한국지리학회지 4권 2호 2015(329~345)

일변량 공간연관성통계량에 대한 비교 연구 (I): 전역적 S 통계량을 중심으로*

이상일** · 조대헌*** · 이민파****

Comparing Univariate Spatial Association Statistics (I): Focusing on Global Lee's S Statistics*

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

요약 : 본 연구의 주된 목적은 전역적 공간연관성통계량으로서의 S 통계량의 특성을 기존의 모런 통계량과 기어리 통계량과의 비교를 통해 밝히는 것이다. 일반 통계량으로서의 S, 그리고 공간근접성행렬에 따른 파생 통계량으로서의 S^0와 S^*가 제안되고, 둘 중 보다 적절한 통계량으로서 후자의 사용이 옹호된다. S^* 통계량은 원 변수의 분산에 대한 공간이동평균 벡터의 분산의 비로 정의되며, 원 변수가 높은 양의 공간적 자기상관을 보이는 경우 원 변수의 분산이 적게 감소하기 때문에 통계치가 1에 가까워진다. S^* 통계량의 특성을 파악하기 위한 분석 방법론으로 중심적률 추출법과 고유치 및 고유벡터 추출법이 사용되었다. 전자는 통계량의 분포특성을 파악하기 위한 것이고, 후자는 통계량의 가능치 범위를 계산하기 위한 것이다. 비교 분석을 위해 세 종류(삼각형, 사각형, 육각형)의 정다각 테셀레이션을 세 가지의 서로 다른 샘플 크기(64, 256, 1,024)로 구성하여 사용하였다. 실질적인 공간분석에서의 함의를 파악하기 위해 우리나라 7대 대도시의 읍면동 공간단위를 분석하였다. 가장 중요한 연구결과는 다음의 두 가지이다. 첫째, S^*는 다른 통계량과 달리 공간단위의 형태, 연접성 유형, 공간근접성행렬의 종류 등에 상관없이 일정한 가능치 범위를 보여주었다. 이것은 S^* 통계량의 가장 큰 장점으로, 다양한 활용성을 기대케 하는 속성이다. 둘째, S^*는 왜도와 첨도가 상대적으로 높아 비록 샘플의 크기가 증가 하더라도 정규근사가 가지는 유의성 검정법으로서의 타당성이 낮은 것으로 판명되었다. 이것은 S^* 통계량의 가장 큰 단점으로, 보다 고차적인 유의성 검정 방법이 도입될 필요가 있음을 보여주고 있다.

주요어 : 공간적 자기상관, 공간연관성통계량, 공간근접성행렬, 모런 통계량, 기어리 통계량, S 통계량, 정다각 테셀레이션, 공간적 고유치와 고유벡터

Abstract : The main objective of this paper is to elucidate the characteristics of a new spatial association statistic, S, in comparison with Moran's I and Geary's c. A general statistic, S, is defined and two derivative statistics, S^0 and S^*, are subsequently proposed with a strong guidance for an exclusive use of the latter. S^* is defined as a rate of the variance of one variable's spatial moving average vector to the original variance suggesting that the presence of a strong positive spatial autocorrelation results in a smaller reduction in variance which leads to a higher S^* with a culminating point of 1 in a theoretical sense. In order to examine the properties of S^*, two methods are introduced; one is to derive the first four central moments and the other is to extract eigenvalues and eigenvectors. The former aims at determine the distributional characteristics of the statistics, and the latter

*본 연구는 국토교통부 국토공간정보연구사업의 연구비지원(과제번호14NSIP-B080144-01)에 의해 수행되었습니다.

**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, dhncho@gmail.com)

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

seeks to the obtain their feasible ranges. Regular tessellations of triangles, squares, and hexagons with three different sample sizes (64, 256, 1,024) are generated and used for an investigation. The smallest administrative spatial units for the 7 big cities in South Korea are also utilized to examine the practical research implications. The major findings are twofold. First, S^*, unlike other spatial association statistics, turns out to yield a constant feasible range of zero to one regardless of different spatial unit shapes, different contiguity types, and different spatial proximity matrices. This is the most important merit of the statistic convincing its usability. Second, the skewness and kurtosis of S^* are considerably deviant form the norms with even a large sample size such that the normal approximation based on the first two moments may not be valid. This is the most important defect of the statistic precipitating the use of more advanced significance testing procedures.
Key Words : Spatial autocorrelation, Spatial association statistics (SAS), Spatial proximity matrix (SPM), Moran's I, Geary's c, Lee's S statistics (S^0 and S^*), Regular tessellations, Spatial eigenvalues and eigenvectors

I. 서론

공간적 의존성(spatial dependence) 혹은 공간적 자기상관(spatial autocorrelation)은 토블러(Tobler, 1970)가 '지리학의 제1법칙(The first law of geography)'이라고 부른, "모든 것은 다른 모든 것과 연관되어 있다. 그러나 가까이 있는 것은 멀리 떨어져 있는 것보다 더 많이 연관되어 있다"라는 공리가 의미하는 바이다. 간단히 말하면 관측개체들이 공간상에 인접해 있으면 그 관측개체가 보유한 속성도 유사한 경향이 있다는 것이다(Anselin, 1988; Anselin and Griffith, 1988). 그러나 공간적 의존성과 공간적 자기상관은 개념적으로 다소간 차이가 있다. 공간적 의존성은 '독립관측 가정(independent observations assumption)'의 위배 혹은 결여를 지적하는 것이라면, 공간적 자기상관은 가까이 있는 관측개체 간의 연관성이 더 높은 이유, 즉 인접한 관측개체들 사이에서 발생하는 공간적 상호작용 프로세스에 보다 더 주목한다. 그러나 실질적으로 두 개념은 서로 교체 가능한 것으로 사용되고 있기 때문에 여기서는 공간적 자기상관을 대표 개념으로 사용하고자 한다. 공간적 자기상관은 "관측개체들의 위치 유사성과 속성 유사성 간의 특정한 관련성"으로 정의될 수 있다(Hubert *et al.*, 1981; Getis, 2008; 이상일 등 역, 2009; Lee, 2012; 2015). 다른 말로 표현하면, 공간적 자기상관은 "공간단위 간의 지리적 근접성과 공간단위가 보유한 속성값 간의 수치적 유사성, 이 둘 간의 특정한 관련성"을 의미한다. 만일 지리적 근접성과 수치적 유사성이 양의 관련성을 가진다면, 다시 말해 가까이 있을수록 값이 유사하다면 '양의 공간적 자기상관'이 있다고 말할 수 있고, 그 반대의 경우라면 '음의 공간적 자기상관'이 있다고 말할 수 있다.

이런 의미에서 공간 데이터에 내재되어 있는 공간적 자기상관의 정도를 계측하는 모든 종류의 측도 혹은 통계량은 반드시 두 가지 요소, 즉 공간단위 간의 지리적 관련성을 정의하는 요소와 공간단위가 보유한 속성값 간의 수치적 관련성을 정의하는 요소로 구성되어야만 한다. 이렇게 정의되는 공간통계량을 본 연구에서는 '공간연관성통계량(spatial association statistics, SAS)'이라고 부르고자 한다.[1] 속성값이 등간/비율 척도인 경우, 가장 널리 사용되어 온 SAS는 모런 통계량(Moran, 1948)과 기어리 통계량(Geary, 1954)이다. 이 두 통계량이 제시될 당시에는 '연접비(contiguity ratio)'라는 이름으로 제안된, 단순한 지수 혹은 측도에 불과했으나, 클리프와 오드의 기념비적인 연구(Cliff and Ord, 1969; 1973; 1981)로 통계량으로서의 면모를 갖추게 되었다. 공간적 자기상관이라는 용어가 처음으로 만들어진 것도 클리프와 오드의 1969년 논문의 근간이 된 1968년 학술대회 발표문에서 비롯되었다고 한다(Getis, 2008). 그런데 현재의 활용 상황을 볼 때, 모런 통계량이 다른 어떤 통계량과도 비교가 안될 정도의 독점적 지위를 누리고 있는 것으로 보인다.

이러한 모런 통계량의 독주는 두 가지 점에서 문제가 있는 것으로 판단된다. 첫째, 기어리 통계량은 속성 유사성을 측정하기 위해 차이의 제곱을 사용하는데, 이 점이 공분산을 사용하는 모런 통계량에 비해 상대적인 강점으로 작용할 수 있다는 점이다. 지구통계학(geostatistics)에서 공간적 자기상관을 모델링하기 위해 널리 사용되고 있는 베리오그램(variogram)이 모런 통계량이 아니라 오히려 기어리 통계량과 상동성이 있다는 점을 상기할 필요가 있다. 둘째, 모런 통계량은 '공간 모델링(modeling)'의 관점에서는 강점이 있지만 '공간 탐색(exploration)'의 관점에서는 다른 통계량이 더 강점이 있을 수 있다. 모

런 통계량은 공간적 자기상관을 측정하는 데 있어 해당 공간단위와 그 이웃 공간단위 간의 엄격한 이분법에 기초하고 있다. 여기에 수치적 유사성을 위한 공분산의 사용이라는 특성이 결합됨으로써 회귀분석의 잔차에 대한 공간적 자기상관의 여부를 판정하는 지배적인 통계량으로 활용되고 있다(Upton and Fingleton, 1985; Anselin, 1988; Tiefelsdorf, 2000). 그런데 이러한 특성이 공간 분포의 단순한 '공간 군집성(spatial clustering)'의 정도를 측정하거나 공간 클러스터와 같은 특정 패턴을 탐지하는 목적에서는 약점이 될 수 있다. 이는 국지적 모런 통계량(Moran's I_i)에 비해 게티스-오드의 G_i^* 통계량이 공간 탐색의 목적에서 더 빈번히 활용되는 이유이기도 하다.

이러한 관점에서 게티스-오드가 제안한 일반 G 통계량(Getis and Ord, 1992; Ord and Getis, 1995)이 대안이 될 수 있다. 그러나 이 전역적 통계량의 가장 큰 단점 중의 하나는 G_i^*가 자신의 LISA(local indicators of spatial association, 국지적 공간연관성 지수)가 아니라는 점이다. 왜냐하면 Anselin이 말한 LISA의 두 번째 조건, 즉 "LISA의 합은 전역적 통계량과 비례 관계에 있어야 한다"(Anselin, 1995, 94)를 충족시키지 못하기 때문이다. 이러한 의미에서 Lee(2001a; 2001b; 2004; 2008; 2009)가 제안한 S 통계량은 중요한 대안이 될 수 있다. S 통계량은 중심 공간단위와 주변 공간단위 간의 완고한 구분에 기초하고 있지 않으며, LISA의 정의를 만족시키는 국지 통계량 S_i로 쉽게 분해된다(Lee, 2008; 2009). 그런데 S 통계량이 어떤 특성을 갖고 있는지에 대해서는 거의 논의되지 않았다. 따라서 본 논문의 주된 연구목적은 전역적 SAS로서의 S 통계량의 특성을 기존의 모런 통계량과 기어리 통계량과의 비교를 통해 밝히는 것이다. 분석 방법론으로 중심적률 추출법과 고유치 및 고유벡터 추출법을 사용하고자 한다. 비교 분석은 주로 가상 데이터에 대해 이루어지지만, 실제 연구 상황에서의 함의를 살펴보기 위해 우리나라 7대 대도시의 읍면동 단위도 사용하고자 한다.

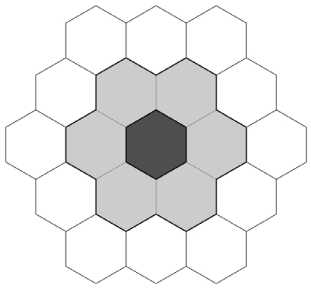
II. 개념적 명료화와 분석 방법론

1. 공간근접성행렬과 공간연관성통계량

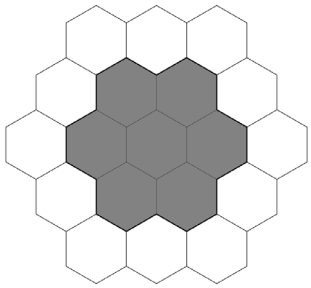
앞에서도 언급한 것처럼, SAS는 공간단위 간의 지리적 관련성을 정의하는 요소와 공간단위가 보유한 속성값 간의 수치적 관련성을 정의하는 요소로 구성되어야만 한다. 그러므로 일반화된 SAS는 이 두 요소를 표현하는 두 매트릭스 간의 '교차곱 통계량(cross-product statistic)'의 형태를 띠게 된다(Mantel, 1967; Hubert *et al.*, 1981; Getis, 1991; Anselin, 1995). 여기서는 전자의 매트릭스에 집중하고자 한다. 관측개체의 공간적 관련성은 '공간근접성행렬(spatial proximity matrix, SPM)'을 통해 정의된다(Bailey and Gatrell, 1995). SPM은 공간단위 간의 위상적, 원근적, 방향적 관련성 등에 의해 정의될 수 있기 때문에 매우 다양한 형태가 가능하다(Bailey and Gatrell, 1995; Getis, 2010). SPM은 연접성(contiguity)에 기반할 수도 있고, 거리(distance)에 기반할 수도 있다. SPM은 1과 0으로만 구성된 이항 매트릭스일 수도 있고, 행-표준화된(row-standardized) 확률론적 매트릭스일 수도 있다.[2)]

그런데 공간데이터분석 혹은 공간통계학이 전통적으로 무시해온 SPM 다양성의 원천이 있는 데, 그것이 바로 주대각(main diagonal) 요소가 0인지 아닌지의 여부이다(Lee, 2004). 주대각 요소가 0이라는 것은 중심 공간단위와 주변 공간단위를 엄격히 구분한다는 것을 의미한다(그림 1(a)). 이에 반해 주대각 요소가 0이 아니라는 것은 중심 공간단위와 주변 공간단위를 합친 하나의 국지세트(local set)를 분석의 단위로 상정한다는 것을 의미한다(그림 1(b)).[3)] 그런데 지배적인 관례는 SPM에 대한 문자로 $\mathbf{W}$를 사용하고, 주대각 요소는 무조건 0으로 간주한다는 것이다(Hubert *et al.*, 1981; Cliff and Ord, 1981; Getis, 2010).

그러나 다양한 공간통계학적 기법들은 주대각 요소가 0이 아닌 SPM의 사용을 권장하고 있다. 우선, 다양한 종류의 공간평활화(spatial smoothing) 기법은 한 공간단위의 값을 추정하기 위해 그 공간단위의 값과 주변 공간단위의 값을 동시에 고려한다(Haining, 2003). 또한 이산적 카운트 데이터에 대한 SAS로 제안된 탱고 통계량(Tango, 1995)이나 로저슨의 공간적 카이-스퀘어 통계량(spatial chi-square statistic)(Rogerson, 1999)도 모두 주대각 요소가 0이 아닌 SPM를 상정하고 있다. 또한 '지리가중회귀분석(geographically weighted regression, 이하 GWR)'의 가중치 설정 과정에서 볼 수 있듯이(Fotheringham *et al.*, 2002), 거리-기반 SPM을 커널 함수(kernel functions)를 통해 재정의하는 경우 대부분 주대각 요소가 0이 아닌 SPM이 활용되고 있다. 그런데 이러한 논의에서 가장



(a) 주대각 요소가 0인 경우



(b) 주대각 요소가 0이 아닌 경우

그림 1. 주대각 요소에 따른 두 가지 공간근접성행렬 구성 방식

중요한 것은 게티스와 오드가 제안한 G_i와 G_i^* 통계량의 구분이다(Getis and Ord, 1992; Ord and Getis, 1995). 전자는 주대각 요소가 0인 SPM을, 후자는 주대각 요소가 0이 아닌 SPM을 활용한다. 제안의 초기에 저자들이 이미 후자가 전자에 비해 클러스터 탐지에 더 효과적이라고 얘기했던 바처럼(Getis and Ord, 1996), 이후 전자는 모든 논의에서 사라지고, 후자만이 널리 사용되고 있다.

SPM의 주대각 요소는 무조건 0이어야 한다는 근거 없는 관행을 철폐하고, SPM의 일반성을 고양시킨다는 의미에서 SPM에 대한 기호로 $\mathbf{V}$를 사용하고자 한다(Tiefelsdorf, 2000; Lee, 2004; 2009). 그리고 SAS에 본질적인 차이를 발생시키는 것이 SPM의 주대각 요소라는 점을 강조하기 위해 $\mathbf{V}$를 $\mathbf{V}^0$와 $\mathbf{V}^*$로 구분하고자 한다. 예를 들어, 모런 통계량의 SPM이 이항의 연접성 기반 SPM이건, 그것의 행-표준화 SPM이건 $\mathbf{V}^0$의 형태를 취하는 한 모런 통계량의 본질은 변하지 않는다. 그러나 모런 통계량에 주대각 요소가 0이 아닌 $\mathbf{V}^*$ 형태 SPM이 적용되면 통계량의 특성은 현저하게 달라진다. 따라서, 이론적으로 말하면, 일반 모런 통계량은 I이지만 실질적으로는 I^0와 I^*로 구분될 수 있고, 기어리 통계량도 마찬가지로 일반 통계량 c가 c^0와 c^*로 구분될 수 있다. 단지 모런 통계량과 기어리 통계량이 명시적으로 $\mathbf{V}^0$ 형태를 선험적으로 취하고 있기 때문에 다른 설명이 제시되지 않는 한 I와 c가 각각 I^0와 c^0만을 의미하게 되는 것이다. 이러한 논의를 바탕으로 전역적 일변량 SAS를 정리하면 표 1과 같다. 세 수식은 각각 일반 모런 통계량(General Moran's I), 일반 기어리 통계량(General Geary's c), 그리고 일반 S 통계량(General Lee's S)이다.

표 1. 전역적 일변량 SAS

SAS	수식	매트릭스 표현	
		1	2
모런 통계량	$I = \frac{n}{\sum_i \sum_j v_{ij}} \frac{\sum_i \sum_j v_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$	$\frac{\mathbf{z}^T \mathbf{V} \mathbf{z}}{\mathbf{1}^T \mathbf{V} \mathbf{1}}$	$\frac{n}{\mathbf{1}^T \mathbf{V} \mathbf{1}} \frac{\boldsymbol{\delta}^T \mathbf{V} \boldsymbol{\delta}}{\boldsymbol{\delta}^T \boldsymbol{\delta}}$
기어리 통계량	$c = \frac{n-1}{2\sum_i \sum_j v_{ij}} \frac{\sum_i \sum_j v_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2}$	$\frac{n-1}{n} \frac{\mathbf{z}^T (\boldsymbol{\Omega} - \mathbf{V}) \mathbf{z}}{\mathbf{1}^T \mathbf{V} \mathbf{1}}$	$\frac{n-1}{\mathbf{1}^T \mathbf{V} \mathbf{1}} \frac{\boldsymbol{\delta}^T (\boldsymbol{\Omega} - \mathbf{V}) \boldsymbol{\delta}}{\boldsymbol{\delta}^T \boldsymbol{\delta}}$
S 통계량	$S = \frac{n}{\sum_i \left(\sum_j v_{ij} \right)^2} \frac{\sum_i \left(\sum_j v_{ij} (x_j - \bar{x}) \right)^2}{\sum_i (x_i - \bar{x})^2}$	$\frac{\mathbf{z}^T (\mathbf{V}^T \mathbf{V}) \mathbf{z}}{\mathbf{1}^T (\mathbf{V}^T \mathbf{V}) \mathbf{1}}$	$\frac{n}{\mathbf{1}^T (\mathbf{V}^T \mathbf{V}) \mathbf{1}} \frac{\boldsymbol{\delta}^T (\mathbf{V}^T \mathbf{V}) \boldsymbol{\delta}}{\boldsymbol{\delta}^T \boldsymbol{\delta}}$

$\mathbf{V}^0$와 $\mathbf{V}^*$의 구분을 이용하면 S 통계량은 S^0와 S^*의 두 통계량으로 세분화 된다.

$$S^0 = \frac{n}{\sum_i \left(\sum_j v_{ij}^0 \right)^2} \frac{\sum_i \left(\sum_j v_{ij}^0 (x_j - \bar{x}) \right)^2}{\sum_i (x_i - \bar{x})^2} \tag{1}$$

$$S^* = \frac{n}{\sum_i \left(\sum_j v_{ij}^* \right)^2} \frac{\sum_i \left(\sum_j v_{ij}^* (x_j - \bar{x}) \right)^2}{\sum_i (x_i - \bar{x})^2} \tag{2}$$

Lee(2001a; 2001b)는 S 통계량이 일종의 '분산감소계수(variance reducing factor)'라는 점을 밝힌 바 있다. 즉, 원 변수를 공간지체(spatial lag) 벡터나 공간이동평균(spatial moving average) 벡터로 변환하면 평활화 효과로 인하여 분산이 감소하게 되는데, 이 때 원 변수가 높은 양의 공간적 자기상관을 보유하고 있다면 원 변수의 분산이 적게 감소하고, 반대로 높은 음의 공간적 자기상관을 보유하고 있다면 분산은 크게 감소할 것이다(Lee(2001a)의 그림 2 참조). 따라서 일반 S 통계량은 원 변수의 분산에 대한 공간지체 혹은 공간이동평균 벡터의 분산의 비로 정의될 수 있다. 이 때 전자가 사용되면 S^0가 되고, 후자가 사용되면 S^*가 되는 것이다. 그러므로 높은 양의 공간적 자기상관이 있을 경우 통계량은 1에 가까워지고, 반대로 높은 음의 공간적 자기상관이 있을 경우는 0에 가까워지는 특성을 가지게 된다.

SAS 간 비교를 보다 분명하게 하기 위해 이항 연접성 SPM의 행-표준화 버전을 의미하는 $\mathbf{W}^0$와 $\mathbf{W}^*$를 적용하고자 한다. 그렇게 하면 세 SAS는 다음과 같이 재정의 될 수 있다.

$$I = \frac{1}{n} \sum_i z_i \sum_j w_{ij}^0 z_j = \frac{1}{n} \sum_i z_i \tilde{z}_i^0 \tag{3}$$

$$S^0 = \frac{1}{n} \sum_i \left(\sum_j w_{ij}^0 z_j \right)^2 = \frac{1}{n} \sum_i \left(\tilde{z}_i^0 \right)^2 \tag{4}$$

$$S^* = \frac{1}{n} \sum_i \left(\sum_j w_{ij}^* z_j \right)^2 = \frac{1}{n} \sum_i \left(\tilde{z}_i^* \right)^2 \tag{5}$$

식 (3)을 보면, 모런 통계량은 공간단위의 표준화점수(z_i)와 그것의 공간지체(주변 공간단위의 표준화점수들의 가중평균값, $\tilde{z}_i^0$) 간의 곱을 모든 공간단위에 대해 구하고, 그것의 평균값을 취한 것임을 알 수 있다. 같은 방식으로 보면, S^0는 표준화점수의 공간지체의 제곱의 평균으로 정의되고, S^*는 표준화점수의 공간이동평균(해당 공간단위와 주변 공간단위의 표준화점수들의 가중평균값, $\tilde{z}_i^*$)의 제곱의 평균으로 정의됨을 알 수 있다. 본 논자는 게티스-오드의 국지 통계량에서 G_i^*가 G_i에 비해 훨씬 더 선호되는 것과 동일한 이유에서 S^*의 사용을 강력히 제안한다. 실질적으로 S^*는 Leung *et al.*(2003)이 수정 G_i^*라고 부른 것의 전역적 통계량과 거의 동일하다. 그러므로 일반 모런 통계량이 항상 I^0를 의미하는 것과 동일한 방식으로 일반 S 통계량도 별 다른 언급이 없다면 늘 S^*를 의미하는 것으로 규정한다.

표 1에는 세 SAS를 매트릭스를 표현하는 두 가지 방법이 함께 제시되어 있다. 첫 번째 방법은 Lee(2004)가 제안한 것으로, '표준화 벡터(standardized vector)($\mathbf{z}$)'(원 벡터에서 평균을 빼고 표준편차로 나눈 것)를 이용하여 표현하는 방법이고, 두 번째 방법은 '편도 벡터(deviate vector)($\boldsymbol{\delta}$)'(원 벡터에서 평균만 뺀 것)를 통해 '이차형식의 비(ratio of quadratic forms)'로 표현한 것이다(Tiefelsdorf, 2000; Lee, 2008). 전자는 SAS를 두 개의 매트릭스로 분해하는데 유리하고, 후자는 SAS를 회귀분석의 잔차 형식으로 표현하는 데 유리하다. 이 중 후자를 이용하여 다음의 방법론을 정의할 것이다.

2. 중심적률 추출과 고유치 및 고유벡터 추출

식 (1)에 나타나 있는 세 SAS는 모두 다음과 같은 이차형식의 비로 표현될 수 있다.

$$\Gamma = \frac{\boldsymbol{\delta}^T \mathbf{T} \boldsymbol{\delta}}{\boldsymbol{\delta}^T \boldsymbol{\delta}} \tag{6}$$

이 때 $\mathbf{T}$는 표준화된 SPM로 정의될 수 있는데, 표 1의 두 번째 매트릭스 표현으로부터 각 SAS에 대한 해당 매

트릭스를 도출하면 다음과 같다(Lee, 2009).

$$\mathbf{T}(I) \equiv n\frac{\mathbf{V}}{\mathbf{1}^T\mathbf{V}\mathbf{1}},\ \mathbf{T}(c) \equiv (n-1)\frac{(\mathbf{\Omega}-\mathbf{V})}{\mathbf{1}^T\mathbf{V}\mathbf{1}}$$

$$\mathbf{T}(S) \equiv n\frac{\mathbf{V}^T\mathbf{V}}{\mathbf{1}^T(\mathbf{V}^T\mathbf{V})\mathbf{1}} \quad (7)$$

또한 이것으로부터 $\mathbf{K}$ 매트릭스를 다음과 같은 방식으로 도출할 수 있다(Tiefelsdorf, 2000).

$$\mathbf{K} \equiv \mathbf{M}_{(1)}\frac{1}{2}(\mathbf{T}+\mathbf{T}^T)\mathbf{M}_{(1)}$$

$$\mathbf{M}_{(1)} \equiv \mathbf{I} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T \quad (8)$$

여기서 $\mathbf{I}$는 단위행렬(unit matrix)이고, $\mathbf{1}$은 요소가 모두 1인 열벡터(column vector)이다. 이렇게 도출된 $\mathbf{K}$를 이용하면 각 SAS의 통계학적 특성 파악을 위한 두 가지 방법론을 도출할 수 있다.

첫 번째 방법론은 중심적률(central moments) 추출법이다. 각 SAS에 대해 다음의 공식을 적용함으로써 네 개의 중심적률을 구할 수 있다(Henshaw, 1966; 1968; Hepple, 1998; Tiefelsdorf, 2000).

$$\mu_1 = \mathrm{E}(\Gamma) = \frac{\mathrm{tr}(\mathbf{K})}{n-1} \quad (9)$$

$$\mu_2 = \mathrm{Var}(\Gamma) = 2\frac{\left[(n-1)\mathrm{tr}(\mathbf{K}^2) - \mathrm{tr}(\mathbf{K})^2\right]}{(n-1)^2(n+1)} \quad (10)$$

$$\mu_3 = 8\frac{\left[(n-1)^2\mathrm{tr}(\mathbf{K}^3) - 3(n-1)\mathrm{tr}(\mathbf{K})\mathrm{tr}(\mathbf{K}^2) + 2\mathrm{tr}(\mathbf{K})^3\right]}{(n-1)^3(n+1)(n+3)} \quad (11)$$

$$\mu_4 = \frac{12}{(n-1)^4(n+1)(n+3)(n+5)}\begin{Bmatrix}(n-1)^3\left[4\mathrm{tr}(\mathbf{K}^4) + \mathrm{tr}(\mathbf{K}^2)^2\right] \\ -2(n-1)^2\left[8\mathrm{tr}(\mathbf{K})\mathrm{tr}(\mathbf{K}^3) + \mathrm{tr}(\mathbf{K}^2)\mathrm{tr}(\mathbf{K})^2\right] \\ +(n-1)\left[24\mathrm{tr}(\mathbf{K}^2)\mathrm{tr}(\mathbf{K})^2 + \mathrm{tr}(\mathbf{K})^4\right] \\ -12\mathrm{tr}(\mathbf{K})^4\end{Bmatrix} \quad (12)$$

식 (9)와 식 (10)은 각 SAS의 기댓값과 분산을 계산해 주고, 식 (11)과 식 (12)를 이용하면 왜도(skewness)와 첨도(kurtosis)를 다음과 같이 구할 수 있다(Tiefelsdorf, 2000).

$$\beta_1 = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}},\ \beta_2 = \frac{\mu_4}{(\mu_2)^2} \quad (13)$$

이 적률 추출법은 전역적 SAS의 가설검정을 위한 두 가정 중 하나인 정규성 가정(normality assumption)에 의거한 것이고, 또 다른 가정인 랜덤화 가정(randomization assumption)에 의거한 적률 추출에 대해서는 Lee(2004)를 참조할 수 있다.

두 번째 방법론은 고유치 및 고유벡터 추출법이다. $\mathbf{K}$ 매트릭스를 분해하면 고유치(eigenvalues)와 그에 상응하는 고유벡터(eigenvectors)를 추출할 수 있는데, 고유치 스펙트럼은 각 SAS의 '가능치 범위(feasible range)'를 알려주고(de Jong *et al.*, 1984), 각 고유치에 상응하는 고유벡터는 동일 수준의 공간적 자기상관을 가진 공간분포를 산출한다(Boots and Tiefelsdorf, 2000; Griffith, 2003; Tiefelsdorf and Griffith, 2007). $\mathbf{K}$ 매트릭스로부터 n개의 고유치, $\{\lambda_1, \lambda_2, \cdots \lambda_i, \cdots \lambda_n\}(\lambda_i \geq \lambda_{i+1})$를 추출할 수 있는데, 첫 번째 고유치와 마지막 고유치는 해당 SAS의 최대 가능치와 최소 가능치를 나타낸다. 또한 $\mathbf{K}$ 매트릭스로부터 n개의 고유벡터, $\{\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_i, \ldots \mathbf{e}_n\}$를 추출할 수 있는데, 각 고유벡터는 고유치에 상응하는 정도의 공간적 자기상관을 보여주는 벡터이다. 이 고유벡터들은 서로 직교하기 때문에 상관관계가 존재하지 않는다. 따라서 첫 번째 고유치와 고유벡터는 SAS의 최대 가능치와 그것을 나타내는 공간분포이고, 두 번째 고유치와 고유벡터는 첫 번째 고유치와 상관관계가 없는, 최대 가능치와 그것에 상응하는 공간분포이다. 그리고 마지막 고유치와 고유벡터는 앞의 $(n-1)$개의 고유벡터와 상관관계가 없는, 최대 가능치(그러므로 전체적으로 보면 최소 가능치)와 그것에 상응하는 공간분포이다. 따라서 고유벡터를 지도화하면 고유한 특성을 보유한 공간적 자기상관의 패턴을 확인할 수 있다(Boots and Tiefelsdorf, 2000). SAS는 서로 다른 특성을 갖기 때문에(즉, 서로 다른 $\mathbf{T}$ 매트릭스, 그러므로 서로 다른 $\mathbf{K}$ 매트릭스를 가지기 때문에), 도출되는 패턴도 서로 다르게 나타날 것이다.

3. 분석 환경

전역적 일변량 SAS의 비교 분석을 위해 본 연구에서는 오픈소스 GIS 기반의 공간통계분석 도구 및 R을 함께 사용하였다. 구체적으로 오픈소스 GIS 기반의 공간통계분석 도구는 2014년부터 진행 중인 국토교통부 국토공간정보연구 사업을 통해 개발된 것이다(국토교통과학기술진흥원, 2015). 이 도구의 토대가 되는 오픈소스 GIS는 GeoTools 라이브러리(http://geotools.org/) 및 uDig 데스크톱(http://udig.refractions.net)이며, 본 도구 역시 오픈소스 소프트웨어(https://github.com/MapPlus/spatial_statistics_for_geotools_udig)로 개발되어 누구나 이용 가능하다. 또한 이 도구는 데스크톱 GIS인 uDig에 탑재되는 플러그인의 형태를 취함으로써 그래픽 사용자 인터페이스(GUI) 기반의 데스크톱 GIS가 가진 강점을 그대로 유지하고 있다(그림 2).

하지만 보다 중요한 것은 공간통계분석에서 요구되는 다양한 편의 기능과 분석 기능을 포함하고 있다는 점이다. 예를 들어 이 도구에 포함된 유틸리티는 공간통계분석에 흔히 사용되는 가상의 모의 데이터 생성을 지원하는데, 삼각형, 사각형, 육각형 등 다양한 형태의 공간단위의 테셀레이션을 산출한다. 본 연구에서 사용된 가상 데이터 역시 본 공간통계분석 도구를 통해 생성되었다(그림 3 참조). 분석 기능으로는 요약 통계와 같은 비교적 간단한 통계 분석은 물론 최근린 분석과 같은 포인트 패턴 분석이나 다양한 공간적 자기상관 분석을 포함하고 있다. 구체적으로는 모런 통계량, 기어리 통계량, 게티스-오드 통계량, S 통계량과 같은 SAS를 지원하고 있다.

III. 전역적 공간연관성통계량의 비교 분석

1. 정다각 테셀레이션 분석

본 연구에서는 모런 통계량, 기어리 통계량, 그리고 S^* 통계량의 세 통계량을 비교하고자 한다. 비교 분석을 위해 Boots and Tiefelsdorf(2000)가 사용한 방법을 원용하고자 한다. 우선 세 종류의 정다각(삼각형, 사각형, 육각형) 테셀레이션을 세 가지의 샘플 크기(64, 256, 1,024)로 만들어 낸다(그림 3). 이 9개의 가상적 공간적 형상에 대해 앞에서 언급한 방법론을 적용하여 SAS의 특성을 비교 분석하고자 한다. 서로 다른 다각형을 살펴보는 것은 주변 공간단위와의 연결성의 차이에 SAS가 어떻게 반응하는지를 살펴보기 위함이고, 서로 다른 샘플 크기를 살펴보는 것은 점근적 정규성(asymptotic normality)을 검토함으로써 기댓값과 분산 만을 이용하여 가설검정을 행하는 정규근사(normal approximation)가 얼마나 타당한가를 알아보기 위한 것이다. 본 연구는 여기에다 다음

그림 2. 오픈소스 GIS 기반의 공간통계분석 도구

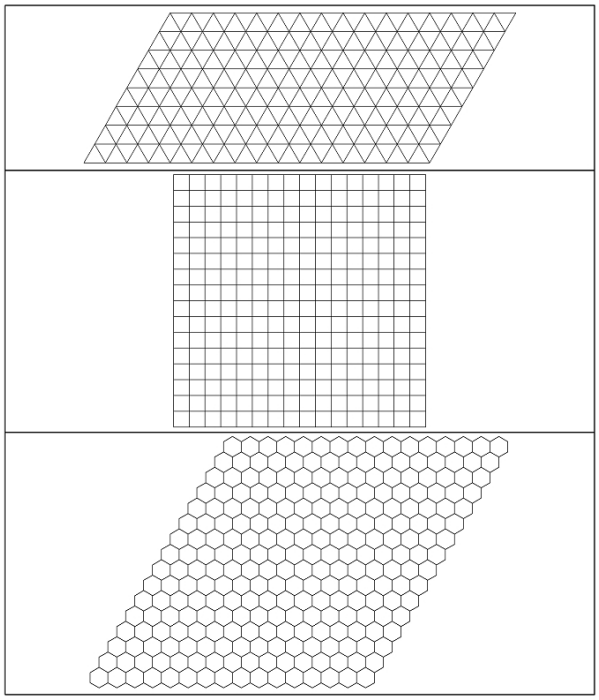


그림 3. 정다각 테셀레이션(삼각형, 사각형, 육각형)(n=256)

출처 : Boots and Tiefelsdorf, 2000:321.

의 두 가지 사항을 더 부가적으로 다루고자 한다. 첫째는 SPM의 영향력을 알아보기 위해 네 종류의 SPM를 사용하고자 한다(SPM이 공간통계적 결론에 미치는 영향에 대해서는 Tiefelsdorf *et al.*(1999) 참조). $\mathbf{C}^0$는 주대각요소가 0인 이항 연접성 SPM, $\mathbf{C}^*$은 $\mathbf{C}^0$의 주대각선에 1을 넣은 SPM, $\mathbf{W}^0$는 $\mathbf{C}^0$의 행-표준화 SPM, $\mathbf{W}^*$은 $\mathbf{C}^*$의 행-표준화 SPM이다. $\mathbf{C}^0$와 $\mathbf{W}^0$은 모런과 기어리 통계량에 대해, $\mathbf{C}^*$와 $\mathbf{W}^*$는 S^* 통계량에 대해 사용하도록 한다. 둘째, 연접성 유형의 효과를 살펴보기 위해 루크(rook) 방식과 퀸(queen) 방식을 비교하고자 한다. 루크 방식은 변이 접하는 경우에만 이웃으로 간주하는 것이고, 퀸은 점으로 연결되어도 이웃으로 간주하는 방식이다. 따라서 평균이웃수는 루크 방식에 비해 퀸 방식에서 훨씬 더 많다.

첫 번째 분석은 세 SAS에 대한 상관관계 분석이다. 이를 위해 256개 육각형 테셀레이션에 대해, 1,000개의 정규 벡터를 무작위로 생성하였다. 각 벡터에 대해 세 SAS를 계산함으로써, 1,000개씩의 통계치를 산출할 수 있었다. 결과는 그림 4에 나타나 있다. 상관관계의 절대 크기 순서는 모런-기어리(-0.9326), 모런-S^*(0.8543), 기어리-S^*(-0.8032)로 나타났다. 이는 S 통계량이 모런이나 기어리 통계량과는 다소 다른 특성을 보이고 있음을 함축하고 있다. 이는 보완적, 혹은 대안적 SAS로서 S^*가 충분한 가치를 가지고 있음을 보여주는 것이다.

두 번째는 세 SAS의 가능치 범위를 비교하는 것이다. 샘플링 크기 256개에 대해, 세 개의 공간단위 형태, 두 개의 연접성 유형, 두 종류의 SPM를 적용하여 비교 분석하였다(표 2). 주요 연구 결과는 다음과 같다. 첫째, 삼각형과 사각형의 루크 유형의 경우에는, 모런 통계량이 거의 -1~1의 값을 가지는 것으로 드러났다. 기어리 통계량의 경우도 동일한 경우에 대해 거의 0~2의 값을 가지는 것으로 드러났다. 이는 이 두 통계량의 특성에 대한 일반적으로 알려진 사실에 정확히 부합하는 것이다. 그러나 상대적으로 평균이웃수가 많은 정육각형의 경우(정육각형의 경우는 루크와 퀸의 구분이 없음)에는 음의 공간적 자기상관을 나타내는 최소치가 절댓값 기준으로 거의 절반 수준으로 떨어진다. 이는 기어리 통계량에서도 동일하게 드러난다. 둘째, 모런과 기어리 통계량의 경우,

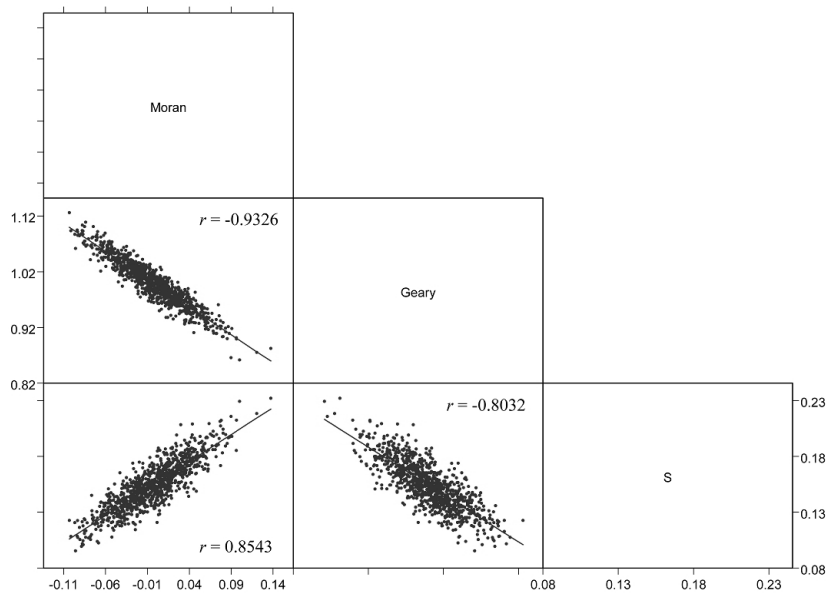


그림 4. 일변량 공간연관성통계량 간의 상관관계 매트릭스(육각형, n=256)

표 2. 일변량 공간연관성통계량의 가능치 범위 비교(n=256)

공간단위 형태	연접성 유형	SPM	통계량 별 가능치 범위					
			모런 통계량		기어리 통계량		S^* 통계량	
			최대	최소	최대	최소	최대	최소
삼각형	루크 (2.813)	$\mathbf{C}^0$ 혹은 $\mathbf{C}^*$	1.0375	-1.0477	2.1031	0.0000	1.0440	0.0000
		$\mathbf{W}^0$ 혹은 $\mathbf{W}^*$	1.0122	-1.0124	2.1357	0.0000	1.0224	0.0000
	퀸 (10.539)	$\mathbf{C}^0$ 혹은 $\mathbf{C}^*$	1.0579	-0.3655	1.4924	0.0000	1.0711	0.0000
		$\mathbf{W}^0$ 혹은 $\mathbf{W}^*$	0.9990	-0.4233	1.4260	0.0000	1.0033	0.0000
사각형	루크 (3.750)	$\mathbf{C}^0$ 혹은 $\mathbf{C}^*$	1.0216	-1.0485	2.1046	0.0000	1.0244	0.0000
		$\mathbf{W}^0$ 혹은 $\mathbf{W}^*$	0.9919	-1.0016	1.9987	0.0000	0.9865	0.0000
	퀸 (7.266)	$\mathbf{C}^0$ 혹은 $\mathbf{C}^*$	1.0319	-0.5319	1.6239	0.0000	1.0294	0.0000
		$\mathbf{W}^0$ 혹은 $\mathbf{W}^*$	0.9929	-0.5233	1.5417	0.0000	0.9936	0.0000
육각형	루크 (5.508)	$\mathbf{C}^0$ 혹은 $\mathbf{C}^*$	1.0403	-0.5330	1.6140	0.0000	1.0489	0.0000
		$\mathbf{W}^0$ 혹은 $\mathbf{W}^*$	0.9988	-0.5414	1.5420	0.0000	0.9986	0.0000

* 연접성 유형 칼럼의 괄호 안의 숫자는 평균이웃수인데, 각 공간단위가 평균적으로 가지는 이웃 공간단위의 수를 의미함. 모런과 기어리 통계량에 대해서는 $\mathbf{C}^0$와 $\mathbf{W}^0$ SPM을, S^* 통계량에 대해서는 $\mathbf{C}^*$와 $\mathbf{W}^*$ SPM을 적용함.

루크와 퀸 유형의 차이가 음의 공간적 자기상관에 대한 가능치 범위에 큰 영향을 끼치는 것으로 드러났다. 보다 평균이웃수가 많은 퀸 방식의 경우 음의 공간적 자기상관의 가능치가 절댓값 기준으로 절반 이하의 수준으로 떨어졌다. 예를 들어 모런 통계량의 경우, 삼각형의 루크 방식 최솟값은 -1.0477이었는데, 퀸 방식의 경우는 약 1/3 정도에 해당하는 -0.3655로 떨어졌다. 이는 특히 음의 공간적 자기상관의 경우 서로 다른 지역에 대한 모런 혹은 기어리 통계치를 단순 비교하는 것이 매우 위험한 것임을 함축하고 있다. 셋째, S^*는 모든 경우에 대해 거

표 3. 일변량 공간연관성통계량의 표본분포 특성 비교

통계량	공간단위 형태	공간단위 수	표본분포 특성			
			기댓값	분산	왜도	첨도
모런 통계량	삼각형	64	-0.015873	0.011369	-0.000158	2.942996
		256	-0.003922	0.002746	-0.000008	2.985230
		1024	-0.000978	0.000670	-0.000000	2.996209
	사각형	64	-0.015873	0.008405	-0.011953	3.038531
		256	-0.003922	0.002052	-0.001414	3.011604
		1024	-0.000978	0.000502	-0.000174	3.002970
	육각형	64	-0.015873	0.005665	0.270432	3.081494
		256	-0.003922	0.001386	0.142315	3.023869
		1024	-0.000978	0.000338	0.071931	3.006016
기어리 통계량	삼각형	64	1.000000	0.012363	0.046237	2.938631
		256	1.000000	0.002890	0.012415	2.984447
		1024	1.000000	0.000689	0.003166	2.996097
	사각형	64	1.000000	0.009100	0.047528	3.041928
		256	1.000000	0.002157	0.011860	3.011332
		1024	1.000000	0.000516	0.002896	3.002901
	육각형	64	1.000000	0.007146	-0.130323	2.997791
		256	1.000000	0.001589	-0.100680	3.007335
		1024	1.000000	0.000364	-0.060712	3.003531
S^* 통계량	삼각형	64	0.272487	0.002590	0.359990	3.159855
		256	0.263399	0.000613	0.189330	3.049962
		1024	0.257250	0.000146	0.096331	3.013630
	사각형	64	0.214815	0.001956	0.492083	3.355581
		256	0.209935	0.000485	0.262867	3.107633
		1024	0.205604	0.000118	0.134854	3.028990
	육각형	64	0.160470	0.001908	0.635339	3.565055
		256	0.154379	0.000462	0.340711	3.171266
		1024	0.149304	0.000110	0.175791	3.046531

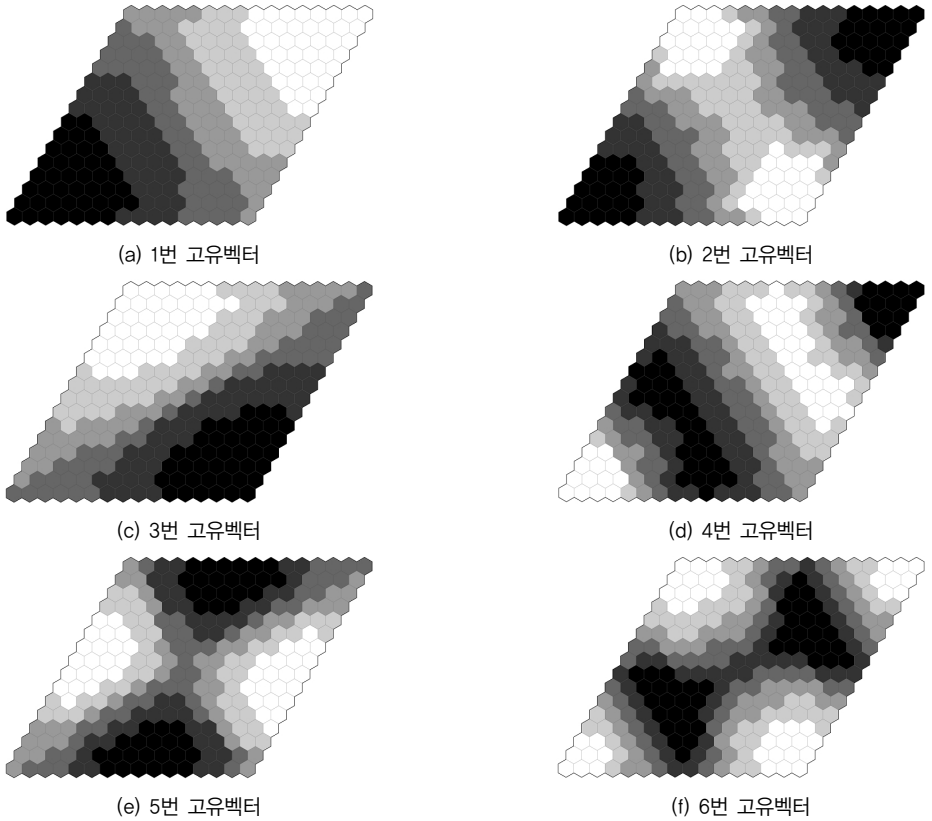
* 연접성 유형은 모두 루크임. 모런 통계량과 기어리 통계량에 대해서는 C^0를, S^* 통계량에 대해서는 W^*를 적용함.

의 일정한 0~1의 범위를 나타내었다. 이는 매우 고무적인 사실이 아닐 수 없는데, 서로 다른 연구지역 간에 통계치를 절대 비교하는데 큰 문제가 없음을 함축하고 있기 때문이다. 넷째, 이항 연접성 SPM과 행-표준화 SPM 사이에는 큰 차이는 없는 것으로 드러났다.

세 번째 분석은 식 (9)~(12)에 의거해 각 SAS의 기댓값, 분산, 왜도, 첨도를 공간단위 형태와 공간단위 수를 달리하면서 비교한 것이다(표 3). 주요 분석 결과는 다음과 같다. 첫째, 모든 경우에 대해 공간단위 수가 증가할수록 분산과 왜도는 0에, 첨도는 3에 가까워진다. 이는 어느 정도 점근적 정규성이 존재함을 의미한다. Boots and Tiefelsdorf(2000)에 따르면, 공간단위 수가 100개 정도를 넘어 가면 정규근사를 통한 유의성 검정에 타당성이 있다고 보고한 바 있다. 그런데, 이러한 점근적 정규성은 모런과 기어리 통계량에 비해 S^* 통계량에서 상대적으로 취약한 것으로 드러났다. 특히 왜도가 심각한데, 1,024개인 경우에도 거의 0.1이 넘는 값을 보여주고 있다. 둘째, 세 SAS는 기댓값에서 서로 다른 특성을 보인다. 모런 통계량의 경우는 기댓값이 공간단위 개수의 함수이고, 기어리 통계량의 경우는 항상 1이며, S^*의 경우는 평균이웃수가 많아질수록, 공간단위의 수가 많아질수록 기댓값이 낮아진다. 즉, 기어리 통계량은 9가지 경우 모두에 대해 기댓값이 1인 반면, 모런 통계량은 오로지 공간단위의 수에 따라 달라지는 3개의 기댓값을 가진

표 4. S^* 통계량에 의거한 공간적 자기상관 상하위 각 6개 고유벡터(육각형, n=256)

	S^*		모런 통계량		기어리 통계량	
	등위	통계치	통계치	등위	통계치	등위
상위 6개	1	0.9986	0.9559	7	0.0180	254
	2	0.9814	0.9294	9	0.0300	253
	3	0.9782	0.9739	6	0.0276	253
	4	0.9445	0.9239	9	0.0490	251
	5	0.9249	0.9394	7	0.0587	250
	6	0.9187	0.9038	9	0.0662	250
하위 6개	251	0.0000	-0.1758	135	1.1508	115
	252	0.0000	-0.1767	135	1.1663	114
	253	0.0000	-0.1770	137	1.1685	114
	254	0.0000	-0.1850	139	1.1718	114
	255	0.0000	-0.1850	139	1.1744	114
	256	0.0000	-0.1785	137	1.1639	114



(a) 1번 고유벡터 (b) 2번 고유벡터

(c) 3번 고유벡터 (d) 4번 고유벡터

(e) 5번 고유벡터 (f) 6번 고유벡터

그림 5. S^* 통계량에 의거한 공간적 자기상관 상위 6개 고유벡터의 공간패턴(육각형, n=256)

다. 하지만 S^*는 9개 경우에 대해 모두 다른 기댓값을 가진다.

네 번째 분석은 S^*의 상하위 각 6개의 고유벡터를 모런과 기어리 통계량이 어떻게 측정하는 지를 알아본 것

이다(표 4). 모런과 기어리 통계량에서의 등위는 S^*에 의거해 상하위 6대 패턴으로 선정된 고유벡터에 대해 모런과 기어리 통계량을 적용해 측정하고, 그것이 각 SAS의 고유치 스펙트럼에서 몇 위에 해당하는지를 따져본 것이다. 이 분석의 가장 중요한 결과는 상위 6개 패턴에 대해서는 모런과 기어리 통계량에 의해서도 높은 양의 공간적 자기상관을 보이는 것으로 측정되었지만, 하위 6개 패턴은 모런과 기어리 통계량에 의하면 중위 정도에 그치는 것으로 드러났다. 그림 5는 상위 6개 고유벡터의 공간패턴을 보여주고 있다.

2. 우리나라 7대 대도시의 읍면동 단위 분석

실질적인 공간분석에서의 함의를 파악하기 위해 우리나라 7대 대도시에 적용하여 각 SAS 별 가능치 범위를 파악하였다(표 5). 평균이웃수는 정육각형의 5.508과 유사한 값을 보여주었다. 인천과 울산이 가장 작은 5.123과 5.179의 값을 보였으며, 서울과 대구가 가장 높은 5.858과 5.799의 값을 보였다. 주요한 분석 결과를 요약하면 다음과 같다. 첫째, 모런 통계량의 경우 최대 가능치는 거의 1정도이지만, 최소 가능치는 C^0를 기준으로 했을 때 대구의 -0.5447에서 인천의 -0.6171에 이르기 까지 다양하다. 앞에서 살펴본 것처럼, 음의 공간적 자기상관의 경우, 도시간 단순 비교에는 위험성이 도사리고 있음을 함축하고 있다. 둘째, 기어리 통계량의 경우, 최대 가능치(음의 공간적 자기상관)에서 매우 큰 편차를 보였다. C^0를 기준으로 했을 때 대전의 1.9415에서 부산의 2.7280까지 그 차이가 매우 크다. 특히 부산의 극단치는 상상하기 어려울 정도의 낮은 공간적 자기상관이 부산에 대해 도출될 수 있음을 시사하는 것이다. 셋째, 표 2의 결과와는 달리 SPM의 종류에 따라 최소 가능치에서 상당한 차이가 있는 것으로 드러났다. 대체적으로 이항 연접성 SMP에 비해 행-표준화 SPM이 더 낮은 값을 보여주었는데, 특히 부산과 대구의 경우 그 차이가 0.15에 달했다. 셋째, 표 2에 나타난 결과와 마찬가지로, S^* 통계량은 모든 도시의 모든 SPM에 대해 0~1의 값을 보여 가장 안정적인 값을 산출하는 것으로 드러났다.

상이한 SPM의 적용은 가능치 범위에 영향을 미칠 뿐만 아니라 고유벡터의 공간패턴에도 영향을 미치게 된다. 이러한 사실을 살펴보기 위해 서울을 사례로 분석하였는데, 원 SPM(한강을 고려하지 않고 행정동의 연접성에만 의존하여 SPM를 구성한 경우)과 한강을 고려하여 연접성을 재정의한 SPM의 경우를 비교하였다. 그림 6에 나타나 있는 패턴은 S^* 기준으로 각각 1.0454와 1.0569

표 5. 우리나라 7대 대도시에 대한 일변량 공간연관성통계량의 가능치 범위 비교

도시	평균이웃수	SPM	통계량 별 가능치 범위					
			모런 통계량		기어리 통계량		S^* 통계량	
			최대	최소	최대	최소	최대	최소
서울 (423)	5.858	C^0 혹은 C^*	1.0959	-0.5770	2.4380	0.0000	1.1116	0.0000
		W^0 혹은 W^*	1.0139	-0.6181	1.8694	0.0000	1.0454	0.0000
부산 (214)	5.579	C^0 혹은 C^*	1.1433	-0.6162	2.7280	0.0000	1.1649	0.0000
		W^0 혹은 W^*	1.0394	-0.7734	1.9351	0.0000	1.0607	0.0000
대구 (139)	5.799	C^0 혹은 C^*	1.0402	-0.5447	2.1172	0.0000	1.0134	0.0000
		W^0 혹은 W^*	0.9895	-0.6934	1.9761	0.0000	1.0003	0.0000
인천 (146)	5.123	C^0 혹은 C^*	1.1737	-0.6171	2.2656	0.0000	1.1845	0.0000
		W^0 혹은 W^*	1.0987	-1.1002	2.4206	0.0000	1.1093	0.0000
광주 (94)	5.638	C^0 혹은 C^*	1.0411	-0.5908	2.3749	0.0000	0.9976	0.0000
		W^0 혹은 W^*	0.9946	-0.5832	1.9034	0.0000	1.0334	0.0000
대전 (77)	5.584	C^0 혹은 C^*	0.9918	-0.5606	1.9415	0.0000	0.9297	0.0000
		W^0 혹은 W^*	0.9591	-0.5572	1.7412	0.0000	0.9548	0.0000
울산 (56)	5.179	C^0 혹은 C^*	1.0121	-0.5924	2.0728	0.0000	0.9460	0.0000
		W^0 혹은 W^*	0.9904	-0.6154	1.7049	0.0000	1.0056	0.0000

* 도시 칼럼의 괄호 안의 숫자는 공간단위의 수임. 연접성 유형은 모두 루크임.

(a) 한강을 고려하지 않은 SPM을 적용한 경우 (b) 한강을 고려한 SPM을 적용한 경우

그림 6. 최대 공간적 자기상관을 보이는 고유벡터의 공간분포

로 거의 동일하지만, 공간패턴 상에서는 확실한 차이가 있음을 보여준다. (a)의 경우는 동북에서 서남으로 높은 값에서 낮은 값이 자연스럽게 이어지지만, (b)의 경우는 한강 이북에는 높은 값이, 한강 이남에는 낮은 값이 분포하는 단절적 패턴을 보여주고 있다.

IV. 결론

본 연구는 전역적 SAS로서의 S 통계량의 특성을 기존의 모런 통계량과 기어리 통계량과의 비교를 통해 밝히는 것이었다. 개념적 차원에서 SPM을 둘러싼 개념적 혼동을 우선적으로 정리하고자 했다. 일반 SPM에 대한 문자기호로서 $\mathbf{V}$를 제안했으며, 주대각 요소가 0인 경우와 그렇지 않은 경우를 구분하기 위해 $\mathbf{V}^0$와 $\mathbf{V}^*$의 사용을 제안했다. 이렇게 함으로써 일반 SAS로서의 모런 통계량, 기어리 통계량, S 통계량을 정의할 수 있었다. S 통계량에 대해서는 S^0와 S^* 중 후자의 사용을 강력히 주장하였다. S^* 통계량은 원 변수의 분산에 대한 공간이동평균 벡터의 분산의 비로 정의되는데, 원 변수가 높은 양의 공간적 자기상관을 보유하고 있다면 원 변수의 분산이 적게 감소하여 통계치가 증대하여 1에 가까워지고, 반대로 높은 음의 공간적 자기상관을 보유하고 있다면 통계치가 0에 가까워지는 특성을 가지는 것으로 드러났다. 그리고 분석 방법론으로 중심적률 추출법과 고유치 및 고유벡터 추출법이 사용되었다. 전자는 SAS의 분포 특성을 비교하기 위한 것이었고, 후자는 SAS별 가능치 범위를 계산하기 위한 것이었다. 비교 분석은 주로 가상적인 데이터에 대해 이루어졌는데, 세 종류의 정다각 테셀레이션을 세 가지의 샘플 크기로 구성하여 사용했다. 실질적인 공간분석에서의 함의를 파악하기 위해 우리나라 7대 대도시의 읍면동 공간단위를 분석하였다.

중요한 연구결과는 다음과 같다. 첫째, S^* 통계량은 모런과 기어리 통계량과 절대값 0.8이 넘는 높은 상관관계를 보였지만, 극단적으로 높은 상관성으로 보이지 않은 것에서 볼 수 있듯이 모런과 기어리 통계량이 측정하지 못하는 공간적 자기상관의 다른 측면도 측정하고 있는 것으로 판단된다. 둘째, 모런과 기어리 통계량은 평균이웃수가 상대적으로 많아 지는 경우(삼각형이나 사각형이 아니라 육각형인 경우, 루크 방식이 아니라 퀸 방식의 경우), 음의 공간적 자기상관에 대한 값이 절댓값 기준으로 절반 이하의 수준으로 떨어지는 것으로 드러났다. 이에 비해 S^*는 모든 경우에 대해 0~1의 안정적 범위를 보여주었다. 이는 통계치의 직접적인 지역 간 비교에서 S^*가 상대적인 강점을 가지고 있음을 함축하고 있는 것이다. 셋째, 공간단위 수가 증가할수록 정규근사에 기반한 유의성 검정의 합리성이 높아지는 것으로 드러났지만, 이러한 점근적 정규성은 S^* 통계량에서 상대적으로 취약한 것으로 드러났다. 넷째, S^*의 상하위 각 6개의 고유벡터를 모런과 기어리 통계량이 어떻게 측정하는 지를 알아본 결과 상위 6개 패턴에 대해서는 두 통계량 모두에서도 높은 양의 공간적 자기상관을 보이는 것으로 측정되었지만, 하위 6개 패턴에 대해서는 모런과 기어리 통계량에 의하면 중위 정도에 해당하는 것으로 드러났다. 다섯째, 우리나라 7대 대도시의 읍면동 수준을 분석한 결과, 특히 음의 공간적 자기상관에 대한 모런

과 기어리 통계량의 가능치에서 높은 수준의 지역간 차이와 불안정성이 노출되었다. 이에 비해 S^* 통계량은 모든 도시의 모든 SPM에 대해 0~1의 값을 보여 가장 안정적인 값을 산출하는 것으로 드러났다. 마지막으로 한강의 존재를 감안한 SPM의 경우, 그렇지 않은 SPM에 비해 매우 다른 고유벡터의 공간패턴이 도출됨을 알 수 있었다.

본 연구를 통해 S^*의 장단점이 극명하게 드러났다. 가장 큰 정점은 공간단위의 형태, 연접성 유형, SPM의 종류 등에 상관없이 일정한 가능치 범위를 보여준다는 것이다. 이는 S^*이 다른 통계량에 비해 상대적으로 MAUP (modifiable areal unit problem, 공간단위 임의성의 문제)이나 말단효과(edge effect)의 영향을 적게 받는 통계량임을 함축하고 있는 것이기 때문에 매우 중요한 속성이 아닐 수 없다. 가장 큰 단점은 왜도나 첨도가 상대적으로 높아 비록 샘플의 크기가 증가한다 하더라도 정규근사가 가지는 유의성 검정법으로서의 타당성이 상대적으로 낮다는 점이다. 이를 극복하기 위해서는 '정확분포(exact distribution)' 접근(Tiefelsdorf and Boots, 1995; Hepple, 1998), '안장점근사(saddlepoint approximation)' (Tiefelsdorf, 2002), 혹은 고차 적률을 사용하여 정규분포가 아닌 다른 이론적 확률분포에 맞추어 근사하는 방법(Costanzo *et al.*, 1983; Hepple, 1998) 등이 고려되어야 함을 의미하는 것이다. 이는 차후의 연구과제로 남겨두고자 한다.

S^*가 안정적인 0~1의 값을 가진다는 사실은 공간적 다변량분석 기법의 개발에 중대한 함의를 내포하고 있는 것으로 보인다. 예를 들어 공간적 주성분분석(spatial principal components analysis)의 경우 공간적 상관관계 매트릭스를 분해하는 방법론이 Wartenberg(1985)에 의해 이미 오래 전에 제안되었지만, 널리 사용되지 못한 가장 큰 이유로 음의 고유치가 산출될 가능성이 있기 때문이었다(Griffith, 1988). 음의 고유치가 생산되는 이유에 공간적 상관관계 매트릭스가 비대칭적일 수 있다거나 주대각 요소에 음수 값이 있을 수 있다는 것 등이 있다. 여기서 특히 후자의 경우, Wartenberg의 공간적 상관관계 매트릭스의 주대각선 상에는 모런 통계량이 놓이기 때문에 발생하는 것이다. 공간적 상관관계 매트릭스를 다르게 정의하면 주대각선 상에 S^*가 놓이게 할 수 있다(Lee and Cho, 2014; Lee, 2015). 이렇게 하면 음의 고유치 산출의 문제는 간단히 해결된다.

가장 시급한 향후 연구 과제는 국지적 S_i^*의 특성을 파악하는 것이다. S_i^*는 명백히 S^*의 LISA이다. 이 국지 통계량의 특징을 국지적 모런 통계량(I_i), 국지적 기어리 통계량(c_i), 그리고 게티스-오드의 G_i^* 통계량과의 비교분석을 통해 밝혀 냄으로써 그 가능성과 한계에 대한 명백한 결론에 도달할 수 있을 것이다. 본 연구가 시사하는 바처럼, S_i^*는 기존의 국지 통계량과는 다른 특성을 보여줄 것으로 보이며, 공간 클러스터의 탐색에 새로운 통찰력을 제공할 수 있을 것으로 기대된다.

사사

본 논문 속에 포함되어 있는 아이디어를 발전시키는데 있어 그 크기를 계측할 수 없을 정도의 영감을 준 두 분께 감사의 마음을 전합니다. 한 분은 미국 오하이오주립대학교 지리학과의 고(故) Lawrence "Larry" Alan Brown (1935~2014) 교수이고, 또 다른 한 분은 미국 텍사스대학교(댈러스) 지리공간정보과학(Geospatial Information Sciences) 프로그램의 Michael Tiefelsdorf 교수입니다.

註

1) '공간적 자기상관 통계량' 대신 공간연관성통계량이라는 용어를 사용하는데는 다음과 같은 이유가 있다. 특정한 공간적 변동(spatial variation)은 전역적 트랜드와 관련된 '1차 효과(first order effect)'와 국지적 상호작용과 관련된 '2차 효과(second order effect)' 모두에 의해 발생하는데(Bailey and Gatrell, 1995, 32), 공간적 자기상관은 원론적으로는 2차 효과에 의한 공간적 의존성만을 의미한다. 이러한 관점에서 보면, 공간적 변동 전체에 대한 패턴 탐색을 위해 사용되고 있는 다양한 통계량이 과연 공간적 자기상관 통계량인가, 혹은 본래 개념에 충실한 형태로 사용되고 있는가에 대한 의문이 제기될 수 있다. 따라서 본 논문에서는 다소 중립적으로 보이는 공간연관성통계량이라는 용어를 사용하고자 한다.
2) 행-표준화된 경우에만 SPM의 요소가 일종의 가중치의 의미를 띠기 때문에 SPM을 공간가중행렬(spatial weights matrix)(Anselin and Smirnov, 1996)이라고 부르는 것은 SPM의 다양성을 지나치게 제약하는 것이다.

3) 그림 1은 뒤이어 소개되는 '공간지체'와 '공간이동평균' 개념과 연관지어 해석할 필요가 있다. 주대각 요소가 0인 (a)의 경우는 중심 공간단위의 값과 주변 공간단위의 대푯값으로서의 '공간지체'를 비교하게 되지만, 주대각 요소가 0이 아닌 (b)의 경우는 국지 세트 전체에 대한 대푯값으로서의 '공간이동평균' 하나 만을 상정한다.

참고문헌

국토교통과학기술진흥원, 2015, 공간정보 SW활용을 위한 오픈소스 가공기술개발 2차년도 연차실적계획서 (내부자료).

이상일·신정엽·김현미·홍일영·김감영·전용완·조대헌·김종근·이건학 역, 2009, 「지리정보시스템과 지리정보과학」, 시그마프레스(Longley, P.A., Goodchild, M.F., Maguire, D.J., and Rhind, D.W., 2005, *Geographic Information Systems and Science*, 2nd edition, Hoboken, NJ: John Wiley & Sons).

Anselin, L., 1988, *Spatial Econometrics: Methods and Models*, Boston: Kluwer Academic Publishers.

Anselin, L., 1995, Local indicators of spatial association-LISA, *Geographical Analysis*, 27(2), 93-115.

Anselin, L. and Griffith, D.A., 1988, Do spatial effects really matter in regression analysis?, *Papers in Regional Science*, 65(1), 11-34.

Anselin, L. and Smirnov, O., 1996, Efficient algorithms for constructing proper higher order spatial lag operators, *Journal of Regional Science*, 36(1), 67-89.

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

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

Cliff, A.D. and Ord, J.K., 1969, The problem of spatial autocorrelation, in Scott, A.J., ed., *Studies in Regional Science*, London: Pion, 25-55.

Cliff, A.D. and Ord, J.K., 1973, *Spatial Autocorrelation*, London: Pion.

Cliff, A.D. and Ord, J.K., 1981, *Spatial Processes: Models and Applications*, London: Pion.

Costanzo, C.M., Hubert, L.J., and Golledge, R.G., 1983, A higher moment for spatial statistics, *Geographical Analysis*, 15(4), 347-351.

de Jong, P.D., Sprenger, C., and Veen, F.V., 1984, On extreme values of Moran's I and Geary's c, *Geographical Analysis*, 16(1), 17-24.

Fotheringham, A.S., Brunsdon, C., and Charlton, M. 2002, *Geographically Weighted Regression*, Chichester, West Sussex: John Wiley & Sons.

Geary, R.C., 1954, The contiguity ratio and statistical mapping, *The Incorporated Statistician*, 5(3), 115-145.

Getis, A., 1991, Spatial interaction and spatial autocorrelation: A cross-product approach, *Environment and Planning A*, 23(9), 1269-1277.

Getis, A., 2008, A history of the concept of spatial autocorrelation: A geographer's perspective, *Geographical Analysis*, 40(3), 297-309.

Getis, A., 2010, Spatial autocorrelation, in Fischer, M. M. and Getis, A., eds., *Handbook of Applied Spatial Analysis*, New York: Springer, 255-278.

Getis, A. and Ord, J.K., 1992, The analysis of spatial association by use of distance statistics, *Geographical Analysis*, 24(3), 189-206.

Getis, A. and Ord, J.K., 1996, Local spatial statistics: an overview, in Longley, P. and Batty, M., eds., *Spatial Analysis: Modelling in a GIS Environment*, Cambridge: GeoInformation International, 261-277.

Griffith, D.A, 1988, *Advanced Spatial Statistics: Special Topics in the Exploration of Quantitative Spatial Data Series*, Dordrecht: Kluwer Academic Publishers.

Griffith, D.A., 2003, *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization*, Berlin: Springer.

Henshaw, R.C., Jr., 1966, Testing single-equation least squares regression models for autocorrelated disturbances, *Econometrica: Journal of the Econometric Society*, 34(3), 646-660.

Henshaw, R.C., Jr., 1968, Errata: Testing single-equation least squares regression models for autocorrelated

disturbances, *Econometrica: Journal of the Econometric Society*, 36(3/4), 646-660.

Hepple, L.W., 1998, Exact testing for spatial autocorrelation among regression residuals, *Environment and Planning A*, 30(1), 85-107.

Haining, R.P., 2003. *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.

Hubert, L.J., Golledge, R.G., and Costanzo, C.M., 1981, Generalized procedures for evaluating spatial autocorrelation, *Geographical Analysis*, 13(3), 224-233.

Lee, S.-I., 2001a, 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., 2001b, Spatial Association Measures for an ESDA-GIS Framework: Developments, Significance Tests, and Applications to Spatio-temporal Income Dynamics of US Labor Market Areas, 1969-1999, Ph.D. Dissertation, The Ohio State University.

Lee, S.-I., 2004, 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., 2008, A generalized procedure to extract higher order moments of univariate spatial association measures for statistical testing under the normality assumption, *Journal of the Korean Geographical Society*, 43(2), 253-262.

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.

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.

Leung, Y., Mei, C.L., and Zhang, W.X., 2003, Statistical test for local patterns of spatial association, *Environment and Planning A*, 35(4), 725-744.

Mantel, N., 1967, The detection of disease clustering and a generalized regression approach, *Cancer Research*, 27(2 Part 1), 209-220.

Moran, P.A., 1948, The interpretation of statistical maps, *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2), 243-251.

Ord, J.K., and Getis, A., 1995, Local spatial autocorrelation statistics: distributional issues and an application, *Geographical Analysis*, 27(4), 286-306.

Rogerson, P.A., 1999, The detection of clusters using a spatial version of the chi-square goodness-of-fit statistics, *Geographical Analysis*, 31(1), 130-147.

Tango, T., 1995, A class of tests for detecting 'general' and 'focused' clustering of rare diseases, *Statistics in Medicine*, 14(21-22), 2323-2334.

Tiefelsdorf, M., 2000, *Modelling Spatial Processes: The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I*, New York: Springer.

Tiefelsdorf, M., 2002, The saddlepoint approximation of Moran's *I*'s and local Moran's I_i's reference distributions and their numerical evaluation, *Geographical Analysis*, 34(3), 187-206.

Tiefelsdorf, M. and Boots, B., 1995, The exact distribution of Moran's *I*, *Environment and Planning A*, 27(6), 985-999.

Tiefelsdorf, M., Griffith, D.A., and Boots, B., 1999, A variance-stabilizing coding scheme for spatial link matrices, *Environment and Planning A*, 31(1), 165-180.

Tiefelsdorf, M. and Griffith, D.A., 2007, Semiparametric filtering of spatial autocorrelation: The eigenvector approach, *Environment and Planning A*, 39(5), 1193-1221.

Tobler, W.R., 1970, A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46(2), 234-240.

Upton, G.J.G. and Fingleton, B., 1985, *Spatial Data Analysis by Example. Volume 1: Point Pattern and Quantitative Data*, New York: John Wiley.

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)

Correspondence: Sang-Il Lee, 08826, 1 Gwanak-ro, Gwanak-gu, Seoul, Korea, Department of Geography Education, College of Education, Seoul National University (Email: si_lee@snu.ac.kr)

투 고 일: 2015년 11월 23일
심사완료일: 2015년 12월 6일
투고확정일: 2015년 12월 7일