대한지리학회지 제45권 제4호 2010(502~520)

공간 클러스터의 범역 설정을 위한 GIS-기반 방법론 연구 -수정 AMOEBA 기법-

이상일* · 조대헌** · 손학기*** · 채미옥****

A GIS-Based Method for Delineating Spatial Clusters: A Modified AMOEBA Technique

Sang-Il Lee* · Daeheon Cho** · Hakgi Sohn*** · Miok Chae****

요약 : 이 연구의 주된 목적은 공간 클러스터의 범역을 설정하는 GIS-기반 방법론을 개발하는 것이다. 주요 과제는 지리적 경계 분석과 LISA-기반 클러스터 탐지에 대한 기존 방법론을 비교 검토함으로써 진일보한 방법론을 고안하고, 그것을 실행하는 GIS-기반 프로그램을 개발하는 것이다. 주요 연구 결과는 다음과 같다. 첫째, 기존 방법론을 검토한 결과, LISA를 이용한 AMOEBA 기법이 가장 타당한 것으로 판단되었다. 둘째, 수정 AMOEBA 기법의 알고리즘을 확립했으며 실행 소프트웨어를 상용 GIS 프로그램의 확장 기능 형태로 개발하였다. 셋째, 수정 AMOEBA 기법을 실험 데이터와 실 데이터에 적용한 결과 제안된 기법의 유용성이 확인되었다.
주요어 : 지리적 경계 분석, 클러스터 탐지, 공간 클러스터의 범역 설정, 웜블링, 국지적 공간 연관성 지표, 아메바 기법

Abstract : The main objective of the paper is to develop a GIS-based method for delineating spatial clusters. Major tasks are: (i) to devise a sustainable algorithm with reference to various methods developed in the fields of geographic boundary analysis and cluster detection; (ii) to develop a GIS-based program to implement the algorithm. The main results are as follows. First, it is recognized that the AMOEBA technique utilizing LISA is the best candidate. Second, a modified version of the AMOEBA technique is proposed and implemented in a GIS environment. Third, the validity and usefulness of the modified AMOEBA algorithm is assured by its applications to test and real data sets.
Key Words : geographic boundary analysis, cluster detection, delineation of spatial clusters, wombling, local indicators of spatial association (LISA), AMOEBA technique

본 연구는 2009년 국토연구원에서 수행된 「선진적 국토관리를 위한 용도지역제 개선과 손익조정제도 도입방안 연구(Ⅰ)」의 일환으로 이루어졌음.

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

** 이화여자대학교 교육대학원 특임교수(Professor for Special Appointment, The Graduate School of Education, Ewha Womans University), dhncho@gmail.com

*** 국토연구원 책임연구원(Associate Research Fellow, Korea Research Institute for Human and Settlements), hgsohn@krihs.re.kr

**** 국토연구원 선임연구위원(Senior Research Fellow, Korea Research Institute for Human and Settlements), mochae@krihs.re.kr

1. 서론

지리적 속성의 분포 패턴으로부터 해당 현상이 두드러지게 드러나는 구역을 찾아내어 그것의 경계를 획정하는 일은 순수 학문적 맥락에서나 실질적인 계획 실행의 맥락에서나 중요한 사안이다. 예를 들어, 다변량 자료에 기반한 도심성 지표의 공간 분포로부터 도심의 범위를 획정하는 것(예, Thurstain-Goodwin and Unwin, 2000; Office of the Deputy Prime Minister, 2002), 원격 탐사 이미지나 인구 그리드를 이용해 인구 클러스터 혹은 도시 지역의 범위를 확인하는 것(예, Sutton, 2003; Balk *et al*., 2006), 주택 매매 거래 데이터를 바탕으로 주택 과열 지구의 범위를 설정하는 것(Sohn, 2008; Sohn and Park, 2008) 등은 연구자나 계획가에게 중요한 관심 사항이다. 이 모든 과제에서 핵심적인 것은 특정한 공간 분포로부터 특정한 규준을 만족시키는 관심 지역의 범위를 결정하는 것이다. 보다 GIS적으로 표현하자면, 속성의 분포로부터 역(域)형 객체(area objects)를 추출하는 것이다.

이러한 관심 지역의 범위를 결정하는 것, 즉 범역을 설정하는 것은 개념적인 측면에서 볼 때 두 가지의 상호 연관된 공간적 과제의 결합을 의미한다. 하나는 관심 지역의 규준을 결정하는 것이며, 또 다른 하나는 그러한 관심 지역의 지리적 경계를 결정하는 것이다. 관심 지역 중에서도 아주 높은 값들이 집중해 있거나 아주 낮은 값들이 집중해 있는 경우, 우리는 그것을 공간 클러스터라고 부르며, 많은 연구 상황에서 연구자가 확인하고자 하는 관심 지역은 이러한 공간 클러스터이다. 하나의 공간 클러스터가 존재하기 위해서는 매우 특징적인 속성값을 보유한 공간 단위가 존재해야 할 뿐만 아니라 그러한 공간 단위가 서로 연접해 있어야 한다. 그런데, 공간 클러스터는 핵심부에서 주변부로 이동해감에 따라 이질성이 점차 증대하므로 공간 클러스터로서의 최소한의 응집성을 유지해주는 최대한의 지리적 한계를 설정하는 것이 중요하다. 결국 공간적 클러스터의 경계를 결정하는 것이 공간 클러스터 범역 설정의 또 다른 차원이 되는 것이다.

이러한 공간 클러스터의 범역 설정이라는 과제를 수행하는 연구자는 항상 방법론적인 도전에 직면해 왔다. 왜냐하면 그러한 종류의 과제는 다양한 맥락에서 빈번하게 요구되고 있지만 그것의 수행을 위한 방법론적인 대안의 폭은 그리 넓지 않기 때문이다. 실질적으로 말해서 이러한 과제를 수행할 수 있게 해주는 보편화된 방법론이란 존재하지 않을뿐더러, 실행 소프트웨어 역시 존재하지 않는다. 따라서 본 연구의 주된 목적은 공간 클러스터의 범역을 설정하는 GIS-기반 방법론을 개발하는 것이다. 주요 과제는 기존 방법론을 비교 검토하여 진일보한 방법론을 고안하고, 그것을 실행해주는 GIS-기반 프로그램을 개발하여 그 적용성을 검토하는 것이다. 기존 방법론의 비교 검토를 위해 '지리적 경계 분석' 분야에 대한 집중적인 리뷰가 이루어지며, 이를 바탕으로 준거가 될 수 있는 기법이 선정된다. 준거가 되는 기법의 알고리즘을 일부 수정하여 실행 알고리즘을 확립하고 분석 도구를 개발한다. 이 때 분석 도구는 일반 GIS 엔진에 추가하여 사용할 수 있는 확장 기능 형태로 디자인된다. 끝으로 실험 데이터와 실 데이터를 이용하여 실행 알고리즘이 얼마나 합리적으로 작동하는지가 평가된다.

2. 지리적 경계 분석

1) 지리적 경계의 개념과 범주

공간분석 혹은 공간데이터분석이라고 불리는 연구 영역에서 공간 클러스터의 범역 설정과 가장 관련이 깊은 것은 '지리적 경계 분석'이라 불리는 분야이다. 지리적 경계 분석은 연속적인 필드(continuous field)로부터 '경계'라고 하는 이산적 객체(discrete objects)를 추출하는 기법을 의미한다(Jacquez *et al*., 2000, 222). 여기서 연속적 필드와 이산적 객체 간의 이분법은 지리정보과학 분야에서 지리를 재현하는 두 가지 상반된 개념적 프레임워크와 관련되어 있다(Lee *et al*., 2009, 86-90). 이산적 객체의 관점에서 보면, 세상은 더 이상 나누어질 수 없는 명확한 경계를 갖는 개체로 구성되어 있고, 그 개체들 사이는 비어 있다. 이에 반

해 연속적 필드의 관점에서 보면, 세상은 속성값이 부여되어 있는 무수히 많은 위치들로 채워져 있다. 하나의 속성에 대한 값들의 분포를 필드라고 부르는데, 무수히 많은 속성을 상정할 수 있기 때문에 세상은 무수히 많은 필드들의 층으로 이루어져 있다고 할 수 있다.

보통 연속적 필드의 관점은 래스터 데이터를 통해 구현된다고 볼 수 있기 때문에 규칙적 그리드 데이터(regular grid data)에 한정되는 것으로 생각하기 쉽다. 그러나 연속적 필드 관점의 핵심은 모든 지점에서 존재하는 속성의 분포이기 때문에 비록 불규칙적인 역형 객체로 구성된 벡터 데이터의 경우라 하더라도 공간적으로 내포적인 변수(spatially intensive variables)(Goodchild and Lam, 1980), 즉 정규화된(normalized) 속성의 분포인 경우에는 연속적 필드의 관점을 구현하고 있는 것으로 볼 수 있다. 그러나 여기서는 설명의 편의를 위해, 그리드 셀에 등간 및 비율 척도의 속성값이 부여되어 있는 래스터 데이터를 다루는 상황을 상정하고 논의를 진행해 나갈 것이다.

연속적 필드로부터 경계라고 하는 이산적 객체를 추출한다는 것의 의미를 보다 명확히 하기 위해서는 우선 '경계'라고 하는 객체가 무엇을 의미하는지를 논의해 보아야 한다. '경계'는 상식적인 수준에서 생각해 볼 때, 1차원의 선(線)형 객체(line objects)이다. 이 경우 경계는 그것을 중심으로 양쪽이 매우 다른, 즉 "값의 변화가 빠르게 발생하는 곳을 연결한 선"으로 정의될 수 있다. 예를 들어 연구 대상 지역의 한 하위 지역에 매우 높은 값이 집중해 있고, 근방의 하위 지역에 매우 낮은 값이 집중해 있다면, 우리는 두 하위 지역 사이에서 값이 매우 빠르게 변하는 지대를 발견할 수 있고, 그것을 대표하는 선형 객체를 추출할 수 있다. 한편, '경계'는 1차원의 선형 객체가 아니라 2차원의 역형 객체의 외연을 의미할 수도 있다. 우리는 연구 대상 지역 전체를 값의 유사성에 근거하여 몇 개의 하위 지역으로 분할할 수 있으며, 특정한 공간 클러스터를 나머지 지역과 구분해낼 수도 있다. 이 경우 경계는 "상대적으로 등질적인 지역의 말단"(Jacquez *et al.*, 2000, 224-225)이라고 정의될 수 있다.

이러한 두 가지의 개념 규정을 바탕으로 지리적 경계 추출의 상황을 범주화하면 Figure 1에 나타나 있는 바와 같다. Figure 1의 (a)는 첫 번째 개념 규정을 반영하고 있는 것으로, 선형 객체인 경계를 중심으로 그 양쪽 편에 매우 이질적인 값들이 분포하고 있다는 것을 보여준다. 이러한 개방 경계(open boundary)를 '차이 경계(difference boundary)'라고 부른다(Jacquez *et al.*, 2000, 225). 이에 반해 (b)와 (c)에 나타나 있는 경계는 2차원의 역형 객체의 외연으로, 위에서 언급한 두 번째 정의에 해당하는 것이다. 이러한 폐쇄 경계(closed boundary)를 '구역 경계(areal boundary)'라

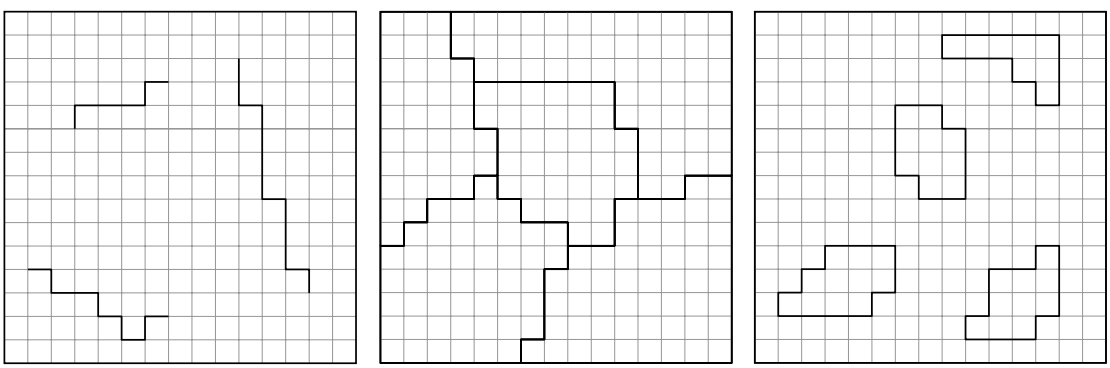


Figure 1. Three situations in geographic boundary analysis.
지리적 경계 분석의 세 가지 상황. (Source: Jacquez *et al*., 2000)

부른다(Jacquez *et al.*, 2000, 225).

그런데 여기서 (b)와 (c)를 구분하는 것은 매우 중요하다. 왜냐하면 두 경우는 지리적 경계를 분석하는 서로 다른 목적과 관련되어 있기 때문이다. (b)는 연구 지역 전체를 내적 동질성을 보이는 몇 개의 하위 지역으로 분할하는 것으로 일종의 지역구분을 실행하는 것이다. 여기서 모든 그리드 셀은 특정한 지역의 멤버십을 가져야 하며 동일 멤버십을 가지는 그리드 셀들은 서로서로 연접해야 한다. 이에 반해 (c)는 역형 객체를 추출한다는 의미에서는 (b)와 다를 바 없지만, 모든 그리드 셀이 특정한 지역의 멤버십을 가질 필요는 없다. 이것은 공간 클러스터 혹은 패치(patch)의 존재를 확인하고 그것의 외연을 설정하는 것과 관련되어 있다. (b)에서 확인된 하위 지역 내부에서 공간적 응집성이 높은 핵심 구역이 공간 클러스터일 가능성이 높지만 모든 하위 지역이 항상 공간 클러스터와 관련되어야 할 필연성은 존재하지 않는다.

Figure 1에 나타난 세 가지 상황 각각에 대해 다양한 분야에서 방법론들이 개발되어 왔다. (a)는 주로 생태학 분야에서 발달해 온 웜블링(wombling) 기법과 관련되어 있고, (b)는 군집분석의 특수한 형태로서의 공간제약 군집분석(spatially constrained cluster analysis)과 관련되어 있으며, (c)는 공간데이터분석 분야의 클러스터 탐지(cluster detection)와 관련되어 있다. 그러나 클러스터 탐지 분야는 주로 클러스터의 규준을 만족시키는 개별 그리드 셀이나 공간 단위의 확인에 집중할 뿐 그것들의 공간적 연합과 그것에 기반한 역형 객체의 도출에는 많은 관심을 두지 않았다. 따라서 본 연구의 목적에 부합하기 위해서는 부가적인 조건이 충족되어야만 한다.

2) 격자 웜블링과 공간제약 군집분석

웜블링은 값에서 빠른 변동이 발생하고 있는 지대를 확인하는 방법으로 Womble(1951)의 연구에서 비롯된 것이다. 주로 생태학의 연구에서 사용되고 있는데, 두 개의 생태적 커뮤니티 혹은 생태계의 구분선을 의미하는 이행대(移行帶, ecotone)의 확인이 그 학문 분야에서는 중요하기 때문이다(Fortin and Dale, 2005, 184). 그러나 구분선의 확인은 다른 많은 학문 영역에서도 중요할 수 있는데, 그것은 경계라는 실체가 과학적, 실제적 중요성을 갖는 경우가 많기 때문이다. 예를 들어, 문화 현상의 분포에서 흔히 발견되는 점이지대의 확인은 그 자체가 하나의 연구 목적일 수 있다.

다양한 웜블링 방법론이 존재하지만 본 연구와 가장 큰 유관성을 가지는 것은 격자 웜블링(lattice-wombling)이다. 격자-웜블링 기법의 핵심적인 사항은 그리드 데이터로부터 규칙적으로 분포하는 지점들의 집합을 생성하고 각 지점에서 값의 '변화율'을 계산하는 것이다. 높은 변화율을 보이면서 인접하고 있는 지점들을 연결하면 선형 객체인 경계가 추출된다(Fortin and Dale, 2005, 190).

계산된 변화율이 높은 지점들은 추출될 경계의 구성원이 될 가능성이 높다는 것을 의미하고, 변화율이 낮다는 것은 그 지점을 중심으로 등질성이 높은 공간적 클러스터나 패치가 존재할 가능성이 높다는 것을 의미한다. 다변량 상황일 경우는 모든 변수에 대해 변화율을 계산하고 각 셀에 대해 그 변화율들의 평균을 구함으로써 최종적인 변화율이 계산된다. 높은 변화율을 보이는 지점들(보통 상위 10%에 포함되는 지점)을 '경계 요소'라고 부르며, 추출될 경계의 부분이 될 수 있는 후보자의 지위를 획득하게 된다. 경계 요소가 서로 인접하면서 동시에 향이 30도를 초과하지 않으면 하나의 하위경계로 확인된다(Jacquez *et al.*, 2000, 228). 최근 베이지언 공리에 기반한 웜블링 기법이 제안되기도 했다(Lu and Carlin, 2005)

공간제약 군집분석은 일반 통계학에서 널리 사용되고 있는 군집분석의 공간적 확장판이다. 보다 단순하게 말하면, 그리드 셀 값이 유사한 것끼리 합쳐 지역을 형성하되 공간적 연접성 요구(adjacency requirement)를 만족시키도록 제약을 가하는 것이다. 통상적으로 군집분석에서 사용되고 있는 계층적 응집 방법과 비계층적 *k*-평균 방법 모두가 사용될 수 있다(다양한 군집화 알고리즘에 대해서는 Legendre and Legendre, 1998 참조).

공간제약 군집분석은 종국적으로 모든 그리드 셀에 멤버십을 부여하며, 동일한 멤버십을 가진 그리드 셀은 하나의 응집체로 드러난다. 모든 그리드 셀이 이러

한 속성을 갖기 때문에 종국적으로는 연구 대상 전체가 몇 개의 하위 지역으로 분할된다. 이러한 지역 구분의 하나의 부산물이 바로 지역들 사이에 존재하게 되는 경계인 것이다(Fortin and Dale, 2005, 177).

공간제약 군집분석의 가장 중요한 단점은 실질적으로 의미 있는 경계가 존재하지 않는 곳에서도 지역 구분의 부산물로 경계가 생성된다는 것이다(Fortin and Dale, 2005, 180). 또한 각 하위지역이 마치 하나의 클러스터로 인식되거나, 최소한 그 핵심 부에 클러스터가 존재하는 것으로 잘못 인식될 수 있다. 다시 말해서 하위 지역의 생성과 클러스터의 존재 사이에는 논리적인 필연성이 존재하지 않는다.

3) LISA와 지리적 경계 분석

지리적 경계 분석의 마지막 상황(Figure 1의 (c))을 다루는 방법론은 앞의 두 가지 방법론과는 다소 다른 학문적 맥락 속에서 발달해 왔다. 앞에서도 설명한 바와 같이 지리적 경계 분석은 공간 클러스터를 확인하고 그것의 범역을 설정하는 것과 관련되어 있다. 따라서 이 분야는 클러스터 탐지 관련 연구와 깊이 관련되어 있는데, 지리학과 보건통계학 혹은 의료통계학 등 다양한 학문 분야에서의 발달과 관련되어 있다(Waller and Gotway, 2004; Lawson and Kleinman, 2005; Rogerson and Yamada, 2009; Tango, 2010). 여기서는 클러스터 탐지와 관련된 연구들 중, 공간 단위로 집계된 데이터, 혹은 역형 객체가 보유한 속성들로부터 클러스터를 발견해내는 방법론에 집중하고자 한다. 그렇게 하는 데는 두 가지 이유가 있다. 첫째, 역형 객체에 대해 개발된 기법은 손쉽게 그리드 데이터로 확장될 수 있다. 공간데이터분석 혹은 공간통계학의 입장에서 보면, 두 종류의 데이터는 구조적으로 동일하다. 둘째, 역형 객체의 클러스터 탐지를 위해 개발된 방법론 중 특히, LISA(local indicators of spatial association, 국지적 공간 연관성 지표)(Anselin, 1995)를 활용한 방법론이 본 연구의 맥락에 가장 부합된다.

LISA는 국지적 공간적 자기상관 통계량을 의미하는 것으로 국지적 Moran의 I_i와 국지적 Geary의 c_i (Anselin, 1995), 그리고 Getis-Ord의 G_i와 G_i^*(Getis and Ord, 1992; Ord and Getis, 1995; Getis and Ord, 1996)를 의미한다. LISA가 클러스터 탐지를 포함한 다양한 ESDA(exploratory spatial data analysis, 탐색적 공간데이터분석) 연구에 많은 잠재력을 가지고 있다는 것이 1990년대 중반부터 인식되었고 다양한 기법들이 제안되었다(Anselin, 1996; Unwin, 1996; Anselin and Bao, 1997; Anselin, 1998; Brunsdon, 1998; Dykes, 1998; Unwin and Unwin, 1998). LISA는 Figure 1의 세 가지 경우 모두와 관련되어 있다. 그 연관성 각각을 좀 더 상세하게 다루면 다음과 같다.

(1) LISA와 웜블링

LISA의 관점은 Figure 1(a)에 나타나 있는 '차이 경계(difference boundary)'의 확인과 관련되어 있다. 개념적으로 보면, 경계는 양의 공간적 자기상관이 사라지는 지점들(혹은 지점들의 연결선 혹은 지대), 혹은 음의 공간적 자기상관이 나타나는 지점들(혹은 지점들의 연결선 혹은 지대)을 의미한다. 특히 음의 공간적 자기상관이 두드러진 구역의 확인은 '차이 경계'의 확인과 밀접히 관련되어 있다. LISA에서 음의 공간적 자기상관을 확인하는데 가장 뛰어난 것은 국지적 Moran의 I_i이다. Anselin(1996)이 제시한 Moran 산포도에서 H-L(높은 값이 낮은 값에 둘러싸여 있는 경우)나 L-H(낮은 값이 높은 값에 둘러싸여 있는 경우) 연관이 두드러진 곳은 바로 차이 경계의 확인에 중요한 단서를 제공할 수 있다.

그러나 실제로 LISA를 이용해 차이 경계의 확인이 시도된 것은 Getis-Ord의 G_i^*를 이용한 경우였다. Boots(2001)는 행정구역과 같은 공간 단위로 구성된 데이터 셋에서 구역과 구역의 쌍에 의해 규정되는 경계들의 상대적 강도를 측정하는 방법을 제시함으로써 일종의 에어리어 웜블링 기법을 도출하였다. 두 구역 사이의 경계 강도를 계산하는 방법은 다음과 같다. 각 구역에 대해 이웃의 세트를 결정한다. 통상적으로 각 구역과 경계를 직접 접하고 있는 인접 구역이 선정된다. 그리고 나서 Getis-Ord의 G_i^*를 계산한다. 주지하는 바처럼, 그 값이 양의 높은 절대치를 가지면 그 구역을 중심으로 높은 값이 집중해 있다는 것을 의미하

며, 음의 높은 절대치를 가지면 낮은 값이 집중해 있다는 것을 의미한다. 두 구역 각각에 대해 구해진 G_i^*값의 차이를 구함으로써 경계 강도를 산출한다. 만약 그 차이가 크다면 그 만큼 이질적이라는 것을 의미하게 되고 그 경계가 견고하다는 것을 의미하게 된다.

(2) LISA와 군집분석

LISA는 군집분석과도 관련성이 있는데, 특히 Anselin의 Moran 산포도 지도(Anselin and Bao, 1997)와 밀접히 관련되어 있다. Moran 산포도는 국지적 공간 연관의 양상을 네 가지로 범주화 하는데, H-H(높은 값이 높은 값에 의해 둘러싸여 있음), L-L(낮은 값이 낮은 값에 의해 둘러싸여 있음), H-L(높은 값이 낮은 값에 의해 둘러싸여 있음), L-H(낮은 값이 높은 값에 의해 둘러싸여 있음)가 그것이다. 모든 공간 단위는 이 네 가지 양상 중 하나에 포함되기 때문에 이러한 분류를 바탕으로 일종의 명목 지도를 작성할 수 있는데, 그것을 Moran 산포도 지도라고 한다. 따라서 이 기법은 국지적 공간 연관의 양상을 바탕으로 지역구분을 행한 것이라 할 수 있다.

이것은 일종의 공간적 체제를 보여주는 것으로 공간적 패턴 분석에 새로운 통찰력을 제공해 준다. 그러나 이 기법은 앞에서 설명한 공간제약 군집분석과는 달리 동일한 국지적 공간 연관을 보여주는 폴리곤이 반드시 공간적으로 연접해야 하는 제약의 적용을 받지 않는다. 비록 LISA의 계산 과정에 일종의 '공간적 제약'이 부여되기 때문에 산포도 지도 상에 지역 구분의 양상이 두드러진다고 해도, 본질적으로 등질지역의 도출을 목적으로 한 것이 아니다. 따라서 Moran 산포도 지도는 지역 구분도라기 보다는 일종의 지역 분류도의 성격을 갖는다.

(3) LISA와 공간 클러스터 탐지

LISA를 이용해 공간 클러스터를 탐지하는 방법은 이미 표준화 되어 있다. 특정 LISA를 산출한 후 그 값에 통계적 유의성 검정을 적용시키면 통계적으로 유의한 공간적 응집체를 추출할 수 있고 그것을 공간 클러스터(핫스팟 혹은 콜드스팟)라 부를 수 있다. 예를 들어 GeoDa라는 공간데이터분석 프로그램에서는 유의성 검정을 통해 산출되는 유의 확률만으로 그린 지도를 '유의성 지도'라 부르고, 그것과 값의 높낮이에 대한 정보를 결합함으로써 핫스팟과 콜드스팟을 확인하게 해주는 지도를 '클러스터 지도'라고 부른다(Anselin, 2003). 그러나 앞에서도 언급한 것처럼 클러스터 탐지 기법들이 클러스터 범역 설정을 위한 많은 정보를 제공하고는 있지만 그것을 위해 개발된 기법이 아니기 때문에 본 연구에 곧 바로 적용될 수는 없다.

이러한 측면에서 볼 때 Wulder and Boots(1998)의 연구는 클러스터 범역 설정을 위한 방법론에 한걸음 더 다가간 기법을 제시하고 있다. 그들은 Getis-Ord의 G_i^*를 그리드 데이터의 상황으로 확장하여 흥미로운 분석 기법을 제안했다. 그들은 공간 클러스터의 스케일에 주목하여 다양한 탐색 반경을 설정한 후 모든 그리드 셀에 대해 탐색 반경 별로 G_i^* 값을 산출하였다. 그렇게 함으로써 각 그리드 셀 마다 산출 가능한 최대 G_i^* 값(maximum G_i^*)과 그 때의 최대 탐색 반경(maximum G_i^* distance)에 대한 정보를 추출하여 공간적 패턴 분석에 사용하였다.

Fortin and Dale(2005, 159)의 해석에 따르면, 최대 G_i^* 지도는 공간적 클러스터의 핵심부를 파악하는데 도움을 주고, 최대 G_i^* 거리 지도는 공간적 클러스터의 범역을 확인하는데 도움을 줄 수 있다고 한다. 이러한 공리를 보다 확장하여 알고리즘화한 것이 AMOEBA이고, 본 연구를 위해 최종적으로 선택하였다. 이는 다음에서 자세하게 다루도록 한다.

3. AMOEBA 기법

1) LISA와 AMOEBA 기법

AMOEBA(A Multidirectional Optimal Ecotope-Based Algorithm)는 LISA를 이용하여 공간 클러스터의 범역을 설정하는 기법이다(Aldstadt and Getis, 2006). AMOEBA에서 활용되는 LISA는 Getis-Ord의 G_i^*이며 그것은 다음과 같은 수식에 의해 주어진다.

$$G_i^* = \frac{\sum_{j=1} w_{ij}x_j - \bar{x}\sum_{j=1} w_{ij}}{s\sqrt{\frac{n\sum_{j=1} w_{ij}^2 - \left(\sum_{j=1} w_{ij}\right)^2}{n-1}}} \quad \text{(식 1)}$$

여기에서 s는 표준편차를, w_{ij}는 공간 가중 행렬의 요소 값을, n은 전체 케이스 수를 의미한다. i와 j는 개별 공간단위(혹은 위치)를 의미하는 것으로 두 공간단위가 이웃으로 정의되면 w_{ij}=1이, 그렇지 않으면 w_{ij}=0이 되며, 자신을 이웃으로 간주하기 때문에 w_{ii}=1로 주어진다. 이 통계량의 기대값은 0이고, 분산은 거의 1이다(Aldstadt and Getis, 2006, 330). 따라서 이 통계량의 유의성 검정은 정규분포를 상정한 표준화 점수에 대한 것과 거의 동일하게 이루어진다.

그런데 전술한 것처럼 LISA에 포함되는 통계량은 다양해서 서로서로 다른 특징을 가진다. 예를 들어 Moran의 I_i와 Geary의 c_i는 중심 셀과 주변 셀 간을 '비교' 한다는 의미에서 유사성을 가진다. 하지만 그 비교의 방식에서 이 두 통계량은 서로 다른 특징을 보인다. 우선 Moran의 I_i는 공변동을 비교의 준거로 삼기 때문에 중심 셀과 주변 셀이 높은 유사성을 보이기 위해서는 두 값 모두 평균으로부터 멀리 벗어나야만 한다. 이에 비해 Geary의 c_i는 차이(엄밀히 말해, 차이의 제곱)를 계산함으로써 중심 셀과 주변 셀을 비교한다. 결국 중심 셀과 주변 셀의 값이 큰 차이가 없다 하더라도 두 값이 평균과 유사하면 Moran의 I_i에 의해서는 높은 유사성이 없는 것이 된다. 이 두 개의 통계량과 달리 Getis-Ord의 G_i^*는 중심 셀과 주변 셀로 이루어진 로컬 영역 전체를 한꺼번에 평가한다.

이러한 특성으로 인해 연구 목적에 잘 부합하는 LISA를 선정하는 것이 중요해진다. Moran의 I_i는 우선 양의 공간적 자기상관과 음의 공간적 자기상관을 매우 잘 구분해 낸다. 단순화해 말한다면, 통계치가 0보다 크면 양의 공간적 자기상관을, 통계치가 0보다 작으면 음의 공간적 자기상관을 나타낸다. 따라서 Moran의 I_i는 공간적 이례치(spatial outlier)를 찾아내는 데 탁월하다. 물론 공간 클러스터의 탐지도 우수하지만 핫스팟과 콜드스팟을 수치상으로는 구분하여 보여주지 못한다는 단점이 있다.

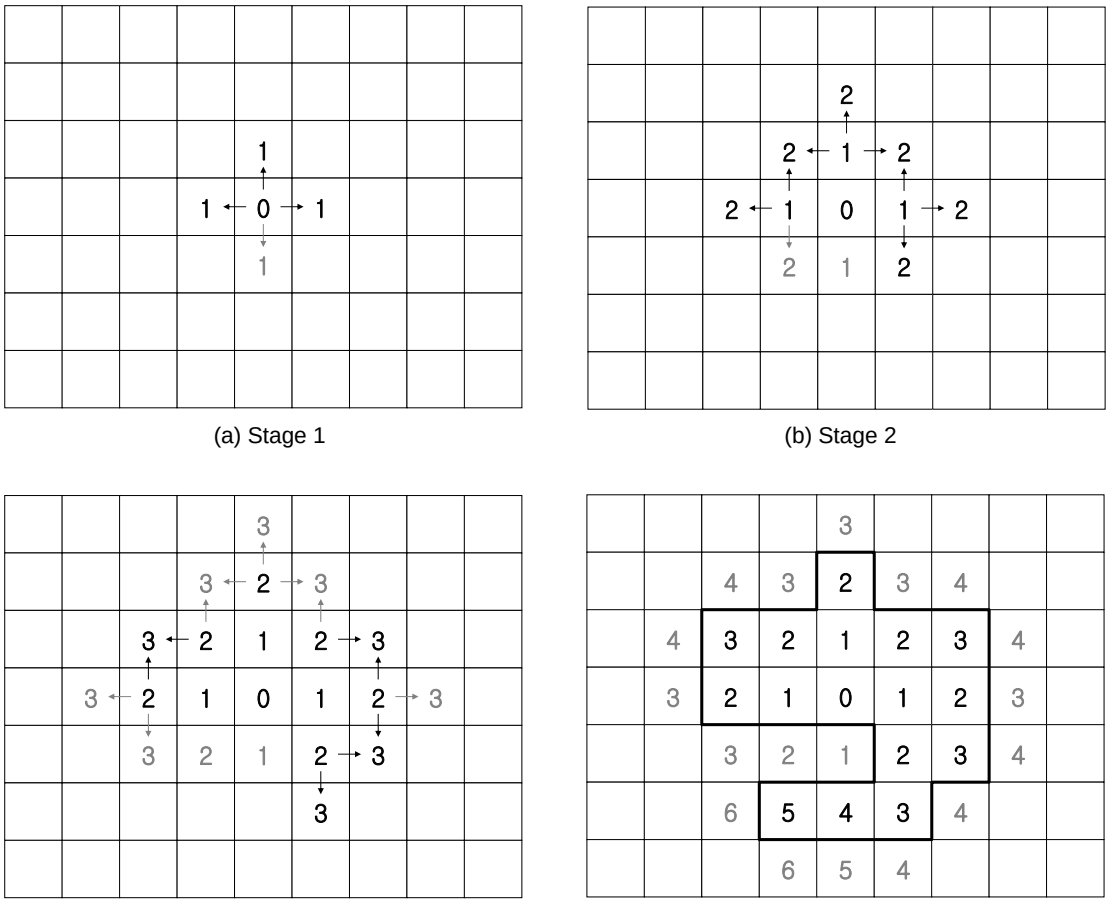
Geary의 c_i는 통념적인 의미에서의 공간 클러스터, 즉 매우 높은 값이 집중해 있거나 매우 낮은 값이 집중해 있는 하위지역을 잘 찾아내지 못하는 경향이 있다. 왜냐하면 오로지 차이에 기반하고 있어 그 속성 값들이 매우 높거나, 낮거나, 평균적이거나 상관없이 차이의 제곱의 합이 충분히 작기만 하면 클러스터로 탐지하기 때문이다. 즉 핫스팟, 콜드스팟, 평균스팟 등이 뒤섞여 찾아지게 되는 것이다. 그러나 이러한 특징이 Geary의 c_i에 고유한 장점을 부여하기도 하는데, '국지적 분산' 의 개념을 측정하는데 탁월하다고 말할 수 있다. Moran의 I_i에 의해 탐지된 핫스팟은 국지적 분산이 작다고 볼 수는 없는데, 값들의 차이가 큼에도 불구하고 중심 셀이나 주변 셀 중 몇몇 셀이 매우 높은 값을 가지게 되면 통상 핫스팟으로 탐지되기 때문이다. 어떤 연구 상황에서는 이것이 환영 받을 만한 일이 아닐 수도 있다.

Getis-Ord의 G_i^*가 가지는 가장 큰 장점은 통계치로부터 직관적으로 핫스팟인지 콜드스팟인지 확인할 수 있다는 점이다. 즉, 통계치가 양수면 핫스팟과 관련되고, 음수면 콜드스팟과 관련된다. 이러한 특성은 다른 어떤 통계량에서도 찾아볼 수 없는 G_i^*만의 장점이고, 다음에서 소개할 AMOEBA 알고리즘에서 채택된 이유이기도 하다. 물론 단점도 있다. Getis-Ord의 G_i^*는 Moran의 I_i가 가지고 있는 공간적 이례치의 탐지 능력이나, Geary의 c_i가 가지고 있는 국지적 분산의 측정 능력은 가지고 있지 않다. 그러나 연구 목적이 핫스팟과 콜드스팟의 확인인 상황이라면, 현재까지 제안된 LISA 중 Getis-Ord의 G_i^*를 선택하는 것이 가장 합리적인 것으로 판단된다.

2) AMOEBA 기법의 원리

Figure 2는 AMOEBA 알고리즘을 단계별로 보여주고 있다. 1단계에서, 우선적으로 한 셀에 대해 그 셀만을 대상으로 G_i^*를 계산하고, 그 값을 $G_i^*(0)$이라고 부른다. G_i^* 통계량 수식(식 1)에 오직 한 셀만이 고려된다는 사실을 적용하면, $G_i^*(0)$ 값은 그 셀에서의 표준점수(z-값)와 동일하다는 사실을 알 수 있다. 따라서

그 셀의 $G_i^*(0)$ 값이 0보다 크다면, 그 셀의 원래 값이 평균보다 크다는 것을 의미하고 잠재적인 핫스팟의 일원이 되며, $G_i^*(0)$ 값이 0보다 작다면 그 반대의 경우를 의미한다. 그 셀을 중심으로 상하좌우에 위치한 네 개의 그리드 셀을 이웃으로 규정하고, 그 네 개 셀로 가능한 모든 조합 각각에 중심 셀을 합해 규정되는 '구역'에 대해 G_i^*를 계산하는 것이 1단계의 핵심이다. 문제는 네 개 셀로 모두 몇 개의 조합이 만들어 질 수 있느냐이다. 이것은 네 개에서 하나를 뽑는 방법, 네 개에서 두 개를 뽑는 방법, 네 개에서 세 개를 뽑는 방법, 네 개에서 네 개를 뽑는 방법을 모두 합하는 것과 같다. 이 경우 15가지(4+6+4+1) 조합이 가능하다. 이 각각의 조합에 중심 셀을 연합시키면 서로 다른 15개의 구역을 갖게 된다. 이 때 각 구역에 대해 계산된 G_i^*값을 $G_i^*(1)$이라고 부른다. $G_i^*(0)$가 0보다 큰 경우, 이 $G_i^*(1)$들 중 최대값이 $G_i^*(0)$보다 크면, 그 최대값을 보인 구역이 1차적인 핫스팟으로 간주 된다. $G_i^*(0)$가 0보다 작은 경우는, $G_i^*(1)$들 중 최소값이 $G_i^*(0)$보다 작으면, 그 최소값을 보인 구역이 1차적인 콜드스팟으로 간주 된다. Figure 2의 (a)는 중심 셀과 상, 좌, 우에 위치한 세 셀을 합한 것이 1단계의 클러스터로 확정되었음을 보여주고 있다. 이 때 누락된 하변의 셀은 다음의



(a) Stage 1 (b) Stage 2

(c) Stage 3 (d) Final Stage

Figure 2. The AMOEBA algorithm. AMOEBA 알고리즘. (Source: Aldstadt and Getis, 2006)

모든 과정에서 배제된다. 즉, 절대로 어떤 클러스터의 멤버십도 가질 수 없게 된다.

2단계에서는 1단계에서 포함된 이웃 셀을 중심으로 상하좌우의 가능한 위치에 두 번째 이웃 셀의 세트를 설정한다. Figure 2의 (b)는 (a)를 전재할 때 7개의 이웃 셀이 설정될 수 있음을 보여주고 있다. 앞의 단계와 마찬가지로 모든 가능한 조합을 통해 구역을 생성하고 모든 구역에 대해 G_i^*를 산출하면 $G_i^*(2)$의 집합을 얻게 된다. 이 때 $G_i^*(2)$의 최대값이 $G_i^*(1)$보다 크면 ($G_i^*(0)$이 0보다 큰 경우), 그 최대값을 산출한 구역이 2차적인 클러스터가 된다. 이러한 과정을 새로운 셀의 첨가를 통해 더 이상의 G_i^* 값의 증가($G_i^*(0)$이 0보다 큰 경우)가 없을 때까지 지속한다. Figure 2에는 나타나 있지 않지만 4~6단계를 거치면 (d)에 나타나 있는 최종 결과가 도출된다. 이 때 그 구역을 모두 포괄하는 외곽 경계선이 클러스터의 범역을 획정하는 선이 되는 것이다.

위의 과정은 하나의 셀을 중심으로 어떻게 공간 클러스터의 범역이 설정되는 지를 보여주는 것이다. 따라서 논리적인 의미에서 본다면, 이러한 과정이 모든 그리드 셀에 대해 반복적으로 수행되어야 한다. 그러면 셀 개수만큼의 공간 클러스터가 형성될 것이고 범역들이 서로 겹치는 문제가 발생할 것이므로 공간 클러스터를 선별하는 이차적인 규준이 필요하게 된다. Aldstadt and Getis(2006, 336)는 다음과 같은 규준을 제시한다. 우선 셀 개수만큼 생성된 클러스터들을 G_i^* 값의 순서에 따라 내림차순으로 정리하고 리스트를 작성한다. 그 다음 가장 큰 G_i^* 값을 보이는 클러스터를 우선 선택하고, 이 클러스터와 범위를 공유하는 모든 클러스터는 제거한다(리스트에서도 삭제한다). 그러고 나서 리스트에서 두 번째로 높은 G_i^* 값을 보유한 클러스터를 선택한다. 마찬가지로 이 두 번째 클러스터와 범위가 겹치는 다른 클러스터는 제거한다. 이러한 과정을 반복 수행하게 되면 최종적으로 공간 클러스터들의 세트가 도출된다.

4. 수정 AMOEBA 기법

1) AMOEBA의 수정

전술한 바와 같이 Aldstatdt and Getis(2006)의 연구에서는 모든 셀들에 대해 각 셀을 기준으로 하는 클러스터들을 찾은 후, 클러스터 간에 중복이 발생할 경우 G_i^* 통계치가 더 작은 클러스터를 제거함으로써 최종적인 클러스터 분포를 확정하였다. 하지만 원 방법과 같이 각 셀에서 산출된 클러스터의 G_i^* 통계치에 따라 중복 클러스터를 제거하게 될 경우 예기치 못한 결과가 도출될 수 있어 본 연구에서는 이를 수정한 알고리즘을 사용하고자 한다.

평균 보다 높은 값들이 분포하는 지역일 경우 G_i^* 통계량은 중심 셀의 값이 보다 크고, 인접 셀의 값이 중심 셀보다 더 크거나 유사한 경우가 많을수록 그렇지 않은 경우 보다 더 높게 산출된다(평균 보다 낮은 값들이 분포하는 지역에서는 반대가 된다). 즉, 중심 셀보다 크거나 유사한 인접 셀이 늘어나면 G_i^*를 구성하는 수식에서 분자의 증가 효과가 분모의 증가 효과에 비해 더 커지면서 G_i^* 통계치도 증가하는 경향성을 보이게 된다. 하지만 중심 보다는 낮은 값이 이웃으로 계속 추가되면 분자 값의 증대 효과가 사라지면서 G_i^* 통계치는 낮아질 수 있다.

이러한 속성으로 인해 G_i^* 통계치에 기초한 클러스터의 범역 설정은 개별 지점의 상황에 따라 그 결과가 다르게 산출되는 특성을 보이게 된다. 예를 들어 상대적으로 낮은 값들이 몰려 있는 어느 지점에서 시작하여 탐지한 클러스터와 비교적 높은 값들이 몰려 있는 어느 지점에서 시작하여 탐지한 클러스터는 상당히 다른 형태로 도출될 수 있다. 값이 매우 큰 셀부터 클러스터를 찾기 시작하는 경우는 이웃으로 추가되는 셀들이 평균 보다는 값이 크더라도 중심 셀보다는 작을 가능성인 높기 때문에 근린 수 증가에 따른 G_i^* 통계량 크기의 증가 효과가 비교적 빠른 속도로 감소하며, 확정된 클러스터 내에서의 셀 값들의 편차도 상대적으로 낮게 나타난다. 하지만 값이 평균과 유사한 셀부터 클러스터를 찾기 시작한 경우는 이웃으로 추가되는 셀들

의 값이 중심 셀의 값보다 더 클 가능성이 높아 G_i^* 통계치가 지속적으로 증가할 수 있는데, 완변하는 양의 공간적 자기상관이 존재하는 경우라면 결과적으로 상당히 넓은 범역이 하나의 클러스터로 탐지된다.

문제는 값이 매우 큰 셀부터 시작하여 탐색한 클러스터와 평균과 유사한 셀부터 시작하여 탐색한 이 두 클러스터가 중복되는 경우에서 분명히 드러난다. 보다 작은 값에서 시작한 클러스터는 높은 값들이 몰려있는 지점을 향해 확장해가게 되는데, G_i^* 통계치의 최대화를 추구하다 보면 결과적으로 큰 값에서 시작한 클러스터의 영역을 대부분 혹은 모두 포괄하게 되고, G_i^* 통계치의 값도 더 커지게 된다. 샘플 자료를 이용하여 값이 가장 큰 지점(표준 점수 5.97)과 평균과 유사한 지점(표준 점수 0.35)에 대해 각각 클러스터를 탐색한 Figure 3은 위와 같은 사실을 잘 보여 준다. 두 클러스터는 서로 다른 셀에서 시작하였으나 서로 중복되었는데(높은 값에서 시작한 클러스터가 낮은 값에서 시작한 클러스터에 포함됨), 값이 가장 큰 지점에서 탐색한

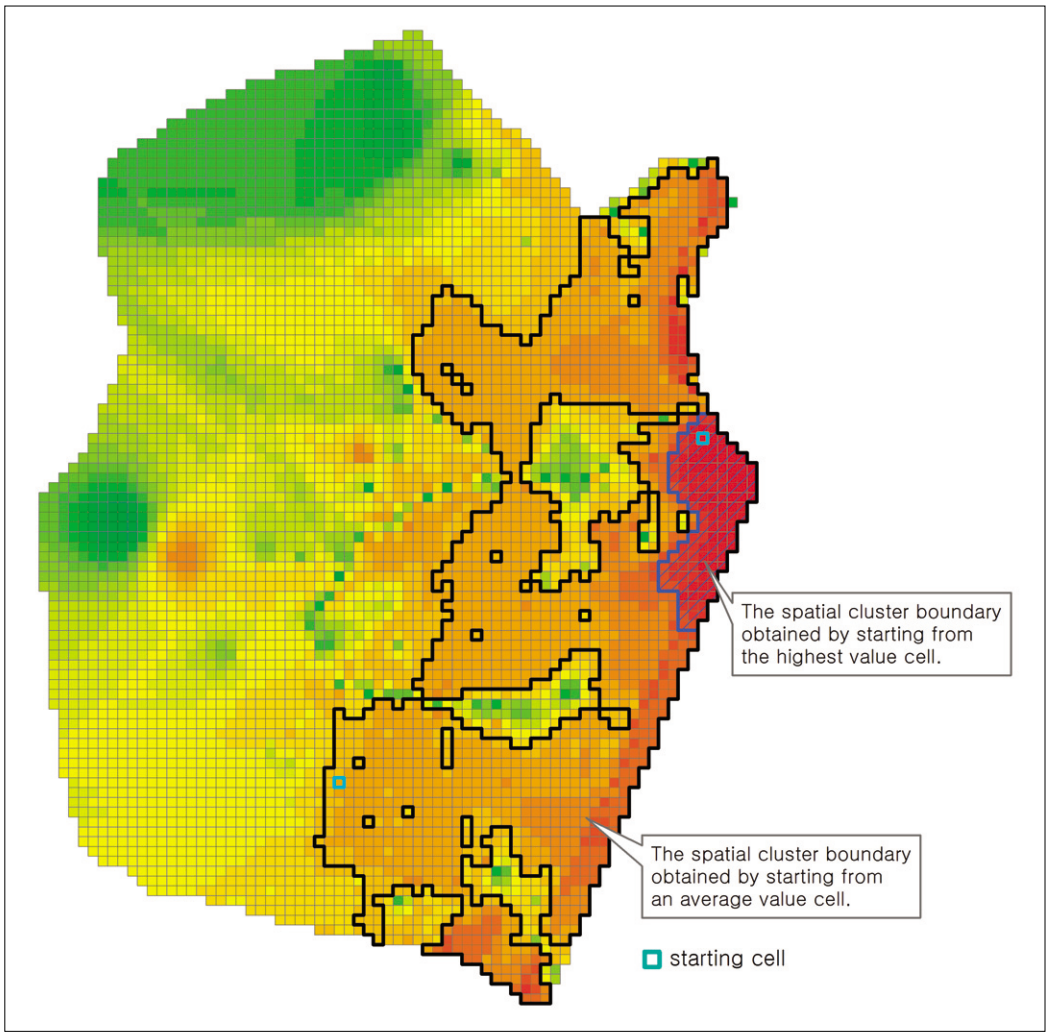


Figure 3. A comparison of different spatial cluster boundaries due to different starting cells.
시작 셀의 위치에 따라 상이하게 추출된 공간 클러스터 경계의 비교.

클러스터의 G_i^* 통계량은 46.56, 평균과 유사한 지점에서 탐색한 통계량은 47.57로 나타났다.

위와 같은 경우 원 방법론에서와 마찬가지로 G_i^* 통계치 기준에 의거하여 중복을 제거하게 되면 거대한 클러스터가 채택되는 대신 보다 높은 값의 클러스터는 사라지게 될 것이다. 바꾸어 말하면 원 AMOEBA 방식을 택할 경우 Aldstadt and Getis(2006, 335)가 이야기한 바와 마찬가지로 최종적으로 확정된 클러스터 내에서는 어느 지점에서 시작하더라도 결과로 도출되는 클러스터의 경계는 동일할 것이다. 즉, 원 AMOEBA에서는 클러스터에 대한 해가 하나 밖에 존재하지 않는데, 이는 학술적인 측면이나 도시계획적 측면에서 더욱 큰 관심의 대상이 되는 보다 높은 값 혹은 보다 낮은 값들의 클러스터의 탐지에는 한계로 작용할 수 있다.

따라서 본 연구에서는 원 AMOEBA에서 사용한 중복 제거의 과정을 다소 수정한 알고리즘을 사용하였다. 두 가지 면에서 원 AMOEBA와 차이를 보이는데, 보다 핵심적인 차이는 동일하게 모든 셀들을 대상으로 클러스터를 찾되, 변수의 값(절대값)을 내림차순으로 정렬하여 값이 높은 위치부터 클러스터를 찾기 시작한다는 점이다. 이때 변수 값이 최대인 셀에서 파악된 영역이 가장 우선적인 클러스터로 확정된다. 다음으로는 이 클러스터에 포함되지 않는 셀 중에서 값이 가장 큰 셀에 대해 클러스터 탐색을 시작한다. 그런데 여기서 탐색을 시작한 클러스터가 먼저 탐색하여 채택된 클러스터와 중복될 경우에는 원 AMOEBA와 동일한 문제상황에 직면하게 된다.

위와 같은 문제 상황에 대응하기 위해 수정 AMOEBA에서는 다음과 같은 방안을 제시한다. 값이 보다 낮은 셀에서 시작한 클러스터가 높은 셀에서 시작하여 이미 채택된 클러스터와 중복하게 될 경우 그 클러스터 전체를 제거하는 방법을 사용할 수 있는데, 이를 '수정 AMOEBA 1'이라 부르고자 한다. 다음으로는 값이 낮은 셀에서 시작한 클러스터에서 중복되는 영역만을 제거하는 방안을 사용할 수 있는데, 이는 '수정 AMOEBA 2'라 부르고자 한다.

수정 AMOEBA 2에서는 중복 부분만 제거하고 남는 영역을 별도의 클러스터로 파악해야 할 것인지에 대한 이슈가 존재할 수 있는데, 원 AMOEBA의 G_i^* 통계치 최대화의 논리를 수용한다면 중복이 되는 두 클러스터를 병합하는 것이 합리적일 것이다. 즉, 중복이 되는 두 클러스터를 병합한 영역의 G_i^* 통계치가 병합하지 않은 경우보다 더 커진다면 병합을 하는데, 이때 클러스터로서의 요건을 평가하는 최소 기준을 통과한 클러스터만 병합하였다. 최소 기준의 설정은 원 AMOEBA와의 두 번째 차이점으로 큰 의미 없는 클러스터를 사전에 제거하는 역할을 한다. 결과적으로 최소 요건을 갖춘 클러스터들만의 분포로부터 중복 클러스터를 병합하였는데, 그 요건으로는 G_i^* 통계치를 사용하였다. 전술한 바와 같이 G_i^* 통계치는 평균 0, 분산 1의 분포를 따르므로 표준 점수와 같이 해석될 수 있어 본 연구에서는 편의상 신뢰도 99%의 임계치에 해당하는 2.58

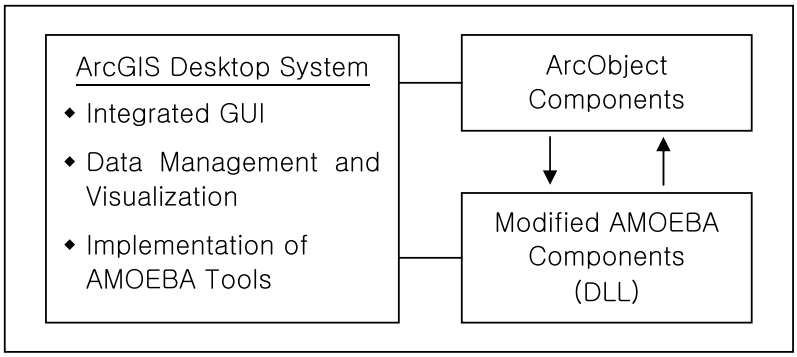


Figure 4. Conceptual framework for a GIS-based program for delineating spatial clusters.
GIS-기반 공간 클러스터 범역 설정 도구 개발 방법.

을 적용하였다.

2) GIS-기반 실행 프로그램의 개발

지금까지 기술한 내용을 바탕으로 공간 클러스터의 범역을 설정할 수 있는 도구를 개발하였다. 분석 도구는 일반 GIS 엔진에 추가하여 사용할 수 있는 확장 기능(컴포넌트 혹은 익스텐션) 형태로 개발하였으며, GIS 엔진으로는 ESRI사의 ArcGIS(9.x)를 대상으로 하였다. 프로그램은 Microsoft사의 Visual Basic 6.0을 통해 작성하였으며, ArcGIS 환경과의 통합이 용이하도록 COM DLL 파일 형태로 구현하였다(Figure 4).

ArcGIS 환경에서 본 분석 도구를 추가한 결과 및 실행 화면은 Figure 5와 같다. 본 분석 도구는 입력 변수에 대한 공간 클러스터를 탐색하여 경계 레이어를 생성한다. 클러스터 경계 레이어는 해당 클러스터의 탐색 시작 위치(셀 ID), G_i^* 통계치를 함께 기록함으로써 분석 결과를 용이하게 파악할 수 있다. 또한 본 분석 도구는 다양한 시드(즉, 클러스터 탐색에 사용될 선택적 시작 위치) 선택 옵션을 적용함으로써 사용자가 원하는 위치에 대해서만 편리하게 클러스터를 분석할 수 있도록 하였다.

3) 수정 AMOEBA 알고리즘의 평가

GIS 기반의 분석 도구를 개발한 후 알고리즘의 평가를 위해 모의 데이터 및 실 데이터를 이용하여 두 가지 실험을 실시하였다. 하나는 공간적 자기상관이 전형적으로 드러나 공간적 클러스터의 범역에 대해 직관적인 판단이 가능한 가상의 공간 패턴을 대상으로 수정

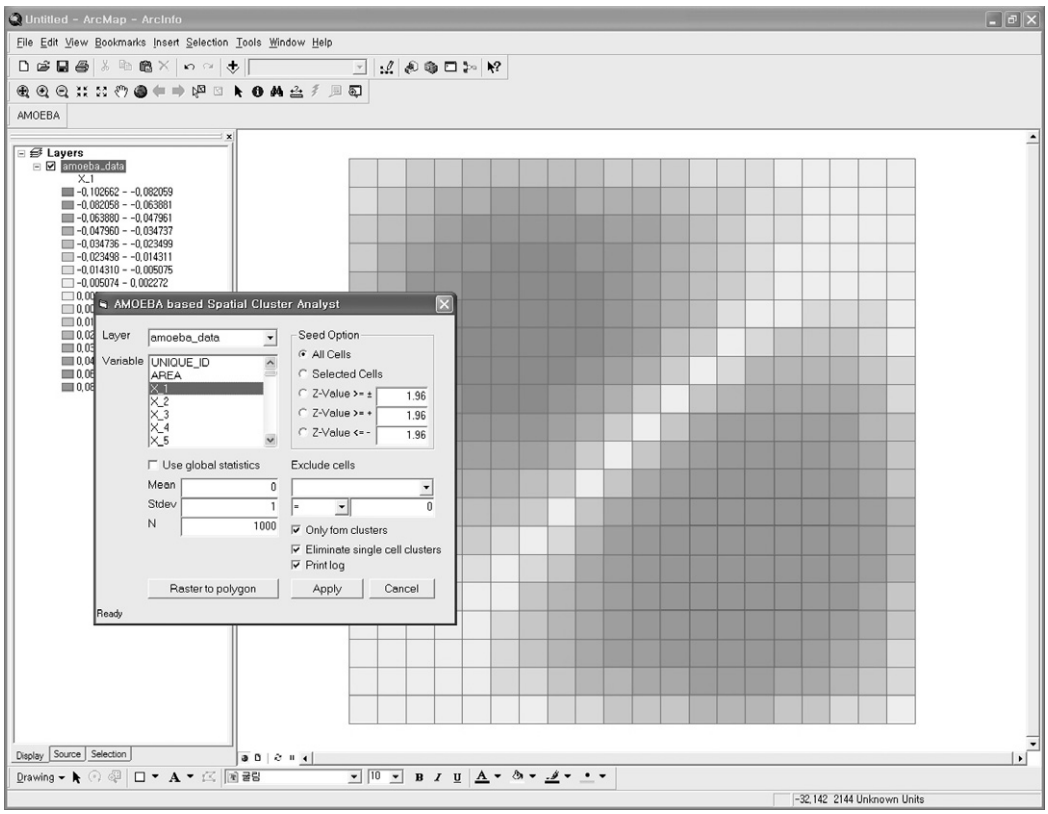


Figure 5. A GIS-based program for delineating spatial cluster. GIS-기반 공간 클러스터 범역 설정 도구의 실행 화면.

AMOEBA를 적용하는 것이고, 또 하나는 실 세계의 인구밀도 데이터에 적용하는 것이다. 두 실험 모두 원 AMOEBA와 수정 AMOEBA를 비교하였는데, 수정 AMOEBA에 대해서는 두 가지 수정 방안 모두의 결과를 제시하였다.

우선 모의 데이터 상에서 전형적인 공간적 자기상관을 보이는 패턴을 생성하기 위해 Boots and Tiefelsdorf(2000)와 Boots(2001)가 사용한 방법을 원용했다. 프로젝션 매트릭스(projection matrix)와 공간 가중 행렬을 사용하면 공간적 자기상관 통계량을 2차 형식(quadratic form)으로 표현할 수 있고, 이 때 프로젝션 매트릭스와 공간 가중 행렬을 곱한 후, 도출된 매트릭스를 고유치와 고유벡터로 분해하면 해당 공간적 자기상관 통계량의 최대 범위(최대값~최소값)에 해당하는 공간적 패턴을 추출할 수 있다. Moran 통계량을 예로 든다면, 양의 공간적 자기 상관이 가장 높은 패턴에서부터, 공간적 자기상관이 없는 패턴을 거쳐, 음의 공간적 자기상관이 가장 높은 패턴을 도출할 수 있게 된다.

테스트에 사용된 샘플 데이터는 20×20개의 셀로 구성된 그리드로 공간적 자기상관의 수준에 따라 클러스터의 분포를 달리하는 400개의 변수들로 구성되어 있다. 각 변수들은 평균 0, 표준 편차 0.05를 갖는 분포에서 추출되었다. 테스트의 결과는 공간 클러스터의 경향을 전형적으로 나타내는 4개 변수를 대상으로 제시하였는데, 일반적인 통계 특성은 Table 1에 나타나 있다. 표에 나타난 4개의 변수는 양의 공간적 자기상관이 강하게 나타나지만 각기 클러스터의 크기나 분포에서 서로 구별되는 패턴이다. 모든 셀에 대해 클러스터를 파악하되, 원 AMOEBA와 수정 AMOEBA의 결과를 비교하여 적용력을 검토하였는데, 각 방법에 의해 도출된 클러스터 중 G_i^* 통계치가 2.58 미만인 클러스터는 제거하였다.

Figure 6은 이러한 4개의 공간 패턴에 대해 원 AMOEBA, 수정 AMOEBA 1, 수정 AMOEBA 2의 알고리즘을 각각 적용하여 공간 클러스터의 범역을 추출한 결과를 보여주고 있다. 패턴 1은 공간적 자기상관의 수준이 최대화된 분포로 모든 방법에서 직관적인 인식과 같은 두 개의 클러스터가 도출되었다. 원 AMOEBA에 비해 수정 AMOEBA의 클러스터가 다소 작게 도출되었으나 수정 AMOEBA의 두 결과는 서로 일치하였다. 패턴 2는 국지적인 클러스터가 늘어나면서 Moran의 I_i 값이 다소 감소하였는데, 높은 값의 클러스터와 낮은 값의 클러스터가 서로 독립하여 분포하고 있다. 직관적으로 볼 때 9개의 클러스터가 존재하나 원 AMOEBA에서는 7개가, 수정 AMOEBA에서는 두 경우 모두 동일하게 9개가 탐지되어 보다 큰 차이가 드러나기 시작했다.

패턴 3은 국지적인 클러스터가 더 늘어나면서 Moran의 I_i 값이 더 감소하고 있지만 높은 값의 클러스터와 낮은 값의 클러스터가 일부 서로 근접하는 패턴이다. 핫스팟 간에, 그리고 콜드스팟 간에 근접하는 이 패턴에서 원 AMOEBA와 수정 AMOEBA의 차이는 상당히 분명하게 드러나고 있다. 직관적으로 16개의 클러스터가 파악되나 원 AMOEBA에서는 9개의 클러스터가 탐지되었는데, 핫스팟과 콜드스팟이 근접한 곳에서는 거대 클러스터가 도출되었다. 반면 수정 AMOEBA에서는 두 경우 모두 16개의 클러스터가 동일하게 파악되고 있음을 알 수 있다.

마지막 패턴 4는 패턴 3과 유사한 분포이지만 서로

Table 1. Statistical summary for test data. 테스트 데이터의 통계 요약.

Descriptive Statistics	Pattern 1	Pattern 2	Pattern 3	Pattern 4
Mean	0.000	0.000	0.000	0.000
Standard Deviation	0.050	0.050	0.050	0.050
Maximum	0.102	0.097	0.091	0.091
Minimum	-0.102	-0.095	-0.091	-0.091
Moran's *I*	1.035	0.926	0.881	0.875

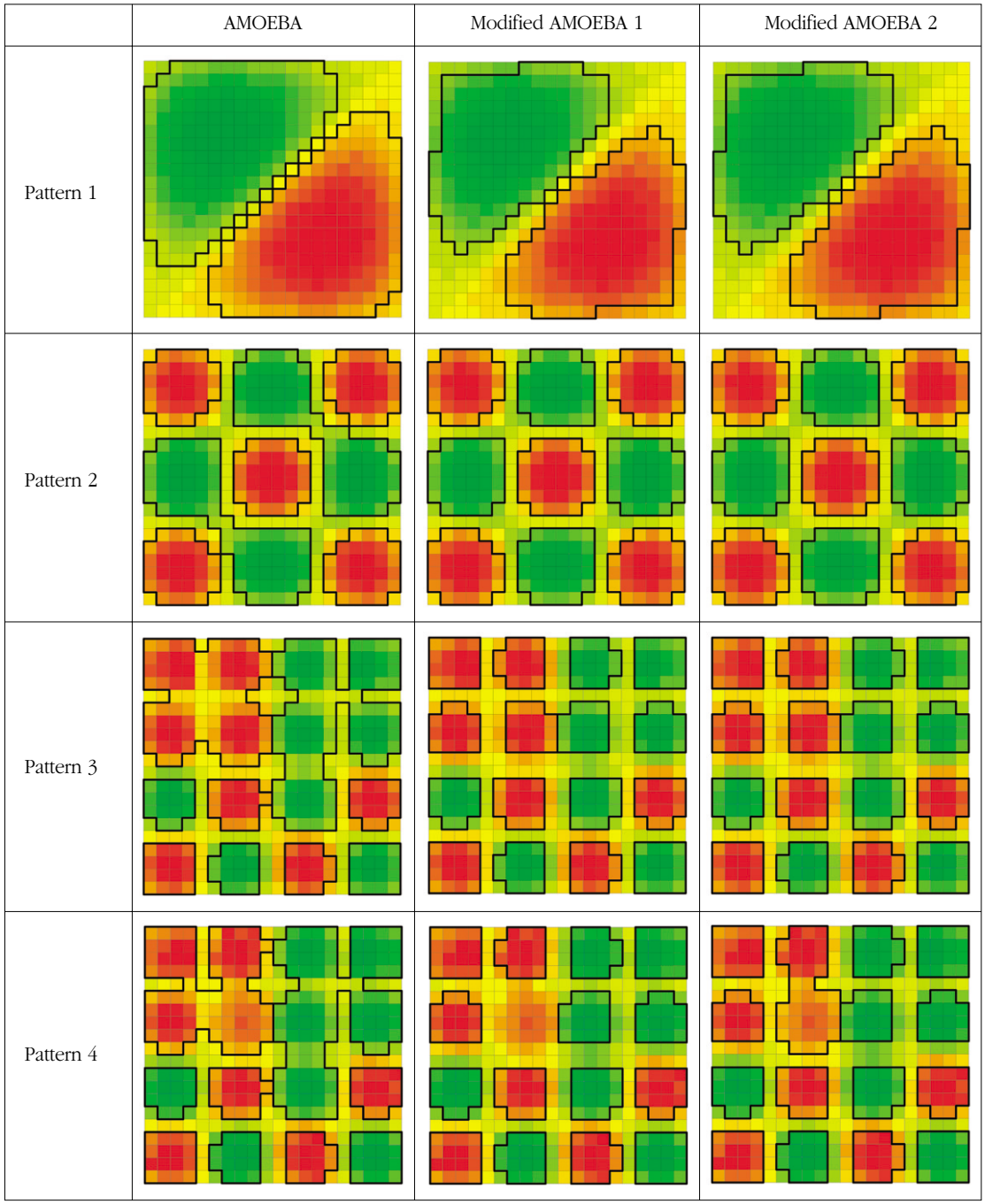


Figure 6. Results of spatial cluster delineation for test data (Black lines are cluster boundaries).
테스트 데이터에 대한 공간 클러스터 범역 설정 결과.

근접한 높은 값의 클러스터들 중 하나가 다른 클러스터에 비해 값이 다소 낮은 패턴이다. 즉, 핫스팟으로 추정되는 지역들이 서로 근접하지만 변수 값에서는 차이가 있는 패턴인데, 여기에서는 세 가지 방법에서 모두 다른 결과가 도출되었다. 직관적으로 15개의 클러스터 및 1개의 잠재 클러스터가 존재하는데, 원 AMOEBA에서는 9개의 클러스터가 도출되었고 그 중 일부는 거대 클러스터로 탐색되었다. 수정 AMOEBA 1에서는 서로 근접한 높은 값의 클러스터 중 값이 상대적으로 낮은 클러스터는 다른 클러스터와 중복되어 제거되는 결과를 나타낸 반면, 수정 AMOEBA 2에서는 잠재 클러스터가 인접한 클러스터에 병합되어 탐색되는 결과를 나타내었다. 수정 AMOEBA 1에서는 제거되었으나 수정 AMOEBA 2에서는 클러스터로 포함된

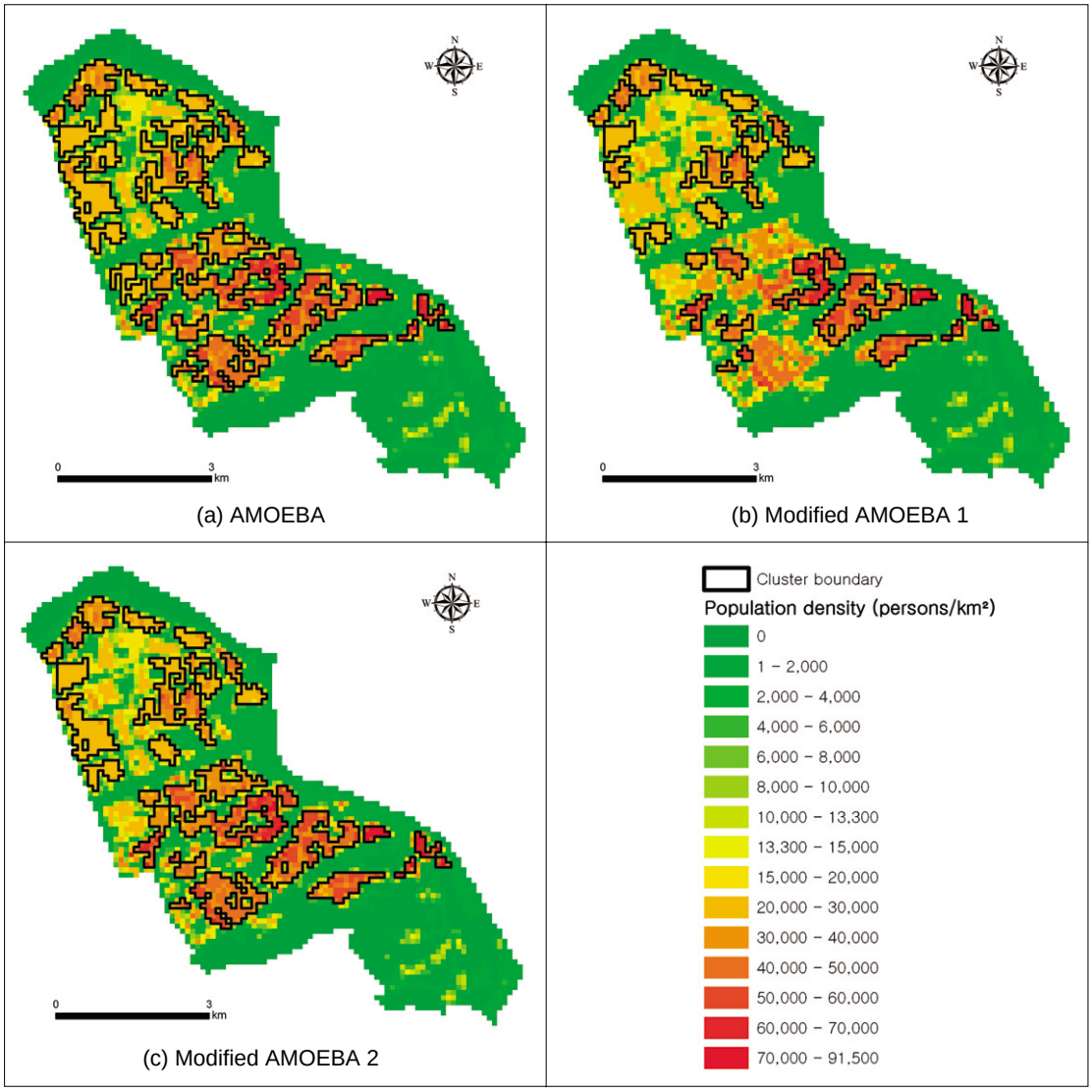


Figure 7. Results of spatial cluster delineation for population density pattern of Gangnam-gu, Seoul.
서울 강남구의 인구밀도 패턴에 대한 공간 클러스터 범역 설정 결과.

영역만을 대상으로 하면 G_i^* 통계치는 3.55, 변수의 평균 값은 0.035(전체 셀 값의 약 70 퍼센타일)를 나타내어 제거되지 않은 수정 AMOEBA 2의 결과가 더 합리적인 것으로 판단된다.

이상과 같은 모의 데이터와는 달리 실 데이터의 분포 패턴은 상당히 비정형적으로 나타날 가능성이 높다. 비정형적인 실 데이터에 대한 수정 AMOEBA 기법의 적용성을 검토해 보기 위해 서울 강남구의 인구밀도 분포에 대해 위와 동일한 방식으로 실험을 해 보았는데, 편의상 높은 값들의 클러스터만을 탐색하였다. 분석에 사용된 실 데이터는 대시메트릭 에어리얼 인터폴레이션(dasymetric areal interpolation) 기법을 통해 생성된 100×100m 인구 밀도 그리드이다(Lee and Kim, 2007). 분석 결과는 Figure 7과 Table 2에 나타나 있다.

우선 원 AMOEBA의 경우 모두 25개의 공간 클러스터가 탐지되었으나 수정 AMOEBA 1과 AMOEBA 2에서는 각각 31개와 30개가 탐색되었다. 공간 클러스터들의 총면적과 평균면적을 비교해보면 원 AMOEBA가 가장 넓고, 수정 AMOEBA 1이 가장 좁으며, 수정 AMOEBA 2가 평균적인 경향을 보였다. 원 AMOEBA는 클러스터의 수가 더 적지만 면적은 더 넓어 클러스터의 크기가 상대적으로 더 넓게 탐색됨을 알 수 있다. 이러한 경향은 지도 상에서도 확인할 수 있는데 수정 AMOEBA의 결과는 원 AMOEBA의 결과 보다 더 정비된 모습으로 나타나고 있다.

또한 공간 클러스터들의 인구밀도를 평균해 보면 원 AMOEBA에 비해 수정 AMOEBA에서 값이 높게 나타나 값이 더 큰 셀들이 클러스터에 포함되고 있음을 알 수 있다. 뿐만 아니라 공간 클러스터 중 인구밀도가 최소인 경우는 값이 서로 유사하지만 최대인 경우는 수정 AMOEBA에서 훨씬 크게 나타나고 있어 더 높은 값들의 클러스터가 유지되고 있음을 짐작할 수 있다. 실제 원 AMOEBA에 의한 클러스터 중 인구밀도가 최대인 경우 보다 더 큰 값을 갖는 클러스터가 수정 AMOEBA에서는 5개 정도 더 포함된 것으로 나타났다. 공간 클러스터들의 인구밀도에 대한 표준편차 또한 수정 AMOEBA에서 더 크게 나타나고 있는데 이는 수정 AMOEBA의 경우 더 높은 값의 클러스터들과 더 낮은 값의 클러스터들 간에 차이가 더 분명하게 나타나고 있음을 보여준다.

수정 AMOEBA 간에도 차이가 발견되고 있는데, 두 방법은 비슷한 수의 클러스터를 도출하였으나 전체 면적에서는 수정 AMOEBA 2가 수정 AMOEBA 1에 비해

Table 2. Statistical summary for results of spatial cluster delineation. 공간 클러스터 범역 설정 결과에 대한 통계 요약.

Summary Statistics	AMOEBA	Modified AMOEBA 1	Modified AMOEBA 2
Number of spatial clusters	25	31	30
Average area of spatial clusters(km^2)	0.42	0.20	0.30
Total area of spatial clusters(km^2)	10.39	6.23	9.11
Average population density of spatial clusters(persons/km^2)	33,572	41,781	39,906
Standard deviation of population density of spatial clusters	10,126	13,854	12,887
Maximum population density of spatial clusters(persons/km^2)	54,672	77,836	77,836
Minimum population density of spatial clusters(persons/km^2)	20,900	24,700	24,700

훨씬 넓은 면적을 나타냈다. 수정 AMOEBA 1의 경우 중복 클러스터가 제거되면서 더 단순해 보이는 결과를 나타냈다. 특히 중앙부에는 보다 높은 값들의 잠재적 클러스터가 존재하는 것으로 보이지만 중복으로 인해 모두 제거되면서 실 데이터에서도 적용력의 한계를 노출하였다. 즉, 이 지역은 주변 핫스팟에 비해서는 다소 값이 낮지만 북쪽에 위치한 클러스터들에 비해 더 높은 값들을 가지고 있으므로 단순히 인접 클러스터와의 중복으로 인해 제거하는 것은 비합리적이라 판단된다. 이러한 결과는 앞서 모의 데이터에서 잠재적인 핫스팟들이 서로 근접하지만 핫스팟 간에 변수 값이 차이를 보이는 경우에서와 동일한 결과라고 할 수 있다.

이상과 같은 테스트 데이터와 실 데이터에 대한 실험 결과를 정리해 보면, 우선 양의 공간적 자기상관이 존재하면서 국지적인 클러스터가 많지 않은 경우는 원 AMOEBA와 수정 AMOEBA 간에 큰 차이가 발견되지 않았다. 하지만 국지적인 클러스터가 늘어나면서 원 AMOEBA와 수정 AMOEBA 간에 차이가 발견되었는데, 특히 핫스팟이나 콜드스팟이 서로 독립적으로 분포하기 보다는 서로 근접 분포하는 경우 원 AMOEBA는 제한적임에 비해 수정 AMOEBA는 효과적인 탐지 결과를 보였다. 나아가 핫스팟이나 콜드스팟이 서로 근접하되 상호 간에 변수 값의 차이가 존재하는 경우는 수정 AMOEBA 간에도 차이를 보였는데, 수정 AMOEBA 1은 잠재적인 클러스터를 제거해버리는 반면 수정 AMOEBA 2는 효과적으로 대응하는 결과를 나타내었다.

5. 결론

본 연구는 기본적으로 특정 현상이 두드러지게 드러나는 구역을 찾아내어 그것의 경계를 획정하는 기법을 개발하는 방법론 연구이다. 주요한 결과를 요약하면 다음과 같다. 첫째, 기존 방법론을 검토한 결과, LISA를 이용한 AMOEBA 기법이 가장 타당성이 높은 것으로 판단되었다. 둘째, 수정 AMOEBA 기법의 알고리즘을 확립했으며 실행 소프트웨어를 상용 GIS 프로그램의 확장 기능 형태로 개발하였다. 셋째, 실험 및 실 데이터에 적용한 결과 수정 AMOEBA 2라고 명명한 기법의 유용성이 가장 뛰어난 것으로 확인되었다.

방법론적인 측면에서 몇 가지 향후 연구 과제를 제시할 수 있다.

첫째, AMOEBA 프레임워크 속에서 다양한 공간적 자기상관 지수가 실험될 필요가 있다. Aldstadt and Getis(2006) 역시 Getis-Ord의 G_i^* 외에 Moran의 I_i나 공간적 스캔 통계량(Kulldorff, 1997; 2009)을 사용하는 것이 가능하다고 말하고 있다. 이 외에 Geary의 c_i나 Lee의 S_i (Lee, 2001; 2004; 2009) 역시 좋은 후보자가 될 수 있다. 특히 Lee의 S_i는 표준 점수의 가중평균의 제곱으로 표현되기 때문에 알고리즘이나 분석 결과를 직관적으로 이해하기 쉽다는 장점이 있다.

둘째, 도출된 공간 클러스터의 통계적 유의성을 검정할 수 있는 방법론의 개발이 시급하다. 왜냐하면 도출된 모든 공간 클러스터가 통계적으로 유의한 것도 아니고, 통계적 유의 확률이 동일한 것도 아닐 것이기 때문이다. 또한 클러스터 탐색 과정의 매 단계에서 공간적 자기상관 통계량의 최대값만을 따지는 것 외에 유의 확률을 고려하는 방법도 좋은 대안이 될 수 있을 것이다.

셋째, 수정 AMOEBA 알고리즘은 그 성격상 단 한 세트의 클러스터 경계만을 도출해 준다. 그러나 논리적으로는 어떤 기준을 달리함으로써(예를 들어 서로 다른 유의 수준) 서로 다른 세트의 클러스터 경계들이 생성될 수 있다. 이렇게 되면 일종의 '경계 등고선'이 생성될 것이고, 이를 이용하면 패턴 탐색성뿐만 아니라 경계 설정의 유연성을 고양할 수 있을 것이다.

넷째, 전혀 새로운 방식의 공간 클러스터 범역 설정 기법을 시도할 필요 역시 존재한다. 최근 제안된 인공신경망(artificial neural network)을 이용한 기법(Moreira *et al.*, 2007)도 하나의 대안이 될 수 있을 것이다.

참고문헌

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 tool to assess local instability in spatial association, in Fisher, M., Scholter, H., and Unwin, D. (eds.), *Spatial Analytical Perspectives on GIS*, Taylor & Francis, London, 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*, John Wiley & Sons, Chichester, West Sussex, 77-94.

Anselin, L., 2003, *GeoDa 0.9 User's Guide*, Spatial Analysis Laboratory, Department of Agricultural and Consumer Economics, University of Illinois.

Anselin, L. and Bao, S., 1997, Exploratory spatial data analysis linking SpaceStat and ArcView, in Fischer, M. and Getis, G. (eds.), *Recent Development in Spatial Analysis*, Springer-Verlag, Berlin, 35-59.

Balk, D. L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S. I., and Nelson, A., 2006, Determining global population distribution: Methods, applications and data, *Advances in Parasitology*, 62, 119-157.

Boots, B., 2001, Using local statistics for boundary characterization, *GeoJournal*, 53(4), 339-345.

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

Brunsdon, C., 1998, Exploratory spatial data analysis and local indicators of spatial association with XLISP-STAT, *Journal of the Royal Statistical Society Series D: The Staistician*, 47(3), 471-484.

Dykes, J., 1998, Cartographic visualization: Exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv, *Journal of the Royal Statistical Society Series D: The Staistician*, 47(3), 485-497.

Fortin, M.-J. and Dale, M., 2005, *Spatial Analysis: A Guide for Ecologists*, Cambridge University Press, Cambridge.

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*, GeoInformation International, Cambridge, 261-277.

Goodchild, M. F. and Lam, N. S.-N., 1980, Areal interpolation: A variant of the traditional spatial problem, *Geoprocessing*, 1, 297-312.

Jacquez, G. M., Maruca, S., and Fortin, M.-J., 2000, From fields to objects: A review of geographic boundary analysis, *Journal of Geographical Systems*, 2(3), 221-241.

Kulldorff, M., 1997, A spatial scan statistic, *Communications in Statistics: Theory and Methods*, 26(6), 1487-1496.

Kulldorff, M., 2009, *SaTScan User Guide* (version 8.0), Available at http://www.satscan.org/.

Lawson, A. B. and Kleinman, K., 2005, *Spatial and Syndromic Surveillance for Public Health*, John Wiley & Sons, Chichester, West Sussex.

Lee, S.-I., 2001, Developing a bivariate spatial association measurer: An integration of Pearson's *r* and Moran's *I*, *Journal of Geographical Systems*, 3(4), 369-385.

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., 2009, A generalized randomization approach to local measures of spatial association, *Geographical Analysis*, 41(2), 221-248.

Lee, S.-I. and Kim, K., 2007, Representing the population density distribution of Seoul using

dasymetric mapping techniques in a GIS environment, *Journal of the Korean Cartographic Association*, 7(2), 53-67 (in Korean).

Lee, S.-I., Shin, J., Kim, H.-M., Hong, I., Kim, K., Chun, Y., Cho, D., Kim, J.-G., and Lee, G. (translation), 2009, *Geographic Information Systems and Science*, 2nd Edition, Sigmapress, Seoul (이상일 · 신정엽 · 김현미 · 홍일영 · 김감영 · 전용완 · 조대헌 · 김종근 · 이건학 역, 2009, 지리정보시스템과 지리정보과학, 제2판, 시그마프레스, 서울; Longley, P. A., Goodchild, M., Maguire, D. J., and Rhind, D. W., 2005, *Geographic Information Systems and Science*, 2nd Edition, John Wiley & Sons, Chichester, West Sussex).

Legendre, P. and Legendre, L., 1998, *Numerical Ecology*, 2nd English Edition, Elsevier, New York.

Lu, H. and Carlin, B. P., 2005, Bayesian areal wombling for geographical boundary analysis, *Geographical Analysis*, 37(3), 265-285.

Moreira, G. J. P., Takahashi, R. H. C., and Duczmal, L., 2007, Delineating spatial clusters with artificial neural networks, *Advances in Disease Surveillance*, 4, 104.

Office of the Deputy Prime Minister, 2002, *Producing Boundaries and Statistics for Town Centres: London Pilot Study Summary Report*, The Stationery Office, UK.

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

Rogerson, P. and Yamada, I., 2009, *Statistical Detection and Surveillance of Geographic Clusters*, Chapman & Hall/CRC, Boca Raton, FL.

Sohn, H., 2008, Modeling spatial patterns of an overheated speculation area, *Journal of the Korean Geographical Society*, 43(1), 104-116 (in Korean).

Sohn, H. and Park, K., 2008, A spatial statistical method for exploring hotspots of house price volatility, *Journal of the Korean Geographical Society*, 43(3), 392-411 (in Korean).

Sutton, P. C., 2003, A scale-adjusted measure of "urban sprawl" using nighttime satellite imagery, *Remote Sensing of Environment*, 86, 353-369.

Tango, T., 2010, *Statistical Methods for Disease Clustering*, Springer, New York.

Thurstain-Goodwin, M. and Unwin, D. J., 2000, Defining and delimiting the central areas of towns for statistical monitoring using continuous surface representations, *Transactions in GIS*, 4(4), 305-317.

Unwin, A., 1996, Exploratory spatial analysis and local statistics, *Computational Statistics*, 11, 387-400.

Unwin, A. and Unwin, D. J., 1998, Exploratory spatial data analysis with local statistics, *Journal of the Royal Statistical Society Series D: The Statistician*, 47(3), 415-421.

Waller, L. A. and Gotway, C. A., 2004, *Applied Spatial Statistics for Public Health Data*, John Wiley & Sons, Hoboken, NJ.

Womble, W. H., 1951, Differential systematics, *Science*, 114, 315-322.

Wulder, M. and Boots, B., 1998, Local spatial autocorrelation characteristics of remotely sensed imagery assessed with the Getis statistics, *International Journal of Remote Sensing*, 19(11), 2223-2231.

교신: 이상일, 151-748, 서울특별시 관악구 관악로 599, 서울대학교 사범대학 지리교육과(이메일: si_lee@snu.ac.kr, 전화: 02-880-9028)

Correspondence: Sang-Il Lee, Department of Geography Education, College of Education, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-748, Korea (e-mail: si_lee@snu.ac.kr, phone: +82-2-880-9028)

최초투고일 2010. 7. 21
수정일 2010. 8. 19
최종접수일 2010. 8. 20