환경영향평가 제20권 제5호(2011) pp.663~673

연구논문

의사결정나무 변수 선정 방법을 적용한 대축적 생물다양성 지도 구축

김도연 · 허준 · 김창재
연세대학교 사회환경시스템공학부
(2010년 7월 14일 접수, 2011년 9월 3일 승인)

Mapping Biodiversity throughoptimized selection of input variables in decision tree models

Kim, Do Yeon · Heo, Joon · Kim, Chang Jae
School of Civil and Environmental Engineering, Yonsei University
(Manuscript received 14 July 2010; accepted 3 September 2011)

Abstract

In the face of accelerating biodiversity loss and its significance in our coexistence with nature, biodiversity is becoming more crucial in sustainable development perspective. To estimate biodiversity in the future which provides valuable information for decision making system especially in the national level, a quantitative approach must be studied forehand as a baseline of the present status. In this study, we developed a large-scale map of Plant Species Richness (PSR, typical indicator of biodiversity) for Young-dong and Pyung-chang provinces. Due to the accessibility of appropriate data and advance of modelling techniques, reduction of variables without deteriorating the predictive power is considered by applying Genetic algorithm. In addition, a number of Correctly Classified Instances (CCI) with 10-fold cross validation which indicates the predictive power, was carried out for evaluation. This study, as a fundamental baseline, will be beneficial in future land work as well as ecosystem restoration business or other relevant decision making agenda.

Keywords : Biodiversity, Plant Species Richness, Decision Tree Algorithm, Genetic Algorithm, Spacial distribution model

Corresponding Author: Joon Heo, School of Civil and Environmental Engineering, Yonsei University, Shinchon-dong 134, Seodaemoon-gu, Seoul, Korea, 120-749 Tel: 82-2-2123-2809 E-mail: jheo@yonsei.ac.kr

1. 서 론

기후변화와 도시화로 인해 생태계 파괴가 빠른 속도로 진행됨에 따라, 생물다양성 또한 감소하는 추세이고 최근에는 그 상황이 점진적으로 악화되고 있다. 생물다양성의 유지는 자연과 공존하는 인간의 삶과 지속가능한 발전을 위한 필수적인 요소이다. 이에 생물다양성 보존은 지구온난화에 이어지는 국제적인 환경이슈로 부상하고 있고, 특히 UN IPCC (Intergovernmental Panel on Climate Change)에 해당하는 UN IPBES (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Service) 구성은 향후 생물다양성 보존을 위한 국제적 규제의 틀을 준비하는 움직임으로 파악할 수 있다. 또한 유럽의 국가들이 주축이 된 TEEB (The Economics of Ecosystem and Biodiversity)는 생물다양성의 보존, 증진, 손실에 대한 가치평가 프레임워크를 제안하고 있으며, 이를 통해 생물다양성에 영향을 미치는 산업 생산 활동에 대한 규제의 틀을 구축할 것으로 예상된다. 이와 같은 여러 국제적 기구들의 움직임에 대응하고 향후 우리나라 지역의 생물다양성 변화를 감지하여 국가적 차원의 생물다양성 기반 의사결정체계에 구축을 가능하게 위해서 무엇보다 정량적인 접근방법이 요구된다. 이에 생물다양성을 나타낼 수 있는 공간 분포 모델을 구현하고 이를 통한 현황 지도를 제작하면 유용한 기초 자료(baseline)로써의 활용이 가능해질 것으로 사료된다.

자연의 보존 및 생물학적인 모니터링과 밀접한 종다양성의 예측 모델의 중요성은 이미 널리 알려져 있다 (Jongman *et al.*, 1995; Fielding and Bell, 1997). 생물다양성이 생태적 지위를 판단하기 위한 척도로 사용되면서 이를 통한 모니터링 목적의 데이터 취득이 가능해졌고 나아가 예측 모델을 제작하기 위한 여러 가지 접근 방법들 또한 제안되었다 (Nogue, 2009; JHA, 2005; Phillips and Dudik, 2008). 그 중에서도 의사결정나무 알고리즘은 크고 복잡한 생태자료를 다루는 생태 예측 모델과 생물종 공간 분포 모델에 활용되고 있다 (Michaelsen *et al.*, 1987; Meentemeyer *et al.*, 2001; Flesch & Hahn, 2005; Garzon *et al.*, 2008). 의사결정나무 알고리즘은 machine learjning 방식 중 한가지로 모델제작에 참여하는 변수들을 이용해서 예측된 의사결정규칙을 도표화시키는 특징을 지니는 알고리즘을 말한다(Witten and Frank, 2000). 최적의 모델 제작 및 변수 선정 작업은 여러 가지 시각에서 고려되어야 하는 복잡한 문제이다 (D' heygere et. al., 2003). 또한 모델 제작의 대상지역에 해당되는 실측 데이터가 현실적으로 부족하기 때문에 연구진행에 있어 많은 제약이 존재한다. 하지만 우리나라의 경우, 세계적으로 드문 1:5,000 수준의 국가생태조사와 전국의 4000점을 표본점으로 갖는 식생종풍부도 (National Forest Inventory, NFI)가 구축되어 있다. 이미 생물다양성 관련 여러 주요 연구들에서 발표된 바와 같이 식생종풍부도는 생물종다양성의 대표지표로 활용되고 있고, 양서류 및 포유류 등의 척추동물과 아주 높은 양의 상관관계를 나타낸다 (Myers et. al., 2000; Kier *et al.*, 2009; Currie, 1991; Andrews and O' brien, 2000; Boone and Krohn, 2000). 따라서 식생종풍부도를 생물다양성 대표지표로써 사용하게 되면 여타 생물다양성 지표와의 상관관계를 고려할 때는 물론 생물다양성의 현황을 나타내는 대축적지도 제작에도 유용하다.

본 연구에서는 평창군과 영동군 두 지역에 대해 취득한 여러 가지 생태적, 지형적, 환경적 변수들의 데이터를 이용하여, 의사결정나무를 통한 식생출현 종수의 공간 분포 모델을 제작하였다. 또한 유전자 알고리즘을 사용하여, 기 제작된 공간 분포 모델의 정확도를 유지시키는 최소한의 변수를 선정하였다. 최종적으로 연구 대상지역의 생물다양성 대축적지도를 구축하고 정확도 평가를 실시하였다.

2. 대상 지역 및 입력 자료

본 연구에서 선정한 연구 대상지는 강원도 평창군지역과 충청북도 영동군지역이다. 평창군의 경우

태백산맥 중에 위치하며 해발고도가 700m이상인 곳이 전체 면적의 약 60%를 차지하고 있다. 특히 임야의 면적 또한 전체 평창군 면적의 84%로 산림의 비중이 상당히 높은 지역이다. 영동군 역시 전체 면적 중 임야 지역이 약 77.8%로, 산림의 우세한 지역이다.

본 연구에서는 생물다양성 지도를 작성하기 위한 공간 분포 모델의 생성과 정확도 평가를 위해서, 산림청 산하의 산림조사부 자료인 NFI를 바탕으로 식생의 출현종수를 획득하였다. NFI는 전국 산림의 실태를 정확히 파악하는 목적으로 구축되며, 계통추출법에 따라 4km 간격으로 전국에 배치된 4000개의 표본점들 중, 산림이 위치한 점에서 산림자원 정보를 수집한다 (그림 1). 그림 2는 하나의 플롯으로 구성된 4개의 표본점으로, 식생 출현종수의 조사는 중앙 표본점에서 실시한다. 환경부에서 제공받은 생태자연도로부터 산림의 임상, 영급, 경급, 밀도, 식생보존등급, 생태, 녹지조성등급별 변수들을 취득하였고 산림청의 산림입지도를 이용하여 1:25,000의 토양습윤도를 취득하였다. 또한 생물다양성에 영향을 미치는 것으로 알려진 기후변화의 요소들과 인위적인 요인을 독립변수로 고려하였다 (Echeverria, 2008; Beaumont, 2009). 기후 변수는 Museum of Vertebrate Zoology, University of California, Berkeley에서 구축한 WorldClim 자료를 사용하였다 (Hijmans *et al.*, 2005). Worldclim 자료는 공간 모델링과 지도 제작 연구에 활용할 수 있도록 전 세계 곳곳의 기상관측소로부터 취득한 월평균 기후데이터를 바탕으로 보간한 격자형 자료이다. Worldclim의 총 19가지의 생물기후적 자료 중 연평균 강수량, 가장 추웠던 분기의 평균온도, 가장 더웠던 달의 최고온도, 평균 낮의 길이, 가장 건조했던 달의 강수량, 강수의 계절적 주기성을 택하였다 (Beaumont, 2009). 평균적인 수치들 이외에 극한값도 선정하였는데, 이는 평균치에 비해 극한값이 토양의 수분함유량과 식생의 생산성에 중요한 영향을 주기 때문이다 (Knapp, 2002). 이 밖에 인위적인 요인으로는 통계청의 행

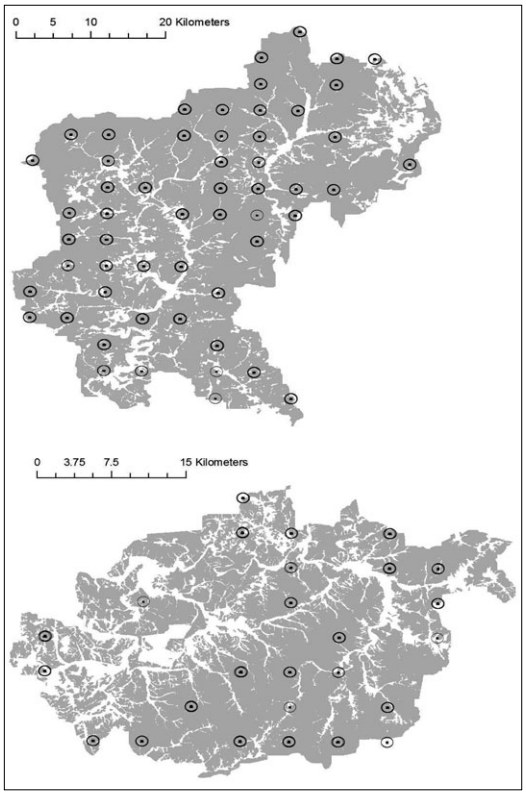

그림 1. 대상지역의 NFI 자료

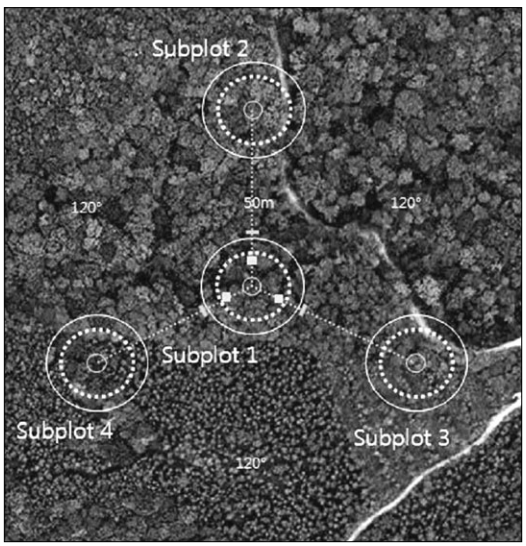

그림 2. NFI 자료 조사점

정구역별 2009년 인구조사 자료를 바탕으로 대상지역의 인구밀도 자료를 구축하였다. 그리하여 총 15가지의 독립변수(임상, 영급, 경급, 밀도, 식생보

표 1. 영동군 지역의 연구에 이용된 데이터

	최저치(등급)	최고치(등급)
토양습윤도	1	4
임상	0	10
생태자연도	1	3
녹지자연도	0	8
밀도	0	3
영급	0	6
경급	0	3
식생보존등급	0	5
인구밀도	17.20	56.32
AP(Annual precipitation)	1284	1721
MTCQ(Mean Temperature of Coldest Quarter)	-19	-101
MTHM(Max Temperature of hottest Month)	202	285
MDR(Mean Diurnal Range)	98	117
PDM(Precipitation of driest month)	29	43
PS(Precipitation seasonality)	73	83
출현종수	19	69

표 2. 평창군 지역의 연구에 이용된 데이터

	최저치(등급)	최고치(등급)
토양습윤도	1	4
임상	0	10
생태자연도	1	3
녹지자연도	0	8
밀도	0	3
영급	0	6
경급	0	3
식생보존등급	0	5
인구밀도	16.74	208.16
AP(Annual precipitation)	1160	1579
MTCQ(Mean Temperature of Coldest Quarter)	-60	-5
MTHM(Max Temperature of hottest Month)	242	300
MDR(Mean Diurnal Range)	104	116
PDM(Precipitation of driest month)	25	34
PS(Precipitation seasonality)	78	81
출현종수	8	52

존등급, 생태, 녹지조성등급, 토양습윤도, 6가지 기후변수 및 인구밀도)와 종속변수(식생 출현종수)로 정의한 후, 공간 분포 모델링 작업을 시행하였다.

표 1-2는 각각 본 연구에서 사용한 영동군, 평창군 지역에 해당하는 변수들로, 그 표기 및 범위를 나타내었다. 환경부의 생태자연도에서 취득한 각 변수들과 토양습윤도의 등급에서, 실제 조사의 진행이 불가하여 데이터가 없는 경우에 한하여 '0' 등급으로 구분하였다.

3. 예측모델의 구축

1) 의사결정나무 및 종속변수의 분류

의사결정나무는 계량적 분석방법 중 하나로 예측모델 제작 시 유용하다. 이는 모델제작에 참여하는 이산형으로 나눠진 종속변수와 이산형 혹은 연속형 독립변수를 이용해서 예측된 의사결정규칙을, 사용자들이 보다 쉽게 이해할 수 있게 도표화하는 특징을 가진 machine learning 방식 중 한가지이다 (Witten and Frank, 2000).

본 연구에서는 식생 출현종수를 종속변수로 설정하고 그 외의 생태정보, 6가지 기후변수 및 인구밀도를 독립변수로 지정하였다. 종속변수는 앞서 설명한대로 이산형의 데이터 형태로 입력이 되어야 하고 또한 의사결정나무 분석의 종료후 생물다양성을 가시화하기 위하여 내츄럴 브레이크 (Natural Break) 방법에 의해 각각의 클래스로 3분류, 4분류, 5분류의 3가지 형태로 출현종수를 구분하였다. 내츄럴 브레이크 분류 방법은 많은 값들로 이루어진 데이터 집합을 분류하는 경우 가장 최적하게 서로 다른 클래스로 층화시키는 방법 중 한가지이다. 즉, 이 방법은 같은 클래스에서의 편차는 최소화 시키면서 서로 다른 클래스들 간의 편차는 최대화 시키는 방법이다. 네츄럴브레이크 방식은 식 (1)에서와 같이 임의의 데이터셋 A의 값과 i부터 j까지에 해당하는 값들의 평균값간의 차이를 제곱한 수들의 총합을 나타내는 식으로 이 값을 최소화시키는 경계(k)를 찾아내는 형태로 데이터를 분류한다 (Jenks, 1967).

$$SSD_{i\ldots j} = \sum_{k=i}^{j} (A[k] - \text{mean}_{i\ldots j})^2 \qquad (1)$$

표 3. 구분된(Natural Break) 평창군 출현종수

분류 수	클래스	출현종수
3 분류	class1	19~29
	class2	29~42
	class3	42~69
4 분류	class1	19~26
	class2	26~38
	class3	38~54
	class4	54~69
5 분류	class1	19~26
	class2	26~32
	class3	32~40
	class4	40~54
	class5	54~69

표 4. 구분된(Natural Break) 영동군 출현종수

분류 수	클래스	출현종수
3 분류	class1	8~22
	class2	22~35
	class3	35~52
4 분류	class1	8~19
	class2	19~27
	class3	27~35
	class4	35~52
5 분류	class1	8~13
	class2	13~22
	class3	22~32
	class4	32~44
	class5	44~52

표 3과 4는 평창군과 영동군 자료를 내츄럴 브레이크를 활용하여 3개, 4개, 5개의 클래스로 나눈 출현종수를 구분한 결과이다.

본 연구에 사용된 의사결정나무는 입력한 데이터를 유사한 특성을 갖는 클래스로 분할하는 방법이다. 의사결정나무의 구성은 노드로 되어있고, 뿌리 노드로부터 잎 노드까지 하나의 데이터가 어떤 클래스에 속하는지 잎 노드까지 조사하는 작업이 진행된다. 조사 작업은 처음 입력된 데이터 집합을 효율적으로 나누는 분리 기준과 잘못된 분류를 제거하는 가지치기를 수반하고, 최하위 노드인 잎 노드가 오직 하나의 변수로 표현되면 정지한다(Franklin, 2009). 본 연구에서는 의사결정나무 알고리즘인 C4.5가 구현된 WEKA 소프트웨어를 사용하였다(Quinlan, 1993). 공간 분포 모델의 검정은 10-fold validation방법을 사용하여 정분류율로 평가하였다.

2) 유전자 알고리즘을 통한 변수 선정기법

이들 복잡하고 많은 변수들 사이에서 식생 공간분포 모델과 관련된 최적의 변수를 찾아낼 수 있다면 작업 절차의 효율성을 극대화 시킬 수 있다. 따라서 최적의 경제적인 공간 분포 모델을 제작하기 위한, 변수 선정 작업으로 유전자 알고리즘이 사용되었다(Goldberg, 1989). 본 연구에서 사용한 유전자 알고리즘은 단계별로 크게 재생산, 교배, 돌연변이 과정으로 나눌 수 있다. 이 단계별로 존재하는 연산자들은 다윈의 자연선택 (natural selection)의 원칙을 그 모델로 삼고 있다. 즉, 우수 형질의 개체는 다음 세대에 더 많이 발현시킬 수 있는 기회를 주고 열등 형질의 개체는 발현이 어렵게 하는 원리를 갖는다.

이와 같은 유전자 알고리즘을 활용하면 모든 변수 집합을 trial and error 방식으로 검정해야 하는 과정이 생략되기 때문에 경제적이다 (Tom D' heygere, 2003). 유전자 알고리즘의 초기값으로 최대 세대수를 20, 돌연변이와 교배확률을 각각 0.033 과 0.6으로 지정하였다(Witten and Frank, 2000). 최적의 식생 공간 분포 모델 제작의 단계적인 수행 절차는 다음과 같다.

1 단계: 연구에 사용된 모든 변수들로 입력 데이터 집합을 구성

2 단계: 의사결정나무 알고리즘을 통한 모델 제작

3 단계: 모델 제작에 사용된 변수들에 유전자 알고리즘을 적용

4 단계: 생존한 변수로 입력 데이터 집합을 재구성

5 단계: 제작된 모델에서 생존한 변수들의 출현빈도가 50%미만(특정 변수의 출현 빈도가 50%를 넘지 못했을 때, 모델 생성에 영향이 작다고 판단함) 일 때까지 2-4 단계 반복 수행

이와 같은 방법을 통하여 선정되 최적의 변수들을 이용하여 평창군과 영동군 각각의 지역을 대상으로 생물다양성 지도를 구축하였다.

4. 실험결과 및 분석

1) 변수의 중요도

그림 3은 앞서 언급한 단계적인 수행 절차를 3 단계까지 실행한 결과로, 모든 내츄럴브레이크 분류 방식에 대해 유전자 알고리즘을 적용하였을 때, 생존한 독립변수들의 빈도이다. 평창군의 경우 연평균 강수량, 가장 더웠던 달의 최고온도와 가장 추웠던 분기의 평균 온도 및 강수의 계절적 주기성이 가장 많이 생존하였다. 반면 식생의 경급과 임상지수를 제외한 나머지 생태적인 변수들은 그에 비해 모델 트레이닝과 검정과정에서 비교적 선택 빈도가 낮게 나타났다. 영동군을 대상 지역으로 한 데이터는 앞선 평창군과 유사한 결과를 보였다. 생태적인 변수들에 비해 가장 추웠던 분기의 평균온도와 가장 더웠던 달의 최고온도가 모델 생성에서 가장 많은 출현 빈도를 보였다. 또한 인구밀도항목의 경우 역시 생태적인 변수들 보다 높은 빈도수를 갖는 것으로 나타났다.

두 대상 지역을 비교했을 때, 대체로 기후 변수가 중요하게 작용하였고 특히 기후 변수의 극한값인 가장 추웠던 분기의 평균 온도는 유일하게 두 지역에서 공통으로 우세한 빈도수를 나타내었다. 이는 식생 공간 분포를 확인하는데 기후 인자의 극한값이 중요한 역할을 함을 의미한다. 식생을 보유하는 토양의 수분함유량과 식생의 생산성간의 상호작용은 기후 변수의 극한값에 민감하게 반응하기 때문에 이와 같은 결과를 보였다고 판단된다 (Knapp, 2002). 인구 밀도의 경우, 평창군과는 달리 영동군 지역에서 높은 빈도수를 보였다. 영동군이 평창군

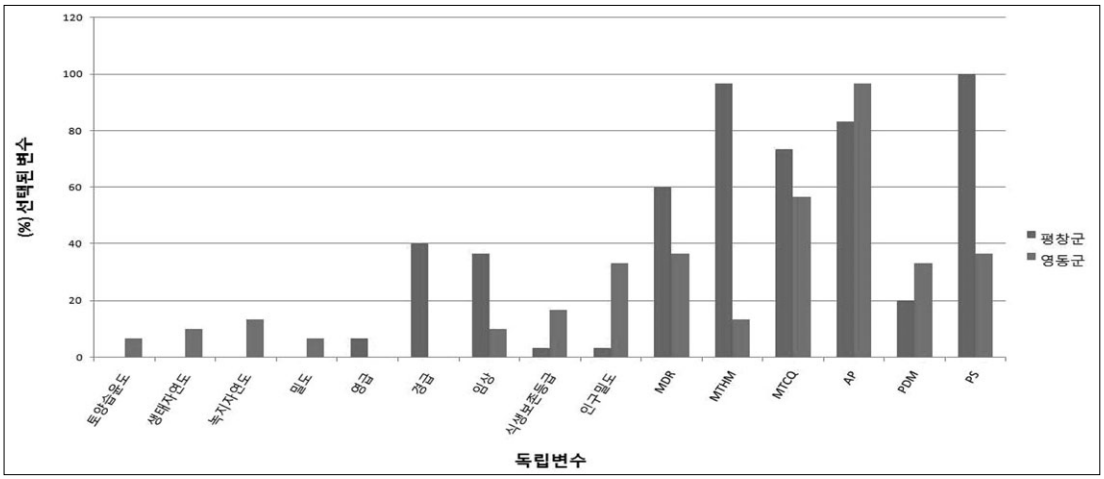

그림 3. 연구대상지역에 적용된 변수들의 유의성

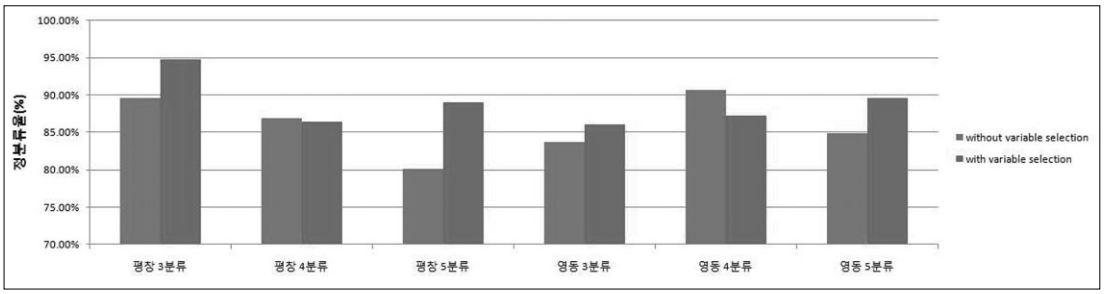

그림 4. 변수선택 전 · 후에 따른 정분류율

보다 많은 행정구역으로 나눠져 있으므로 입력되는 데이터 양에서 기인하는 요인으로 사료된다.

반면 두 지역 모두 산림의 밀도, 영급 등의 생태적인 변수들은 상대적으로 비생태적인 변수들에 비해 중요도가 떨어졌다. 변수 선정 과정에서 나타나는 것과 같이 결국 자연적, 지형적인 특색과 더불어 여러 가지 기후 인자들이나 인간의 인위적인 영향 또한 생물다양성에 큰 영향을 미치는 것이라 판단된다.

2) 변수 선정 작업 및 정확도 분석

변수 선정 작업을 통한 정확도 평가의 결과는 그림 4와 같다. 그림 4는 최적의 변수 선정 작업 전과 후의 정분류율의 차이를 그래프로 나타낸 것이다. 평창군 지역의 출현종수를 내츄럴브레이크로 3분류하여 의사결정나무에 적용한 결과, 초기의 15가지의 독립변수 사용시 정분류율은 89%로 확인되었다. 앞서 언급한 바와 같이 유전자 알고리즘을 사용하여 각 변수가 모델생성에 미치는 기여도를 확인하여 가장 낮은 변수를 우선적으로, 순차적인 변수 제거를 반복 시행하였을 때의 결과는 다음과 같다. 변수 제거의 1회 시행결과 총 15개의 변수 중, 9가지의 변수가 제거되었고 나머지 6가지의 변수가 선택되어 예측 모델을 생성하였고, 그 이후 2회의 추가적인 변수 제거를 한 결과 최종적으로 4개의 변수로 예측 모델을 생성하는 결과를 보였다. 초기 모델의 89% 정분류율에서, 변수가 제거됨에 따라 91%로 정확도가 향상되었고 최종적으로 생존한 4가지 변수들의 공간 분포 모델의 경우 94%의 정확도를 기록하였다. 영동군의 경우도 역시 변수 선정에 따른 정분류율이 비슷한 양상을 띄었다. 식생의 출현종수를 내츄럴브레이크로 3분류 하였을 때, 변수 제거 전의 15가지의 독립변수로 이루어진 모델의 정분류율은 83%로 나타났다. 앞선 방법과 마찬가지로 유전자 알고리즘을 적용시켜 변수 선택을 실시하였고, 그 결과 변수는 6가지, 3가지, 2가지로 순차적으로 줄어들었다. 정분류율의 결과는 각각 90%, 89%, 그리고 86%를 기록하였다. 모델 생성과정에서 출현한 빈도가 낮은 변수들이 차례로 제거되기 때문에 상대적으로 높은 빈도를 보인 기후변수만이 최종적인 모델 생성에 참여하게 된다. 이에 정분류율 산출 시 올바르게 분류된 경우가 많아지는 결과를 보이는 것으로 판단된다. 따라서 변수의 개수가 줄어듦에도 불구하고 모델의 정확도가 향상되거나 유지되는 것을 볼 수 있었다. 초기에 여러 가지 변수들 사이에서 생성된 공간 분포 모델의 정확도는 모델 제작에 투입되는 변수들을 80%이상 제거함에도 불구하고 오히려 상승하였다. 이 외에도 남은 네 가지 지역과 분류 방식에 따른 경우(평창군과 영동군 지역의 출현종수의 각각 4분류와 5분류)에 대해서도 앞선 방법과 같이 반복적인 실험을 한 결과, 모델의 정확도는 대체로 기존의 정확도보다 상승된 값을 보이거나 혹은 정확도가 떨어진 경우에는 그 정도가 미세한 것으로 나타났다. 두 연구 대상지의 출현종수를 내츄럴브레이크로 4분류하였을 때를 제외한 나머지의 결과들이 모두 최소 2.3%에서 최고 8.9% 상승하는 결과 값을 보여주었다. 네츄럴브레이크로 4분류한 각 대상 지역의 데이터의 경우, 기존의 정확도에 비해 감소하였지만 그 정도가 미비한 것으로 나타났다.

그림 5는 초기의 15가지의 모든 독립변수의 사용

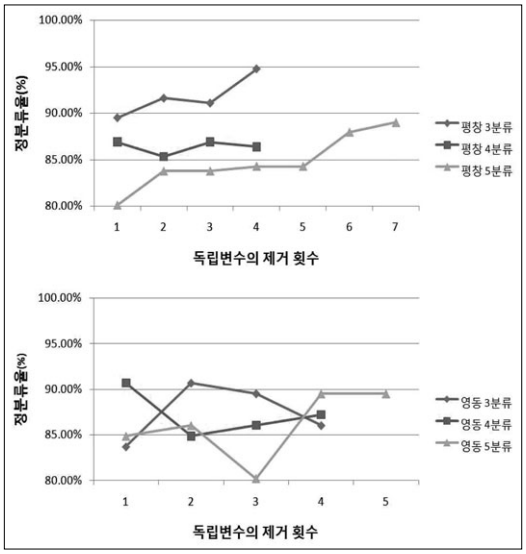

그림 5. 최적변수 설정과정에서의 분류율

으로부터 유전자 알고리즘을 통해 선정된 변수들을 순차적으로 제거하였을 때의 정확도의 변화를 나타낸 것이다. 그래프의 변화에서 볼 수 있듯이 독립변수의 제거가 진행됨에 따라 평창의 출현종수 5분류를 제외한 모든 경우에 정분류율이 최소 1회 하락하는 사례가 존재하였다. 하지만 초기의 15가지의 독립변수 중 최적의 변수들을 선택하는 과정이 단순히 예측모델의 정확도를 저하시켰다고 판단하기는 힘들다.

결과적으로 변수 선정 절차가 종료된 후 제작된 총 6가지의 모델에서 4가지의 경우(평창군 3분류, 평창군 5분류, 영동군 3분류, 영동군 5분류)에서 오히려 초기 모델 보다 높은 정확도를 갖는 결과를 보였다.

공간 분포 모델 형성에 참여하는 변수들이 줄어들게 되면 정보의 손실이 생기고 이에 따른 영향(Auger *et al.*, 2000)이 존재하지만 오히려 정확도가 향상되는 결과를 보이는 이유는 다음과 같이 이해할 수 있다. 많은 변수들로 모델을 제작하게 되면, 각 변수들 사이에 복잡한 관계로 인하여 결국 모델의 설명력을 충분히 제시해주지 못하고 오히려 변수의 중복도(redundancy)의 영향이 증가한다(John, 1997). 따라서 많은 변수들 중에서 최적의 변수를 선정하여 중복도를 줄이는 것이 보다 효율적이고 예측모델을 이해하는데 도움을 준다 (D'heygere, 2003).

본 연구에서는 공간 분포 모델을 제작한 후 이를 통해 생물다양성 지도를 구축하였다. 그림 6와 7은 제작된 생물다양성 지도의 예로써 각각 출현종수를 3단계와 4단계로 분류한 결과이다. 앞서 수행한 정분류율의 정량적인 분석과 제작된 지도를 통한 정성적인 분석을 토대로 최적의 변수 선정 전·후의 생물다양성 지도를 비교한 결과, 각 지역에 분포하는 식생 출현종수의 현황이 흡사한 것으로 나타났다. 한편, 모델 제작에 사용된 자료간의 해상도 차이로 인해 최적의 변수로 제작된 지도가 다소 격자 모양이 두드러져 보이나, 대개의 경우 더 높은 정확도를 가지는 것으로 나타났다.

5. 결 론

본 연구에서는 향후 생물다양성의 변화를 모니터링하기 위한 정량적인 수단을 제시하였다. 이에 생물다양성의 대표적 지표인 식생종풍부도를 바탕으로 식생의 공간 분포 모델을 구축하였고 이를 통해 생물다양성 지도를 제작하였다. 식생 공간 분포 모델 제작에는 의사결정나무 알고리즘과 16가지의 자연 및 환경적 변수들이 고려되었으며, 이들 중 최적의 변수를 선정하기 위해서 유전자 알고리즘을 사용하였다. 본 알고리즘을 통한 변수 제거 과정에 있어서 모델의 정확도는 대체로 향상되었다. 또한 최적의 변수 선택으로 구축된 모델을 통해, 투입된 변

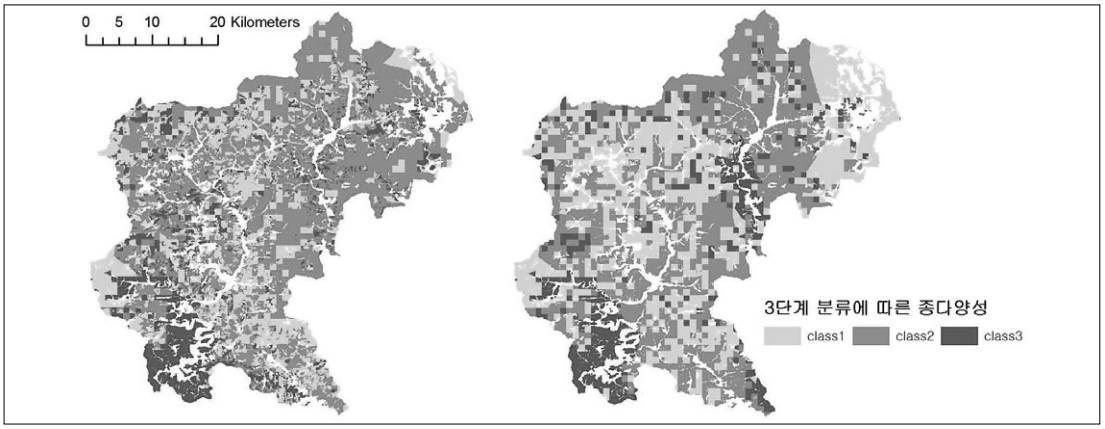

그림 6. 평창군의 네츄럴 브레이크 3단계에 대한 변수선택 전(좌)과 후(우)의 생물다양성 지도

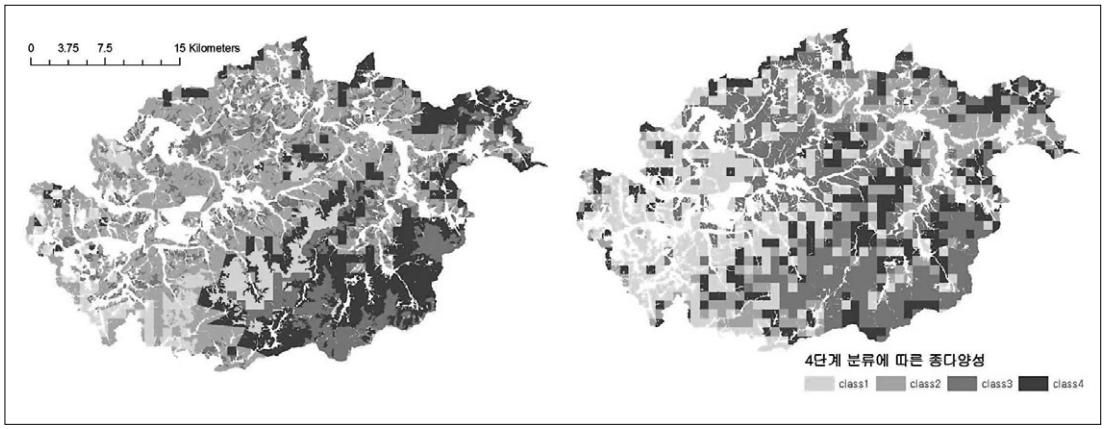

그림 7. 영동군의 네츄럴브레이크 4단계에 대한 변수선택 전(좌)과 후(우)의 생물다양성 지도

수들이 미치는 중요도의 경향을 판단할 수 있었다. 최종적으로 최적의 변수들만 활용하여 평창군과 영동군지역의 생물다양성 지도를 구축하였다.

본 연구에서 제안한 변수 선정에 따른 생물다양성 지도는 대상지역의 모든 곳을 방문하여 면밀히 조사하지 않더라도

비교적 높은 정확도를 갖는 생물다양성 현황정보를 제공해 준다는 점에서 그 장점이 부각된다.

또한 국제적인 환경이슈로 부상하고 있는 생물다양성 기반 의사결정체계에 기여하기 위한 수단으로써, 또한 생물종의 보존우선순위를 판단하거나 개발제한구역을 설정하는 등의 관련 정책결정 및 가치판단 시에도 유용한 기초자료로써 이용할 수 있다. 이 밖에 실측된 현황 데이터를 제공받을 수 없거나 데이터의 양이 제한되어있는 경우와 같이 환경적으로 제약적인 조건하에서도 최적의 변수선정을 통한 공간 분포 모델의 제작이 가능하다.

하지만 생물다양성의 현황 및 지위를 판단하기 위한 필수적인 변수를 일반화시키는 연구 및 식생의 출현종수를 최적으로 등급화하여 구분할 수 있는 분류(classification) 방식에 대해서는 추가적인 고찰이 필요하다. 나아가 다양한 해상도에서 지도를 구현하여, 남한 전체를 대상으로 하는 최적의 생물다양성지도를 만드는 연구가 진행된다면 관련된 의사결정의 중요한 참고자료로 활용될 수 있을 것이다.

참고문헌

통계청, 2009. 인구주택총조사. http://kosis.kr/

Andrews, P. & O' Brien, E.M. (2000) Climate, vegetation, and predictable gradients in mammal species richness in southern Africa, Journal of Zoology, 251, 205-231.

Auger, P., Charles, S., Viala, M., Poggiale, J., c2000. Aggregation and emergence in ecological modelling: intergration of ecological levels. Ecol. Model. 127, 11-20.

Beaumont, L. J., Gallagher, R. V., Thuiller, W., Downey, P. O., Leishman, M. R. and Hughes L. 2009. Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. Diversity and Distribution, 15, 409-420.

Boone, R.B. & Krohn, W.B. (2000) Relationship between avian range limits and plant transition zones in Maine, Journal of Biogeography, 27, 471-482.

Currie, D.J. (1991) Energy and large-scale patterns of animal and plant species richness. American Natualist, 137, 27-49.

D' heygere, T., Goethals, P. L. M., De Pauw,

N. 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecological Modelling. 160, 291-300.

Echeverria, C., Coomes, D. A., Hall, M. and Newton, A., 2008. Spatially Explicit Models to Analyze Forest Loss and Fragmentation Between 1976 and 2020 in Southern Chile. Ecological Modelling, 212, pp. 439-449.

European Communities, 2008, TEEB(the Economices of Ecosystem and Biodiversity) an interim Report.

Fielding, A.H., Bell, J.F., 1997. A review method for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38-49.

Flesch, A. D. & Hahn, L. A. 2005. Distribution of birds and plants at the western and southern edges of the Madrean Sky Islands in Sonora, Mexico. U S Forest Service Rocky Mountain Research Station Proceedings RMRS-P, 36, 80-87.

Franklin, J. 2009. Mapping Species Distribution. Cambridge, p. 320.

Garzon-Orduna, I., Miranda-Esquivel, D., Donato, M. 2008. Parsimony analysis of endemicity describes but does not explain: an illustrated critique. Journal of Biogeography. 35, 903-913.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Pulishing Company, Reading, MA, p. 412.

Jenks, G. F., 1967. The Data Model Concept in Statistical Mapping. International Yearbook of Cartography. 7. pp. 186-190.

Jha, C. S., Goparaju, L., Tripathi, A., Gharai, B., Raghubanshi, A. S., Singh, J. S. 2005. Forest fragmentation and its impact on species diversity: an analysis using remote sensing and GIS. Biodiversity and Conservation. 14, 1681-1698.

John, G.H., 1997. Enhancement to the data mining process. PHD Dissertation, Computer Science Department, Standford University.

Jongman, R.H.G., Ter Braak, C.J.F., Van tongeren, O.F.R., 1995. Data Analysis in Community and Landscape Ecology, 2nd ed.. Cambridge University Press, Cambridge, p. 299.

Kier, G., Kreft, H., Lee, T. M., Jetz, W., Ibisch, P. L., Nowicki, C., Mutke, J. & Barthlott, W. 2009. PNAS. Vol. 106. pp. 9322-9327.

Knapp, A. K., *et al.* 2002. Rainfall variability, carbon cycling, and plant species diversity in a mesic grassland. Science 298. pp. 2202-2205.

Meentemeyer, R. K., Moody, A., & Franklin, J. 2001. Landscape-scale patterns of shrub-species abundance in California chaparral: the role of topographically mediated resource gradient. Plant Ecology, 156, 19-41.

Michaelsen, J., Davis, F. W., & Borchert, M. 1987, Anon-parametric method for analyzing hierarchical relationships in ecological data. Coenoses, 2, 39-48.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B. & Kent, J. 2000. Biodiversity hotspots for conservation priorities. Nature. Vol. 403. pp. 853-858.

Nogue, S., Rull, V., Vegas-Vilarrubia, T. 2009. Modeling biodiversity loss by global warming on Pantepui, northern South America: projected upward migration and potential habitat loss. Climate Change. 94, 77-85.

Phillips, S. J., and Dudik, M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography. 31, 161-175.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco.

Witten, I. H. & Frank, E. 2000. Data Mining, Morgan Kaufmann, p. 525

The University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/~ml/index.html

최종원고채택 11. ??. ??