAIR1과 PAIR3를 비교해 보면 보다 명확해 지는데, PAIR1은 약한 음의 공간적 자기상관을 대변하고 있고 PAIR3은 약한 양의 공간적 자기상관을 대변하고 있지만, 둘을 수정 t -검정이 구분하지 못하고 있는 것으로 보인다. 또한 본 연구에서 사용된 R 패키지는 연결성에 기반한 공간근접성행렬을 사용하지 않고 거리-기반의 공간근접성행렬을 사용하는데, 비일관적인 연구 결과가 이와 관련되어 있을 수도 있다.

4. 결론

본 연구는 두 변수 간의 상관성을 측정하는데 지배적인 통계기법으로 사용되어 온 피어슨 상관계수를 공간화하는 방식에 대해 다루었다. 이변량 공간적 자기상관이 존재할 경우, 피어슨 상관계수값과 그것에 대한 유의성 검정 결과가 갖는 통계학적 의미는 훼손될 수 밖에 없다. 본 연구는 이변량 상관관계에서의 공간적 자기상관의 문제를 해결하기 위해 제시된 세 가지 연구 기법(수정 t -검정, 공간필터 상관계수, 이변량 공간적 자기상관 통계량)에 대한 상세한 리뷰를 제공했으며, 다소 독립적으로 발전해 온 세 기법이 얼마나 일관성 있는 결과를 보여주는지를 실험 연구를 통해 살펴보고자 했다. 주요 결과는 다음의 두 가지이다. 첫째, 몇몇 예외를 제외한다면, 세 가지 접근법은 상당한 정도의 상호 일관성을 갖는 결과를 보여주었다. 즉, L^* 에 의거해 높은 이변량 공간적 자기상관을 보여주는 패턴 쌍일수록 낮은 공간필터 상관계수, 작은 유효표본크기(자유도), 높은 유의확률을 보여주었

다. 둘째, L^* 와 가장 일관성 있는 결과를 보여준 것은 ESF에 기반한 공간필터 상관계수 기법이었다. 즉, L^* 가 커질수록 공간필터 상관계수가 감소하는 거의 완벽한 경향성을 보여주었다.

본 연구의 가장 큰 의미는 피어슨 상관계수가 본질적으로 비공간적인 통계량임을 명확히 하고, 이 문제점을 해결하기 위해 제안되어 온 세 가지 접근법이 개별적 특성에도 불구하고 서로 일관성 있는 결과를 보여준다는 점을 실험 연구를 통해 밝혔다는 점이다. 부수적인 가치로는, 피어슨 상관계수 자체를 사용하거나 그것에 기반한 다변량 통계기법을 적용하는 연구의 경우, 그것의 통계학적 결과를 해석할 때 보다 더 많은 주의를 기울여야 한다는 점을 다시 한 번 부각시켰다는 점을 들 수 있다. 그러나, 이러한 가치에도 불구하고 본 연구는 기본적으로 시론적인 성격을 가질 수 밖에 없는 명백한 한계를 지니고 있다. 즉, 본 연구의 결과는 제한된 실험 환경에서 도출된 것이기 때문에 그 해석에서 보다 세심한 주의를 기울여야 할 명백한 필요성이 존재하는 것이다.

이를 극복하기 위해 후속 연구가 필수적인데, 두 가지 정도를 생각해 볼 수 있다. 첫째, 분석 디자인을 완전한 형태의 시뮬레이션을 포용하는 방식으로 확장할 필요가 있다. 본 연구에서처럼 8개의 전형적인 이변량 공간적 자기상관 수준을 설정하고 연구를 진행할 것이 아니라 랜덤화 과정을 통해 모든 가능한 이변량 공간적 자기상관 수준을 보이는 패턴쌍을 도출함으로써 보다 일반화가 가능한 연구로 나아갈 필요가 있는 것이다. 특히, L^* 통계량과 ESF 방식에 의거한 공간필터 상관계수의 관계를 정식화할 수 있다면 이 분야의 연구 수준을 한단계 진일보 시킬 수도 있을 것이다. 둘째, ESF 방법론에 기반한 ‘회귀계수분해(correlation coefficients decomposition)’ 기법(Griffith, 2010; Griffith and Paelinck, 2011; Chun and Griffith, 2013)의 가능성을 보다 완전한 형태의 시뮬레이션 디자인 속에서 탐색하는 것이다. ESF 방법론은 본 연구에서 다루어진 공간필터 상관계수의 산출 방식을 제공해 줄 뿐만 아니라 피어슨 상관계수를 다섯 개의 서로 다른 하위 상관계수로 분해하는 방식을 제공해 준다. 이를 통해 어느 요소가 피어슨 상관계수를 어느

방향으로 얼마만큼 팽창시키거나 위축시키는지에 대한 보다 면밀한 분석이 가능해질 것이다.

주

- 1) 최근에는 공간적 회귀분석 외에 공간적 주성분분석(spatial principal components analysis) 기법의 발달이 눈에 띈다 (Demšar *et al.*, 2013; Lee and Cho, 2014; Lee, 2015).
- 2) 표준점수를 계산하는 과정에서 모 표준편차가 아닌 표본 표준편차를 사용하면 수식의 n 이 $n-1$ 로 바뀌어야 하지만, 산출값에 변화가 없고 피어슨 상관계수가 표준점수 곱의 '평균'이라는 점을 강조한다는 측면에서 모 표준편차를 사용한 이 수식을 사용하고자 함.
- 3) 공간적 자기상관이 없다면 1이고 양의 공간적 자기상관이 있다면 1 이상의 값을 가짐.
- 4 여기서 '=' 기호는 이변량 연관성을 나타내는 것으로 사용하고자 함.
- 5) 37개의 지점에서 이변량 연관성이 모두 다르다면 가능한 모든 쌍의 개수는 37!이어야 한다. 그런데 1=1(변수 A와 변수 B가 모두 1인 경우) 연관을 보인 공간단위가 7개, 1=2가 6개, 2=1이 5개, 2=2가 8개, 2=3이 4개, 3=2가 3개, 3=3이 3개 이므로 37!/(7!6!5!8!4!3!3!)이 됨.
- 6) 가상 데이터가 모두 37개의 육각형으로 이루어져 있으므로 총 37개의 고유벡터가 추출되는데, 모런 통계량에 의거했을 때 양의 공간적 자기상관이 가장 높은 고유벡터를 1번 고유벡터, 음의 공간적 자기상관이 가장 높은 고유벡터를 37번 고유벡터라 부르기로 함.
- 7) 모런과 기어리 통계량을 기준으로 할 경우이며, S^* 를 기준으로 하면 PAIR6-X도 유의함($p=0.016$).

참고문헌

국토교통과학기술진흥원, 2018, 공간정보 SW활용을 위한 오픈소스 가공기술개발 5차년도 연차실적계획서(내부자료).

이상일, 2007, "거주지 분화에 대한 공간통계학적 접근 (I): 공간 분리성 측도의 개발," *대한지리학회지*, 42(4), 616-631.

이상일, 2008, "거주지 분화에 대한 공간통계학적 접근 (II): 국지적 공간 분리성 측도를 이용한 탐색적

공간데이터 분석," *대한지리학회지*, 43(1), 134-153.

이상일·조대현·이민파, 2015, "일변량 공간연관성통계량에 대한 비교 연구 (I): 전역적 S 통계량을 중심으로," *한국지리학회지*, 4(2), 329-345.

이상일·조대현·이민파, 2016, "일변량 공간연관성통계량에 대한 비교 연구 (II): 국지적 S_i 통계량을 중심으로," *한국지리학회지*, 5(3), 375-396.

이상일·조대현·이민파, 2017, "공간적 자기상관 통계량의 고유벡터 간 비교 연구: 공간근접성행렬의 효과와 공간적 회귀분석에의 함의를 중심으로," *대한지리학회지*, 52(5), 645-660.

이화정·이상일·조대현, 2013, "거주지 이동을 통한 학교 선택의 공간성에 관한 연구: 서울시 초등학교의 전학 양상을 사례로 한 시론적 분석," *대한지리학회지*, 48(6), 897-913.

Anselin, L., 2009, Spatial regression, in Fotheringham, A. S. and Rogerson, P. (eds.), *The SAGE Handbook of Spatial Analysis*, SAGE, London, 255-275.

Anselin, L. and Rey, S. J., 2014, *Modern Spatial Econometrics in Practice*, GeoDa Press, Chicago.

Bailey, T. C. and Gatrell, A. C., 1995, *Interactive Spatial Data Analysis*, Longman, Harlow.

Bivand, R., 1980, A Monte Carlo study of correlation coefficient estimation with spatially correlate observations, *Questiones Geographical*, 6, 5-10.

Boots, B. and Tiefelsdorf, M., 2000, Global and local spatial autocorrelation in bounded regular tessellations, *Journal of Geographical Systems*, 2(4), 319-348.

Chun, Y. and Griffith, D. A., 2013, *Spatial Statistics & Geostatistics*, SAGE, Los Angeles.

Chun, Y., Griffith, D. A., Lee, M., and Sinha, P., 2016, Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters, *Journal of Geographical Systems*, 18(1), 67-85.

Clifford, P. and Richardson, S., 1985, Testing the association between two spatial processes, *Statistics and Decisions*, 2 (Supplementary issue), 155-160.

Clifford, P., Richardson, S., and Hémon, D., 1989, Assessing the significance of the correlation between two spatial processes, *Biometrics*, 45(1), 123-134.

Demšar, U., Harris, P., Brunson, C., Fotheringham, A.

- S., and McLoone, S., 2013, Principal component analysis on spatial data: an overview, *Annals of the Association of American Geographers*, 103(1), 106-128.
- Dutilleul, P., 1993, Modifying the t test for assessing the correlation between two spatial processes, *Biometrics*, 49(1), 305-314.
- Fotheringham, A. S., Brunsdon, C., and Charlton, M., 2002, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, John Wiley & Sons, Hoboken.
- Getis, A. and Griffith, D. A., 2002, Comparative spatial filtering in regression analysis, *Geographical Analysis*, 34(2), 130-140.
- Griffith, D. A., 1980, Towards a theory of spatial statistics, *Geographical Analysis*, 12(4), 325-339.
- Griffith, D. A., 1993, Which spatial statistics techniques should be converted to GIS functions? in Fischer, M. and Nijkamp, P. (eds.), *New Directions in Spatial Econometrics*, Springer, Berlin, 101-114.
- Griffith, D. A., 1996, Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data, *The Canadian Geographer*, 40(4), 351-367.
- Griffith, D. A., 2000, A linear regression solution to the spatial autocorrelation problem, *Journal of Geographical Systems*, 2(2), 141-156.
- Griffith, D. A., 2003, *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization*, Springer, Berlin.
- Griffith, D. A., 2010, Spatial filtering, in Fischer, M. M. and Getis, A. (eds.), *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, Springer, New York, 301-318.
- Griffith, D. A., 2017, Spatial filtering, in Shekhar, S., Xiona, H., and Zhou, X. (eds.), *Encyclopedia of GIS*, Volume 3, 2nd edition, Springer, New York, 2018-2031.
- Griffith, D. A. and Amrhein, C. G., 1997, *Multivariate Statistical Analysis for Geographers*, Prentice Hall, Upper Saddle River.
- Griffith, D. A. and Chun, Y., 2014, Spatial autocorrelation and spatial filtering, in Fischer, M. M. and Nijkamp, P. (eds.), *Handbook of Regional Science*, Springer, Berlin, 1477-1507.
- Griffith, D. A. and Paelinck, J. H. P., 2011, *Non-standard Spatial Statistics and Spatial Econometrics*, Springer, Berlin.
- Haining, R. P., 1980, Spatial autocorrelation problems, in Herbert, D. T. and Johnston, R. J. (eds.), *Geography and the Urban Environment*, Wiley, New York, 1-44.
- Haining, R. P., 1991, Bivariate correlation with spatial data, *Geographical Analysis*, 23(3), 210-227.
- Lee, S.-I., 2001a, Spatial Association Measures for an ES-DA-GIS Framework: Developments, Significance Tests, and Applications to Spatio-Temporal Income Dynamics of U.S. Labor Market Areas, 1969-1999, Ph.D. Dissertation, The Ohio State University.
- Lee, S.-I., 2001b, Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I , *Journal of Geographical Systems*, 3(4), 369-385.
- Lee, S.-I., 2004a, Spatial data analysis for the U.S. regional income convergence, 1969-1999: A critical appraisal of β -convergence, *Journal of the Korean Geographical Society*, 39(2), 212-228.
- Lee, S.-I., 2004b, A generalized significance testing method for global measures of spatial association: an extension of the Mantel test, *Environment and Planning A*, 36(9), 1687-1703.
- Lee, S.-I., 2009, A generalized randomization approach to local measures of spatial association, *Geographical Analysis*, 41(2), 221-248.
- Lee, S.-I., 2015, Some elaborations on spatial principal components analysis, Annual Meeting of the Association of American Geographers, April 21-25, Chicago, USA (April 25 presented), p.406.
- Lee, S.-I., 2017, Correlation and spatial autocorrelation, in Shekhar, S., Xiona, H., and Zhou, X. (eds.), *Encyclopedia of GIS*, Volume 1, 2nd edition, Springer, New York, 360-368.
- Lee, S.-I. and Cho, D., 2014, Developing a spatial principal components analysis, Annual Meeting of the Association of American Geographers, April 8-12,

- Tampa, Florida, USA (April 11 presented), p.267.
- Osorio, F., Vallejos, R., Cuevas, F., and Mancilla, D., 2018, Package 'SpatialPack' (<https://cran.r-project.org/web/packages/SpatialPack/SpatialPack.pdf>)
- Reich, R. M., Czaplewski, R. L., and Bechtold, W. A., 1994, Spatial cross-correlation of undisturbed, natural shortleaf pine stands in northern Georgia, *Environmental and Ecological Statistics*, 1(3), 201-217.
- Richardson, S. and Hémon, D., 1981, On the variance of the sample correlation between two independent lattice processes, *Journal of Applied Probability*, 18(4), 943-948.
- Tiefelsdorf, M. and Griffith, D. A., 2007, Semiparametric filtering of spatial autocorrelation: The eigenvector approach, *Environment and Planning A*, 39(5), 1193-1221.
- Vallejos, R., Osorio, F., and Cuevas, F., 2013, SpatialPack – An R package for computing spatial association between two stochastic processes defined on the plane. (<http://rvallejos.mat.utfsm.cl/Time%20Series%20I%202013/paper3.pdf>)
- Vallejos, R., Osorio, F., and Bevilacqua, M., 2019, *Spatial Relationships between Two Georeferenced Variables with Applications in R*, Springer, New York (forthcoming).
- Wartenberg, D., 1985, Multivariate spatial correlation: a method for exploratory geographical analysis, *Geographical Analysis*, 17(4), 263-283.
- 교신: 이상일, 08826, 서울특별시 관악구 관악로 1, 서울대학교 사범대학 지리교육과 (이메일: si_lee@snu.ac.kr, 전화: 02-880-9028)
- Correspondence: Sang-Il Lee, Department of Geography Education, College of Education, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea (e-mail: si_lee@snu.ac.kr, phone: +82-2-880-9028)
- 최초투고일 2018. 9. 27
수정일 2018. 10. 12
최종접수일 2018. 10. 17