The Journal of Engineering Geology, Vol.18, No.1, March, 2008, pp. 45-54

의사결정나무모형을 이용한 편마암 지역에서의 급경사지재해 예측기법 개발

송영석* · 채병곤

한국지질자원연구원 지질환경재해연구부

Development to Prediction Technique of Slope Hazards in Gneiss Area using Decision Tree Model

Young-Suk Song* and Byung-Gon Chae

Geological & Environ. Hazards Division, Korea Inst. of Geoscience and Mineral Res.

본 연구에서는 기 조사된 편마암 지역에서의 급경사지재해 발생지역 및 미발생지역에 대한 현장조사자료 및 토질시험자료를 토대로 통계적인 분석방법인 의사결정나무모형을 이용하여 급경사지재해 예측기법을 개발하였다. 편마암 지역에서의 조사된 급경사지재해 자료는 서울 및 경기지역에서 1998년 집중호우로 발생된 104개소구간이다. 이 가운데 예측모델 개발에 활용된 자료수는 결측치를 제외한 61개소로서, 급경사지재해 발생구간 34개소와 미발생구간 27개소이다. 의사결정나무모형을 이용한 통계적인 분석은 카이제곱 통계량, 지니 지수 및 엔트로피 지수를 적용하여 실시하였다. 분석결과 사면경사, 포화도 및 사면고도가 분리기준으로 선택되었으며, 엔트로피 지수를 이용한 의사결정나무모형 예측모델이 정확도가 가장 높은 것으로 나타났다. 선정된 급경사지재해 예측모델의 분리기준은 최상위부터 사면경사, 포화도 및 사면고도의 순서로 선택되었으며, 각각의 분리기준치는 사면경사의 경우 17.9°, 포화도의 경우 52.1%, 사면고도의 경우 320 m로 결정되었다.

주요어 : 급경사지재해, 편마암지역, 예측기법, 의사결정나무모형, 엔트로피 지수

Based on the data obtained from field investigation and soil testing to slope hazards occurrence section and non-occurrence section in gneiss area, a prediction technique was developed by the use of a decision tree model, which is one of the statistical analysis methods. The slope hazards data of Seoul and Kyonggi Province, which were induced by heavy rainfall in 1998, were 104 sections in gneiss area. The number of data applied in developing prediction model was 61 sections except a vacant value. Among these data, the number of data occurred slope hazards was 34 sections and the number of data non-occurred slope hazards was 27 sections. The statistical analyses using the decision tree model were applied to chi-square statistics, gini index and entrophy index. As the results of analyses, a slope angle, a degree of saturation and an elevation were selected as the classification standard. The prediction model of decision tree using entrophy index is most likely accurate. The classification standard of the selected prediction model is composed of the slope angle, the degree of saturation and the elevation from the first choice stage. The classification standard values of the slope angle, the degree of saturation and elevation are 17.9°, 52.1% and 320 m, respectively.

Key words : slope hazards, gneiss area, prediction technique, decision tree model, entrophy index

서 론

급경사지재해는 사면지역의 지질학적, 지형학적 및 지반공학적 특성에 따라서 그 규모, 형태 및 발생빈도 등이 다르게 나타난다. 국내의 자연사면에서 발생하는 급경사지재해는 대부분 잔류토, 붕적토 및 충적토 등의 미고결층 즉, 토층에서 강우로 인한 간극수압 상승, 지표침식, 단위중량 증가 등에 의하여 발생되고 있으며, 대부분의 급경자지재해는 토석류급경사지재해(debris flow landslide)에 해당한다. 이와 같은 토석류급경사지재해는

*Corresponding author: yssong@kigam.re.kr

발생위치가 비록 인간의 생활권과 떨어져 있고 소규모이지만 사태물질의 많은 부분이 모래입자보다 큰 암편들로 구성되어 있고 빠른 속도로 사면하부로 이동된다. 따라서 사면의 하부에 위치하고 있는 인간의 생활권에 큰 피해를 줄 수 있다(채병곤 등, 2005).

우리나라 연평균 강우량중 절반이상이 7월과 8월에 집중되며, 이 시기에 토석류급경사지재해가 대부분 발생한다. Olivier(1994)는 24시간 동안의 강우량이 연평균 강우량의 20%를 초과할 경우 대형 급경사지재해가 일어날 수 있다고 보고한 바 있다. 그리고 Brand(1981)는 짧은 시간에 내리는 집중강우는 지질조건이나 수문지질조건과 관계없이 대형 급경사지재해를 일으킬 수 있다고 보고한 바 있는데 이는 집중강우가 지표물질을 완전히 포화시킬 수 있는 상태의 강우량을 의미한다.

그런데 동일한 강우량을 갖는 지역에서도 급경사지재해가 발생되는 지역과 발생되지 않는 지역으로 구분된다. 이는 강우량이 급경사지재해를 발생시키는 가장 큰 요인임에도 불구하고 지반 및 지질매체의 특성에 따라 급경사지재해 발생정도가 다름을 의미한다. 즉, 지반 및 지질매체의 공학적 특성에 따라 동일한 강우조건에서도 급경사지 재해가 발생되는 경우와 발생되지 않는 경우로 나눌수 있다. 따라서 일정 강우조건하에서 대상지역의 어떤 지반조건 및 지질조건일 때 과연 급경사지재해가 발생하며 정량적으로 급경사지재해 발생가능성을 예측하는 것은 매우 중요한 사항이다.

김원영 등(2003)은 지질조건별 국내에서 발생된 자연사면의 산사태 발생특성 및 원인을 규명하고, 이에 대한 자료를 조사하였다. 이를 토대로 광역적인 지역을 대상으로 산사태 발생가능성을 예측하기 위하여 로지스틱 회귀모델을 이용한 산사태 예측모델을 개발하여 일부지역에 대한 산사태 예측지도를 작성한 바 있다.

그러나, 김원영 등(2003)의 산사태 예측모델은 전문적인 지식을 가진 지질 및 GIS전문가에 의해서만 수행이 가능하므로, 본 기술을 범용화 및 실용화하기에는 여러가지 문제점이 있다. 그러므로 일반 지질 및 토목기술자가 쉽게 활용할 수 있는 단순하고 정확한 예측모델의 개발이 필요하다.

따라서 본 연구에서는 기 조사된 편마암 지역에서의 급경사지재해 발생지역 및 미발생지역에 대한 현장조사 자료 및 토질시험자료를 토대로 통계적인 분석방법인 의사결정나무모형(decision tree model)을 이용하여 급경사지재해 정밀예측모델을 개발하고자 한다. 이를 위하여 의사결정나무모형의 분석방법 가운데 카이제곱 통계량, 지니 지수 및 엔트로피 지수를 활용하고자 한다. 그리고 이상의 예측모델들에 대한 정확성 검증을 수행한 후 가장 정확도가 높은 예측모델을 선정하고자 한다.

급경사지재해 자료조사 및 분석

급경사지재해 발생 및 미발생 자료수집

새로운 급경사지재해 예측모델 개발하기 위하여 가장 먼저 수행되어야 할 사항은 현재까지 발생된 급경사지재해에 대한 자료수집 및 발생특성을 분석하는 것이다. 본 연구에서는 최근 10년간 급경사지재해가 발생된 지역가운데 서울 및 경기지역을 대상으로 조사된 자료를 활용하였다. 이들 자료는 자연사면에서의 급경사지재해 발생지역에 대한 야외 정밀조사 및 토질시험결과를 토대로 수집된 것이다.

대상지역인 서울 및 경기지역은 서울, 포천, 성동, 문산 등으로서 1998년 8월 4일부터 7일까지 최고 588.5 mm의 집중호우가 내렸으며, 이로 인하여 많은 급경사지 재해가 발생된 지역이다. 대상지역의 지질조건은 모두 편마암이며, 총 104개소의 현장정밀 조사자료 및 토질시험자료를 활용하였다(김원영 등, 2000).

Table 1은 본 연구에서 수집된 지역별 급경사지재해 발생구간 및 미발생구간의 개소수를 정리한 것이다. 표에서 보는 바와 같이 급경사지재해 발생구간은 77개소이고, 미발생구간은 27개소로서 총 104개소에 대한 자료를 수집하였다. 이들 정밀조사된 자료를 토대로 의사결정나무모형을 이용한 새로운 급경사지재해 정밀예측모델을 개발하였다.

급경사지재해 현장조사 방법

급경사지재해가 발생된 지역에서 실시되는 현장조사는 크게 네가지 방법으로 구분할 수 있는데, 첫째는 개략조사, 둘째는 정밀조사, 셋째는 전수조사, 그리고 넷째

Table 1. Number of slope hazards occurrence section and non-occurrence section.

지역	급경사지재해		합계
	발생구간	미발생구간	
서울 및 경기지역	77	27	104

는 토질조사이다. 개략조사에서는 급경사지재해가 발생된 즉시 1 : 50,000 축척의 지형도와 개략조사용 시트(landslide field survey sheet)를 이용하여 주로 광역적인 급경사지재해의 분포현황을 파악하게 된다. 정밀조사에서는 1 : 5,000 축척의 지형도와 정밀조사용 시트를 이용하여 급경사지재해위치, 노두의 발달상태, 암반의 풍화도, 미고결층의 분포형태 및 급경사지재해의 유형 등 세부적인 급경사지재해 특성을 조사하는 것이다. 그리고 전수조사는 정밀조사와 병행하여 수행되며 주로 급경사지재해의 규모 등 기하특성을 조사하는 것이다. 마지막으로 토질조사는 역시 정밀조사시 수행되는 것으로 급경사지재해지역 토층의 분포상태를 파악하고 물리적 성질 및 공학적 특성평가를 위한 토질시험용 시료를 채취하게 된다.

개략조사

개략조사의 목적은 급경사지재해 위치와 규모, 파괴유형, 추가적인 급경사지재해 발생가능성 등을 개략적으로 파악하여 정밀조사 계획을 수립하는 것이다. 지형도, 지질도 및 과거에 발생되었던 급경사지재해이력 등을 문헌조사, 자료조사 및 현장조사를 통하여 분석하고, 지형 및 암반상태, 급경사지재해 징후 등을 파악한다. 그리고 이러한 모든 자료를 종합하여 급경사지재해의 유형과 발생지점 등을 예측하여 정밀조사 계획을 수립한다. 개략조사의 범위는 추가적인 사면파괴의 가능성과 지질조건이 동일한 곳에서 유사한 사면파괴가 발생하기 쉬운 점을 고려하여 인접지역을 포함한 보다 넓은 지역을 대상으로 설정할 필요가 있다. 이와 같은 목적의 개략조사에서는 소축척의 지도를 사용하게 되고 기재를 위한 조사야장 또한 꼭 필요한 요소만을 함축적으로 기록할 수 있도록 간략화 하는 것이 일반적이다. 우리나라에서 발행되고 있는 지형도 중에서 비교적 소축척에 해당하며 현장조사시 지형요소의 반영이 가능한 것으로는 1 : 50,000 축척의 지형도가 있으며, 이를 이용하면 광역적인 급경사지재해 분포조사와 개략적인 기하양상의 파악이 가능하다. 그리고 이를 토대로 하여 정밀조사를 목적으로 한 대상지역과 개별 급경사지재해의 선정 및 토질평가를 위해 적정한 시료채취 위치를 선정할 수 있다.

개략조사의 경우 1 : 50,000 축척의 지형도와 조사시트를 이용하여 현장조사를 수행한다. 개략조사에서는 먼저 1 : 50,000 축척의 지형도에 급경사지재해의 위치, 방향 및 크기 등을 개략적으로 표시하고, 모든 개소에 대해 각각의 일련번호를 부여한다.

정밀조사

정밀조사는 일반적으로 급경사지재해에 대한 제반사항을 자세히 조사하고 분석함으로써 급경사지재해의 징후, 형태 및 현황 등을 파악하기 위한 것으로서 안정성 해석과 대책공법 선정 등의 기초자료로 사용된다. 그리고 정밀조사의 결과에 따라 급경사지재해가 발생할 가능성이 있거나 위험도가 매우 높은 경우에는 정밀조사의 결과를 근거로 하여 보다 세밀한 추가조사가 수행되기도 한다.

개략조사 결과를 토대로 한 급경사지재해지역의 정밀조사에서는 급경사지재해의 위치, 범위, 규모 및 형태 등을 확인하고 급경사지재해 발생요인의 규명과 추가로 발생할 가능성을 검토한다. 이를 위해서 지형의 고도나 경사, 사면의 높이와 경사, 슬라이드나 유동형태 등의 붕괴양상과 토질특성 등을 검토하고, 안정성해석프로그램이나 급경사지재해 예측모델 등을 이용한 안정성평가와 급경사지재해 가능성을 분석한다. 그리고 급경사지재해는 주로 사면 내부의 연약대나 불연속면을 따라서 발생하게 되므로, 사면의 연약대나 불연속면의 특성을 정확하게 조사할 필요가 있다. 불연속면의 특성으로는 연장, 간격, 틈새 및 지하수 유동상태 등을 세밀히 조사한다. 특히, 사면의 상부에 형성되는 인장균열은 추가적인 사면파괴에 큰 영향을 미치는 것이 일반적이므로 이러한 인장균열의 발생유무 및 상태 등을 자세하게 조사할 필요가 있다. 이와 같은 목적의 정밀조사에서는 대축척의 지도를 사용하고 조사야장 또한 급경사지재해와 관련되는 다양한 요소들을 자세하게 조사할 수 있도록 세분화하는 것이 일반적이다. 우리나라에서 발행되고 있는 지형도 중 비교적 대축척의 지형도로서 정밀조사시 지형요소와 급경사지재해의 형태를 자세하게 나타낼 수 있는 것으로는 1 : 5,000 축척의 지형도가 적합하다.

정밀조사의 경우 1 : 5,000 축척의 지형도와 조사시트를 이용하여 현장조사를 수행한다. 정밀조사에서는 먼저 1 : 5,000 축척의 지형도에 급경사지재해의 양상을 실제의 모양대로 자세하게 표현한다. 그리고 정밀조사용 시트에는 지질, 풍화도, 불연속면, 사태종류, 변화양상, 지형고도, 사면경사, 사면방향, 급경사지재해가 시작된 지점의 위치, 급경사지재해의 고도와 길이, 폭, 깊이, 체적 및 붕괴가 시작된 지점부터 사태물질이 쌓인 곳까지의 형태 등 급경사지재해와 관련되는 다양한 요소들을 기재한다. 한편, 정밀조사와 함께 전수조사와 토질조사를 병행하여 급경사지재해의 기하특성과 토질특성을 평가한다.

전수조사

전수조사는 일반적으로 급경사지재해의 발생형태 및 규모와 사태물질의 이동경로 등 사면붕괴가 시작된 상부지점부터 사태물질이 흘러내려 쌓인 하부지점까지의 급경사지재해 기하양상을 측정할 목적으로 수행한다. 급경사지재해의 기하측정은 개별 급경사지재해지역에서 길이방향 즉, 종단면을 기준으로 하여 각 지점별 사면경사, 폭 및 방향 등을 측정하고, 횡단면을 기준으로 하여 지형이 변화하는 지점들마다에서 사태물질이 붕괴되어 유실된 깊이 및 퇴적된 두께를 측정한다. 이를 통하여 파괴가 시작된 사면의 상부로부터 사태물질이 이동되어 쌓인 사면하부까지의 총길이와 지형에 따른 급경사지재해의 변화폭을 측정함으로써 급경사지재해가 시작되어 끝날 때까지의 모든 기하상태를 파악할 수 있다. 뿐만 아니라 개별 급경사지재해의 형태변화와 사태물질의 거동특성을 지형 및 지질조건과 연관시켜 파악함으로써 광역적인 통계분석이 가능하다.

급경사지재해의 기하양상을 보다 정밀하게 측정하기 위해서는 지형측량에 의한 방법도 하나의 방안으로 고려될 수 있다. 그러나 다수의 급경사지재해를 대상으로 조사를 실시해야하기 때문에 급경사지재해지역에 대한 효율적인 전수조사를 위하여 줄자(tape)를 이용해서 직접 측정하며, 이 때에 미리 작성된 전수조사용 기록지에 측정치를 기재하는 방법으로 급경사지재해 전수조사를 실시한다. 줄자를 이용해서 급경사지재해의 기하양상을 측정하는 방법은, 먼저 급경사지재해가 시작된 지점부터 사태물질이 쌓인 곳까지 급경사지재해가 진행된 길이방향을 따라 줄자를 길게 늘어뜨린 후 줄자의 방향과 경사를 측정한다. 그리고 상부사면으로부터 하부지점으로 내려가면서 급경사지재해의 깊이나 폭이 변화되는 지점들마다 길이방향의 줄자에 좌우방향으로 다른 테이프를 직각으로 교차하게 늘이고 교차점의 거리와 좌우방향의 연장 및 깊이를 측정한다. 이렇게 측정된 각각의 기록들을 이용하여 실내에서 종·횡단도를 작성하여 급경사지재해의 기하특성을 파악할 수 있다.

토질조사

자연사면을 구성하고 있는 물질들 중 암반의 상부에 위치한 토층의 여러 물성과 공학특성을 이해하기 위해서는 현장이나 실내에서 수행한 토질시험으로부터의 측정결과가 필요하다. 토층에서 발생되는 급경사지재해들은 그 사면을 구성하고 있는 토층물질의 토질특성과 직접적으로 관계가 있으며, 특히, 우리나라와 같이 주로 집중호우에 의해 발생되는 급경사지재해들의 경우 토층의 기본적인 물성들 외에 간극이나 밀도, 투수성 및 전단강도와 같은 공학특성과 관련되므로 이들에 대한 시험방법과 그 결과들의 적용방법 등을 이해하는 것은 매우 중요한 사항이다. 또한, 급경사지재해발생과 토질특성간의 상관성에 주안점을 두고 지질조건별로 급경사지재해 발생지역과 미발생지역의 토질특성을 비교분석하기 위하여 집중호우로 인해 급경사지재해가 집중적으로 발생한 대상지역별로 급경사지재해 발생지역과 미발생지역으로 구분한 각각의 토층시료를 채취한다. 그리고 토층시료는 급경사지재해빈도, 지형, 지질조건, 토층분포 및 단위면적당의 빈도 등을 고려함으로써 토층의 특성이 균등하게 평가될 수 있도록 한다.

토층시료는 표토를 제거한 후 40-60 cm 깊이에서 교란 및 불교란시료를 각각 채취한다. 교란시료는 비닐팩을 이용하여 채취하였으며, 불교란시료는 스테인레스(stainless)로 제작한 직경 10 cm, 높이 6 cm 크기의 원통형 몰드(ring sampler) 및 상하부 캡을 이용한다. 특히, 투수시험용 불교란시료는 직경 10 cm, 높이 13 cm 크기의 원통형 몰드를 사용하여 채취한다. 한편, 모든 시료는 현장조건이 최대한 유지되도록 밀봉한 상태로 실험실로 운반하여 실내시험에 이용된다. Fig. 1은 원통형 몰드를 이용하여 시료를 채취하는 모습을 나타낸 것이다.

토질시험은 현장으로부터 운반된 토층시료를 대상으로 비중(specific gravity), 함수비(moisture content), 입도(grain size), 액성한계(liquid limit) 및 소성한계(plastic limit), 간극비(void ratio), 간극율(porosity), 포화도(degree of saturation) 등의 물리적 특성, 그리고 밀도(density), 투수계수(coefficient of permeability) 및 전단강도(shear strength) 등의 공학적 특성을 평가하기

Fig. 1. Soil sampling work using cylindrical mold.

위한 총 10여종의 토질시험을 실시하며, 시험방법은 모두 한국산업규격(KS)에 준하여 시험한다.

급경사지재해 정밀예측모델 개발

방대한 지리정보는 공간 데이터베이스 또는 공간 데이터 웨어하우스 등에 저장되어 공간 데이터 마이닝에 이용된다. 공간 데이터 마이닝이란 공간 데이터 저장소로부터 함축적인 지식, 공간적 관계 또는 명시적으로 저장되어 있지 않은 패턴들의 추출을 의미한다. 이러한 유용한 패턴들을 발견하기 위해 연관분석, 분류, 군집 등 여러 가지 데이터 마이닝 기법이 소개되었다(장윤경 외, 2006). 이러한 데이터 마이닝 기법으로는 의사결정나무 기법, 로지스틱회귀분석기법, 신경망기법, 베이시안 기법 등이 있다.

의사결정나무는 데이터 마이닝의 분류와 예측 작업에 주로 사용되는 기법으로 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴을 분류 모형트리의 형태로 만드는 것이다. 이 기법은 신경망기법보다 훈련시간이 짧기 때문에 데이터의 규모가 클 경우 유리하고 결과에 대해 분류나 예측의 근거를 알려주기 때문에 이해하기 쉽다.

의사결정나무는 분석대상에 대한 분류나 예측을 수행하기 위해서 사용되는 분석기법으로 대용량의 데이터 내에 존재하는 관계, 패턴 및 규칙 등을 탐색하고 모형화하는 역할을 수행하며, 신경망이나 판별분석 등에 의한 방법과는 달리 적용결과에 의해 규칙을 명확하게 나타낼 수 있다. 또한, 예측모형 자체뿐만 아니라 최적의 결과를 검색하거나 분석에 필요한 변수 간의 교호효과, 즉 두 개 이상의 입력변수가 결합하여 목표변수에 어떻게 영향을 주는지를 찾아내는데 이용된다(김종규 외, 2006).

특히, 나무모형구조로 표현되기 때문에 다른 기법들과 비교하여 쉽게 이해되고 설명할 수 있으며, 임의의 데이터 범주에서 동일한 특성을 갖는 집합으로 구분하여 특성을 정의하고, 목표변수에 대한 규칙을 추론하여 미래에 대한 예측을 할 경우 유용하게 활용할 수 있다(최기헌, 1995).

본 연구에서는 서울북부 편마암 지역에서의 급경사지재해 발생지역 및 미발생지역에 대한 현장조사자료 및 토질시험자료를 토대로 의사결정나무모형을 이용하여 급경사지재해 정밀예측모델을 개발하였다. 그리고 의사결정나무모형을 이용한 급경사지재해 예측결과의 정확성을 검증하기 위하여 정오분류표를 적용하였다.

의사결정나무모형 이론

의사결정나무모형의 정의

의사결정나무모형은 의사결정규칙(decision rule)을 나무구조로 도표화하여 관심대상이 되는 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다. 이 방법은 분류 또는 예측의 과정이 나무구조에 의한 추론규칙에 의해 표현되기 때문에 신경망, 판별분석 등에 비해 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있는 분석방법이다.

데이터마이닝에서의 의사결정나무모형은 탐색(exploration)과 모형화(modeling)라는 두 가지 특성을 모두 가지고 있다고 할 수 있다. 차원축소 및 변수선택, 교호작용 효과의 파악, 범주의 병합 또는 연속형 변수의 이산화는 탐색단계에 포함된다고 할 수 있고 세분화, 분류 및 예측은 모형화 단계에 포함된다고 할 수 있다.

의사결정나무구조는 마디(node)로 구성되며, 뿌리마디(root node)로부터 시작하여 분리기준(splitting criterion), 정지규칙(stopping rule), 가지치기(pruning) 등에 의해 각 가지(branch)가 끝마디(종단마디, terminal node)에 이를 때까지 자식마디(child node)를 계속 형성해 나감으로써 완성된다. 뿌리마디와 반대로 트리의 가장 끝에 위치하여 가지가 분리되지 않는 마디를 끝마디라고 하며, 뿌리마디부터 종단마디까지의 분리단계를 깊이(depth)라고 한다.

Fig. 2는 의사결정나무모형의 분석과정을 나타낸 것으로 일반적으로 의사결정나무의 형성, 가지치기, 타당성 평가, 해석 및 예측의 과정을 거쳐 수행된다.

(1) 의사결정나무의 형성 : 분석의 목적과 자료구조에 따라서 적절한 분리기준과 정지규칙을 지정하여 의사결정나무를 얻는다.

(2) 가지치기 : 분류오류(classification error)를 크게 할 위험(risk)이 높거나 부적절한 추론규칙(induction rule)을 가지고 있는 가지를 제거한다.

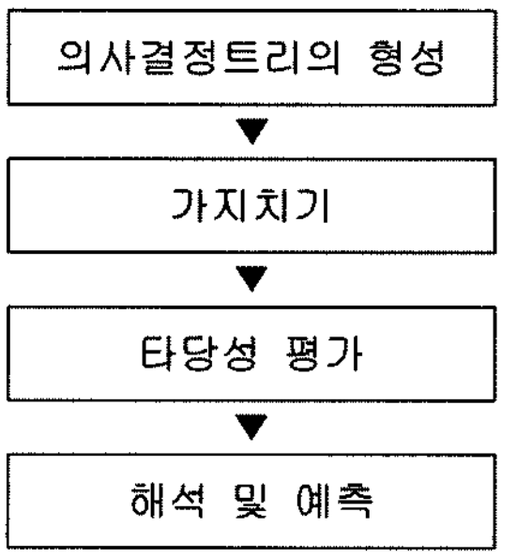

Fig. 2. Analysis process of decision tree model.

(3) 타당성 평가 : 이익도표(gains chart)나 위험도표(risk chart) 또는 검증용 자료(test data)에 의한 교차타당성(cross validation) 등을 이용하여 의사결정나무를 평가한다.

(4) 해석 및 예측 : 의사결정나무를 해석하고 분류 및 예측모형을 설정한다.

의사결정나무모형의 알고리즘

의사결정나무분석을 위해서 CHAID, CART, QUEST 등과 같은 다양한 알고리즘이 제안되어 있으며 최근에는 이들의 장점을 결합하여 보다 개선된 알고리즘들이 제안되고 상용화되고 있다. 의사결정나무모형의 대표적인 알고리즘은 CHAID (Chi-squared Automatic Interaction Detection) 알고리즘(Kass, 1980)으로 명목형, 순서형, 연속형 등 모든 종류의 목표변수와 분류변수에 적용이 가능하며, Exhaustive CHAID 알고리즘(Biggs et al, 1991)으로 발전하였다. 그 밖에 CART(Classification and Regression Tree), QUEST(Quick, Unbiased, Efficient, Statistical), C5.0, C4.5 알고리즘 등이 있다.

순수도(purity) 또는 불순도(impurity)를 기준으로 자식마디를 형성해 나가는 순수도 지수(purity index)중 목표변수가 이산형인 경우에는 목표변수의 각 범주에 속하는 빈도(frequency)에 기초하여 분리가 일어난다. 이때 사용되는 주요 분리기준(partitioning criterion)으로는 카이제곱 통계량(chi-square statistic)의 p-값, 지니 지수(gini index), 엔트로피 지수(entropy index) 등이 있다. 특히 엔트로피 지수는 식 (1)과 같이 표현되며, 다항분포에서의 우도비 검정통계량을 사용하는 것과 같은 것으로 알려져 있고 최근에 널리 알려진 알고리즘인 C4.5는 엔트로피 지수를 분리기준으로 사용한다.

$$E=-\sum_{j=1}^{c}(P(j)\ln P(j)) \quad (1)$$

여기서, j : 1, 2, …, c로서, c는 목표변수의 범주수

P(j) : 해당마디에서의 번째 그룹에 속하는 자료의 비율을 추정치로 사용

예측모형의 평가방법

일반적인 모형평가(model assessment)의 기준으로는 모형이 얼마나 효과적으로 구축되었는가 즉, 얼마나 적은 입력변수로 모형을 구축했는가의 문제나 혹은 같은 모집단 내의 다른 데이터에 적용하는 경우 얼마나 안정적인 결과를 제공해 주는가 즉 일반화의 가능성 등 여러 각도에서 생각할 수 있다. 그러나 무엇보다도 우선적으로 고려되어야 할 사항은 구축된 모형이 얼마나 예측과 분류에서 뛰어난 성능을 보이는가를 알아보는 것이다. 이는 아무리 안정적이고 효과적인 모형도 실제 문제에 적용했을 경우 빗나간 결과만을 양산한다면 아무런 의미가 없기 때문이다.

따라서 모형의 평가는 예측을 위해 만든 모형이 임의의 모형보다 우수한지, 고려된 다른 모형과 비교하여 어느 것이 가장 우수한 예측력을 보유하고 있는지를 비교분석하는 과정이라 할 수 있다. 모형의 평가방법으로는 정오분류표((mis)classification table), Lift Chart의 %Response 이익도표, ROC 도표 등이 있다.

오분류표 평가방법은 목표변수가 범주형인 경우에 적용할 수 있을 것이다. 통계모형의 평가분석을 위해 사후확률(posterior probability)을 비교할 수 있다. 일반적으로 분류의 기준으로 삼는 사후확률(posterior probability)의 경계는 "1/(목표변수의 범주 개수)" 로 삼는 것이 보통이다.

또한 구축된 모형에 대하여 예측과 분류가 얼마나 뛰어난 성능을 보이는가, 그리고 얼마나 안정적인가를 비교하기 위해서 training data(분석용 자료)와 validation data (평가용 자료)의 정분류율(판별력)을 비교하여 validation data의 오분류율을 선정한다. 추가적으로 구축된 모형별 오분류율에 대해서도 검토한다. 식 (2) 및 식 (3)은 정분류율 및 오분류율을 산정하는 방법이다.

$$\text{정분류율}=\frac{(\text{실제0, 예측0})\text{의 빈도}+(\text{실제1, 예측1})\text{의 빈도}}{\text{관찰지의 빈도}}\times 100(\%) \quad (2)$$

$$\text{오분류율}=\frac{(\text{실제0, 예측0})\text{의 빈도}+(\text{실제1, 예측1})\text{의 빈도}}{\text{관찰지의 빈도}}\times 100(\%) \quad (3)$$

의사결정나무모형을 이용한 급경사지재해 정밀예측모델

분석자료

본 연구에서는 기 조사된 서울 및 경기지역의 급경사지재해 발생지역 및 미발생지역에 대한 현장조사자료 및 토질시험자료를 토대로 의사결정나무모형(decision tree model)을 이용하여 급경사지재해 정밀예측모델을 개발하였다.

Table 2는 예측모델개발을 위해 서울 및 경기지역을 대상으로 조사된 자료수를 나타낸 것으로 총 104개소의 조사자료 가운데 결측치를 제외한 61개소의 자료를 활용하였다. 그리고, Table 3은 각 지역별로 의사결정나무모형 분석에 포함된 변수를 나타낸 것이다. 표에서 *표시는 범주형 변수를 나타낸 것이며, 그 외는 연속형의 변수를 나타낸 것이다.

의사결정나무모형을 이용한 예측모델

편마암지역에서의 급경사지재해 예측모델을 개발하기 위하여 전술한 분석자료(n = 61)를 토대로 의사결정나무모형을 이용한 통계적인 분석을 실시하였다. 의사결정나무모형을 이용한 통계적인 분석은 카이제곱 통계량, 지니 지수 및 엔트로피 지수를 적용하였다. 본 의사결정나무모형에 대한 통계분석에는 SAS와 SAS/E-miner 프로그램을 사용하였다.

Fig. 3은 카이제곱 통계량을 이용하여 의사결정나무모형 예측모델을 분석한 결과이다. 그림에서 보는 바와 같이 예측모델의 분리기준변수로는 사면경사만 선택되었으며, 급경사지재해 발생을 일으키는 사면경사의 기준은 24.7°인 것으로 나타났다. 이와 같은 의사결정나무모형을 이용한 예측모델을 평가하기 위하여 정오분류표를 활용하였다. Table 4는 카이제곱 통계량을 이용한 의사결정나무모형 예측모델의 정오분류표를 나타낸 것이다. 그리고 식(2) 및 식(3)을 이용하여 정오분류표의 정분류율을 계산해보면 86.89%, 오분류율을 계산해보면 13.11%로 나타났다.

$$- \text{정분류율} = \frac{(21+32)}{61} \times 100(\%) = 86.89(\%)$$

$$- \text{오분류율} = \frac{(6+2)}{61} \times 100(\%) = 13.11(\%)$$

Fig. 4는 지니 지수를 이용하여 의사결정나무모형 예측모델을 분석한 결과이다. 그림에서 보는 바와 같이 예측모델의 최상위 및 하위 분리기준변수로는 사면경사가

Table 2. Number of analysis data.

지 역	총자료	분석용 자료
서울 및 경기지역	104	61

Table 3. Variable items.

구분	변수	설명	서울 및 경기지역
목표변수	산사태 발생 여부	1 : 발생. 0 : 미발생	○
입력변수	lithology (*)	암석종류	○
	weathering (*)	풍화정도	×
	elevation	지형고도	○
	slope direction	사면방향	×
	slope angle	사면경사	○
	length	산사태 길이	×
	width	산사태 폭	×
	thickness	토층두께	×
	specific gravity	비중	○
	moisture content	함수비	○
	void ratio	간극비	○
	porosity	공극률	○
	degree of saturation	포화도	○
	wet(bulk) density	전체밀도	○
	saturation density	포화밀도	○
	dry density	건조밀도	○
	USCS (*)	입도분포	×
	permeability	투수계수	○
	triggering position	산사태발생위치	×
	gravel	자갈	×
	sand	모래	×
	silt / clay	실트/점토	×
	liquid limit	액성한계	○
	plastic limit	소성한계	○
	plasticity index	소성지수	○
	shear strength-cohesion	점착력	×
	shear strength-friction angle	내부마찰각	×

Table 4. Classification table of analysis result using chi-square statistics.

실제 관측된 값 \ 예측값	급경사지재해 미발생	급경사지재해 발생	합 계
급경사지재해 미발생	21	6	27
급경사지재해 발생	2	32	34
합 계	23	38	61

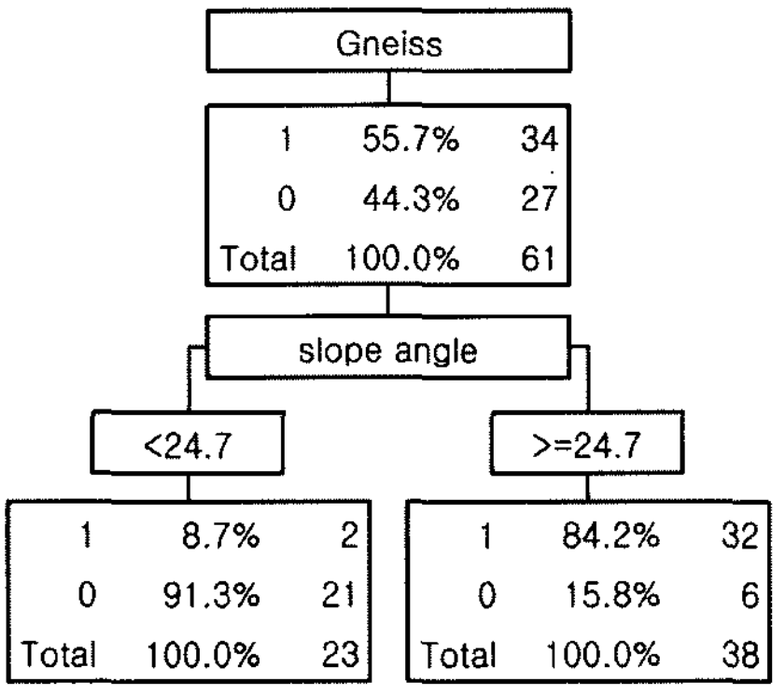

Fig. 3. Decision tree model resulting from chi-square statistics.

선택되었으며, 최하위 분리기준변수로는 포화도가 선택되었다. 급경사지재해 발생을 일으키는 사면경사의 기준은 24.7°-33.3°인 것으로 나타났으며, 사면경사가 33.3° 이상인 경우 급경사지재해 발생을 일으키는 토층의 포화도 기준은 39.8%인 것으로 나타났다. 이와 같은 의사결정나무모형을 이용한 예측모델을 평가하기 위하여 정오분류표를 활용하였다. Table 5는 지니 지수를 이용한 의사결정나무모형 예측모델의 정오분류표를 나타낸 것이다. 그리고 식(2) 및 식(3)을 이용하여 정오분류표의 정분류율을 계산해보면 88.52%, 오분류율을 계산해보면 11.48%로 나타났다.

$$- \text{정분류율} = \frac{(25+29)}{61} \times 100(\%) = 88.52(\%)$$

$$- \text{오분류율} = \frac{(2+5)}{61} \times 100(\%) = 11.48(\%)$$

그리고 Fig. 5는 엔트로피 지수를 이용하여 의사결정나무모형 예측모델을 분석한 결과이다. 그림에서 보는 바와 같이 예측모델의 최상위 분리기준변수로는 사면경사가 선택되었으며, 하위 분리기준변수로는 포화도가 선택되었고, 최하위 분리기준변수로는 사면고도가 선택되었다. 급경사지재해 발생을 일으키는 사면경사의 기준은 17.9°인 것으로 나타났으며, 사면경사가 17.9°이상인 경우 급경사지재해 발생을 일으키는 토층의 포화도 기준은 52.1%인 것으로 나타났다. 그리고 토층의 포화도가 52.1%이상인 경우 급경사지재해 발생을 일으키는 사면고도의 기준은 320 m인 것으로 나타났다. 이와 같은 의사결정나무모형을 이용한 예측모델을 평가하기 위하여 정오분류표를 활용하였다. Table 6은 엔트로피 지수를

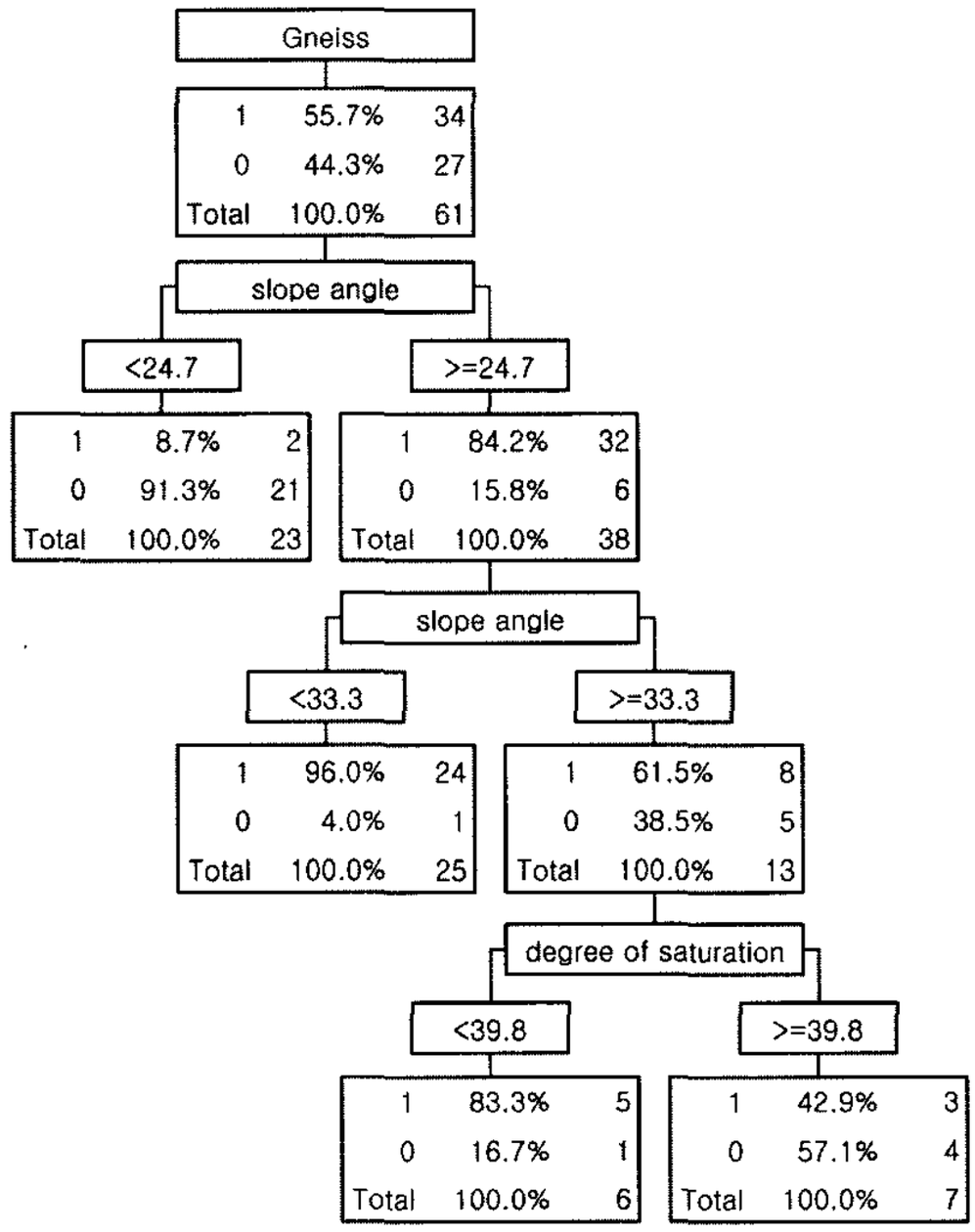

Fig. 4. Decision tree model resulting from gini index.

Table 5. Classification table of analysis result using gini index.

실제 관측된 값 \ 예측값	급경사지재해 미발생	급경사지재해 발생	합 계
급경사지재해 미발생	25	2	27
급경사지재해 발생	5	29	34
합 계	30	31	61

Table 6. Classification table of analysis result using entrophy index.

실제 관측된 값 \ 예측값	급경사지재해 미발생	급경사지재해 발생	합 계
급경사지재해 미발생	23	4	27
급경사지재해 발생	1	33	34
합 계	24	37	61

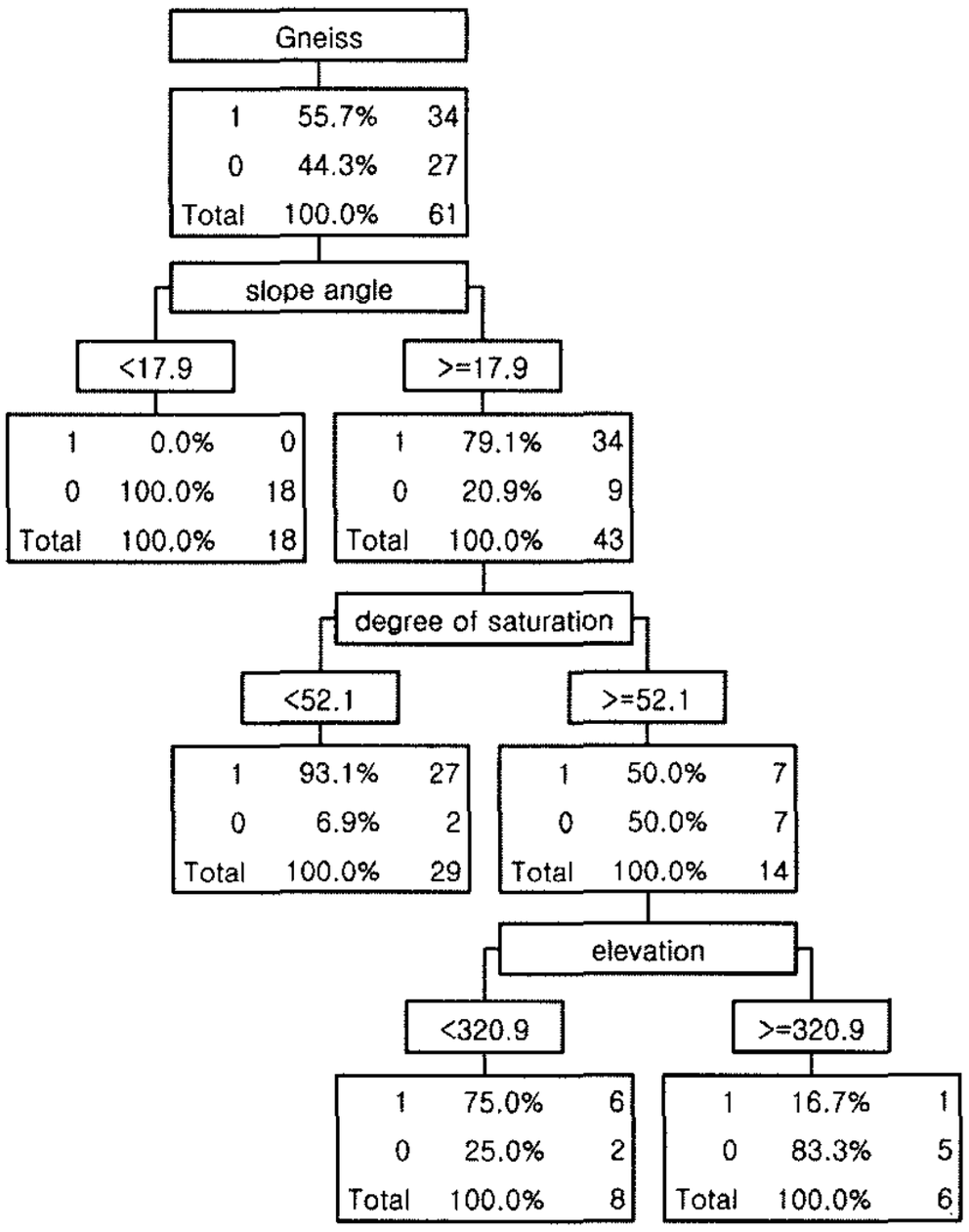

Fig. 5. Decision tree model resulting from entrophy index.

이용한 의사결정나무모형 예측모델의 정오분류표를 나타낸 것이다. 그리고 식(2) 및 식(3)을 이용하여 정오분류표의 정분류율을 계산해보면 91.80%, 오분류율을 계산해보면 8.20%로 나타났다.

$$- \text{정분류율} = \frac{(23+33)}{61} \times 100(\%) = 91.80(\%)$$

$$- \text{오분류율} = \frac{(4+1)}{61} \times 100(\%) = 80.20(\%)$$

이상의 분석결과를 살펴보면 Fig. 5의 급경사지재해 예측모델이 가장 정확도가 높은 것으로 평가되었으며, 분리기준도 합리적이라고 판단되었다. 따라서 이를 급경사지재해 예측모델로 선정 및 제안하였다.

결론 및 요약

본 연구에서는 의사결정나무모형을 이용한 편마암지역에서의 급경사지재해 예측모델을 개발하였다. 먼저 한국지질자원연구원에서 기 조사된 편마암 지역에서의 급경사지재해 발생지역 및 미발생지역에 대한 현장조사자료 및 토질시험자료를 토대로 통계적인 분석방법인 의사결정나무모형을 이용하여 급경사지재해 예측모델을 개발하였다. 이를 위하여 카이제곱 통계량, 지니 지수 및 엔트로피 지수를 활용하여 분석을 실시하였으며, 이들 결과를 정리하면 다음과 같다.

(1) 대상지역은 서울, 포천, 성동, 문산 등 서울 및 경기지역으로서 1998년 8월 4일부터 7일까지 최고 588.5 mm의 집중호우로 인하여 급경사지재해가 발생된 구간이다. 대상지역의 지질조건은 모두 편마암으로서, 총 104개소의 조사자료 가운데 급경사지재해 발생구간은 77개소이고, 미발생구간은 27개소이다.

(2) 편마암지역에서의 예측모델을 개발하기 위하여 활용된 조사자료수는 총 104개소 가운데 현장조사 및 토질시험 결측치를 제외한 61개소이다. 이 가운데 급경사지재해 발생구간은 34개소이고, 미발생구간은 27개소이다.

(3) 의사결정나무모형을 이용한 통계적인 분석은 카이제곱 통계량, 지니 지수 및 엔트로피 지수를 적용하여 실시하였다. 카이제곱 통계량을 이용한 분석결과의 경우 사면경사, 지니지수를 이용한 분석결과의 경우 사면경사 및 포화도, 그리고 엔트로피 지수를 이용한 분석결과의 경우 사면경사, 포화도 및 사면고도가 분리기준으로 선택되었다.

(4) 정오차분류법에 의한 예측모델의 정확성을 평가한 결과 엔트로피 지수를 이용한 의사결정나무모형 예측모델의 정분류율은 91.80%로서 가장 높게 나타났다. 따라서 해당 예측모델을 편마암지역 급경사지재해 예측모델로 선정하였다.

(5) 선정된 급경사지재해 예측모델의 분리기준은 최상위부터 사면경사, 포화도 및 사면고도의 순서로 선택되었으며, 각각의 분리기준치는 사면경사의 경우 17.9°, 포화도의 경우 52.1%, 사면고도의 경우 320 m로 결정되었다.

사 사

본 연구는 2006 건설기술혁신사업인 '국가 주요시설물 안

전관리 네트워크 시범구축 및 운영시스템 개발' 의 세부협동과제인 'GIS기반 급경사지 재해위험 취약지구 선정기법 연구' 의 일환으로 수행되었습니다.

참 고 문 헌

김원영, 채병곤, 김경수, 기원서, 조용찬, 최영섭, 이사로, 이봉주, 2000, 산사태 예측 및 방지기술연구, 과학기술부, 한국자원연구소, KR-00-(T)-09, 642p.

김원영, 채병곤, 김경수, 조용찬, 최영섭, 이춘오, 이철우, 김구영, 김정환, 김준모, 2003, 산사태 예측 및 방지기술연구, 과학기술부, 한국지질자원연구원, KR-03-(T)-03, 339p.

김종규, 사공명, 이준석, 이용주, 2006, 의사결정트리 기법을 이용한 터널 보조공법 선정방안 연구, 대한토목학회논문집, 26(4C), 255-264.

장윤경, 유병섭, 이동욱, 조숙경, 배해영, 2006, 공간 데이터의 분포를 고려한 공간 엔트로피 기반의 의사결정트리 기법, 정보처리학회논문지, 13-B(7), 643-652.

채병곤, 김원영, 이춘오, 김경수, 조용찬, 송영석, 2005, 지질조건에 따른 사태물질 이동특성 고찰, 지질공학, 15(2), 185-199.

최기헌, 1995, 데이터 마이닝: 개념 및 기법, 자유아카데미.

Biggs, D., de Ville, B. and Ville, E., 1991, A method of choosing multiway partitions for classification and decision tree, Journal of Applied Statistics, 18, 46-62.

Brand, E. W., 1981, Some thoughts on rainfall-induced slope failures, Proceedings of 10th International Conference on Soil Mechanics Foundation Engineering, Stockholm, The Netherlands, 373-376.

Kass, G., 1980, An exploratory technique for investigating large quantities of categorical data, Applies Statistics.

Olivier, M. Bell, F. G. and Jemy, C. A., 1994, The effect of rainfall on slope failure, with examples from the Greater Durban area, Proceedings of 7th intern. Cong. IAEG., 3, 1629-1636.

2008년 2월 2일 원고접수, 2008년 3월 10일 게재승인

송영석

한국지질자원연구원 지질환경재해연구부

305-350, 대전광역시 유성구 가정동 30

Tel: 042-868-3035

Fax: 042-868-3415

E-mail: yssong@kigam.re.kr

채병곤

한국지질자원연구원 지질환경재해연구부

305-350, 대전광역시 유성구 가정동 30

Tel: 042-868-3052

Fax: 042-868-3415

E-mail: bgchae@kigam.re.kr