공간 연관규칙을 이용한 도시성장 확률모형의 구현

조성휘*, 박수홍**
* 인하대학교 공과대학 지리정보공학과 대학원
** 인하대학교 공과대학 지리정보공학과 조교수

Seong-Hwi Cho, Soo-Hong Park

요 약

GIS는 공간과 관련된 문제를 해결하는데 있어 좋은 도구가 되며 도시성장 예측과 같은 문제에 사용될 수 있다. 본 연구에서는 수도권 내에 위치한 수원지역을 대상으로 1960년대부터 1990년대까지의 도시성장에 관한 데이터베이스를 구축하고 도시의 물리적인 확산에 초점을 맞추어 모형의 핵심이 되는 공간 연관규칙을 추출하였다. 공간 연관규칙의 추출을 위해 GIS 공간분석 기능과 데이터마이닝 기법을 이용하였으며, 규칙을 기반으로 모형을 작성하여 도시성장을 분석 및 예측하고 UGM(Urban Growth Model)과 비교하였다.

Ⅰ. 서론

1.1 연구배경과 목적

공간과 관련된 여러 문제들 중에서 인구의 증가와 이로 인해 광역화되어 가는 도시역에 대한 분석 및 예측은 많은 관련 연구가 진행되어 왔으며 인문·사회, 공학 등 분야도 매우 다양하다. 특히 도시성장에 영향을 미치는 물리적인 요소, 즉 지형자료를 통한 연구가 진행되어 왔으나 미리 정의되어진 특정한 패턴에 맞추어 인자를 설정해 가는 방식이 대부분을 차지하고 있다. 이는 범용적이라 할 수 있으나 대상지역이 갖는 구체적인 특성을 반영하는데는 한계를 가지고 있으며 모형이 매우 복잡하다.

본 연구에서는 GIS의 데이터 저장·관리·분석 기능을 이용하여 특정한 공간에 숨어있는 사상들의 공간연관규칙을 발견하는 것을 목적으로 한다. GIS가 갖는 강력한 공간분석 기능과 공간자료와 속성자료의 동시처리 기능은 공간 연관규칙을 추출하는 좋은 도구가 될 수 있다. 이를 통해 도시성장 확률모형의 핵심이 되는 공간 연관규칙을 추출하고 모형의 구현을 통해 규칙의 검증을 수행한다. 또한 UGM 결과와의 비교·분석을 통해 효율성을 증명하고자 한다.

1.2 연구범위와 방법

본 연구에서는 수도권 내에 위치한 대도시 중 경기도 수원시(121,093,951.3 m^2, 2003년 1월 1일 기준)를 연구대상지역으로 선정하고 1960년대부터 1990년대까지 약 30년간 구축된 다

양한 지형자료를 이용하여 공간 연관규칙을 추출하고 도시성장 확률모형에 적용하였다.

공간 연관규칙의 추출은 두 가지 접근방법을 통해 수행하였다. 우선 GIS 소프트웨어가 제공하는 분석기능을 이용하여 전처리 과정을 수행하고 다시 데이터 마이닝 기법을 이용하여 최종적으로 일련의 규칙들을 추출하였다. 추출된 규칙은 모형에 적용할 수 있는 알고리즘으로 작성하였다.

도시성장 확률모형은 CA(Cellular Automata)기반의 시뮬레이터인 CAS(Cellular Automata Systems)를 사용하여 구현하였다. Cellular Automata Systems는 Cellang이라는 셀룰라 언어를 제공하여 작성된 알고리즘을 쉽게 반영할 수 있도록 한다(박수홍, 1997). 모형을 이용한 시뮬레이션 결과는 Clarke Keith의 UGM 결과와 비교·분석하였다.

II. 실험데이터 구축

연구지역에 대해 활용 가능한 다양한 축척의 지형도 및 주제도, 수치지도에 대한 조사획득 및 자료의 특성분석을 하였다. 수집된 여러 유형의 공간 데이터를 연구지역에 대해 전체 시기(약 30여년간)에 걸쳐 일관성을 유지하고 연속적인 형태의 GIS데이터베이스로 구축하였다. 주제별 공간 데이터의 입력은 종이 형태로 발간된 지형도와 주제도의 경우에 수동적인 디지타이징(digitizing) 방법과 스캐닝(scanning) 및 벡터화 방법(vectorizing)의 병행에 의해 이루어졌으며, 수치지형도와 기타 주제도는 데이터의 변환과정을 통해 수행하였다. 구축한 시공간 GIS 데이터베이스의 주요 내용은 표1과 같으며 대상 시기는 1960년, 1970년, 1980년, 1990년 등 총 네 시기이다.

표 1. 구축된 GIS 데이터베이스의 주요 내용

구분	주데이터	보조데이터	시기(시기수)
도시역	지형도	토지이용현황도	1960-1990(4)
도로	지형도	도로망도, 지세도	1960-1990(4)
배제지역	지형도(수계)	지세도(수계)	1960-1990(4)
	수도권개발제한구역도 (개발제한구역)	지형도(개발제한구역)	1970-1990(3)
경사도	수치지형도	수치데이터	1990(1)

데이터베이스 구축초기에는 지금보다 많은 항목의 데이터가 존재했으나 연구의 목적 및 방법에 맞추어 이를 통합하였다. 농림지역과 기타지역을 비도시화 지역으로 통합하였고 교통망은 도시화에 실질적인 영향을 미치는 2, 4차선 도로를 이용하고 철도와 고속도로는 제외하였다. 수계망은 폴리곤 형태만을 이용하였는데 라인 형태의 데이터는 100m×100m 크기를 갖는 셀의 범주내에 들어오는 데이터가 아니므로 제외하였다. 개발제한구역은 1970년대부터 대도시의 시가지가 무제한적으로 확산되는 것을 방지하기 위해 설치된 지역이므로 1970년대부터 데이터를

구축하였다. 경사도는 수치지형도에서 등고선 레이어만 추출한 후 TIN으로 보간하고 퍼센트형식의 DEM으로 변형하여 이용하였다. 구축된 데이터베이스의 포맷은 mdb의 형태의 데이터로 되어있는데, 이를 shape → grid → ascii 형태의 데이터로 변환하였다. ascii 포맷은 각각의 레이어를 중첩하여 공간연산을 적용하거나 다른 애플리케이션으로부터의 접근이 용이하기 때문이다.

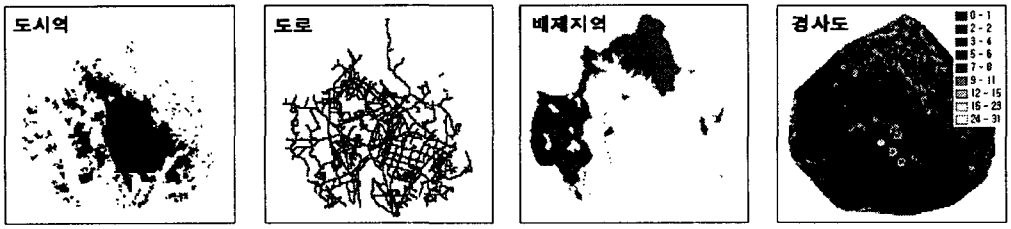

그림 1. 수원시 지형자료 - 1990년대

III. 공간 연관규칙 추출과정

3.1 규칙추출을 위한 전처리 과정

도시역의 확장과 관련된 규칙을 찾기 위해 우선적으로 네 시기의 도시역 레이어를 중첩하여 각 시기별(1960-1970, 1970-1980, 1980-1990)로 신규 생성된 도시셀들을 추출하였다. 각 시기별 신규 도시셀들은 이전 시기에서 비도시화 지역에 속하므로 이전과 당해년대의 도시역 및 도로 레이어와 중첩하여 주변지역과의 연관성을 분석하였다.

그림 2. 시기별 신규 도시셀 추출

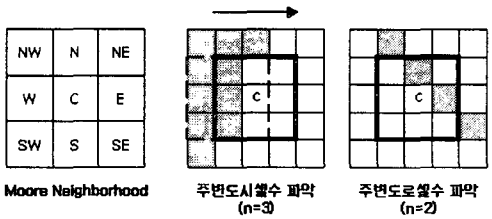

그림 3. Moving Window를 이용한 주변지역 분석

비도시화 지역은 주변에 위치한 기존의 도시역과 도로셀에 의해 영향을 받아 도시로 변환된다. 일반적으로 CA에서는 대상 셀에 영향을 미치는 주변의 셀들을 가리켜 인접(neighborhood)이라 하며, Moore의 인접 개념은 GIS에 사용되는 공간필터(spatial filter)의 마스크(mask)나 윈도우(window)와 개념적으로 매우 유사하다(박수홍, 1997). 많은 경우에 규칙적인 주변지역의 설정보다는 불규칙적이고 다양한 형태의 경계설정이 요구되나 본 연구에서는 일정한 패턴의 추출을 위해 동일한 형태와 크기의 윈도우를 적용하였다.

주변도시역과 도로에 대한 분석결과는 각각 레이어로 생성되며 배제지역 및 경사도 레이어와의 중첩을 통하여 도시로의 성장이 가능한 셀의 후보를 선정하게 된다. 또한 각 시기별로 선정된 셀들을 하나로 통합하여 대상지역의 시공간 데이터를 하나의 집합 또는 테이블의 형태로 일반화하였다.

3.2 데이터마이닝 기법을 이용한 공간연관규칙 추출

GIS의 중요한 기능 중의 하나는 새로운 패턴을 인식하는 것이다(김계현, 2000). GIS는 속성자료를 관계형 데이터베이스의 형태로 가지고 있어 데이터마이닝 기법 중 하나인 attribute oriented induction(AOI)의 적용이 가능하다. AOI는 자료의 분류방식 중 상향식 요약 방식으로 각각의 속성이 원하는 단계로 일반화될 때까지 불필요한 속성의 제거, 튜플의 통합을 반복해 나가는 방법이며 일련의 튜플들로 자료가 가진 특징 또는 규칙이 요약된다.

AOI는 일반화된 상위개념을 따라 공간 혹은 비공간 데이터를 요약할 수 있으므로 복잡한 공간관계를 갖는 도시현상을 몇 개의 규칙들로 압축할 수 있다. 자료의 압축과정에서 필연적으로 자료의 손실이 발생하게 되지만 모든 조건을 만족하는 모형의 구축이나 이를 위한 방대한 자료의 수집이 현실적으로 어려운 점을 감안할 때 상대적으로 효율적인 방법이 될 수 있다. 대상지역에 대하여 일차적으로 전처리를 과정을 수행하하여 도시성장이 가능한 후보셀, 즉 도시성장과 관련한 공간 연관규칙을 내포하는 공간자료들을 분류하였다. 분류된 각 셀에 대한 속성자료는 여러 가지 형태의 분석결과를 포함하고 있다. 본 연구에서는 물리적인 도시역의 확장에 초점을 맞추어 인접도시와 도로, 배제지역과 경사도를 입력데이터로 고려하였다. 각각의 입력데이터는 하나의 속성으로서 일반화를 위한 대상이 된다. 경사도의 경우 AOI를 적용하기에 앞서 제약조건으로서 분리해내었다. 먼저 경사도의 통계치를 분석한 결과 대부분의 값이 표준편차 3σ (약 99.74%) 범위에 해당되므로 경사도의 제약조건을 11이하로 결정하였다.

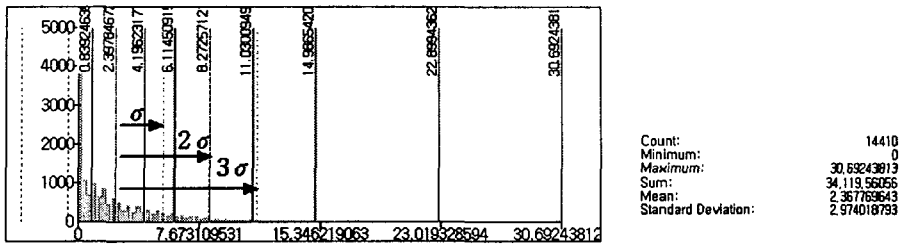

그림 4. 입력데이터가 갖는 경사도의 통계치

AOI의 적용에서 우선 일반화를 위해 대상이 되는 속성을 인접도시셀 수와 도로셀 수, 배제지역으로 정의하였다. 제거할 속성(attribute removal)으로 배제지역을 선택하여 배제지역에 속하는 튜플들을 제거하면 배제지역의 속하지 않은 튜플들이 남게 되고 배제지역 속성을 제거하게 된다. 남은 두 개의 속성, 즉 도로와 도시가 서로 상호연관성이 있다는 가정하에 주변도시셀 수의 값은 9가지(0-8), 주변도로셀 수 9가지(0-8)로 총 81가지의 조합을 만들 수 있다. 각 조합에 속하는 튜플들의 개수를 이용하여 지지도(support)와 신뢰도(confidence)를 구하였다. 무의미한 값 또는 규칙의 제거를 위해서는 사용자에 의한 임계치(threshold)의 명시가 필요하므로 우선 지지도에 대한 임계치를 0.01이상으로 설정하였다. 걸러진 조합들에 대해 다시 신뢰도에 대한 임계치 0.10을 설정하여 13개의 조합을 선정하였다. 나머지 조합들에 대해서는 도시를 기준으로 같은 행에 속하는 도로값들에 대하여 ∑(지지도*신뢰도)/∑신뢰도의 식을 적용하여 9개의 값을 구하였다(표 2). 주변도시셀과 도로의 조합, 배제지역과 경사도를 이용한 제약조건을 통합하여 최종적인 공간 연관규칙을 추출하였으며 표 3과 같은 알고리즘을 제안하였다.

표 2. 주변도시셀 수/도로셀 수 조합의 Support와 Confidence

도시/도로	0	1	2	3	4	5	6	7	8
0	0.124563	0.016171	0.018065	0.039773	0.027098	0.01049	0.004516	0.001894	0.001603
1	0.058129	0.007721	0.006702	0.021653	0.018648	0.005828	0.004225	0.002185	0.000146
2	0.049242	0.007284	0.009324	0.015443	0.01588	0.00641	0.004516	0.001311	0.000874
3	0.034528	0.005536	0.00743	0.013988	0.012238	0.007139	0.003497	0.002914	0.00102
4	0.018794	0.004662	0.005828	0.011655	0.013112	0.005828	0.00539	0.002477	0.001457
5	0.014277	0.003788	0.004371	0.009536	0.010344	0.007576	0.00539	0.002914	0.001894
6	0.009596	0.003205	0.00204	0.005539	0.007139	0.00539	0.004953	0.003205	0.002041
7	0.005393	0.002185	0.001748	0.006702	0.007284	0.005973	0.007867	0.003642	0.001311
8	0.005952	0.015734	0.016463	0.039018	0.039918	0.09278	0.028118	0.012675	0.30641

Support = 각 조합 수/전체 조합 수

도시/도로	0	1	2	3	4	5	6	7	8
0	0.510143	0.066229	0.073965	0.162888	0.110979	0.042959	0.018496	0.007757	0.006563
1	0.463415	0.061556	0.053428	0.174216	0.148664	0.046453	0.033682	0.017422	0.001161
2	0.465499	0.06885	0.088544	0.140026	0.143989	0.058124	0.042951	0.011889	0.007928
3	0.391089	0.062706	0.084153	0.158416	0.138514	0.080853	0.039604	0.033003	0.011551
4	0.271579	0.067368	0.084211	0.168421	0.189474	0.084211	0.077895	0.035789	0.021053
5	0.241379	0.064039	0.073892	0.14532	0.174977	0.128079	0.091133	0.049261	0.032032
6	0.206294	0.068923	0.043956	0.122378	0.171329	0.129371	0.118881	0.076923	0.048951
7	0.15	0.05	0.04	0.153333	0.166667	0.136667	0.18	0.083333	0.03
8	0.12358	0.072144	0.075484	0.179025	0.180033	0.150301	0.129925	0.055116	0.029392

Confidence = 각 조합 수/주변도시셀의 수

표 3. 추출된 공간 연관규칙을 이용한 알고리즘

```
If 대상셀 = 도시화가 가능한 비도시셀  then
 If 경사도 >=0 & 경사도 <= 11  then
  If not 배제지역 then

  // 임계치를 만족하는 조합들의 confidence 이용
  규칙 1 : (도시 3 / 도로 0)  then  Int(0.391089109 * 100) >= 난수 1   then  대상셀 → 도시셀
  규칙 2 : (도시 3 / 도로 3)  then  Int(0.158415842 * 100) >= 난수 2   then  대상셀 → 도시셀
                .....
  규칙 n : (도시 8 / 도로 6)  then  Int(0.128924516 * 100) >= 난수 n   then  대상셀 → 도시셀
 // 임계치를 만족하지 못하는 조합의 경우 ∑(confidence*support) / support 값을 이용
  else 규칙 :  ∑(confidence * support) / ∑support >= 난수   then  대상셀 → 도시셀
```

Ⅳ. 도시성장모형의 구현 및 분석

4.1 모형 구현

추출된 공간 연관규칙을 검증하기 위해 모델의 구현을 통해 시뮬레이션을 수행하였다. 본 연구에서는 CA기반의 시뮬레이터인 CAS를 사용하여 도시성장 확률모형을 구현하였다. CAS는 Cellang이라는 셀룰라 언어를 자체적으로 제공하고 있어 공간 연관규칙을 쉽게 모형에 적용하여 구현할 수 있다. 모형의 구현을 위해 수차례의 반복을 통하여 최적의 규칙을 찾아내는 과정인 보정단계를 반복하였다. 1960년대부터 1990년대까지 네 시기 동안의 변화를 가장 잘 반영할 수 있는 규칙을 찾기 위해 1회의 시뮬레이션을 1년으로 간주하고 난수의 영향을 최소화하기 위해 10회를 반복하여 정확도의 평균값을 사용하였다.

4.2 결과 분석 및 비교

본 연구에서는 도시성장의 여러 유형 중 물리적인 도시역 확장에 초점을 맞추었다. 이러한 유형의 대표적인 사례로 UGM을 들 수 있으며, 본 연구의 결과와 비교하였다. UGM은 모델에 대한 정보 및 프로그램 소스가 공개되어 있어 비교·연구가 가능하였다. 공간 연관규칙을 이용한 모형의 정확도를 평가하기 위해 본 연구에서는 Lee-Sallee 지수를 도입하였다. Lee-Sallee 지수는 기준시점의 도시역 이미지와 시뮬레이션한 도시역 이미지간의 일치하는 셀 수를 이용하여 계산한다. UGM 또한 모형의 보정(calibration)을 위해 Lee-Sallee 지수를 사용하고 있으며 3번에 걸쳐 보정 단계를 수행하게 된다. 일반적으로 UGM의 경우 마지막 보정의 단계에서 Lee-Sallee 지수가 가장 높은 10개의 인자(coefficient) 조합을 결정하게 된다. 이 때의 UGM에서 상위 10개의 Lee-Sallee 지수와 본 연구의 보정단계에서 나온 결과값을 비교하였다. 비교하는 값의 개수를 동일하게 하기 위해 10번의 반복을 수행하였다.

표 4. UGM과의 정확도 비교 결과

	본 연구결과	UGM 결과
Lee-Sallee 지수	0.40441 - 0.41228	0.37077 - 0.36998
지수 선택방법	1960년대를 입력데이터로 하여 1990년대를 시뮬레이션한 결과 (10회반복)	보정단계 중 final 단계에서의 상위 10개의 값을 사용

표 4는 1960년대부터 1990년대까지 두 모형간의 Lee-Sallee 지수를 비교한 결과이며 본 연구에서 수행된 결과가 UGM보다 상대적으로 높은 수치를 보였다. UGM의 경우 각 보정단계에서 인자들의 범위를 설정해주는 과정과 본 연구의 속성자료의 일반화 단계 중 임계치를 설정하는 과정에서 모두 사용자의 판단이 요구된다는 점을 감안해야 한다. 그러나 다양한 인자들의 조합을 이용하여 결과를 비교하는 UGM과 비교하였을 때 비교적 단순한 규칙을 통해 도시성장 확률모델에 적용하였다. 정확도의 결과 또한 상대적으로 높은 값을 얻을 수 있었다.

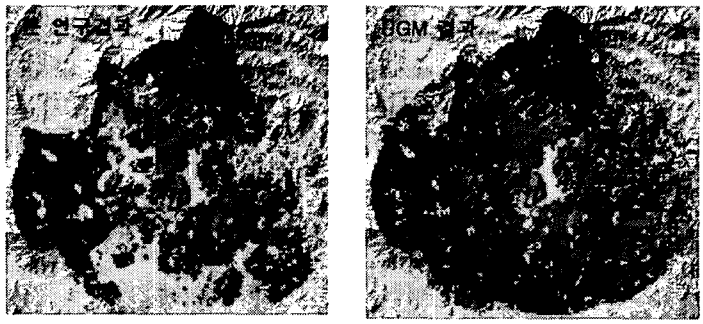

그림 7. 도시성장 분석결과 비교

V. 결론

본 연구에서는 GIS의 공간분석 기능과 데이터 마이닝 기법을 이용하여 복잡한 도시성장의 유형들을 일반화하고 이를 통해 공간 연관규칙을 추출하고자 하였다. 이를 위해 먼저 대상지역에 대해 1960년대에서 1990년대까지 10년 단위의 시공간 데이터베이스를 구축하고 연구에 적합하도록 데이터를 통합하였다. 구축된 데이터에서 공간 연관규칙을 추출하기 위하여 GIS의 공간분석 기능과 데이터마이닝 기법인 AOI를 적용하여 일련의 규칙들을 추출하였다. 추출된 규칙의 알고리즘을 이용하여 CA 시뮬레이터인 CAS으로 도시성장 분석단계를 시뮬레이션하였다.

시뮬레이션 결과는 본 연구와 물리적 도시역의 확장에 맥락을 같이 하는 UGM과 비교하였다. Lee-Sallee 지수를 이용하여 두 모형의 정확도를 비교한 결과 0.40441 - 0.41228로 UGM에 비해 상대적으로 높은 수치를 얻을 수 있었다. GIS의 데이터 처리 및 공간분석 기능을

통해 데이터에 내재되어 있는 공간 연관규칙을 찾아내고 이를 모델링에 이용하는 것은 기존의 CA모형과 비교하였을 경우 시간 및 효율적인 측면을 비교하였을 때 우수하다 할 수 있다. 본 연구에서는 도시성장과 관련된 요소들로 단순히 몇가지 지형적인 요소만을 적용하였는데, 이외의 사회·경제 등 다양한 요소들을 적용한다면 더욱 신뢰할 수 있는 규칙들을 추출할 수 있다고 본다.

반면 공간 연관규칙을 추출하는데 있어 적용된 지식추출기법은 감독분류기법의 하나로써 규칙을 정하는 단계에서 사용자의 판단이 요구되므로 주관적인 요소가 개입될 수 있다. 다른 모형들과는 달리 데이터에 내재된 규칙을 찾아야 하므로 이러한 한계점을 보완할 수 있는 방법이 필요하다. 또다른 문제점은 외부에서 모형을 보정할 수 있는 외부변수의 적용이 어렵다는 점이다. 이는 동적인 모형을 만드는 데 있어 제약사항이 되므로 향후 보완 및 연구가 필요하다고 본다.

참고문헌

강영옥·박수홍, 서울대도시지역 도시성장 예측에 관한 연구, 대한지리학회지 제35권 4호 pp.621-639, 2000

김계현, GIS 개론, 대영사, 2000

박수홍·주용진·신윤호, 도시성장 예측 모델 개발을 위한 시공간 데이터베이스의 구축, 지리학연구 제36권 4호 pp.313-326, 2002

박수홍, CA-GIS 통합 시스템을 이용한 GIS 연산의 구현, 한국GIS학회지 제5권 1호 pp.99-113, 1997

정재준, 도시권의 도시성장 분석 및 예측을 위한 셀룰라 오토마타 모델링, 서울대학교, 2001

조영아, Arc/View를 이용한 광주·전남지역 공간 연관 규칙 탐사, 전남대학교, 1999

Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 2001

W. Lu, J. Han, and B. C. Ooi, Discovery of General Knowledge in Large Spatial Databases, In Proc. Far East Workshop on Geographic Information Systems pp.275-289, Singapore, 1993

www.ncgia.ucsb.edu/projects/gig/index.html

http://www.vbi.vt.edu/~dana/ca/cellular.shtml