의사결정트리 기법을 이용한 인천시 도시성장 예측 모델링

[1]김정엽 , [1]이성규, [2]박수홍
[1]인하대학교 공과대학 지리정보공학과 대학원 · [2]인하대학교 공과대학 지리정보공학과 조교수

1. 서론

일반적으로 도시화(urbanization)란 도시지역으로 인구가 집중되어 결과적으로 국가 전체인구중 도시지역에 거주하는 인구의 비율이 증가되는 과정이라고 정의될 수 있다. 이러한 도시화는 환경문제를 비롯하여 실업, 질병 등의 여러 문제를 야기시키고 있다. 따라서, 도시화가 미치는 영향을 고려해 볼 때, 지속 가능한 도시성장을 위해 도시가 어떻게 성장해 왔는가를 분석하고, 도시성장을 예측하며, 이에 대한 적절한 대안을 모색하는 것은 도시관리를 위해 매우 중요한 문제이다(강영옥 · 박수홍, 2000). 이러한 이유로, 도시성장에 관한 분석과 예측은 향후 도시정책에 있어서 매우 중요한 자료가 된다.

본 연구는 데이터마이닝의 기법중 하나인 의사결정트리(Decision Tree)를 이용하여, 인천시가 과거에 어떠한 패턴으로 도시화가 이루어졌는지를 도시성장 규칙으로 추출하고, 이 규칙들을 이용하여 향후 30년간 인천시의 도시성장을 예측하였다. 도시성장 규칙으로 추출하기 위해 데이터마이닝 도구중에 하나인 SPSS사의 Clementine 6.5를 이용하였으며, Visulal Basic을 통해 CA(Cellular Automata) 모델을 만들었다. 대상지역은 1960년대 인천시 행정구역을 기준으로 하였으며, 시간적 범위로는 1960년대에서 1990년대까지 약 30년간을 시간적 범위로 하였다. 모델은 인천시 토지이용 상황(도시/비도시/수계), 주변 도시셀 수, 도로로부터의 거리, 경사도, 개발제한구역을 데이터를 이용하였다.

2. 도시성장 모델의 구현과 평가

1) 모델의 설계

본 연구는 과거의 시공간 데이터가 어떠한 패턴을 가지고 도시성장을 하였는지 의사결정트리 기

법을 통해 도시성장 규칙으로 추출을 하여 CA를 기반으로 도시성장 모델링을 하는 것이다. 따라서, CA의 구성요소들에 맞게 모델을 설계하였다.

먼저, 격자공간(Grid Space)은 확산의 상호과정이 일어나는 기본적인 공간을 의미한다. 이러한 공간은 여러가지 형태로 있을 수 있으나 일반적으로 2차원 정방형의 셀로 이루어진다. 본 연구에서는 기타 다른 래스터 자료와의 교환을 용이하게 하기 위해 기본적인 정방향 셀로 구성하였다. 격자의 크기는 100m×100m의 셀 사이즈로 결정하였다. 셀의 지역상태는 셀의 특성을 나타내는 것으로 각각의 셀에 부여되는 값이라고 할 수 있다. 연구에서는 토지의 이용상태(도시/비도시/수계), 도시화여부, 주변 도로로부터의 거리, 주변 도시셀 수, 경사도, 개발제한구역에 따르는 특성을 나타내는 각각의 레이어로 구성하였다. 주변 셀은 중심 셀(focus cell)주위에 있는 셀들을 말하며 변이규칙의 요인을 제공하게 된다. 주변 셀들의 범위는 잡는 방법은 여러 가지가 있으나 본 연구에서는 주위 8개 셀을 대상으로 하는 개념을 이용하였다. 변이 규칙은 주변 셀의 상태와 제약 조건 등에 의해 시간 t에서 t+1로 변화함에 따라 중심 셀이 어떻게 변화할지 규정하는 것으로 CA의 중심 요소이다. 본 연구에서는 이러한 변이 규칙을 찾아내기 위하여 의사결정트리 기법을 제공하는 SPSS사의 clementine6.5를 이용하였다.

의사결정트리란 데이터마이닝의 분류 작업에 주로 사용되는 기법으로, 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 부류별 특성을 속성의 조합으로 나타내는 분류모형을 나무의 형태로 만드는 것이다. 그리고 이렇게 만들어진 분류모형은 새로운 레코드를 분류하고 해당 부류의 값을 예측하는데 사용된다. 의사결정트리는 새로운 레코드의 해당 부류값을 예측하기 위해 이미 만들어진 분류모형(의사결정트리)이 지시하는 바에 따라 레코드의 속성값을 질문하는 작업을 반복적으로 수행한다. 특히 결정적인 질문을 던지게 되면 다른 모든 속성의 값을 묻지 않고도 레코드의 부류값을 정확히 예측할 수 있다. 따라서 레코드를 분류하고 예측할 수 있는 트리(모형)를 얼마나 잘 만드냐가 의사결정트리 기법의 핵심이다.

2) 모델의 실험

일반적으로 도시성장 모델에서는 두 가지 전제조건이 있는데, 첫째는, 수계나 개발제한구역같은 지역에서는 도시성장이 이루어지지 않는다는 것이며, 두 번째는 기존의 도시는 다시 비도시로 변하지 않는다는 조건이다. 그 중 첫 번째 조건은 도시성장이 불가능한 지역을 나타내는데, 인천시는 수계로 분류된 염전이 간척사업으로 인해 도시로 성장해 가는 특수한 경우가 있어서 개발제한구역에 의한 조건만 고려하였다. 두 번째로 과거의 도시가 다시 비도시로 변하는 경우는 현실적으로 불가능하므로 모델의 전제조건으로 두었다.

도시성장 모델링 실험은 1960년대의 데이터를 시작점으로 하고 1990년대를 끝점으로 하였다. 개발제한구역은 1970년대에는 인천시에 영향을 미치지 않았으나 1980년대부터 확대되면서 인천시까

지 영향을 미쳤기 때문에, 20번의 시뮬레이션 사이클이 지난 후에 21번째부터 제약조건으로 모델에 적용을 하였다. 향후 예측은 1990년대를 기준시점으로 하여 2020년까지 30년간을 예측하였다. 예측은 10번의 반복을 통해 모델이 각 시뮬레이션을 할 때마다 유사한 도시성장을 하는지 분석하였다. 그 이유로는, 도시성장을 하면서 우연적인 요소를 가미하였기 때문에 매 시뮬레이션마다 유사한 패턴의 도시성장을 하는지 혹은 전혀 새로운 패턴으로 도시성장을 하는지를 알아보기 위해서이다. 10번의 시뮬레이션을 반복한 결과 랜덤변수에 의한 약간의 차이는 있지만 전체적인 도시성장 패턴은 큰 변화가 없는 것으로 나타났다. 10번의 반복을 통해 항상 도시가 되는 지역은 짙은 빨간색으로 나타났고, 매번 도시로 성장하지 않는 셀은 옅은 빨간색으로 나타났다. 모델을 통해 나온 결과물인 도시지역 레이어는 각 년대마다 실제 도시지역 레이어와 일치도를 검증하여 얼마나 현실을 반영하는 모델인가를 판단할 수 있어야 한다. 따라서 아래와 같은 통계치를 이용하여 일치도를 구하였다. 여기서, α는 시뮬레이션 결과와 실제 도시를 단순히 비교한 것으로 과성장($\alpha>1$)을 하는지 저성장($\alpha<1$)을 하는지를 판단할 수 있다. β는 시뮬레이션 결과와 실제 도시와의 일치율이며, 모델링을 수행한 결과가 위치적으로 얼마나 정확하게 일치하는지를 나타낸 것이다. Lee-Sallee 지수는 시뮬레이션 결과와 실제 도시간의 공간적 모습이 얼마나 일치하는가를 나타낸 것이다

$$\alpha = \frac{\text{시뮬레이션을 통해 나온 도시셀 수}}{\text{실제 도시셀 수}}$$

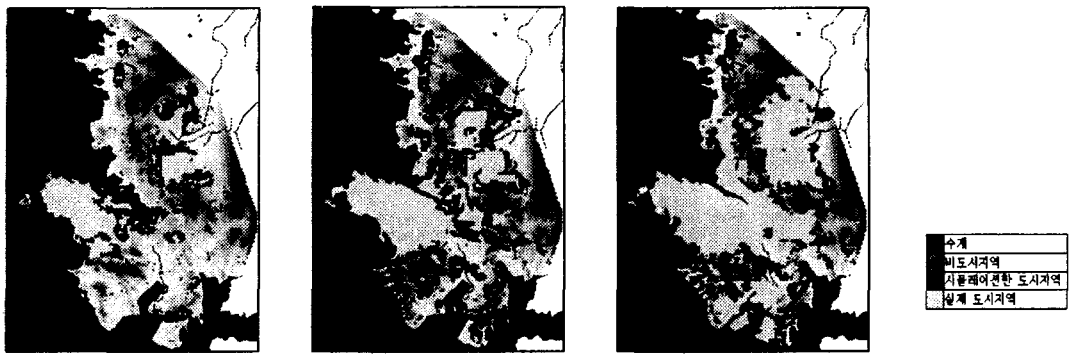
$$\beta = \frac{\text{실제 도시셀과 일치하는 시뮬레이션 결과 도시셀의 수}}{\text{실제 도시셀 수}}$$

$$\text{Lee-Sallee 지수} = \frac{\text{실제 도시셀과 일치하는 시뮬레이션 도시셀 수}}{(\text{실제 도시셀} \cup \text{시뮬레이션 도시셀})\text{의 수}}$$

3) 모델의 평가 및 결과 분석

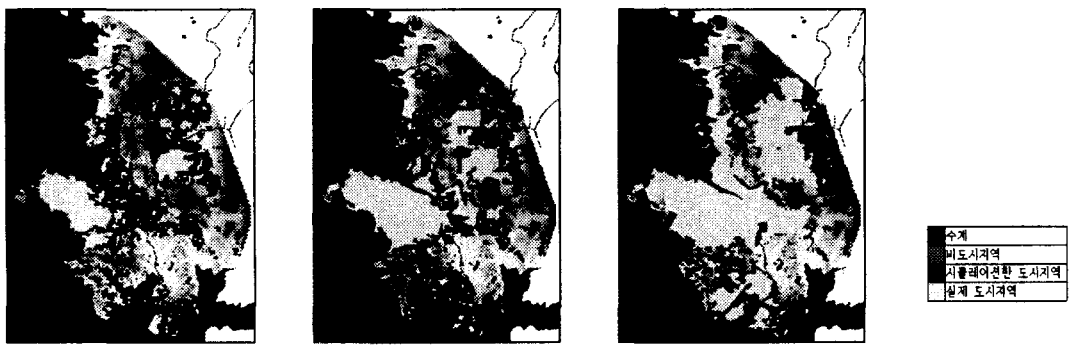
본 연구의 도시성장 모델을 통해 나온 결과를 보면 시기별로 특징을 나타내고 있다. (그림 1) 1970년대까지의 도시성장을 보면, 부평구와 계양구가 실제보다 과성장을 하고 있음을 알 수 있다. 1980년대까지의 도시성장 또한 많은 과성장을 했는데, 이는 남동구와 서구, 계양구가 실제보다 더 도시로 변하였기 때문이다. 1980년대에는 인천의 서쪽 해안지역이 간척사업으로 인해 도시로 많이 변하였는데, 시뮬레이션에서도 이러한 모습을 잘 나타내었다. 1990년대까지의 도시성장은 모델링의 최종 목표 시점이 되는 시기로 α, Lee-Sallee 지수는 이전 시기보다 좋은 결과를 나타내었고, β는 약간 낮은 수치를 나타내지만 85%라는 좋은 결과가 나왔다. 정확도는 도시성장 모델의 대표격인 UGM 방법과 비교해 보았다. UGM은 과거의 데이터로부터 현재까지 도시지역의 성장을 규칙화하고, 미래의 시점까지 도시지역이 어떻게 확산이 되는가를 예측하는 모델로 미국 USGS의 Project Gigalopolis에서 개발한 모델이며, 본 연구 모델과 비교할 수 있는 좋은 모델이다. 결과를

보면 수치적으로는 큰 차이가 없으나 영상을 보면, UGM 모델이 과성장을 심하게 하고 간척사업을 표현하지 못하고 있음을 알 수 있다.



1970년대 1980년대 1990년대

그림 1. 본 연구의 시기별 시뮬레이션 결과와 실제 도시



1970년대 1980년대 1990년대

그림 2. UGM의 시기별 시뮬레이션 결과와 실제 도시

표 1. 시뮬레이션의 정확도와 UGM과의 비교

시기 / 검사항목	1970	1980	1990	평균
α	4630/3324 = 1.39	8845/5002 = 1.77	10491/8391 = 1.25	1.47
β	2465/3324 = 0.74	4406/5002 = 0.88	7136/8391 = 0.85	0.82
Lee-Sallee 지수	2465/5489 = 0.45	4406/9441 = 0.46	7136/11746 = 0.61	0.51
UGM	coarse단계	fine단계	final단계	
Lee-Sallee 지수	50	50	50	50

향후 예측 부분은 개발제한구역을 설정하고 해제하는 두 가지 시나리오로 진행하였는데, 그 이유는 이미 1990년대에 도시화가 거의 포화상태이기 때문이다. 개발제한구역을 설정하였을 경우에는 사실상 더 이상 도시성장이 거의 이루어지지 않는 반면에 해제를 했을 경우에는 도시성장이 2010년까지는 활발히 이루어지는 것을 알 수 있었다. 2010년 이후로 도시성장이 거의 이루어지지 않는다는 것은 현재와 다른 도시성장 개념을 적용해야 할 것으로 보인다. 즉, 현재는 양적인(횡적인) 팽창을 이루고 있지만, 포화상태를 이룰 경우 종적인(질적인) 팽창으로서 도시성장의 개념을 바꿔야 할 것으로 보인다.

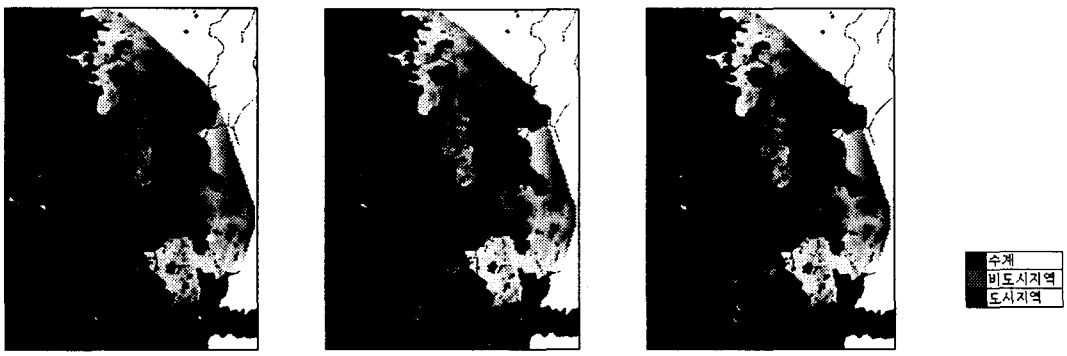
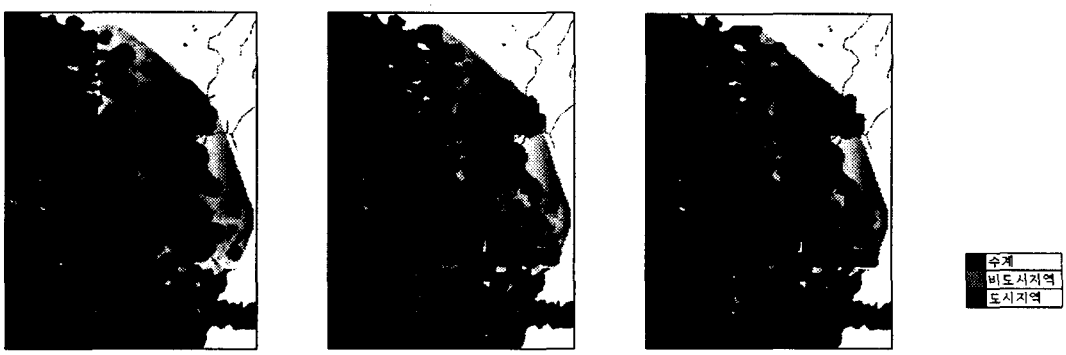


그림 3. 그린벨트 적용시 향후 30년 예측



2000년대 2010년대 2020년대

그림 4. 그린벨트 해제시 향후 30년 예측

3. 결론

도시성장 예측은 도시정책 및 관리에 있어서 매우 중요하다. 본 연구에서는 1960년대에서 1990년대까지 지난 30년간 인천지역의 시공간 데이터를 이용하여 도시성장 모델을 개발한 후 개발제한구역의 설정과 해제라는 두 가지 시나리오로 인천시 도시성장을 예측하여 보았다. 인천시 도시성장 모델은 도시성장 모델의 대표격인 UGM과 비교해 보았을 때, 더 나은 결과를 나타내었다.

도시성장은 한두 가지 요인이 관여하는 것이 아니라 자연적 요인, 인문·사회적 요인, 정책적 요인 등 다양한 요인에 의해 이루어지는 것이다. 하지만 본 연구의 도시성장 모델은 물리적인 요소만 이용하여 인문·사회적인 요소들에 대해서는 실험을 하지 않았다. 특히 우리 나라는 정책적인 요소가 도시성장에 많은 영향을 미치고 있으나 이러한 점을 반영하지 못하였다. 따라서, 본 연구의 결과를 그대로 받아들이기보다는 다른 여러 가지 요소들을 적용하여 좀 더 나은 결과를 얻는 것이 좀 더 정확한 결과를 나타낼 것으로 판단된다.

■참고문헌

Jin Chen, peng Gong, Chunyang He, Wei Luo, Masayuki Tamura, and Peijun Sho, 2002, Assessment of the Urban Development Plan of Beijing by Using a CA-Based Urban Growth Model, Photogrammetric Engineering & Remote Sensing, 2002, Vol. 68, NO. 10, 1063-1071

Jiawei Han · Micheline Kamber, 2001, Data Mining Concepts and Techniques, MORGAN KAUFMANN PUBLISHERS

최창영, 2001, "셀룰러 오토메타를 이용한 도시 성장 모델링", 경상대학교 대학원, 석사학위 논문

강영옥, 박수홍, 2000, "서울대도시지역 도시성장 예측에 관한 연구", 대한지리학회지 제35권 제4호, 621-639

정재준, 이창무, 김용일, 2002, "도시성장 분석 및 예측을 위한 셀룰라 오토마타", 대한국토·도시계획학회지「국토계획」제 37권, 1호