Geomorphology 171–172 (2012) 94–100

Contents lists available at SciVerse ScienceDirect

Geomorphology

journal homepage: www.elsevier.com/locate/geomorph

Assessment of debris flow hazards using a Bayesian Network

Wan-jie Liang [a,b], Da-fang Zhuang [a,b], Dong Jiang [a,b,*], Jian-jun Pan [a], Hong-yan Ren [b]

[a] College of Resources and Environmental Sciences, Nanjing Agricultural University, Nanjing 210095, China
[b] State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

Article history:
Received 11 February 2011
Received in revised form 16 March 2012
Accepted 8 May 2012
Available online 15 May 2012

Keywords:
Debris flow hazard
Bayesian Network
Hazard assessment
Chinese mainland

ABSTRACT

Comprehensive assessment of debris flow hazard risk is challenging due to the complexity and uncertainties of various related factors. A reasonable and reliable assessment should be based on sufficient data and realistic approaches. This study presents a novel approach for assessing debris flow hazard risk using BN (Bayesian Network) and domain knowledge. Based on the records of debris flow hazards and geomorphological/environmental data for the Chinese mainland, approaches based on BN, SVM (Support Vector Machine) and ANN (Artificial Neural Network) were compared. BN provided the highest values of hazard detection probability, precision, and *AUC* (area under the receiver operating characteristic curve). The BN model is useful for mapping and assessing debris flow hazard risk on a national scale.

1. Introduction

A debris flow is a common geological hazard. It often begins with a landslide, and the potential energy of the generated sliding mass can rapidly convert into kinetic energy. Debris flows can induce a series of disasters that may pose a serious threat to lives, properties, and economic development. Many countries suffer from serious debris flow hazards. For example, large-scale debris flows occurred in Uganda on March 1, 2010, resulting in disastrous casualties, with 94 deaths, 320 people missing, and three buried villages.

China is one of the debris-flow prone countries. Debris flows occur in regions that correspond to 45% (10^6 km^2) of the Chinese mainland (Kang et al., 2004). For example, a debris flow hazard occurred in Zhouqu in Gansu Province on August 8, 2010, resulting in 1467 deaths, 298 missing people and direct economic losses of 425,000 RMB. Similar debris flow hazards also occurred in Chuxiong and Gongshan of Yunnan Province on the same day. Therefore, the assessment of regional debris flow hazards is of great significance for the sustainable development of China.

Various factors including topography, geology and climate influence geological hazards, and the measurement of these factors may involve large uncertainties (Kondratyev et al., 2006). Over the past few decades, numerous hazard analyses have employed qualitative and quantitative methods including artificial intelligence (AI). Qualitative approaches were widely used in the 1970s to 1990s and were based on the knowledge and opinions of experts (Carrara and Merenda, 1976; Rupke et al., 1988; Carmassi et al., 1992; Hearn, 1995; Pachauri et al., 1998). They have the following shortcomings: (i) evaluation tends to be subjective and assessment results from different experts are not comparable; (ii) updating assessment using new data is difficult; and (iii) required field experiments and investigations are expensive and time-consuming. In quantitative approaches, statistical analyses are adopted to solve the problem of subjectivity (Baeza and Corominas, 2001; Carrara, 2008). For example, Ayalew and Yamagishi (2005) adopted logistic regression for assessing landslide susceptibility; Guzzetti et al. (2005) introduced a probability approach to assess landslide hazard risk on a basin scale; and Calvo and Savi (2009) conducted Monte Carlo simulations for assessing debris flow risks. However, nonlinear relationships between the variables used cannot be solved by these approaches.

With the recent development of geographical information science, data mining and AI have been adopted in assessing geological hazards (Jiang and Eastman, 2000; Li et al., 2005). The techniques include ANN (Artificial Neural Network; Chang and Chao, 2006a,b; Chang, 2007; Chen et al., 2008; Gomez and Kavzoglu, 2005; Liu et al., 2005; Lu et al., 2007), SVM (Support Vector Machine; Wan and Lei, 2009; Yao et al., 2008), GA (genetic algorithms; Chang and Chien, 2007; Chang et al., 2009), and decision tree models (Saito et al., 2009; Wan, 2009; Wan and Lei, 2009). However, existing AI methods have three shortcomings: (i) limited use of prior knowledge makes it difficult to interpret assessment results; (ii) multiple sources of information cannot be integrated into a consistent system for assessment; and (iii) they are not good at dealing with the uncertainty of assessment.

BN (Bayesian Network) is an effective tool for knowledge representation and reasoning under the influence of uncertainty (Pearl,

* Corresponding author at: College of Resources and Environmental Sciences, Nanjing Agricultural University, Nanjing 210095, China. Tel.: +86 10 64889433; fax: +86 10 64855049.
E-mail address: jiangd@igsnrr.ac.cn (D. Jiang).

0169-555X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.geomorph.2012.05.008

1988; Reckhow, 1999). Because BN can present uncertainty interdependencies among random variables that are used to describe real-world domains, it has great potential for natural hazard assessment. Compared with other assessment methods, BN has several merits: (i) domain knowledge and multi-source information integrated into a consistent system; (ii) many flexible learning algorithms for searching optimal solutions; (iii) flexibility to include additional information; and (iv) decision support using nodes of functions and decisions. In this study, a novel method for assessing debris flow hazard risk based on BN and domain knowledge is proposed. Three debris flow hazard maps of the Chinese mainland from BN, ANN and SVM were produced and compared.

2. Data and methods

2.1. Assessment method based on Bayesian Network

A BN model can be expressed by (N, A, θ), where (N, A) is a directed acyclic graph (DAG) and θ is a parameter for a node. Each node $n \in N$ represents a domain variable (often corresponding to an attribute in the database), and each arc $a \in A$ between nodes represents a probabilistic dependency between the associated nodes. Each node $n_i \in N$ is associated with a conditional probability distribution, collectively represented by $\Xi \in \{\theta_i\}$, which quantifies how strongly a node depends on its parent node (Pearl, 1988). BN has great potential for natural hazard assessment. Compared with other AI methods such as ANN, a major advantage of BN is that they represent knowledge in a semantic way; and individual components such as specific nodes, arcs, or even values in the conditional probability tables have some meaning and can be understood independently (Greiner et al., 2001). This allows us to construct and interpret a network relatively easily.

A naïve BN model is a simple probabilistic classifier based on Bayes' theorem with a strong independence assumption, where all the attributes A_i are conditionally independent given the value of a class called C. By independence, we mean probabilistic independence, that is, A is always independent of B given C whenever $P_r(A|B,C) = P_r(A|C)$ and $P_r(C) > 0$ where P_r is probability (Nir et al., 1997).

Our debris flow hazard assessment using a BN model has the following six steps:

1) Selecting relevant parameters and spatial units;
2) Constructing training sample datasets for the model;
3) Learning and constructing the structure of the model;
4) Learning and determining the parameters for each node of the model;
5) Evaluating the performance and accuracy of the model; and
6) Using the model for assessment.

2.1.1. Learning structure of the BN model

To construct a BN model, the network that best matches a given training set needs to be found. The learning algorithms may be divided into two types: dependency analysis, and a scoring function with a search algorithm. The algorithm of the latter can be subdivided into two types: constraint-based and heuristic. The K2 algorithm (Gregory and Edward, 1992) is typically constraint-based, and conducts search according to the given node order with the limited maximum number of parent nodes. The main drawback of the K2 algorithm is that only the optimal structure within a limited search space can be found. The greedy hill-climbing algorithm (Lim et al., 2006) is heuristic and belongs to the local search family. It tends to fall into local optimization; to avoid this problem, the random mutation hill-climbing algorithm has been put forward (David, 1994). There are many other heuristic search algorithms such as the simulated annealing algorithm and GA (Renner and Ekart, 2003). In our method, an initial BN structure is obtained using the K2 search. The structure is then refined using domain knowledge to obtain a hazard assessment model.

2.1.2. Learning parameters of the BN-based model

Once a BN structure is constructed, parameters of CPT (conditional probability table) for each node can be obtained with two general approaches: using domain knowledge and using parameters learned from sample datasets. If sufficient knowledge regarding the mechanism of debris flow hazards is obtained, the CPT parameters can be determined by an expert. If enough training data are given, the parameters can also be derived from them. The two methods can be combined. The parameter learning algorithms include maximum likelihood estimation or Bayesian estimation. In this study, Bayesian estimation was utilized.

2.2. Factors for debris flow hazard assessment

Debris flow occurrence is affected by complex factors such as climate, geology, topography, and hydrology. Seven environmental factors were selected in this study to construct the assessment model of debris flow hazards for the Chinese mainland: X_1 — annual maximum cumulative rainfall of three consecutive days; X_2 — annual number of days with daily rainfall above 25 mm; X_3 — vegetation coverage index; X_4 — fault length; X_5 — area percentage of slope land with >25° inclination (APL_{25}); X_6 — maximum elevation difference of the basin; and X_7 — Gravelius index.

Rainfall is the main triggering factor of debris flow hazards, and debris flow occurrence is related to both current and antecedent rainfalls. Therefore, effective cumulative rainfall is useful for debris flow hazard assessment (Hsieh and Chen, 1993) although its calculation is difficult. It could be represented by the annual maximum cumulative rainfall of three consecutive days, and the annual number of days with daily rainfall above 25 mm could also indicate the rainfall intensity and concentration.

Slope is an essential and important factor of debris flow occurrence (Johnson and Rodine, 1984; Wang, 1994). According to Liu et al. (2005) and our field reconnaissance, most debris flows in China have initiated on slopes steeper than 25°.

Some land use/cover types especially vegetation with strong and large root systems increase slope stability (Dai and Lee, 2002). Franks (1999) indicated that sparsely vegetated slopes are the most susceptible to failure. Nilaweera and Nutalaya (1999) stated that vegetation provides hydrological and mechanical effects of slope stabilization. To incorporate the effects of land use/cover, we used the following vegetation coverage index, I:

$$I = a\left(\sum_{i=1}^{5} W_i \left(\sum_{j=1}^{n} SW_j S_j\right)\right)/S \tag{1}$$

where a is the normalization coefficient; W_i is the weight of the first class of land use (Table 1); SW_j is the weight of the subclass of land use (Table 1); S_j is the area of the subclass in an assessment unit; and S is the total area of the unit.

Fault zone development may provide weaker rocks and facilitate slope failure and debris production. Therefore, we measured the total length of faults in a basin.

The Gravelius index (Casali et al., 2008), K_g, can be another factor influencing debris flows:

$$K_g = P/2\sqrt{\pi A_b} = 0.28P/\sqrt{A_b} \tag{2}$$

where P is the basin perimeter (m) and A_b is basin area (m^2). K_g represents the ratio of the basin perimeter to the perimeter of a circle with the same area. Circular basins tend to have larger peak flow rates. Therefore, a basin with a Gravelius index close to one may often cause debris flows. Concerning drainage basin form, the maximum elevation difference of a basin was also chosen as a factor of debris flow hazard because it reflects potential energy.