이변량 SAM을 이용한 ESDA-GIS 연구방법론

이상일
강원대학교 지리교육과 교수

1. 서론

지리정보체계(GIS: Geographic Information System)의 급격한 발전은 지리학적 연구의 대상과 접근방법에 있어 광범위한 영향을 끼치고 있다. 특히 공간자료분석(SDA: spatial data analysis) 혹은 계량지리학은 여타 지리학의 분과들에 비해 보다 민감하고 적극적으로 GIS의 도입에 반응해왔다. ESDA-GIS (exploratory spatial data analysis - geographic information system) 연구방법론은 그러한 반응으로서 가장 폭넓게 받아들여지고 있는 것으로 보인다. 즉, SDA의 새로운 패러다임으로 인식되고 있는 ESDA와 일반연구환경(general purpose platform)으로서의 GIS의 결합은 분석적 GIS 혹은 지리정보과학(GISc: Geographic Information Science)의 정립을 위한 하나의 대안으로 폭넓은 호응을 얻고 있다. 이러한 ESDA-GIS 연구방법론 중에서 최근 SAM(spatial association measure)에 기반한 연구틀이 많은 주목을 받고 있다. 수치들의 공간적 집중/분산의 정도에 대한 측정치로 정의될 수 있는 SAM을 이용한 공간적 패턴 분석법은 특히 域單位로 수집된 공간데이터에 대해 공간적 집중지(spatial clusters)·공간적 이상치(spatial outlier)·공간적 체제(spatial regime)를 확인하는 데 있어 매우 유용한 것으로 제시되어 왔다. 그러나 이러한 SAM-기반 ESDA-GIS 연구방법론은 단지 일변량 분석에만 집중해온 한계점을 가지고 있다. 즉, 연구의 초점이, 공간적 자기상관(spatial autocorrelation)이라는 개념으로 보다 널리 알려진 것처럼, 한 변수가 보여주는 공간적 의존성(spatial dependence)과 공간적 이질성(spatial heterogeneity)의 파악에 주어졌고, 이변량 혹은 다변량 분석을 위한 방법론의 개발은 거의 이루어지지 않았다.

본 연구는 SAM-기반 ESDA-GIS 연구방법론을 발전시키기 위해 이변량 SAM을 개발하고, 그것의 유의성 검정법을 제시하며, 나아가 관련된 분석기법을 제공하는 것이다. 이변량 SAM의 존재 근거는, 두 변수간의 상관관계를 계산하는 것과 두 공간패턴을 비교하는 것은 본질적으로 다르다는 점을 인식하는 것으로부터 시작되어야 한다. 즉, 지리적 변수는 공간적 패턴을 의미하고 두 공간적 패턴간의 co-patterning을 측정하는 과제는 단순히 두 변수간의 공변량을 측정하는 것으로는 성취될 수 없다는 것이다. 이러한 과제는 일반통계기법을 공간화하고자 하는 보다 폭넓은 공간통계학적 운동과 관련되어 있으며, 단순히 말해 '공간적 상관계수'를 고안하는 것이다. 다음에서 이변량 SAM의 논리적 토대가 되는 두 개념, 즉 '이변량 공간적 의존성(bivariate spatial dependence)'과 '이변량 공간적 이질성(bivariate spatial dependence)'이 논의될 것이고, 전역적 이변량 SAM으로서의 L과 국지적 이변량 SAM으로서의 L_i가 제시될 것이다. 마지막으로 국지적 이변량 SAM을 이용한 시각화 방법들이 예와 함께 제공될 것이다.

2. 이변량 SAM (bivariate spatial association measure)

1) 이변량 공간적 의존성과 全域的 이변량 SAM, L

'이변량 공간적 의존성'이라는 개념은 공간적 자기상관에서 제시되는 일변량 공간적 의존성의 확장이다. 일반적으로 공간적 의존성이란 '공간단위의 위상학적 유사성과 변량의 수치적 유사성 간의 정(+)적인 결합'을 의미한다. 즉, 공간적으로 인접해 있는 값은 떨어져 있는 값에 비해 상대적으로 유사하다는 점을 지적하는 것이다. 그런데, 이러한 공간적 의존성에 대한 일반정의는 '일변량 공간적 의존성' 혹은

'공간적 자기상관'과 동일한 것으로 오인되어 왔다. 이변량 공간적 의존성은 두 변수간의 상관관계가 보여주는 공간성을 파악하고자 한다. 즉, 모든 공간단위에 대해 두 변수간의 상관관계가 국지적으로 계산되어 지도화되었을 때, 유사한 국지적 상관관계가 공간적으로 집중되어 있다면, 우리는 두 공간적 패턴간에 공간적 의존성이 존재한다고 말할 수 있다. 이러한 이변량 공간적 의존성은 두가지 측면에서 문제를 발생시킨다.

첫째, 이변량 공간적 의존성은 관측치 간의 관계를 정의하는 자유도(degree of freedom)을 현저하게 감소시킴으로써, 일반 통계학에서 제시하고 있는 상관관계 분석의 기본적 가정, 즉, 모든 관측치들은 서로에 대해 독립적이라는 가정을 위배하게 된다. 보다 구체적으로 말하면, 두 변수의 상관관계를 측정하기 위해 널리 사용되고 있는 Pearson's r의 경우, 유의성 검정을 위한 자유도가 n-2의 수식에 의거해 산출될 수 없기 때문에, 이변량 공간적 의존성을 보이는 두 변수간의 관계에 대해서는 신뢰할 수 있는 통계적 진술을 제공해 주지 못한다.

둘째, 동일한 Pearson's r을 가지는 수많은 공간적 패턴의 쌍이 도출될 수 있다. n개의 서로 다른 값을 가지는 두 변수로부터 동일한 Pearson's r값을 가지면서 동시에 서로 다른 co-patterning의 수준을 가진 $n!$개의 공간적 패턴의 쌍이 도출될 수 있다. Co-patterning은 두 공간적 패턴간의 시각적 일치성으로 개념화될 수 있는데, 이것을 측정하기 위해서는 각 측정치에서의 수치적 공변량뿐만 아니라 측정치간의 위상학적 관계에 대한 정보가 주어져야만 한다. 각 측정치에서의 국지적 공변량을 계산하고 그 값들을 임의적으로 공간상에 위치시켰을 때, 모든 패턴에 대해 Peason's r의 값은 동일하겠지만, 그 값들이 보여주는 공간적 의존성의 정도는 상이할 것이다. 이러한 이변량 공간적 의존성을 측정하기 위해서는 새로운 통계치가 개발되어야 하는 것이다.

따라서 전역적 이변량 SAM은 이변량 공간적 의존성을 측정하는 통계치이며 단순히 말해 '공간적 상관계수'이다. 전역적 이변량 SAM을 개발하기 위한 최초의 시도는 Wartenberg (1985)에 의한 것인데, 그가 고안한 통계치는 전역적 일변량 SAM인 Moran's I의 단순한 확장에 지나지 않는 것으로 많은 문제점을 가지고 있다. 이에 Lee(2001)는 비공간적 이변량 상관계수인 Pearson's r과 공간적 일변량 SAM인 Moran's I를 결합하여 새로운 통계치 L을 개발하였다.

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

2) 이변량 공간적 이질성과 局地的 이변량 SAM, L_i

일반적으로 '공간적 이질성'이란 값들의 공간적 분포가 보여주는 지역 간 동질성의 통계적 결여를 의미한다. 현저한 공간적 이질성은 일반 통계학의 또 다른 궁극적 가정, 즉 모든 관측치는 하나의 모집단으로부터 추출된 표본이라는 가정을 위배한다. 이러한 공간적 이질성은 이변량 분석의 맥락에서 다음의 두 가지 의미를 갖는다.

첫째, 두 변수간 상관관계에 대한 전역적 통계치는 국지적 수준에서의 다양한 상관관계를 일원화한다. 예를 들어, 전역적이고 평균적인 상관관계의 경향이 정(+)적인 것이라 하더라도 특정 하위지역에서는 국지적으로 부(-)적인 상관관계가 발생할 수도 있다. 더 나아가 정(+)적인 상관관계라 하더라도 어떤 하위지역에서는 정(+)-정(+)의 공분산 관계이고 다른 하위지역에서는 부(-)-부(-)의 공분산 관계일 수 있다. 이렇게 국지적 상관관계를 계산하고 그것의 공간적 분포를 살펴보는 것이 공간적 분석에서는 필수적임에도 불구하고 최근까지 이러한 접근은 과도하게 제한되어 왔다.

둘째, 이변량 공간적 의존성 자체가 공간적으로 이질적이라는 점이 지적되어야 한다. 한 공간단위에서의 국지적 공변량과 그 주변 공간단위에서의 국지적 공변량들과의 유사성으로 정의될 수 있는 국지

적 이변량 공간적 의존성은 하위지역별로 차별적일 수 있다. 즉, 어떤 하위지역에서는 유사한 공변량의 집중이 강할 수 있지만, 다른 지역에서는 그렇지 않을 수 있다. 따라서 국지적 이변량 SAM의 개발은 필연적이다.

Lee(2001a; 2001b)는 위에서 제시된 L을 각 공간단위로 분할 할당함으로써 국지적 이변량 SAM인 L_i를 제시하였다. 이때 L_i값의 평균은 바로 전역적 L이 된다.

$$L_i = \frac{n^2}{\sum_i \left(\sum_j v_{ij}\right)^2} \cdot \frac{\left(\sum_j v_{ij}(x_j - \bar{x})\right) \cdot \left(\sum_j v_{ij}(y_j - \bar{y})\right)}{\sqrt{\sum_i (x_i - \bar{x})^2} \cdot \sqrt{\sum_i (y_i - \bar{y})^2}}$$

3) 이변량 SAM을 위한 일반화된 유의성 검정법

일변량 SAM에 대한 유의성 검정법은 여러 종류의 접근법에 의해 제시되었다. 그러나 이변량 SAM에 대한 유의성 검정법의 개발은 한번도 시도된 적이 없다. Lee(2001b; 2002a; 2002b)는 randomization assumption에 기반하여 전역적 SAM과 국지적 SAM 모두에서 이용 가능한 유의성 검정법을 제시하였다. 더욱이 이 검정법은 이변량 SAM 뿐만 아니라 일변량 SAM에 대해서도 적용가능하기 때문에 일반화된 유의성 검정법인 것으로 주장된다. 특히 국지적 SAM에 대한 유의성 검정법은 통계적으로 유의한 통계치만을 시각화할 수 있게 해주기 때문에 뒤에서 제시될 ESDA-GIS 분석기법을 구축하는데 지대한 공헌을 하게 된다.

3. 이변량 SAM을 이용한 ESDA-GIS 분석기법

Lee (2001b; 2002c)는 위에서 개발된 이변량 SAM을 이용한 몇가지 분석기법을 개발하였다.

'Local-L 지도'는 L_i을 사용하여 단순하게 그린 단계구분도이다. 전역적 L이 두 공간적 패턴간에 존재하는 이변량 공간적 의존성의 정도를 수치로 보여주는 것이라면, 이 지도는 그러한 공간적 의존성의 전체 값에 대해 각각의 하부지역들이 어떻게 공헌하고 있는지 확인할 수 있게 해준다. 즉, 어떤 하위지역에서 공간적 의존성이 두드러지는지에 대해 개괄적인 파악을 가능하게 해준다. 또한 어떤 하위지역이 전체적인 공간적 상관관계에 위배되는 지를 시각적으로 보여준다. 예를 들어 전역적 통계치가 정(+)적일 때, 어떤 하위지역에서 부(-)적인 국지적 통계치를 보이는지를 확인할 수 있게 해준다.

'Local-L scatterplot'은 앞에서의 L_i값을 네가지 범주로 구분해 주는 산포도이다. 일사분면과 삼사분면에 위치한 측정치는 두 변수간에서 정(+)적인 관계를 보이는 것인데, 전자는 정(+)-정(+) 관계, 후자는 부(-)-부(-) 관계를 나타낸다. 이에 반해 이사분면과 사사분면에 위치한 측정치들은 두 변수간에서 부(-)적인 관계를 보이는 것으로, 전자는 부(-)-정(+) 관계, 후자는 정(+)-부(-) 관계를 나타낸다. 만일 전역적 L이 높은 정(+)적인 값을 보인다면, 일사분면과 삼사분면에 위치한 측정치의 수가 훨씬 많은 것이다. 하지만 공간적 이질성의 관점에서 보면 오히려 이사분면과 사사분면에 위치한 측정치가 더 중요할 수 있다.

'Local-L scatterplot map'은 각 공간단위가 위의 산포도에서 어느 분면에 위치하는가에 의거해 그린 일종의 명목지도이다. 이것은 두 변수간의 관계에서 도출되는 네 가지 관계의 범주가 공간상으로는 어떻게 드러나는지를 명확하게 보여준다. 각 범주가 뚜렷한 공간적 분리를 보여 줄 때 연구자는 다수의 공간적 체제(spatial regime)의 존재를 상정할 수 있으며 차후의 연구과정에서 보다 유의미한 가설을 설정하는데 도움을 받을 수 있다.

'Local-L significance map'은 Local-L scatterplot map과 앞에서 언급한 유의성 검정법을 결합한 지도이다. 즉 Local-L scatterplot map에서 통계적으로 유의한 공간단위만을 골라내어 시각화한 것이다

이 지도로부터 확인되는 정(+)-정(+) 관계의 공간적 집중 지역은 '이변량 hot spot'으로, 부(-)-부(-) 관계의 공간적 집중 지역은 '이변량 cold spot'으로 개념화될 수 있을 것이다. 통계적으로 유의미한 것으로 드러난 부(-)적인 관계의 공간적 집중지는 두 변수간 상관관계에 있어 극명한 이질성을 보여주는 지역이기 때문에 향후 연구 과정에서 보완설명이 시도되어야 할 것이다.

4. 앞으로의 과제

여기서 제시된 이변량 SAM에 기반한 ESDA-GIS 방법론은 다양한 연구주제에 적용됨으로써 연구방법론으로서의 효용성이 검증되어야 한다. 일차적으로 도시 내 거주지 분리 연구에 이용될 수 있을 것이다. 거주지 분리 정도에 대한 전역적 통계치로서 일반적으로 사용되고 있는 상이지수(Index of dissimilarity)는 본질적으로 비공간적임이 밝혀 졌다 (Wong, 1993). 거주지 분리가 기본적으로 두 집단의 거주지 공간 분포를 비교하는 것이기 때문에 앞에서 제시된 전역적 이변량 SAM이 상이지수를 대체할 수 있을 것이다. 더 나아가 거주지 분리의 국지적 분포에 대한 어떠한 함의도 상이지수는 제공해 주지 못하지만, 국지적 이변량 SAM은 두 집단 사이에서 발생하고 있는 배제의 공간적 이질성을 탐구할 수 있게 해줄 것이다.

이변량 SAM에 기반한 ESDA-GIS 방법론은 사회과학적 GIS 혹은 분석적 GIS를 지향하는 데 있어 일정한 공헌을 할 수 있을 것으로 기대된다. 또한 공간을 다루는 제 사회과학 분과들의 정형화된 방법론의 오류를 지적함으로써 지리학적 연구가 가지는 가치를 보다 효과적으로 전달하는 데도 도움을 줄 수 있을 것이다.

참고문헌

Anselin, L., 1995, Local indicators of spatial association: LISA, Geographical Analysis, 27, 93-115

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 U.S. labor market areas, 1969-1999, Unpublished Dissertation, Department of Geography, Ohio State University

Lee, S-I., 2002a, A generalized significance testing methods for spatial association measures: an extension of Mantel Test, Environment and Planning A, forthcoming

Lee, S-I., 2002b, A generalized significance testing method for local spatial association measures: a randomization approach, Environment and Planning A, on review

Lee, S-I., 2002c, Exploring the bivariate spatial dependence and heterogeneity: a local bivariate spatial association measure for exploratory spatial data analysis (ESDA), International Journal of Geographical Information Science, on review

Wong, D. W. S., 1993, Spatial indexes of segregation, Urban Studies, 30, 559-572