

일변량 공간연관성통계량에 대한 비교 연구 (II): 국지적 S_i 통계량을 중심으로*

이상일** · 조대현*** · 이민파****

Comparing Univariate Spatial Association Statistics (II): Focusing on Local Lee's S_i Statistics*

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

요약 : 본 연구는 국지적 공간연관성통계량으로서의 S_i 통계량의 특성을 기존의 국지적 모란 통계량(I_i), 국지적 기어리 통계량(c_i), 국지적 게티스-오드 통계량(G_i^*)과의 비교를 통해 밝히는 것이다. 새로이 제시된 S_i^* 통계량은 각 국지 세트가 평활화 효과로 인한 분산의 감소에 어느 방향으로 얼마나 기여했는가를 측정하는데, 국지 세트가 높은 양의 공간적 자기상관을 보이면 높은 S_i^* 값을, 그 반대라면 낮은 S_i^* 값을 갖게 된다. 네 가지 중요 개념(공간 클러스터, 공간 특이점, 공간 체제, 국지적 안정성)과 두 가지 부가 준거를 바탕으로 국지적 SAS를 비교한 결과, 두 가지 중요한 결론이 도출되었다. 첫째, 국지적 SAS는 중심 공간단위와 주변 공간단위 간의 연관성에 집중하는 SAS(I_i 와 c_i)와 국지 세트 전체의 공간 균집성을 다루는 SAS(S_i^* 와 G_i^*)로 대별된다. 둘째, LISA의 조건을 만족시키고, 유의성 검정 측면에서 우위에 있는 S_i^* 를 G_i^* 에 대한 대체 통계량으로 사용하는 것이 합당하다. 가능치의 범위와 표본분포 상의 특성을 살펴보기 위해 정다각 테셀레이션 분석을 수행한 결과는 다음과 같다. 첫째, 국지적 SAS는 전역적 SAS에 비해 통계량 간 상관관계가 높지 않고, 훨씬 더 넓은 가능치 범위를 가진다. 둘째, SAS의 가능치 범위는 전체 공간단위의 개수, 공간단위 형태의 복잡성, 주변 공간단위의 개수, 공간근접성행렬의 행표준화 여부에 따라 다양하게 나타난다. 셋째, 국지적 SAS의 왜도와 첨도는 전역적 SAS에 비해 훨씬 커 정규근사의 타당성이 훨씬 더 떨어진다. 넷째, 우리나라 7대 대도시에 적용한 결과, 도시별 공간적 형상의 특성에 따라 가능치 범위와 표본분포상의 특성이 다양하게 나타나는 것이 확인되었다. 본 연구는 국지적 SAS의 통계적 특성을 전면적으로 비교·분석했다는 의미를 가질 뿐만 아니라 연구 목적에 따라 어떤 SAS를 사용하는 것이 더 바람직한지 제시하고 있어 그 활용성이 매우 높을 것으로 기대된다.

주요어 : 공간적 자기상관, 국지적 공간연관성통계량, 국지적 S_i^* 통계량, 공간적 고유치와 고유벡터

Abstract : The main objective of this paper is to elucidate the characteristics of a new spatial association statistic, S_i , in comparison with local Moran's I_i , local Geary's c_i , and Getis-Ord's G_i^* . S_i^* as a new local statistic measures how much (and in what direction) a local set contributes to an overall variance reduction resulting from the smoothing effect which occurs when an original variable is represented by its spatial moving average; the presence of a strong positive spatial autocorrelation in a local set results in a higher S_i^* and vice versa. The main findings

*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)

of a comparison of the four local spatial association statistics on the basis of four main concepts (spatial clusters, spatial outliers, spatial regimes, and local stability) and two additional criteria are two fold. First, they are largely divided into two distinct categories, I_i and c_i as more association-centered ones, and S_i^* and G_i^* as more clustering-centered ones. Second, S_i^* can be seen as a substitute for G_i^* in the sense that the former satisfies the two conditions for a LISA and its distributional properties are better known. A regular tessellation analysis is conducted to examine the feasible ranges and distributional properties of the local spatial association statistics. Major findings are as follows. First, correlations among the local statistics are much smaller in amount and their feasible ranges are much narrower when compared to their global counterparts. Second, the feasible ranges and distributional properties vary in accordance to the number and shape of spatial units, the number of neighboring spatial units, and the specification of spatial proximity matrix. Third, both the skewness and kurtosis are much more pronounced when compared to those from global SAS such that normal approximation is proved to be much less reliable for significance testing. Fourth, differences in spatial configuration of the 7 largest cities in South Korea dictate differences in the feasible ranges and distributional characteristics. This study can be viewed as one of the most comprehensive studies to address different pros and cons of the local statistics and is expected to help researchers choose a statistic suitable for their empirical studies.

Key Words : Spatial autocorrelation, Local spatial association statistics, Local Lee's S_i^* statistics, Spatial eigenvalues and eigenvectors

I. 서론

공간데이터분석(spatial data analysis, 이하 SDA)의 발전에 있어, 혹은 넓게는 지리정보과학(geographic information science, 이하 GIS) 전체의 발전에 있어, 1990년대 중반은 매우 중요한 의미를 갖는다. 1970년대부터 시작된 탐색적 데이터분석(exploratory data analysis)의 광범위한 확산,¹⁾ 단순한 저장 및 관리 도구에서 일반 분석 플랫폼으로의 GIS의 발전적 변모, 그리고 소위 ‘국지적 전회(local tum)’²⁾라고 불리는 통계적 연구에서의 새로운 패러다임의 등장 등이 거의 동시적으로, 서로 맞물리면서 발생한 것이다. 이러한 환경 변화 속에서, SDA와 GIS의 통합 가능성이 활발하게 논의 되었고,³⁾ 보다 GIS 친화적인 탐색적 공간데이터분석(exploratory spatial data analysis, 이하 ESDA)이 크게 부각되었다.⁴⁾ 다양한 ESDA의 과제는 GIS의 공간데이터 처리 및 시각화 기능과 결합되었을 때 더욱 원활히 진행될 수 있다는 논의가 활발했을 뿐만 아니라, GIS에 익숙한 수 많은 연구자들에게 ESDA의 방법론을 소개함으로써 저변을 확대하는 계기이기도 했다(Unwin, 1996; Goodchild and Longley, 1999). 이러한 ESDA-GIS 프레임워크의 확립 과정에 지대한 공헌을 한 것이 바로 국지적 공간연관성통계량(spatial association statistics, 이하 SAS)이다.⁵⁾ 기본적으로 국지적 SAS는 ESDA가 추구하는 목적을 달성하기에 적절한

특성을 내재적으로 보유하고 있었다.⁶⁾ 결국 2010년대 중반인 지금은 상용 GIS 소프트웨어를 통해 공간연관성 분석이나 공간 클러스터 탐지 연구를 손쉽게 수행할 수 있는 그런 시대가 도래한 것이다.⁷⁾

국지적 SAS 연구는 게티스와 오드(Getis, 1991; Getis and Ord, 1992; Ord and Getis, 1995), 그리고 안셀린(Anselin, 1995)의 선구적인 연구에 의해 촉발되었다. 게티스와 오드는 공간적 자기상관의 일반 모델에 입각한 국지적 통계량을 고안하고자 했으며, 몇 번의 수정을 거친 끝에 최종적으로 G_i 와 G_i^* 통계량을 제시하였다(Getis and Ord, 1996). 이에 반해 안셀린은 기존에 널리 사용되고 있던 전역적 모런 통계량과 기어리 통계량을 분해함으로써 국지적 모런 통계량(I_i)과 국지적 기어리 통계량(c_i)을 도출해내는데 성공했다(Anselin, 1995). 그런데 안셀린은 이 두 통계량에 LISA(local indicators of spatial association, 국지적 공간연관성 지수)라는 칭호를 부여하면서, 국지적 SAS가 LISA이기 위한 두 가지 조건을 제시했다(Anselin, 1995:94). 하나의 조건은 “LISA는 공간 군집성(spatial clustering)의 정도를 측정해야 한다”는 것이고, 또 다른 하나의 조건은 “LISA의 합은 전역적 통계량과 비례 관계에 있어야 한다”는 것이었다. 안셀린의 관점에서 보면 게티스-오드의 통계량은 후자의 조건을 만족시키지 못하기 때문에 진정한 의미의 LISA라 할 수 없다(이상일 등, 2015:331). 보다 최근에는 Lee

가 전역적 S 통계량(Lee, 2001a; 2001b; 2004b; 2008)과 함께 국지적 S_i 통계량(Lee, 2001b; 2008; 2009)을 제시한 바 있다. 국지적 S_i 통계량은 안셀린의 두 조건을 모두 만족시키기 때문에 또 다른 LISA라 할 수 있다.

앞에서 언급한 것처럼, 이러한 국지적 SAS는 ESDA-GIS 프레임워크의 핵심적인 요소가 되었다. 특히 I_i 와 G_i^* 는 이 과정에서 가장 널리 활용되어 왔는데, 전자는 모란 산포도(Moran scatterplots)(Anselin, 1996)와 함께 주로 공간연관성 유형 분석에, 후자는 주로 공간 클러스터(spatial clusters) 탐지 분석에 사용되어 왔다. 현재 이러한 기법은 다양한 소프트웨어 환경에서 이용 가능한, 공간데이터분석을 위한 표준적인 방법론으로 자리잡고 있다. 그런데, 이러한 기법의 단순 사용자가 아닌 방법론 자체를 탐구하는 연구자의 입장에서 보면, 아직 해결되어야 할 과제가 많이 남아 있다.⁸⁾ 우선, 각 국지적 SAS의 본질적 성격에 대한 전면적인 비교 연구는 거의 행해지지 않았다. 즉, 국지적 상황의 공간적 자기상관은 어떤 다양한 측면을 가지며 다양한 국지적 SAS는 그러한 측면을 어떻게 측정하고 있는지에 대한 정교한 분석이 부족했던 것이다. 또한 가능치 범위와 표본분포에 대한 체계적인 비교 분석 역시 부재했다. 전역적 SAS의 비교 분석(이상일 등, 2015)에서 사용되었던, 정다각 테셀레이션(regular tessellations)을 활용한 기법들이 국지적 SAS에도 적용될 필요가 있다.

본 연구는 기본적으로 이상일 등(2015)의 전역적 SAS에 대한 연구를 국지적 SAS로 확장한 것이다. 따라서 본 논문의 주된 연구목적은 새로운 국지적 SAS로서의 S_i 통계량의 특성을 기존의 I_i , c_i , 그리고 G_i^* 통계량과의 비교 연구를 통해 밝히는 것이다. 연구 내용은 크게 두 가지로 나뉜다. 첫째, 국지적 SAS에 대한 새로운 정식화와 비교 분석을 위한 근거 틀을 제시하고 그것을 바탕으로 국지적 SAS 간의 서로 다른 특성을 파악한다. 이를 통해 S_i 통계량의 상대적 장단점이 드러날 것이다. 둘째, 국지적 SAS가 가지는 통계량으로서의 특성을 중심적률(central moments) 추출법과 고유치(eigenvalues) 및 고유벡터(eigenvectors) 추출법을 통해서 밝히는 것이다. 이를 통해 S_i 통계량의 표본분포 상의 특성이 파악될 것이다. 분석은 기본적으로 가상적인 정다각 테셀레이션 데이터에 대해 이루어지지만, 실제 연구를 위한 함의를 검토하기 위해 우리나라 7대 대도시의 읍면동 단위 데이터도 부가적으로 사용된다.

II. 국지적 공간연관성통계량의 정식화와 특성 비교

1. 국지적 공간연관성통계량의 정식화와 S_i^* 통계량의 성격

표 1에는 본 연구에서 다룬 네 가지 국지적 SAS에 대한 수식이 제시되어 있다. G_i^* 통계량은 LISA가 되기 위한 첫 번째 조건만을 만족시키지만, 나머지는 두 조건 모두를 만족시킨다. 사실상 표 1에 나타나 있는 수식은 안셀린의 두 번째 조건을 보다 강화한 ‘가법성 요구조건(additivity requirement)’, 즉 “국지적 SAS의 평균은 전역적 SAS의 값과 동일하다”(Tiefelsdorf, 1998; 2000; Lee, 2009)를 만족시킨다. 이상일 등(2015:332)의 표 1과 비교해 보면, 전역적 SAS로부터 이 가법성 요구조건을 만족시키는 국지적 SAS를 도출하는 방법이 비교적 간단함을 알 수 있다. 수식의 분자에서 모든 i 공간단위에 적용된 합산 기호를 제거하고, 모든 j 공간단위에 대한 합산 기호만 남긴다. 그리고 나서 전체 수식에 n 을 곱하기만 하면 된다. 이렇게 보다 완고한 요구조건을 적용함으로써 전역적 SAS와 국지적 SAS를 직접 비교하는 것이 가능해진다. 즉, 특정한 공간단위의 SAS 값이 전역적 SAS와 동일하다면, 대상 지역 전체의 공간적 자기상관의 방향과 크기가 그 국지 세트(local set)에서도 동일하게 발생하고 있다고 말할 수 있다.⁹⁾ 국지 세트는 중심 공간단위와 그와 이웃한 주변 공간단위로 이루어진 소규모 하위 지역을 지칭하는 용어로 이 논문에서 지속적으로 사용할 것이다.¹⁰⁾

표 1에는 또한 매트릭스를 표현하는 두 가지 방법이 함께 제시되어 있다. 첫 번째 방법은 Lee(2004b)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)($\mathbf{\delta}$)’(원 벡터에서 평균만 뺀 것)를 통해 ‘이차형식의 비(ratio of quadratic forms)’로 표현한 것이다(Tiefelsdorf, 2000; Lee, 2008; 이상일 등, 2015). 이 매트릭스 표현을 자세히 살펴보면, 전역적 SAS와 달리 국지적 공간근접성행렬(spatial proximity matrix, 이하 SPM)인 \mathbf{V}_i 가 적용되어 있음을 알 수 있다. \mathbf{V}_i 는 일반화된 전역적 SPM인 \mathbf{V} 로부터 도출된 것인데, \mathbf{V} 는 SPM의 주대각 요소는 무조건 0 이어야 한다는 근거 없는 관행으로부터 탈피함으

$$S_i^* = \frac{n^2 \left(\sum_j v_{ij}^* (x_j - \bar{x}) \right)^2}{\sum_i \left(\sum_j v_{ij}^* \right)^2 \sum_i (x_i - \bar{x})^2} \quad (3)$$

Lee(2001a; 2001b)는 전역적 S 통계량을 일종의 ‘분산 감소계수(variance reducing factor)’라고 정의한 바 있다. 원 변수를 ‘공간적 파생변수(spatially derived variables)’인 공간지체(spatial lag, 이하 SL)나 공간이동평균(spatial moving average, 이하 SMA)으로 변환하면 평활화 효과로 인해 분산이 감소한다. 원 변수가 높은 양의 공간적 자기상관을 보일수록 파생변수의 분산이 적게 감소하고, 원 변수가 높은 음의 공간적 자기상관을 보일수록 파생변수의 분산이 많이 감소하는 성질을 이용한 것이다. 따라서 전역적 S 는 원 변수 분산에 대한 SL 분산의 비(S^0 의 경우) 혹은 SMA 벡터의 분산의 비(S^* 의 경우)로 정의되며, 이론적으로 0~1의 값을 갖는다(이상일 등, 2015).¹⁴⁾ 이것을 국지적 S_i 통계량의 해석에 적용하면 다음과 같다. S_i 는 각 국지 세트가 평활화 효과로 인한 분산의 감소에 어느 방향으로 얼마나 기여하는지를 측정한다. 다른 말로 표현하면 국지적 S_i 는 SL 벡터나 SMA 벡터의 분산에 대한 각 국지 세트의 상대적 기여도를 의미한다.

보다 자세한 설명을 위해 두 파생 SAS 중 S_i^* 에 집중하고자 한다.¹⁵⁾ 만일 평균 보다 매우 높은(혹은 낮은) 값이 그와 유사한 값에 의해 둘러싸여 있다면(양의 공간적 자기상관), 이 국지 세트의 SMA 역시 높을(낮을) 것이고(적은 분산 감소), 따라서 높은 S_i^* 값을 갖게 된다. 반대로 평균 보다 매우 높은 값이 평균 보다 매우 낮은 값에 의해 둘러싸여 있거나 그 반대의 경우라면(음의 공간적 자기상관), 이 국지 세트의 SMA는 상쇄효과로 인해 평균에 근접할 것이고(많은 분산 감소), 따라서 낮은 S_i^* 값을 갖게 된다. 다른 식으로 표현하면, 최소의 분산 감소는 매우 높은 값이 매우 높은 값에(혹은 매우 낮은 값이 매우 낮은 값에) 의해 둘러싸여 있어 여전히 매우 높은 SMA를 유지하는 국지 세트에서 발생할 것이고(양의 공간적 자기상관), 최고의 분산 감소는 매우 높은 값과 매우 낮은 값이 뒤섞여 있어 SMA가 평균에 근접하게 되는 국지 세트에서 발생할 것이다(음의 공간적 자기상관).

국지적 SAS 간 비교를 보다 분명하게 하기 위해 이항연접성(binary contiguity) SPM의 행표준화 버전을 의미하는 \mathbf{w}^0 와 \mathbf{w}^* 를 적용했을 때를 상정하고자 한다. 우선

국지적 모런 통계량과 두 개의 S_i 통계량은 다음과 같이 단순화된다.

$$I_i = z_i \sum_j w_{ij}^0 z_j = z_i \tilde{z}_i^0 \quad (4)$$

$$S_i^0 = \left(\sum_j w_{ij}^0 z_j \right)^2 = \left(\tilde{z}_i^0 \right)^2 \quad (5)$$

$$S_i^* = \left(\sum_j w_{ij}^* z_j \right)^2 = \left(\tilde{z}_i^* \right)^2 \quad (6)$$

식 (4)를 보면, I_i 는 중심 공간단위의 표준화점수(z_i)와 그것의 표준화 SL 값(주변 공간단위의 표준화점수들의 가중평균값, \tilde{z}_i^0) 간의 곱임을 알 수 있다. 같은 방식으로 보면, S_i^0 는 표준화 SL 값의 제곱으로 정의되고, S_i^* 는 국지 세트 전체의 표준화 SMA 값(중심 공간단위와 주변 공간단위의 표준화점수들의 가중평균, \tilde{z}_i^*)의 제곱으로 정의됨을 알 수 있다. 본 연구는 게티스-오드의 국지 통계량에서 G_i^* 가 G_i^0 에 비해 훨씬 더 선호되는 것과 동일한 이유에서 S_i^* 의 사용을 강력히 제안한다. 그러므로 일반 국지적 모런 통계량이 항상 I_i^0 를, 일반 국지적 기어리 통계량이 항상 c_i^0 를, 일반 게티스-오드 통계량이 항상 G_i^* 를 의미하는 것과 동일한 방식으로 일반 S_i 통계량도 별 다른 언급이 없다면 늘 S_i^* 를 의미하는 것으로 규정한다.

c_i 와 G_i^* 는 앞의 세 SAS에 비해 단순화가 상대적으로 쉽지 않다. 앞에서와 마찬가지로 이항연접성 SPM의 행표준화 버전을 적용했을 때, 두 SAS는 다음과 같이 단순화 된다.

$$c_i = \frac{n-1}{2n} \sum_j w_{ij}^0 (z_i - z_j)^2 = \frac{n-1}{2n} \frac{\sum_j (z_i - z_j)^2}{n_i^0} \quad (7)$$

$$G_i^* = \frac{1}{\sqrt{(n/n_i^* - 1)/(n-1)}} \tilde{z}_i^* = \sqrt{\frac{n-1}{n/n_i^* - 1}} \tilde{z}_i^* \quad (8)$$

식 (7)의 n_i^0 는 주변 공간단위의 개수이고, 식 (8)의 n_i^* 는 자신을 포함하는 국지 세트의 공간단위 개수이다. 따라서 후자는 항상 전자보다 1 단위가 많다. c_i 는, 앞에 붙어 있는 상수를 제외할 경우, 중심 공간단위와 주변 공간단위의 (표준화점수 간의) 차이의 제곱합의 국지적

평균 정도로 해석될 수 있다. G_i^* 의 단순화는 보다 주의 깊게 살펴볼 필요가 있다. 이러한 단순화는 Lee와 Cho (2013; 2015)에 의해 제시된 것으로, 앞 부분의 상수를 제외한다면 G_i^* 는 기본적으로 표준화 SMA, 그 자체임을 알 수 있다. S_i^* 가 표준화 SMA 값의 제곱으로 정의되다는 점을 염두에 둔다면 기본적으로 G_i^* 와 S_i^* 가 동일한 통계량임을 어렵지 않게 짐작할 수 있다. 이것은 뒤에서 상세히 다루도록 한다.

2. 국지적 공간연관성통계량의 특성 비교

좀 더 구체적으로 국지적 SAS의 특성을 비교하기 위해 가상적인 국지 세트 9개를 설정하고 SAS 값을 산출하였다(표 2). 국지 세트는 1개의 중심 공간단위와 6개의 주변 공간단위가 연결해 있는 형태를 취한다. 우선 중심 공간단위의 표준화점수가 각각 1, 0, -1인 세 경우를 상정한다. 각각은 평균 이상, 평균, 평균 이하를 상징한다. 각각에 대해 주변 공간단위의 성격을 대변하는 세 가지 경우를 상정하였는데, 6개 주변 공간단위의 값이 모두 1인 경우, 모두 0인 경우, 모두 -1인 경우이다. 결국 모두 9가지의 국지 세트의 공간 패턴이 마련되었다. 각각의 패턴은 몇 가지 전형적인 국지 패턴을 재현하고 있음을 주목해야 한다. 국지적 SAS가 이러한 대표적 패턴에 대해 어떠한 방식으로 측정하는지를 살펴볼 것이다. 분석의 실행을 위해서는 국토교통부 국토공간정보연구 사업의 공간정보 SW활용을 위한 오픈소스 가공기술개발 과제로 개발되고 있는 분석 도구와 R을 함께 사용하였다(이상일 등, 2015; 국토교통과학기술진흥원, 2016)

국지적 SAS 간의 비교를 위한 준거로 가장 중요한 것은 특징적인 국지 패턴을 묘사하는 개념들인데, 공간 클러스터, 공간적 특이점, 공간 체제, 국지적 안정성이 그것들이다. 이 개념들은 국지 세트에서 발생하는 공간적 자기상관의 다양한 측면을 묘사하고 있는데, 국지적 SAS 간의 상대적 특성 비교에서 가장 중요한 준거로 사용할 것이다.

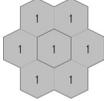
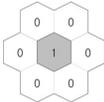
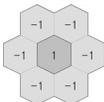
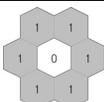
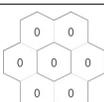
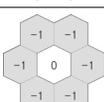
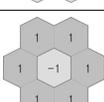
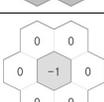
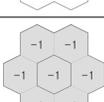
첫 번째 준거는 ‘공간 클러스터(spatial clusters)’ 개념이다. 공간 클러스터란 “유의미하게 높은 양의 공간적 자기상관을 보이는 공간단위들의 응집체” 혹은 “유의미하게 유사한 값을 보이는 공간단위들의 응집체” 정도로 정의할 수 있다(Lee, 2001b). 더 나아가 유의미하게 높은 값의 공간 클러스터는 ‘핫스팟(hot spot)’으로, 유의미하

게 낮은 값의 공간 클러스터는 ‘콜드스팟(cold spot)’으로 구분하고자 한다(Sokal *et al.*, 1998). 그러나 이와 달리 다소 높은, 다소 낮은, 혹은 아예 평균값의 공간 클러스터도 생각해 볼 수 있는데, 이를 여기서는 ‘중간스팟(intermediate spot)’이라고 부르하고자 한다. 이런 의미에서 보면 A1은 핫스팟을, B2는 중간스팟을, C3는 콜드스팟을 재현한다. 표 2를 보면, 핫스팟(A1)과 콜드스팟(C3)의 경우는 네 SAS가 거의 동일한 방식으로 측정하고 있음을 확인할 수 있다. 여기서 G_i^* 만이 공간 클러스터의 방향성(핫스팟 혹은 콜드스팟의 여부)을 부호를 통해 보여주고 있다. 그런데 중간스팟(B2)의 경우는 오로지 c_i 만이 높은 양의 공간적 자기상관이 존재하는 것으로 탐지하고 있다.

두 번째 준거는 ‘공간 특이점(spatial outliers)’ 개념이다. 공간 특이점은 “주변 공간단위와 유의미하게 다른 성격을 보이는 공간단위(혹은 그 공간단위가 보유한 특이값)”를 의미한다(Wartenberg, 1990; Fotheringham and Charlton, 1994; Anselin, 1998; Lee, 2001b). 공간 특이점이 음의 공간적 자기상관과 관련되어 있음을 명심할 필요가 있다. 높은 값을 가진 공간단위가 낮은 값을 가진 공간단위에 의해 둘러싸이거나, 낮은 값을 가진 공간단위가 높은 값을 가진 공간단위에 의해 둘러싸인 경우, 해당 중심 공간단위를 공간 특이점이라 부를 수 있다. 공간 특이점(A3와 C1)의 경우는 I_i 만이 효과적으로 탐지하고 있음을 알 수 있다. 즉, 음의 부호를 통해 음의 공간적 자기상관이 존재함을 명백히 보여주고 있는 것이다. c_i 역시 음의 공간적 자기상관이 존재함을 보여주고 있는데, 양수로만 표현되고 있어 직관적으로 확인하기 어렵다는 단점이 있다. 이와는 대조적으로 G_i^* 와 S_i^* 는 A3와 C1 모두를 높은 값 혹은 낮은 값들이 집중되어 있는 양의 공간적 자기상관과 관련되어 있는 국지 세트로 판단한다.

세 번째 준거는 ‘공간 체제(spatial regimes)’ 개념이다. 공간 체제는 ‘공간적 이질성(spatial heterogeneity)’이 드러나는 한 형식으로 볼 수 있는데(Anselin, 1998), “유사한 통계적 속성을 갖는 공간단위의 연속체 혹은 하위지역”으로 정의할 수 있다. 국지적 공간적 자기상관의 맥락에서 보면, 공간 체제는 유사한 공간연관성 유형이 집중적으로 나타나는 하위지역으로 볼 수 있다. 모런 산포도에 의거하면(Anselin, 1995; 1996), 일변량 공간연관성은 네 개의 유형으로 구분되는 데, 일사분면의 H-H(high-

표 2. 국지적 공간연관성통계량의 통계치 산출 비교

국지 세트의 공간 패턴	z_i	공간적 파생변수		통계량			
		\tilde{z}_i^0	\tilde{z}_i^*	I_i	c_i	G_i^*	S_i^*
A1 	1	1	1	1	0	2,730	1
A2 	1	0	0.143	0	0.495	0.390	0.020
A3 	1	-1	-0.714	-1	1,980	-1,950	0.510
B1 	0	1	0.857	0	0.495	2,340	0.735
B2 	0	0	0	0	0	0	0
B3 	0	-1	-0.857	0	0.495	-2,340	0.735
C1 	-1	1	0.714	-1	1,980	1,950	0.510
C2 	-1	0	-0.143	0	0.495	-0.390	0.020
C3 	-1	-1	-1	1	0	-2,730	1

* Z_i : 중심 공간단위의 표준화점수, \tilde{Z}_i^0 : 주변 공간단위의 표준화 SL, \tilde{Z}_i^* : 국지 세트 전체의 표준화 SMA, c_i 와 G_i^* 의 통계치는 $n = 100$ 을 가정하고 계산한 것임.

high), 이사분면의 L-H(low-high), 삼사분면의 L-L(low-low), 사사분면의 H-L(high-low) 유형으로 구분된다. 이런 측면에서 공간연관성의 네 가지 유형을 가장 잘 탐지하는 것은 당연히 모던 통계량이다. 내부적으로 네 가지 유형을 모두 구분할 뿐만 아니라 외적으로도 부호를 통해 H-H/L-L과 H-L/L-H를 효과적으로 분류한다. G_i^* 는 내부적으로는 공간연관성 유형을 따지지 않지만, H-H와 L-L

의 두 유형은 부호로 확실히 구분하고 있다는 측면에서 기어리 통계량과 S_i^* 보다는 우수한 것으로 평가할 수 있다.

네 번째 준거는 ‘국지적 안정성(local stability)’ 개념이다. 이것은 실지로 국지 세트 내부의 값들이 얼마나 동질적인가를 평가한다. 가장 주목할 만한 양의 공간적 자기상관은 평균으로부터 멀리 떨어져 있는 값이 그와 동

표 3. 국지적 공간연관성통계량의 상대적 특성 요약

비교 준거		통계량			
		I_i	c_i	G_i^*	S_i^*
공간 클러스터	핫스팟과 콜드스팟	△	△	○	○
	중간스팟	×	○	×	×
공간 특이점		○	△	×	×
공간 체제		○	×	△	×
국지적 안정성		×	○	×	×
중심 공간단위의 의존도		×	△	○	○
'이차형식의 비'로의 표현가능성		○	○	×	○

* ○ : 우수(혹은 심하지 않음), △ : 보통, × : 열등(혹은 심함).

일하게 평균으로부터 멀리 떨어져 있는 값들로 둘러 싸여 있는 국지 세트에서 발생한다. 이는 국지적 공간적 자기상관의 크기는 큰 편도(deviates)와 유사성이라는 두 조건에 의해 결정된다는 것을 의미한다. 이렇게 보면 기어리 통계량이 이 조건을 가장 잘 만족시킨다. 나머지 SAS는 두 조건 중 전자에 보다 집중한다. 극단적으로 높은 값이 상당히 높은 값에 의해 둘러싸인 경우, 기어리 통계량은 음의 공간적 자기상관이 있는 것으로 측정하겠지만, 나머지 통계량은 양의 공간적 자기상관이 있는 것으로 측정할 것이다. 이 개념을 다른 방식으로 설명할 수도 있다. 예를 들어 A1 패턴에서 주변 공간단위의 값을 3개의 3과 3개의 -1로 교체한다고 하자. 이 새로운 패턴은 원 패턴과 비교해 봤을 때 국지적 동질성의 정도가 훨씬 낮은 것이다. 그럼에도 불구하고, 모런 통계량, 게티스-오드 통계량, S_i^* 의 수치는 변하지 않을 것이다. 왜냐하면 공간적 파생변수의 값이 변하지 않을 것이기 때문이다. 그러나 기어리 통계량의 수치는 0에서 1.98로 변하며, 이는 국지 세트의 공간적 자기상관이 양에서 음으로 변했음을 나타낸다. 이는 기어리 통계량이 가진 독특한 성격 때문인데, 중심 값에 이웃의 가중평균을 먼저 빼고 나중에 그것의 제곱값을 구하는 방식이 아니라, 중심 값과 개별 이웃값의 차이를 우선 제곱하고, 그것들의 가중평균을 나중에 구하는 방식을 취하고 있기 때문에 주변 공간단위의 개별 특성이 고스란히 국지 통계치에 반영된다.

위의 네 가지 개념 외에, SAS의 장단점을 평가하기 위한 준거로 두 가지 사항을 부가적으로 다루고자 한다. 따라서 다섯 번째 준거는 '중심 공간단위에의 의존성'이다. 이것은 중심 공간단위의 상황이 전체 SAS 값에 얼마

나 큰 영향을 끼치는지와 관련된다. 이를 B1-B3의 예를 통해 살펴보면 다음과 같다. 모런 통계량은 주변 공간단위의 상황과 관계없이 중심 공간단위가 0이면 전체적으로 공간적 자기상관이 없다고 측정한다. 기어리 통계량도 여전히 중심 공간단위에 의존적이긴 하지만 모런 통계량 만큼 심하지는 않다. 즉, B1과 B3을 다소간의 양의 공간적 자기상관이 있는 것으로 측정한다. 그러나 주변 공간단위의 값(절대값)이 더 커지면 음의 공간적 자기상관이 있는 것으로 측정할 것이다. 이에 비해 G_i^* 와 S_i^* 는 둘 다 상당히 높은 양의 공간적 자기상관이 있는 것으로 측정한다. 여섯 번째 준거는 '이차형식의 비'로의 표현가능성이다. 이차형식의 비로 표현될 수 있어야만 해당 통계량의 표본분포에 대한 다양한 사항을 손쉽게 파악할 수 있다. 이미 표 1에 나타나 있는 바처럼, G_i^* 만 유일하게 이차 곱의 형태로 표현될 수 없다.

표 3은 여섯 가지 준거에 기초한 네 종류의 SAS에 대한 비교 검토의 결과를 보여주고 있다. 이 표를 요약하면, I_i 는 공간 특이점과 공간 체제, c_i 는 중간스팟과 국지적 안정성, G_i^* 와 S_i^* 는 핫스팟과 콜드스팟 탐지에 우수한 것으로 판단된다. 또한 전자의 두 통계량은 중심 공간단위에 대한 의존도가 심한 반면, 후자의 두 통계량은 심하지 않다. 이 표를 바탕으로 두 가지 중요한 결과가 도출된다.

첫째, 네 개의 SAS가 성격이 매우 다른 두 부류의 SAS로 범주화될 수 있다. 모런 통계량과 기어리 통계량은 중심 공간단위와 주변 공간단위를 엄격히 구분하고, 중심 공간단위의 값과 주변 공간단위의 대푯값으로서의 SL 사이의 (비)유사도(모런 통계량의 경우는 공분산율, 기어리 통계량의 경우는 차이의 제곱)를 측정한다. 이에

반해 게티스-오드 통계량과 S_i^* 은 중심 공간단위와 주변 공간단위를 결합하여 하나의 국지 세트를 상징하고 공간적 자기상관을 국지 세트 전체의 대푯값으로서의 SMA에 의거해 측정한다. 따라서 엄밀한 의미에서 G_i^* 와 S_i^* 은 공간연관성통계량이라기 보다는 ‘공간군집성통계량 (spatial clustering statistics)’이다.¹⁶⁾ 이러한 두 범주 간의 차이는 표 3에서 핫스팟과 콜드스팟 준거에 대해 모런과 기어리 통계량에게는 보통, G_i^* 과 S_i^* 에게는 우수 등급을 부여한 것과 직접적으로 관련되어 있다. 공간 클러스터는 2차원 객체이며, 통상적으로는 연결한 공간단위의 연속체로 드러난다. 이런 측면에서 볼 때 한 클러스터를 구성하는 모든 공간단위가 나머지 공간단위와 반드시 동일한 성격을 가져야 할 필요는 없다. 즉, 한 공간단위가 전체 클러스터의 성격을 잘 대변하지 못한다 해도 공간 클러스터의 일원이 될 수는 있는 것이다. 예를 들어 A3의 경우 중심 공간단위는 주변 공간단위와 이질적이긴 하지만 국지 세트 전체로 보면 낮은 값의 클러스터가 두드러진 것이 사실이다. 주변 공간단위의 값이 모두 -2로 바뀐다면 어떤 측면에서는 중심 공간단위의 공간 특이점으로서의 특성이 강화되었다고 말할 수도 있겠지만, 다른 측면에서는 국지 세트 전체의 콜드스팟으로서의 특성이 강화되었다고도 말할 수 있다. 따라서 공간 클러스터의 탐지가 주목적이라면 후자의 SAS가 전자의 SAS에 비해 더 높은 가능성을 보인다고 말할 수 있다.

둘째, G_i^* 의 장점에도 불구하고 S_i^* 가 G_i^* 를 대체할 충분한 이유가 있다. G_i^* 은 S_i^* 에 비해 다음의 두 가지 정도의 장점이 있다. 우선 앞서서도 언급한 것처럼, G_i^* 은 표준화점수 그 자체이기 때문에 부호를 통해 핫스팟과 콜드스팟을 직관적으로 구분할 수 있게 해준다. 둘째, G_i^* 의 상수 부분으로 인해 중심 공간단위로부터 점차적으로 확장하면서 보다 많은 이웃을 포함시켜 계산하더라도 통계량이 증가할 수 있다. 식 (8)에서 보면 n_i^* 가 증가하면 상수 부분이 증가하게 되는데, 더 많은 이웃을 받아들이는 과정에서 표준화 SMA가 감소하더라도 그것을 별충할 만큼의 상수에서의 증가가 있으면 전체 G_i^* 은 증가하게 된다. 이는 공간 클러스터의 범역을 설정하는 연구에서 강점일 수 있다(Geity and Aldstadt, 2004; Aldstadt and Geity, 2006; 이상일 등, 2010; Lee and Cho, 2013; 2015). 반대로, S_i^* 은 G_i^* 에 비해 몇 가지 중요한 장점을 가진다. 첫째, S_i^* 은 G_i^* 와 달리 LISA의 두 조건을 모두 만족시킨다. S_i^* 의 평균은 전역적 S^* 이다. 둘째,

S_i^* 은 G_i^* 와 달리 이차형식의 비로 표현된다. 이 속성은 국지적 모런 통계량과 국지적 기어리 통계량도 함께 공유하고 있는 것이다. 뒤에서 살펴 볼 것이지만 이차형식의 비로 표현될 수 있다는 것은 그 통계량의 표본분포 (기댓값, 분산, 왜도, 첨도)에 대한 자세한 사항을 손쉽게 얻을 수 있다는 점을 의미한다. 실질적으로, 게티스와 오드는 그들이 최종적으로 제안한 G_i^* 에 대해서는 기댓값과 분산을 포함한 표본분포에 대한 어떠한 논의도 제공하지 않았다. 따라서, 엄밀한 의미에서는, G_i^* 에 대한 유의성검정을 행할 수 없다.¹⁷⁾ 결론적으로, S_i^* 가 G_i^* 와 동일한 것을 측정하고, 전역적 통계량과 쌍을 이루고 있으며, 표본분포에 대한 사항이 선명하게 알려져 있기 때문에, S_i^* 가 G_i^* 를 대체하는 것은 매우 합당한 일이다.

III. 국지적 공간연관성통계량의 가능치 범위와 표본분포 특성 비교

1. 분석 방법론

여기에서 사용될 분석 방법은 이상일 등(2015)이 전역적 SAS 연구에서 사용했던 중심적률 추출법과 고유치 및 고유벡터 추출법이다. 표 1에 나타나 있는 네 가지 국지적 SAS 중 G_i^* 를 제외한 나머지 SAS는 모두 다음과 같이 이차형식의 비로 표현될 수 있다(Tiefelsdorf, 2000; Lee, 2008; 2009).

$$\Gamma_i = \frac{\delta^T \mathbf{T}_i \delta}{\delta^T \delta} \quad (9)$$

여기서 Γ_i 는 국지적 SAS를 의미하고, \mathbf{T}_i 는 표준화된 국지적 SPM이다. 표 1의 두 번째 매트릭스 표현으로부터 각 SAS에 해당하는 \mathbf{T}_i 매트릭스를 도출할 수 있다 (Lee, 2008).¹⁸⁾

$$\mathbf{T}_i(I_i) \equiv n^2 \frac{\mathbf{V}_i}{\mathbf{1}^T \mathbf{V}_i \mathbf{1}},$$

$$\mathbf{T}_i(c_i) \equiv \frac{n(n-1)}{2} \frac{[\mathbf{\Omega}_i - (\mathbf{V}_i + \mathbf{V}_i^T)]}{\mathbf{1}^T \mathbf{V}_i \mathbf{1}},$$

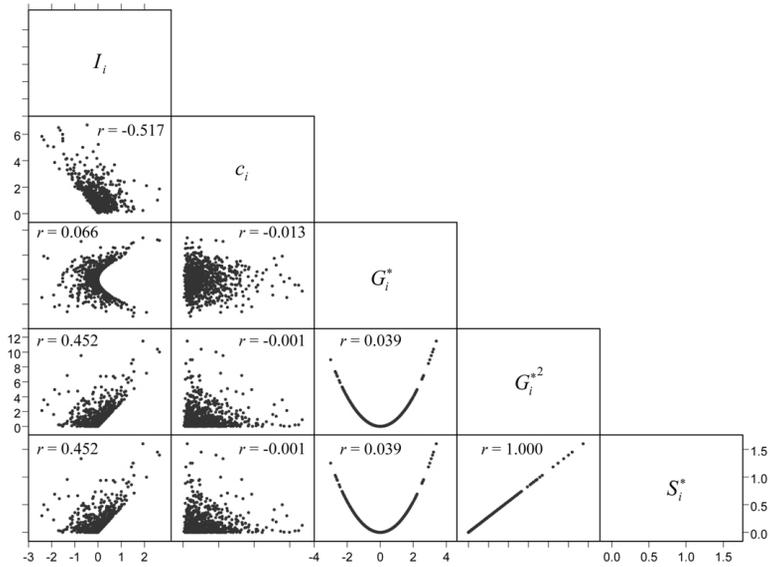


그림 1. 국지적 공간연관성통계량 간의 상관관계 매트릭스

* 육각형(256개) 테셀레이션을 사용하였고, 주변 공간단위 개수가 6인 경우에 대해 분석함.

상관에서 양의 공간적 자기상관에 이르는 스펙트럼에서의 특정 위치를 나타낸다. 이에 반해 G_i^* 통계량은 양의 공간적 자기상관에 대해서, 콜드스팟에서 핫스팟에 이르는 스펙트럼에서의 특정 위치를 나타낸다. 한편 S_i^* 는 국지적 모런 통계량과는 0.452의 다소 높은 상관관계를 보였지만, 국지적 거거리 통계량과는 아무런 관련성이 없는 것으로 드러났다. 이렇게 상대적으로 낮은 SAS간 상관성은 각각의 SAS가 공간적 자기상관의 국지적 상황에 대해 서로 다르게 측정하고 있음을 반증하는 것이다. 전역적 SAS 간의 높은 상관성은 국지적 SAS가 합산되고 평균화되었을 때는 이러한 구체적 차이가 상쇄효과를 통해 드러나지 않게 되었기 때문에 발생한 것으로 해석된다.

둘째, G_i^* 와 S_i^* 는 동일한 통계량인 것으로 밝혀졌다. G_i^* 와 S_i^* 간의 구조적 상동성을 살펴보기 위해 G_i^* 를 제공한 새로운 측도를 산출한 후, 그것과 S_i^* 간의 상관관계를 측정하였는데, 완벽한 정적 상관관계가 나타났다. 이는 G_i^* 와 S_i^* 가 기본적으로 동일한 것을 측정하고 있다는 사실을 입증하는 것이며, S_i^* 가 G_i^* 의 대안이 될 수 있음을 다시 한번 보여주는 것이다.²⁰⁾

2) 가능치 범위 분석

서로 다른 공간단위의 형태(삼각형, 사각형, 육각형)와 서로 다른 공간단위 개수(64, 256, 1024)에 대해 이론적인 가능치 범위, 즉 최댓값과 최솟값을 산출하였다(표 4). 주요 연구 결과는 다음과 같다. 첫째, 국지적 SAS의 가능치 범위는 전역적 SAS의 가능치 범위에 비해 훨씬 넓다. 예를 들어, 256개 사각형의 경우 전역적 모런 통계량의 가능치 범위는 -1.0485~1.0216이었지만, 같은 경우의 국지적 모런 통계량의 가능치 범위는 -68.1321~67.0655에 이른다. 둘째, 전체 공간단위 수가 늘어날수록 가능치의 범위는 급격히 늘어난다. 예를 들어, 육각형의 경우 전체 공간단위 개수가 64개인 경우의 가능치 범위는 -15.3111~14.1185이지만 1,024개인 경우는 -217.8127~216.7696으로 크게 넓어진다. 셋째, 공간단위 형태에 따라 가능치의 범위는 달라진다. 즉, 기본 도형의 변의 개수가 늘어날수록(삼각형에서 육각형으로 갈수록) 가능치의 범위는 줄어든다. 예를 들어 공간단위가 256개인 경우, 국지적 모런 통계량의 가능치 범위는 삼각형일 때의 -78.7444~77.6778에서 육각형일 때의 -56.6891~55.5997로 줄어든다. 넷째, 전역적 SAS와 달리, 국지적 모런 통계량의 가능치 범위가 육각형의 경우에도 0을 기준으로

표 4. 국지적 공간연관성통계량의 가능치 범위: 공간단위의 형태와 개수의 효과

공간단위 형태	공간단위 개수	통계량 별 가능치 범위					
		국지적 모런 통계량		국지적 기어리 통계량		S_i^* 통계량	
		최대	최소	최대	최소	최대	최소
삼각형	64	19,8806	-21,0235	48,0000	0,0000	15,0000	0,0000
	256	77,6778	-78,7444	181,3333	0,0000	63,0000	0,0000
	1,024	304,0267	-305,0589	704,0000	0,0000	255,0000	0,0000
사각형	64	16,9948	-18,1376	45,0000	0,0000	11,8000	0,0000
	256	67,0655	-68,1321	170,0000	0,0000	50,2000	0,0000
	1,024	263,0965	-264,1287	660,0000	0,0000	203,8000	0,0000
육각형	64	14,1185	-15,3111	43,8261	0,0000	8,1429	0,0000
	256	55,5997	-56,6891	162,0426	0,0000	35,5714	0,0000
	1,024	216,7696	-217,8127	622,4842	0,0000	145,2857	0,0000

* 연결성 유형은 루크(rook)임. 공간단위 형태별로 최대의 주변 공간단위 개수(삼각형인 경우는 3개, 사각형인 경우는 4개, 육각형인 경우는 6개)를 갖는 경우에 대해서만 계산함. SPM은 모런 통계량과 기어리 통계량의 경우 C^0 , S_i^* 통계량의 경우는 W^* 를 적용함.

표 5. 국지적 공간연관성통계량의 가능치 범위: 주변 공간단위 개수와 행표준화의 효과

SPM의 종류	주변 공간단위 개수	통계량 별 가능치 범위					
		국지적 모런 통계량		국지적 기어리 통계량		S_i^* 통계량	
		최대	최소	최대	최소	최대	최소
C^0 혹은 C^*	2	32,4917	-32,8548	69,4468	0,0000	17,5777	0,0000
	3	39,6652	-40,2099	92,5957	0,0000	23,3443	0,0000
	4	45,6616	-46,3878	115,7447	0,0000	29,0646	0,0000
	6	55,5997	-56,6891	162,0426	0,0000	40,3662	0,0000
W^0 혹은 W^*	2	89,4792	-90,4792	191,2500	0,0000	84,3333	0,0000
	3	72,8229	-73,8229	170,0000	0,0000	63,0000	0,0000
	4	62,8739	-63,8739	159,3750	0,0000	50,2000	0,0000
	6	51,0388	-52,0388	148,7500	0,0000	35,5714	0,0000

* 256개 육각형의 경우임. 모런과 기어리 통계량의 경우는 C^0 혹은 W^0 를, S_i^* 통계량의 경우는 C^* 혹은 W^* 를 적용함.

음수와 양수 방향으로 대칭을 이루고 있다. 전역적 모런 통계량의 경우, 삼각형 및 사각형과 달리, 육각형의 경우는 음수 방향(음의 공간적 자기상관)으로 일종의 범위 단축이 발생하는 것을 발견할 수 있었는데, 국지적 SAS의 경우는 이러한 현상이 나타나지 않는다. 기어리 통계량의 경우도 동일한 문제점이 발견되었는데 국지적 통계량의 경우는 그렇지 않은 것으로 보인다. 다섯째, 국지적 기어리 통계량과 S_i^* 는 최솟값이 0이고 최댓값이 매우 큰 양의 값인 공통점을 보인다. 그러나 그 방향성은 반대이다. 즉, 국지적 기어리 통계량의 경우는 음의 공간적 자기상관에 대해 극단적인 값을 산출할 수 있지만(양의 공간적 자기상관의 한계치는 0), S_i^* 의 경우는

양의 공간적 자기상관에 대해 극단적인 값을 산출할 수 있다(음의 공간적 자기상관의 한계치는 0).

다음으로 국지적 SAS 통계량의 가능치 범위에 대한 주변 공간단위 개수와 행표준화의 효과를 검토하였다(표 5). 이를 위해 256개 육각형으로 이루어진 테셀레이션에 대상으로 이웃수가 각각 2개, 3개, 4개, 6개인 육각형에 대해 가능치 범위를 구했다. 또한 행표준화의 효과를 알아보기 위해 네 가지 서로 다른 SPM을 사용하였다. 연결성에 기반한 이항 SPM(주대각 요소는 0)을 C^0 , 이것을 행표준화한 SPM을 W^0 , C^* 는 C^0 의 주대각선에 1을 넣은 SPM을, W^* 은 C^* 를 행표준화한 SPM을 의미한다.

가장 중요한 연구결과는, 가능치 범위에 대한 주변 공

표 6. 국지적 공간연관성통계량의 표본분포 특성: 공간단위의 형태와 개수의 효과

통계량	공간단위 형태	공간단위 개수	표본분포 특성			
			기댓값	분산	왜도	첨도
국지적 모런 통계량	삼각형	64	-0.018141	0.408891	-0.154893	7.987376
		256	-0.004183	0.373374	-0.040121	8.726264
		1,024	-0.001009	0.353800	-0.010119	8.930165
	사각형	64	-0.018141	0.301722	-0.180282	7.992798
		256	-0.004183	0.278928	-0.046419	8.726623
		1,024	-0.001009	0.265091	-0.011690	8.930188
	육각형	64	-0.018929	0.211838	-0.224425	8.004183
		256	-0.004272	0.192416	-0.057076	8.727350
		1,024	-0.001020	0.180114	-0.014332	8.930234
국지적 기어리 통계량	삼각형	64	1.142857	1.225746	2.263560	10.839593
		256	1.066667	1.120069	2.397998	12.094830
		1,024	1.032258	1.061398	2.432753	12.438230
	사각형	64	1.142857	1.067504	2.266261	10.914658
		256	1.066667	0.978954	2.397951	12.156292
		1,024	1.032258	0.928464	2.432054	12.496125
	육각형	64	1.192547	0.990047	2.310080	11.212086
		256	1.089362	0.873870	2.436503	12.447291
		1,024	1.043124	0.812365	2.469388	12.785611
S_i^* 통계량	삼각형	64	0.238095	0.108146	2.636701	12.952105
		256	0.247059	0.120651	2.779172	14.450941
		1,024	0.249267	0.123904	2.816027	14.860229
	사각형	64	0.187302	0.066925	2.636701	12.952105
		256	0.196863	0.076605	2.779172	14.450941
		1,024	0.199218	0.079143	2.816027	14.860229
	육각형	64	0.129252	0.031870	2.636701	12.952105
		256	0.139496	0.038464	2.779172	14.450941
		1,024	0.142019	0.040221	2.816027	14.860229

* 연결성 유형은 루크(rook)임. 공간단위 형태별로 최대의 주변 공간단위 개수(삼각형인 경우는 3개, 사각형인 경우는 4개, 육각형인 경우는 6개)를 갖는 경우에 대해서만 계산함. SPM은 모런 통계량과 기어리 통계량의 경우 C^0 , S_i^* 통계량의 경우는 w^* 를 적용함.

간단위 개수의 효과가 행표준화의 여부에 따라 달라진다는 것이다. 즉, 모든 통계량에 대해, C^0 와 C^* 의 경우는 주변 공간단위 개수가 증가할수록 가능치 범위가 증가하지만, w^0 와 w^* 의 경우는 주변 공간단위 개수가 증가할수록 가능치 범위가 감소한다. 이는 국지적 모런 통계량에 대한 기존의 연구 결과(Tiefelsdorf *et al.*, 1999)가 기어리 통계량과 S_i^* 통계량에도 그대로 적용됨을 보여주는 것이다. 앞에서 언급한 것처럼, 전역적 SAS는 국지적 SAS의 평균값과 같다. 이는 극단적인 국지적 SAS 값이 전역적 SAS 값에 지대한 영향을 끼친다는 것을 의미한다. 이 연구 결과는, 이항 SPM을 사용할 경우는 주변 공간단위 개수가 많은 공간단위(즉, 중심부에 위치한 공간단위)의 영향력이 상대적으로 커지고, 행표준화 SPM을 사용할 경우는 이웃수가 적은 공간단위(즉, 주변부에 위치한 공간단위)의 영향력이 상대적으로 커짐을 의미한다.

다. 결국 사용된 SPM의 종류가 SAS 그 자체에 지대한 영향을 끼침을 다시 한번 확인해 볼 수 있는 것이다.

3) 표본분포 특성 분석

공간단위의 형태와 개수가 SAS의 표본분포 상의 특성에 미치는 영향력을 살펴보기 위해, 서로 다른 공간단위의 형태(삼각형, 사각형, 육각형)와 서로 다른 공간단위 개수(64, 256, 1024)에 대해 중심적률 추출법을 적용하여 기댓값, 분산, 왜도, 첨도를 산출하였다(표 6). 주요 분석 결과를 다음과 같다. 첫째, 전역적 SAS에 대한 동일한 분석의 결과와 비교해 볼 때(이상일 등, 2015; 표 3 참조), 왜도와 첨도의 경우 전역적 SAS보다 훨씬 더 큰 값을 보여주고 있다. 이는 다시 한번 기댓값과 분산을 이용한 가설검정, 즉 정규근사(normal approximation)의 타당성이 국지적 SAS에서는 현저히 떨어짐을 보여주

표 7. 국지적 공간연관성통계량의 표본분포 특성: 주변 공간단위 개수와 행표준화의 효과

통계량	SPM의 종류	주변 공간단위 개수	표본분포 특성			
			기댓값	분산	왜도	첨도
국지적 모런 통계량	C^0	2	-0.001424	0.065161	-0.032695	8.725907
		3	-0.002136	0.097358	-0.040121	8.726264
		4	-0.002848	0.129299	-0.046419	8.726623
		6	-0.004272	0.192416	-0.057076	8.727350
	W^0	2	-0.003922	0.494179	-0.032695	8.725907
		3	-0.003922	0.328160	-0.040121	8.726264
		4	-0.003922	0.245151	-0.046419	8.726623
		6	-0.003922	0.162142	-0.057076	8.727350
국지적 기어리 통계량	C^0	2	0.363121	0.162512	2.457029	12.363563
		3	0.544681	0.292060	2.397998	12.094830
		4	0.726241	0.453802	2.397951	12.156292
		6	1.089362	0.873870	2.436503	12.447291
	W^0	2	1.000000	1.232490	2.457029	12.363563
		3	1.000000	0.984436	2.397998	12.094830
		4	1.000000	0.860409	2.397951	12.156292
		6	1.000000	0.736381	2.436503	12.447291
S_i^* 통계량	C^*	2	0.068932	0.009392	2.779172	14.450941
		3	0.091546	0.016566	2.779172	14.450941
		4	0.113979	0.025679	2.779172	14.450941
		6	0.158299	0.049532	2.779172	14.450941
	W^*	2	0.330719	0.216197	2.779172	14.450941
		3	0.247059	0.120651	2.779172	14.450941
		4	0.196863	0.076605	2.770172	14.450941
		6	0.139496	0.038464	2.779172	14.450941

* 256개 육각형의 경우임. 모런과 기어리 통계량의 경우는 C^0 혹은 W^0 를, S_i^* 통계량의 경우는 C^* 혹은 W^* 를 적용함.

는 것이다. 둘째, 기댓값의 경우 공간단위의 형태와 개수의 효과는 SAS 별로 달라진다. 모런과 기어리 통계량은 공간단위의 개수가 많아질수록 기댓값의 절대값이 감소하는 반면, S_i^* 통계량의 경우는 증가한다.²¹⁾ S_i^* 통계량의 경우는 삼각형에서 육각형으로 갈수록 기댓값은 감소하고, 공간단위 개수가 많아질수록 기댓값은 체계적으로 증가한다. 셋째, 분산의 경우는, 모런과 기어리 통계량의 경우는 삼각형에서 육각형으로 갈수록, 공간단위 개수가 많아질수록 감소하지만, S_i^* 통계량의 경우는 삼각형에서 육각형으로 갈수록 분산은 감소하고, 공간단위 개수가 많아질수록 분산은 증가한다. 결론적으로 표본분포 특성이라는 측면에서 보면, S_i^* 는 모런과 기어리 통계량과는 상당히 다른 특성을 보인다.

다음으로는 주변 공간단위 수와 행표준화가 국지적 SAS의 표본분포 특성에 미치는 영향력을 살펴보았는데 (표 7), 주요 분석 결과는 다음과 같다. 첫째, 모런 통계량과 기어리 통계량의 기댓값은 C^0 의 경우 주변 공간단위 개수가 달라짐에 따라 변화하지만, W^0 의 경우에는

일정한 값을 보인다. 사실상 이 값은 전역적 SAS의 기댓값과 일치한다(이상일 등, 2015; 표 3 참조). 이에 반해 S_i^* 통계량의 경우는 C^* 보다 W^* 를 적용한 경우 기댓값이 커진다. 둘째, 분산은 세 SAS 모두에 대해 비-행표준화 SPM의 경우는 주변 공간단위 개수가 늘어날수록 증가하지만, 행표준화 SPM의 경우는 이웃수가 증가할수록 감소한다. 이는 앞에서 살펴본 연구결과와 동일하다. 셋째, 모든 SAS에 대해 왜도와 첨도에 대한 행표준화의 효과는 존재하지 않는다. 왜도와 첨도에 대한 주변 공간단위 개수의 효과는 왜도에 대한 모런 통계량의 경우를 제외하고는 미미하다. 실질적으로 S_i^* 의 경우는 왜도와 첨도에서 주변 공간단위 개수와 행표준화의 효과는 존재하지 않는다.

3. 우리나라 7대 대도시의 읍면동 단위 분석

실질적인 공간분석에서의 함의를 보다 구체적으로 살펴보기 위해 앞에서 제시된 방법론을 우리나라 7대 대도

표 8. 우리나라 7대 대도시에 대한 국지적 공간연관성통계량의 가능치 범위 비교

도시	평균 이웃 수	SPM	통계량 별 가능치 범위					
			국지적 모런 통계량		국지적 기어리 통계량		S_i^* 통계량	
			최대	최소	최대	최소	최대	최소
서울 (423)	5,858	C^0 혹은 C^*	87,1898	-88,2140	252,1271	0,0000	58,7988	0,0000
		W^0 혹은 W^*	85,1286	-86,1286	246,1667	0,0000	59,4286	0,0000
부산 (214)	5,579	C^0 혹은 C^*	45,6660	-46,7414	133,6156	0,0000	31,0024	0,0000
		W^0 혹은 W^*	42,4651	-43,4651	124,2500	0,0000	29,5714	0,0000
대구 (139)	5,799	C^0 혹은 C^*	28,0974	-29,1322	83,2965	0,0000	18,9350	0,0000
		W^0 혹은 W^*	27,1541	-28,1541	80,5000	0,0000	18,8571	0,0000
인천 (146)	5,123	C^0 혹은 C^*	33,4745	-33,6456	99,0575	0,0000	23,4265	0,0000
		W^0 혹은 W^*	28,5832	-29,5832	84,5833	0,0000	19,8571	0,0000
광주 (94)	5,638	C^0 혹은 C^*	19,1187	-20,1829	57,7302	0,0000	12,8701	0,0000
		W^0 혹은 W^*	17,9662	-18,9662	54,2500	0,0000	12,4286	0,0000
대전 (77)	5,584	C^0 혹은 C^*	15,5731	-16,6475	47,6326	0,0000	10,6552	0,0000
		W^0 혹은 W^*	14,4944	-15,4944	44,3333	0,0000	10,0000	0,0000
울산 (56)	5,179	C^0 혹은 C^*	11,8230	-12,9816	37,1724	0,0000	8,2936	0,0000
		W^0 혹은 W^*	10,2044	-11,2044	32,0833	0,0000	7,0000	0,0000

* 도시 칼럼의 괄호 안의 숫자는 전체 공간단위의 개수임. 모런 통계량과 기어리 통계량의 경우는 C^0 혹은 W^0 를, S_i^* 통계량의 경우는 C^* 혹은 W^* 를 적용함.

시의 읍면동 데이터에 적용하였다. 우선, SAS의 가능치 범위가 도시별로 차이가 나는지를 살펴보았는데(표 8), 주요 분석 결과는 다음과 같다. 첫째, 모든 SAS에 있어 가능치 범위가 도시 별로 엄청난 격차를 보여주고 있다. 예를 들어 모런 통계량의 경우, 서울의 가능치 범위는 180 정도인데 반해, 울산의 경우는 24 정도에 불과하다. 이는 표 4에 나타나 있는 연구결과를 그대로 반영하는 것으로, 국지적 SAS 통계치를 도시간에 단순 비교하는 것은 아무런 의미가 없음을 보여주는 것이다. 둘째, 행표준화 SPM을 적용했을 때 가능치 범위가 다소 감소하는 경향이 나타나긴 하지만 무시할 정도인 것으로 나타났다.

두 번째 분석은 SAS의 표본분포 특성이 7대 대도시별로 차이를 보이는지에 대한 것인데(표 9), 주요 분석 결과는 다음과 같다. 첫째, 기댓값의 경우는 전체적으로 표 6에 나타난 일반적인 경향을 따라가지만 분산의 경우는 그렇지 않다. 즉, 기댓값(절대값)의 경우 공간단위 개수가 많은 서울과 부산이 여타 도시들에 비해 모런과 기어리 통계량의 기댓값(절대값)은 낮고, S_i^* 통계량의 기댓값은 높았다.²³⁾ 그러나 서울의 분산이 가장 작고 울산의 분산이 가장 클 것으로 기대되었지만, 실질적으로는

대구의 분산이 가장 작고 인천의 분산이 가장 큰 것으로 드러났다. 둘째, 왜도와 첨도의 경우는 표 6에 나타난 일반적 경향을 잘 따라간다. 즉, 공간단위 개수가 많을수록 왜도와 첨도는 증가한다. 따라서 모든 국지적 SAS에 대해서 서울의 왜도와 첨도가 가장 크다. 이는 정규근사의 한계성이 서울에서 가장 극심하게 나타날 것임을 함축하고 있다.

IV. 결론

본 논문의 주된 연구목적은 새로운 국지적 SAS로서의 S_i 통계량의 특성을 기존의 I_i , e_i , 그리고 G_i^* 통계량과의 비교 연구를 통해 밝히는 것이었다. 일반 S_i 통계량의 두 파생 통계량 중 S_i^* 통계량의 활용성이 훨씬 더 큰 것으로 인정되었다. S_i^* 은 각 국지 세트가 평활화 효과로 인한 분산의 감소에 어느 방향으로 얼마나 기여하는가를 측정하는데, 국지 세트가 높은 양의 공간적 자기상관을 보인다면 공간평활화에 의한 분산의 감소가 최소화되기 때문에 매우 높은 S_i^* 를 갖게 되고, 반대로 국

표 9. 우리나라 7대 대도시에 대한 국지적 공간연관성통계량의 표본분포 특성 비교

통계량	도시	표본분포 특성			
		기댓값	분산	왜도	첨도
국지적 모런 통계량	서울	-0.002427	0.171955	-0.034622	8.832964
	부산	-0.005049	0.186489	-0.068201	8.675806
	대구	-0.007498	0.169580	-0.104610	8.510663
	인천	-0.008077	0.217766	-0.099645	8.532866
	광주	-0.011442	0.174955	-0.153911	8.295532
	대전	-0.014137	0.175319	-0.187250	8.155432
	울산	-0.021066	0.196670	-0.255661	7.881058
국지적 기어리 통계량	서울	1.024213	0.778100	2.453748	12.623804
	부산	1.075377	0.848501	2.427984	12.360823
	대구	1.034739	0.776247	2.400216	12.082287
	인천	1.171123	0.995991	2.403993	12.119871
	광주	1.064151	0.807592	2.362897	11.715780
	대전	1.074419	0.813983	2.337843	11.474686
	울산	1.158621	0.924373	2.286896	10.996404
S_i^* 통계량	서울	0.140826	0.039383	2.798508	14.664539
	부산	0.138833	0.038011	2.769613	14.346257
	대구	0.136646	0.036544	2.738415	14.008816
	인천	0.136946	0.036743	2.742661	14.054369
	광주	0.133641	0.034592	2.696389	13.564239
	울산	0.131579	0.033294	2.668108	13.271415

* 모런 통계량과 기어리 통계량의 경우는 C^0 를, S_i^* 통계량의 경우는 w^* 를 적용함.

지 세트가 높은 음의 공간적 자기상관을 보인다면, 공간 평활화에 의한 분산의 감소가 심대하게 발생하기 때문에 매우 낮은 S_i^* 를 갖게 된다.

SAS간 특성 비교를 위한 6가지 준거, 즉 공간 클러스터(핫스팟, 콜드스팟, 중간스팟), 공간 특이점, 공간 체제, 국지적 인정성, 중심 공간단위에서의 의존성, 이차형식의 비로의 표현가능성에 의거해 국지적 SAS를 비교한 결과 I_i 는 공간 특이점과 공간 체제, c_i 는 중간스팟과 국지적 안정성, G_i^* 와 S_i^* 는 핫스팟과 콜드스팟 탐지에 우수한 것으로 판명되었다. 이와 관련하여 두 가지 중요한 결론이 도출되었다. 첫째는 국지적 SAS는 크게 두 범주로 구분된다는 것인데, 중심 공간단위와 주변 공간단위 간의 연관성에 집중하는 I_i 와 c_i , 그리고 국지 세트 전체를 공간 클러스트의 기본 단위로 취급하는 G_i^* 와 S_i^* 로 나뉜다. 둘째는 S_i^* 가 G_i^* 를 대체하는 통계량으로 사용될 수 있다는 것인데, 그 근거로 S_i^* 가 LISA의 두 조건을

모두 만족시킨다는 점과 이차형식의 비로 정의될 수 있다는 점 등이 제시되었다.

정다각 테셀레이션 분석은 공간단위의 개수와 형태, 인접 공간단위 수, 행표준화 여부 등이 국지적 SAS의 가능치 범위와 표본분포 특성에 어떠한 영향을 끼치는지를 밝혀내기 위해 이루어졌는데, 주요 결과를 요약하면 다음과 같다. 첫째, 전역적 SAS 간의 상관관계에 비해 국지적 SAS 간의 상관관계는 훨씬 낮다. 그리고 G_i^* 와 S_i^* 는 기본적으로 동일한 통계량임이 밝혀졌다. 둘째, 국지적 SAS의 가능치 범위는 전역적 SAS의 가능치 범위에 비해 훨씬 넓다. 그리고 전체 공간단위 수가 늘어날수록 가능치의 범위는 급격히 늘어나며, 공간단위 형태에 따라 가능치 범위가 달라진다. 셋째, 주변 공간단위 수가 가능치 범위에 끼치는 효과는 SPM의 종류에 따라 상이하게 나타난다. 넷째, 국지적 SAS는 전역적 SAS에 비해 훨씬 더 큰 왜도와 첨도를 보인다. 다섯째, 기댓값과 분

산에 대한 공간단위의 형태와 개수의 효과는 SAS 별로 다양하게 나타난다. 여섯째, 기댓값과 분산에 대한 공간 단위 수와 행표준화의 효과는 다양하게 나타나지만, 왜도와 첨도에 대한 효과는 거의 존재하지 않는다. 일곱째, 우리나라 7대 대도시에 적용한 결과, 도시별 공간단위 개수에 따라 가능치 범위가 현저히 달라지지만, 분산의 경우는 공간단위 개수와 체계적인 관련성을 갖지는 않는 것으로 드러났다.

본 연구는 국지적 SAS들에 대한 전면적인 비교 연구라는 측면에서 큰 의의가 있는 것으로 평가된다. 국지적 SAS를 활용한 ESDA 방법론이 제안되고, 다양한 소프트웨어를 통해 그 이용가능성이 증대됨에 따라, 일반 연구자들은 기법을 기계적으로 적용하고, 그 결과를 무비판적으로 받아들이는 경향을 보여왔다.²³⁾ 그러나 본 연구가 밝힌 바처럼, SAS는 국지적 공간적 자기상관의 서로 다른 측면을 측정하고 있으며, 따라서 사용자들은 자신의 연구 목적이 무엇이나에 따라 서로 다른 SAS와 그것을 활용한 ESDA 방법론을 선택할 수 있어야 한다. 또한 가능치 범위와 표본분포 특성에서도 다양한 차이점들이 확인되었다. 이는 국지적 SAS를 통한 연구의 결과를 해석하는데 있어 훨씬 더 높은 수준의 엄정함과 치밀함이 요구된다는 점을 시사하고 있다. 특히 유의성 검정법으로서의 정규근사가 가지는 타당성이 국지적 SAS에서 현저히 떨어진다는 사실은 많은 시사점을 준다. 이를 극복하기 위해서는 ‘정확분포(exact distribution)’ 접근(Tiefelsdorf, 2000; Leung *et al.*, 2003)이나 ‘안장점근사(saddlepoint approximation)’(Tiefelsdorf, 2002)와 같은 방법들이 고려되어야만 한다.

S_i^* 와 G_i^* 가 국지적 공간적 자기상관을 동일한 방식으로 측정하며, 전자가 후자에 비해 LISA의 조건을 만족시키면서 유의성검정의 측면에서 강점이 있다는 사실은 의미하는 바가 크다. Lee(2001b; 2004a)의 연구가 이미 보여주었던 것처럼, S_i^* 를 G_i^* 와 동일한 방식으로 핫스팟과 콜드스팟을 탐지하는데 사용할 수 있다. 이 논문에서 제시된 중심적률 추출법에 기반한 가설검정(정규성 가정에 근거한 유의성 검정) 뿐만 아니라 총체적 랜덤화(total randomization) 및 조건적 랜덤화(conditional randomization) 가정(Sokal *et al.*, 1998; Lee, 2009)에 기반한 가설검정을 활용함으로써, G_i^* 를 사용했을 때보다 통계학적으로 보다 엄밀한 공간 클러스터 탐지 연구를 수행할 수 있다.²⁴⁾ 따라서 본 연구의 가장 중요한 결론 중

의 하나는 S_i^* 가 G_i^* 를 대체하는 것이 합당하다는 사실이다.

일변량 SAS에 대한 두 편의 논문에서 논의된 사항은 이변량 SAS에 대한 연구로 확장될 필요가 있다. 전역적 이변량 SAS로는 Wartenberg(1985)의 논문에서 발견되는 크로스-모런(Cross-Moran) 혹은 이변량 모런 통계량(Czaplewski and Reich, 1993; Reich *et al.*, 1994), Lee가 제안한 L 통계량(Lee, 2001a; 2001b; 2004b; 2012; 2016), 그리고 이변량 기어리 통계량(이상일, 2007) 등이 있다. 또한 국지적 이변량 SAS로 국지적 크로스-모런(Anselin, *et al.*, 2002), 국지적 이변량 기어리(이상일, 2008), Lee의 L_i 통계량(Lee, 2001a; 2001b; 2009; 2012; 2016) 등이 있다. 이변량 SAS에 대한 전면적인 비교 연구는 지금까지 이루어진 적이 없다. 따라서 일변량 SAS의 연구에서 밝혀진 것들을 바탕으로 이변량 SAS를 분석하는 것이 가장 중요한 향후 연구과제가 될 것이다.

註

- 1) EDA는 통계 분석 결과에 집중하던 기존의 ‘확증적 데이터분석(confirmatory data analysis)’에서 벗어나 다양한 과학적 시각화 기법을 통해 데이터 자체에 대한 연구자의 통찰력을 진작시키는데 주안점을 두는 데이터분석 혹은 통계학의 새로운 패러다임을 지칭한다(Lee, 2005:276).
- 2) 이 용어는 넓게는 데이터분석 전체, 좁게는 SDA에서 발생한 새로운 경향을 강조하기 위해 도입된 용어인데(Lee, 2005:277), 데이터 전체의 평균적인 경향성이 아니라, 데이터 포인트 각각의 특이성에 주목하는 새로운 관점을 지칭한다. Fotheringham(1997; 2000)은 공간적 유사성이 아니라 공간적 차이성을, 전역적 일반성이 아니라 국지적 예외성을, 지도 전체에 대한 단일한 통계치가 아니라 지도화 가능한 통계치의 세트를 강조하는 것이 국지적 전회의 중요한 특징이라고 밝힌 바 있다.
- 3) 1990년대 초중반, 미국의 NCGIA(National Center for Geographic Information and Analysis, 국가지리정보분석센터)와 영국의 ESRC(Economic and Social Research Council, 경제사회연구위원회) 등이 중심이 되어 다양한 논의가 진행되었으며, 그 결실이 저널의 특집호와 몇 권의 책으로 출간된 바 있다(Fotheringham and Rogerson, 1994; Fischer, *et al.*, 1996; Longley and Batty, 1997). SDA

와 GIS의 관계에 대한 일반적인 설명에 대해서는 Goodchild and Haining(2004)를 참고할 수 있다.

- 4) SDA-GIS 통합 프레임워크를 구성하는데 있어 ESDA와 CSDA(confirmatory spatial data analysis, 확증적 공간데이터 분석)의 구분은 중요한 역할을 하였다. Bailey(1994: 21)가 말한, SDA에 도움이 되는 GIS의 특성들은 ESDA와 CSDA 모두에 긍정적인 역할을 하겠지만 ESDA와 훨씬 더 높은 친화성을 보일 것이 지명하다. 거꾸로 “GIS와 친화성이 높은 10개의 SDA 기법(The 10 GISable SDA techniques)” (Openshaw, 1990; Bailey, 1994; Openshaw and Clarke, 1996; Unwin, 1996)도 주로 ESDA에서 비롯된 것들이다. ESDA의 일반적 특성에 대한 논의는 Bivand(2010)를 참고할 수 있다.
- 5) 여러 개의 ESDA-GIS 프레임워크가 제시되었는데(Anselin, 1998; Wise *et al.*, 1999; Lee, 2005), 대부분의 경우 국지적 SAS가 핵심적인 역할을 담당하고 있다.
- 6) ESDA는 “공간분포에 대한 묘사 및 시각화, 특이한 값을 보이는 공간단위 혹은 공간 특이점의 확인, 특징적인 공간연관성을 보이는 패턴의 발견(핫스팟 혹은 콜드스팟), 공간체제 혹은 다른 형태의 공간적 이질성의 확인과 같은 과제를 수행하는 분석기법들의 총체”로 정의될 수 있는데(Anselin, 1999), SAS 없이 이러한 ESDA가 수행되는 것은 거의 불가능하다.
- 7) 대표적인 예가 ESRI의 ArcGIS 프로그램에 포함되어 있는 Spatial Statistics 툴박스이다. ESDA-GIS 프로그램의 소프트웨어 상의 다양한 진전은 2006년 *Geographical Analysis* (38(1))의 여러 논문들과 Fischer and Getis(2009)의 편집서적의 여러 논문들을 참조할 수 있다.
- 8) Longley(2000)가 지적한 것처럼, 분석 기법이 표준화되어 소프트웨어 상에서의 이용가능성이 증대되면 사용자 그룹은 확대되겠지만 학문 사회 전체의 ‘탈숙련화(deskilling)’가 촉진될 수도 있다.
- 9) 안셀린은 LISA의 두 번째 조건을 정의하면서 ‘비례 관계의 성립이라는 모호한 표현을 사용했는데(Anselin, 1995) 이는 이 후의 많은 연구에서 발생한 혼란의 빌미가 되었다.
- 10) ‘국지 세트’라는 용어는 설명의 편의를 위해 도입한 것일 뿐, “특정 공간단위와 그것과 지리적 연관성을 갖는 주변 공간단위를 포괄하는 범역”이라는 일반 개념을 훼손할 의도는 전혀 없다. 주변 공간단위가 중심 공간단위와 경계를 공유하고 있는 일차적인 이웃에 한정될 필요도 없으며, 모든 주변 공간단위가 등가적으로 다루어

질 필요도 없으며, 다양한 거리조락함수(distance-decay functions)에 의해 확률적으로 정의되는 이웃 개념으로도 얼마든지 확장될 수 있다. 기술적으로 말하면, 국지 세트는 공간근접성행렬이 어떠한 방식으로 구성되느냐에 의존적이다.

- 11) Tiefelsdorf(1998)과 Leung *et al.*(2003)은 성형(star-shaped)의 국지적 공간근접성행렬을 제안했지만, 이변량 공간연관성통계량에는 적합하지 않은 것으로 판명되었다(Lee, 2004b).
- 12) 국지적 거거리 통계량은 다소 복잡하게 되어 있지만 가법성 요구조건을 정확히 만족시키고 있다. 전역적 Ω 매트릭스의 정의에 대해서는 Lee(2004)를, 국지적 Ω_i 매트릭스의 정의에 대해서는 Lee(2009)를 참고할 수 있다.
- 13) 실질적으로 여기서 정의된 G_i^0 와 게티스-오드의 원 G_i 와는 조금 다르다. 그러나 본질적으로는 동일하다.
- 14) 양의 공간적 자기상관이 지배적일 경우 1에 가까워지고, 음의 공간적 자기상관이 지배적일 경우 0에 가까워진다. 이상일 등(2015)의 연구에 따르면, 최솟값은 0에 일치하지만, 최댓값은 1을 중심으로 소수 둘째 자리에서 반올림을 했을 경우 1.0이 되는 정도의 수준에서 상회 혹은 하회한다.
- 15) S_i^0 는 주변 공간단위의 상황만을 다룰 뿐 그것과 중심 공간단위 간의 연관성은 다루지 않는다. 이런 의미에서 이 통계량은 S_i^* 에 비해 활용성이 훨씬 낮다고 말할 수 있다. 그러나 어떠한 연구 상황에서는 그 중요성이 인정될 수도 있다.
- 16) 이런 관점에서 보면, 공간적 자기상관 통계량은 공간연관성통계량과 공간근접성통계량으로 구분된다. 그러나 개념적 혼동을 막기 위해 본 논문에서는 더 이상의 개념적 확장을 자제하고자 한다.
- 17) 이런 이유로 Leung *et al.*(2003)은 자신들의 일반화된 유의성 검정법을 제안하고 적용하는 시도에서 G_i^* 를 그대로 사용하지 못하고 ‘수정 G_i^* ’라는 것을 상정해야만 했는데, 실질적으로 G_i^* 는 S_i^* 와 구조적으로 동일하다. 또한 그들은 LISA의 조건을 만족시키는 ‘수정 G_i^* ’의 전역적 통계량도 제시했는데, 당연히 S_i^* 와 구조적으로 동일하다.
- 18) 이상일 등(2015:334)의 식 (7)과 비교해 보면, 전역적 SPM이 국지적 SPM으로 교체되었고, 식 전체에 n 을 곱하는 형식을 취하고 있음을 알 수 있다.
- 19) 이상일 등(2015)이 사용한 256개의 육각형 테셀레이션

의 경우, 좌하와 우상의 꼭지점에 해당하는 2개의 육각형은 2개의 이웃을 가지고 있고, 좌상과 우하의 꼭지점에 해당하는 2개의 육각형은 3개의 이웃을 가지고 있으며, 이 꼭지점 육각형을 제외하고 네 변을 따라 위치하고 있는 56개의 육각형은 4개의 이웃을 가지고 있으며, 이 모두를 제외한 196개의 육각형은 6개의 이웃을 가지고 있다. 이 196개 육각형 중 하나를 선택한 것인데, 실질적으로는 196개 중 어느 것을 선정해도 결과는 동일하다.

- 20) 부호가 핫스팟과 콜드스팟을 구분해 주는 G_i^* 의 장점은 S_i^* 의 내부적 계산에서 제곱 이전의 부호를 저장함으로써 손쉽게 계승될 수 있다.
- 21) 흥미로운 것은 모런 통계량과 기어리 통계량의 경우 삼각형과 사각형의 기댓값은 동일한 반면 육각형의 기댓값은 미세하게 다르다는 사실이다. 삼각형과 사각형의 경우 기댓값이 공간단위 개수별로 동일한 것은, 기댓값은 주변 공간단위 개수와 $n/1^T \mathbf{V}1$ 의 곱에 의해 결정되기 때문인데(Lee, 2009), 기어리 통계량을 예를 들어 보면, 삼각형의 경우($3 \times 256/720=1.066667$)와 사각형의 경우($4 \times 256/960=1.066667$)는 동일하다. 그러나 육각형의 경우($6 \times 256/1440=1.089362$)는 조금 달라진다. 여기서의 기댓값은 Lee(2009)가 ‘총체적 랜덤화 가정’이라고 부른 것에 의거한 기댓값과 동일하다. 그러나 ‘조건적 랜덤화 가정’에 의거한 기댓값과는 다르다.
- 22) 기어리 통계량의 경우는 훨씬 더 불명확한 경향성을 보인다. 인천이 울산보다 기댓값이 높고, 대전이 부산보다 오히려 기댓값이 낮다.
- 23) ArcGIS가 제시하는 ‘유의미한’ G_i^* 를 가진 공간 단위를 ‘실질적으로 존재하는’ 공간 클러스터로 그냥 받아들이는 것이다. 그러나 G_i^* 의 통계학적 유의성은 제대로 논의된 적조차 없으며, 그것에 의거한 공간 클러스터의 ‘실질성’도 제대로 평가된 적이 없다.
- 24) 특히 조건적 랜덤화 가정에 기반한 가설검정은, Anselin (1995)에 의거할 때, 국지적 SAS에 가장 적합한 방법이다. 국지적 모런과 기어리 통계량에 대한 서로 다른 랜덤화 가정에 대해서는 Sokal *et al.*(1998)을 참고할 수 있다.

사사

본 논문 속에 포함되어 있는 아이디어를 발전시키는 데 있어 그 크기를 계속할 수 없을 정도의 영감을 준 두

분께 감사의 마음을 전합니다. 한 분은 미국 오하이오주립대학교 지리학과와 고(故) Lawrence “Larry” Alan Brown (1935~2014) 교수이고, 또 다른 한 분은 미국 텍사스대학교(달러스) 지리공간정보과학(Geospatial Information Sciences) 프로그램의 Michael Tiefelsdorf 교수입니다.

참고문헌

- 국토교통과학기술진흥원, 2016, 「공간정보 SW활용을 위한 오픈소스 가공기술개발 3차년도 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·손학기·채미옥, 2010, “공간 클러스터의 범역 설정을 위한 GIS-기반 방법론 연구: 수정 AMOEBA 기법,” *대한지리학회지*, 45(4), 502-520.
- 이상일·조대현·이민파, 2015, “일반량 공간연관성통계량에 대한 비교 연구 (I): 전역적 S 통계량을 중심으로,” *한국지리학회지*, 4(2), 329-345.
- Aldstadt, J. and Getis, A., 2006, Using AMOEBA to create a spatial weights matrix and identify spatial clusters, *Geographical Analysis*, 38(4), 327-343.
- Anselin, L., 1995, Local indicators of spatial association: LISA, *Geographical Analysis*, 27(2), 93-115.
- Anselin, L., 1996, The Moran scatterplot as an ESDA tools to assess local instability in spatial association, in Fischer, M., Scholten, H., and Unwin, D. eds., *Spatial Analytical Perspectives on GIS*, London: Taylor & Francis, 111-125.
- Anselin, L., 1998, Exploratory spatial data analysis in a geocomputational environment, in Longley, P.A., Brooks, S.M., McDonnell, R., and MacMillan, B. eds., *Geocomputation: A Primer*, Chechester: John Wiley & Sons, 77-94.
- Anselin, L., 1999, Interactive techniques and exploratory spatial data analysis, in Longley, P.A., Goodchild,

- M.F., Maguire, D.J., and Rhind, D.W. eds., *Geographical Information Systems: Principles, Techniques, Management, and Applications*, 2nd edition, New York: John Wiley & Sons, 253-266.
- Anselin, L., Syabri, I., and Smirnov, O., 2002, Visualizing multivariate spatial correlation with dynamically linked windows, in Anselin, L. and Rey, S. eds., *New Tools for Spatial Data Analysis*, Proceedings of the Specialist Meeting, Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara.
- Bailey, T.C., 1994, A review of statistical spatial analysis in geographical information systems, in Fotheringham, A.S. and Rogerson, P. eds., *Spatial Analysis and GIS*, London: Taylor & Francis, 13-44.
- Bivand R.S., 2010, Exploratory spatial data analysis, in Fischer, M.M. and Getis, A. eds., *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, New York: Springer, 219-254.
- 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., 1981, *Spatial Processes: Models and Applications*, London: Pion.
- Czaplewski, R.L. and Reich, R.M., 1993, *Expected value and variance of Moran's bivariate spatial autocorrelation statistic for a permutation test*, Research Paper RM-309, U.S. Department of Agriculture, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO.
- Fischer, M., Scholten, H.J., and Unwin, D., 1996, *Spatial Analytical Perspectives on GIS*, Bristol, PA: Taylor & Francis.
- Fischer, M.M. and Getis, A., 2009, eds., *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, New York: Springer.
- Fotheringham, A.S., 1997, Trends in quantitative methods I: Stressing the local, *Progress in Human Geography*, 21(1), 88-96.
- Fotheringham, A.S., 2000, Context-dependent spatial analysis: A role for GIS, *Journal of Geographical Systems*, 2(1), 71-76.
- Fotheringham, A.S. and Charlton, M., 1994, GIS and exploratory spatial data analysis: An overview of some research issues, *Geographical Systems*, 1, 315-327.
- Fotheringham, A.S. and Rogerson, P. eds., 1994, *Spatial Analysis and GIS*, Philadelphia, PA: Taylor & Francis.
- Getis, A. and Aldstadt, J., 2004, Constructing the spatial weights matrix using a local statistic, *Geographical Analysis*, 36(2), 90-104.
- Getis, A., 1991, Spatial interaction and spatial autocorrelation: A cross-product approach, *Environment and Planning A*, 23(9), 1269-1277.
- 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.
- Goodchild, M.F. and Haining, R.P., 2004, GIS and spatial data analysis: Converging perspectives, *Papers in Regional Science*, 83(1), 363-385.
- Goodchild, M.F. and Longley, P.A., 1999, The future of GIS and spatial analysis, in Longley, P.A., Goodchild, M.F., Maguire, D.J., and Rhind, D.W. eds., *Geographical Information Systems: Principles, Techniques, Management, and Applications*, 2nd edition, New York: John Wiley & Sons, 567-580.
- 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), 626-626.
- Hepple, L.W., 1998, Exact testing for spatial autocorrelation among regression residuals, *Environment and Planning A*, 30(1), 85-107.

- 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., 2004a, Exploratory spatial data analysis of σ -convergence in the U.S. regional income distribution, 1969-1999, *Journal of the Korean Urban Geographical Society*, 7(1), 79-95.
- 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., 2005, Between the quantitative and GIS revolutions: Towards an SDA-centered GIScience, *Journal of Geography Education*, 49, 268-284.
- 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., 2012, Exploring bivariate spatial dependence and heterogeneity: A comparison of bivariate measures of spatial association, *Annual Meeting of the Association of American Geographers*, February 24~28, New York, USA.
- Lee, S.-I., 2016, Correlation and spatial autocorrelation, in Shekhar, S., Xiona, H., and Zhou, X. eds., *Encyclopedia of GIS*, 2nd edition, New York: Springer, forthcoming.
- Lee, S.-I. and Cho, D., 2013, Delineating the bivariate spatial clusters: A bivariate AMOEBA technique, *Annual Meeting of the Association of American Geographers*, April 9~13, Los Angeles, USA.
- Lee, S.-I. and Cho, D., 2015, Developing a bivariate AMOEBA technique, *GeoComputation 2015 Conference*, May 20~23, Dallas, Texas, 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.
- Longley, P.A., 2000, The academic success of GIS in geography: Problems and prospects, *Journal of Geographical Systems*, 2(1), 37-42.
- Longley, P. and Batty, M. eds., 1997, *Spatial Analysis: Modelling in a GIS Environment*, Sidney: Halsted Press.
- Openshaw, S., 1990, Spatial analysis and geographical information systems: A review of progress and possibilities, in Scholten, H. and Stillwell, J. eds., *Geographical Information Systems for Urban and Regional Planning*, Dordrecht: Kluwer, 153-163.
- Openshaw, S. and Clarke, G., 1996, Developing spatial analysis functions relevant to GIS environments, in Fischer, M., Scholten, H., and Unwin, D. eds., *Spatial Analytical Perspectives on GIS*, London: Taylor & Francis, 21-37.
- Ord, J.K., and Getis, A., 1995, Local spatial autocorrelation statistics: distributional issues and an application, *Geographical Analysis*, 27(4), 286-306.
- Reich, R., Czaplewski, R.L., and Bechtold, W.A., 1994, spatial cross-correlation of undistributed, natural shortleaf pine stands Northern Georgia, *Environmental and Ecological Statistics*, 1(3), 201-217.
- Sokal, R.R., Oden, N.L., and Thomson, B.A., 1998, Local spatial autocorrelation in a biological model, *Geographical Analysis*, 30(4), 331-354.
- Tiefelsdorf, M., 1998, Some practical applications of Moran's I 's exact conditional distribution, *Papers in Regional Science*, 77(2), 101-129.
- 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 's reference distributions and their numerical evaluation, *Geographical*

Analysis, 34(3), 187-206.

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.

Unwin, D.J., 1996, GIS, spatial analysis and spatial statistics, *Progress in Human Geography*, 20(4), 540-551.

Wartenberg, D., 1985, Multivariate spatial correlation: A method for exploratory geographical analysis, *Geographical Analysis*, 17(4), 263-283.

Wartenberg, D., 1990, Exploratory spatial analyses: Outliers, leverage points, and influence functions, in Griffith, D.A. ed., *Spatial Statistics: Past, Present, and Future*, Ann Arbor, Michigan: Institute of Mathematical Geography, 133-156.

교신 : 이상일, 08826, 서울특별시 관악구 관악로 1, 서울
대학교 사범대학 지리교육과 (이메일 si_lee@snu.ac.kr)

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

투 고 일: 2016년 11월 14일

심사완료일: 2016년 11월 29일

투고확정일: 2016년 12월 6일