Computers & Geosciences 42 (2012) 189–199

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

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

Susceptibility assessment of earthquake-induced landslides using Bayesian network: A case study in Beichuan, China

Yiquan Song [a,b], Jianhua Gong [a,*], Sheng Gao [c], Dongchuan Wang [d], Tiejun Cui [b], Yi Li [a], Baoquan Wei [e]

[a] State Key Lab of Remote Sensing Science, Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing 100101, China
[b] College of Urban and Environmental Science, Tianjin Normal University, Tianjin 300387, China
[c] Forestry Branch, Department of Natural Resources, NL, Canada
[d] Tianjin Institute of Urban Construction, Tianjin 30084, China
[e] National Marine Environmental Monitoring Center, Dalian 116023, China

ARTICLE INFO

Article history:
Received 8 March 2011
Received in revised form
16 September 2011
Accepted 19 September 2011
Available online 4 October 2011

Keywords:
Landslide
Susceptibility assessment
Bayesian network
Wenchuan earthquake

ABSTRACT

Because of the uncertainties and complexities of the factors involved in causing landslides, it is generally difficult to analyze their influences quantitatively and to predict the probability of landslide occurrence. In this work, a hybrid method based on Bayesian network (BN) is proposed to analyze earthquake-induced landslide-causing factors and assess their effects. Our study area is Beichuan, China, where landslides have occurred in recent years, including mass landslides triggered by the 2008 Wenchuan earthquake. To provide a robust assessment of landslide probability, key techniques from landslide susceptibility assessment (LSA) modeling with BN are explored, including data acquisition and processing, BN modeling, and validation. In the study, eight landslide-causing factors were chosen as the independent variables for BN modeling. And this study shows that lithology and Arias intensity are the major factors affecting landslides in the study area. On the basis of the a posteriori probability distribution, the occurrence of a landslide is highly sensitive to relief amplitudes above 116.5 m. Using a 10-fold cross-validation and a receiver operating characteristic (ROC) curve, the resulting accuracy of the BN model was determined to be 93%, which demonstrates that the model achieves a high probability of landslide detection and is a good alternative tool for landslide assessment.

1. Introduction

Landslides are one of the most common secondary disasters caused by earthquakes. Earthquakes with magnitudes greater than 4.0 can trigger landslides on unstable slopes, and earthquakes with magnitudes greater than 6.0 can generate widespread landslides (Jibson, 1993; Keefer, 1984). Based on data from the China Earthquake Networks Center (www.cenc.ac.cn), 316 earthquakes with surface wave magnitudes (Ms) greater than 6.0, 219 earthquakes with Ms greater than 7.0, and 13 earthquakes with Ms greater than 8.0 occurred globally from January 1, 2000 to January 1, 2011.

Analyzing the factors that contribute to landslides is useful for studying the mechanisms of landslide generation, building landslide models, mapping landslide susceptibility, and simulating or predicting their risk in landslide-susceptible areas. Currently, expert evaluation, data mining, and physical models are the three most commonly employed factor-analysis methods for landslide prediction. Based on a landslide inventory and historical information, the expert evaluation method can be used to evaluate and classify landslides, and determine the main contributing factors and disaster levels. The main disadvantage of the expert evaluation method is that many subjective factors are involved in the process. It is also difficult to ensure that the results are consistent (van Westen et al., 1997). The expert evaluation method provides a qualitative assessment, but it is hard to estimate the weight of each factor quantitatively. The data mining method, integrated with machine learning or statistical algorithms, can be used to determine the principal components leading to landslides, calculate the weights of the factors, and construct mathematical models or rules for landslide prediction. This method also provides quantitative or semiquantitative analysis of landslides. Commonly used data mining methods such as frequency ratio, logistic regression, and artificial neural networks (ANNs), have recently been applied to landslide prediction and analysis (Ayalew and Yamagishi, 2005; Chung, 2006; Lee et al., 2008; Regmi et al., 2010a, 2010b; Yilmaz, 2009). The main limitation of the data mining method is that the results are sensitive to the data quality. The physical model method was derived from the slope displacement model, which was first proposed by Newmark (Newmark, 1965) and is still widely used. Based on this model, the most dangerous sliding surface is determined based on the Factor of Safety, and the slope stability is then analyzed (Newmark, 1965). The advantage of the physical model method

* Corresponding author. Tel.: +86 10 64849299; fax: +86 10 64849299.
E-mail address: jhgong@irsa.ac.cn (J. Gong).

0098-3004/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2011.09.011

is that it originates from a Newtonian mechanical model and involves systematic mechanisms and theories to generate high-accuracy results. To calculate the slope stability, the physical model method requires strength parameters, failure depth, and groundwater conditions for every calculation point in the study area. This causes serious problems in data acquisition and in controlling the spatial variability of input parameters (Guzzetti et al., 1999). The physical model method is appropriate for assessing landslides in very small areas, whereas its capability for evaluation over a large area is insufficient.

The Ms 8.0 Wenchuan earthquake occurred on May 12, 2008 in Sichuan Province, China. The earthquake caused more than 15,000 instances of rapid mass movement including landslides, rock falls, and debris flows, resulting in approximately 20,000 deaths (Yin et al., 2009). Analyzing the factors that cause landslides and assessing the landslide susceptibility can prevent future geological disasters and aid in decision making for the reconstruction process, which has become more important and timely following the earthquake. Although many methods have been used in LSA, precision and accuracy remain unmet needs in landslide assessment. Therefore, new methods and mechanisms must be developed.

BN (Bayesian network, also known as belief network or directed acyclic graphical model) was first proposed by Pearl (Alpaydin, 2010; Pearl, 2000) based on Bayes' theorem, and is widely used in the field of complex systems modeling and considered to be excellent tools for the representation and inference of uncertain knowledge. BN models graphically and probabilistically represent correlative and causal relationships among variables (Marcot et al., 2006). Graphically, a Bayesian network is a directed acyclic graph in which the nodes represent variables and the links represent probabilistic dependence or independence between nodes (Aguilera et al., 2010). Compared with other data mining methods, BN have several distinct advantages (Uusitalo, 2007). It provides a natural way of handling missing data, allows for the combination of data with domain knowledge, facilitates learning about causal relationships between variables, and offers a method to avoid overfitting the data. BN can provide good prediction accuracy even with small sample sizes, and it can be easily combined with analytic tools to aid management. BN is widely used in analysis, uncertainty modeling, and decision support for resources and environmental, economic, and social problems (Aguilera et al., 2010, in press; Marcot et al., 2006). Unfortunately, to date, there have been no detailed studies using BN to analyze landslides.

The aim of this study was to develop a BN model that can be used to analyze the factors contributing to earthquake-induced landslides and assess landslide susceptibility. The detailed objectives of this study are as follows: (1) to prepare an inventory map of landslides and identify the factors causing landslides; (2) to select the key landslide-causing factors using a statistical approach; (3) to categorize continuous data using supervised discretization methods; (4) to develop a BN model of landslide susceptibility based on the selected factors; (5) to evaluate the relation between the landslides and the causing factors using marginal probability distributions; (6) to validate an accuracy assessment of the BN model with 10-fold cross-validation and ROC curve; and (7) to create a landslide susceptibility map of the study area.

2. Study area

Beichuan is a county in the jurisdiction of Mianyang Municipality, Sichuan, China. It has an area of 2868 km^2 and a population of 160,156 as of 2006. Our study area is a part of Beichuan County located in the center of the region affected by the Wenchuan earthquake. It is between 31°82′N to 31°87′N and 104°39′E to 104°49′E, with a total area of 1914 km^2 (Fig. 1).

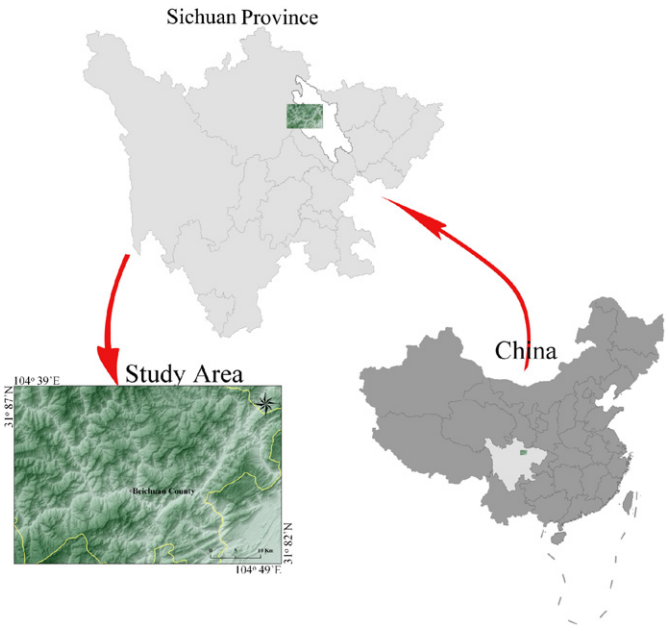

Fig. 1. Location of the study area.

The study area is situated in the transitional belt of the Sichuan Basin of the western Sichuan Plateau, and the topography is dominated by mountains. The elevation of the area generally decreases from the northwest to the southeast. The northwestern part of the area is characterized by high mountains, the central portion is characterized by a medium-relief landscape with rugged mountains and deeply incised valleys, and the southeastern portion has low- and medium-relief mountains. The valley slopes are generally steeper than 25°, with some slopes of 40–50° or more. Elevations in the study area range from 540 to 3033 m. The lowest point is the Jian River, and the highest mountain is Mount Qianfo. The Jian River, the main river in the area, originates from the mountainous area at the northwest of the county, flows to the southeastern corner and finally into the Pei River. The length and drainage area of the Jian River in the county are 48 km and 456 km^2, respectively.

The climate of the study area is predominantly humid subtropical with monsoons and an average annual temperature of 15.6 °C. The study area is in the heavy rainfall area of Lutou Mountain. The average annual precipitation is 1400 mm, with 83% of the annual rainfall between June and September. The average annual precipitation decreases from the southeast to the northwest.

The main strikes of the rock strata have a northeast–southwest orientation. The rocks in the study area are primarily composed of Cambrian sandstones and argillaceous limestones, Silurian slates and phyllites, and Devonian and carboniferous limestones. Loose Quaternary deposits are widely distributed in the form of terraces and alluvial fans. The Beichuan-Yinxiu Fault, which is active and ruptured during the Wenchuan earthquake, runs through the southeast portion of the study area. It is a northwest-dipping thrust fault with dip angles of 60–70°.

3. Methods

3.1. Working procedure

The working procedure of this study consisted of two main steps (Fig. 2): data acquisition and processing, and LSA modeling with BN.

The first step in data acquisition and processing was to collect spatial data and then integrate them into the same spatial reference frame. We applied the geometric correction with a polynomial model using ERDAS IMAGINE software, and then converted the coordinate systems of all data sets into the Universal Transverse Mercator projection, World Geodetic System 1984 (WGS-84-Zone-48N). Next, an event-based landslide inventory was established, and the landslide-causing factors were processed. Each factor was rasterized to a specific grid layer with grid size of 25 m × 25 m. The factors were then evaluated based on the statistical methods. The factor selection was based on the weights which were calculated with the factors related to landslides. After the factor evaluation and selection process, the factors with continuous values were discretized using a minimum description length (MDL) process.

After the data processing in the first step, the data were used to analyze the landslide-causing factors and to assess the study area for landslide susceptibility. Taking the landslide as the root node and the landslide-causing factors as the child nodes in a tree, the BN model was constructed in two steps: structure learning and parameter learning. The K2 algorithm was adopted for the BN structure learning, and the expectation maximization algorithm was adopted for the BN parameter learning. After the BN modeling, the a posteriori probability of landslide-causing factors and the Landslide Susceptibility Index (LSI) of each pixel in the study area were calculated with the BN Junction Tree inference engine. Based on the a posteriori probability, the influence of landslide-causing factors was analyzed both qualitatively and quantitatively. With a coordinate transformation, landslide-susceptibility mapping based on the LSI was obtained with spatial information and the format of it can be converted to Geographic Information System (GIS) grid images. Finally, the BN model used for the LSA was validated by combining the 10-fold cross-validation method with an ROC curve.

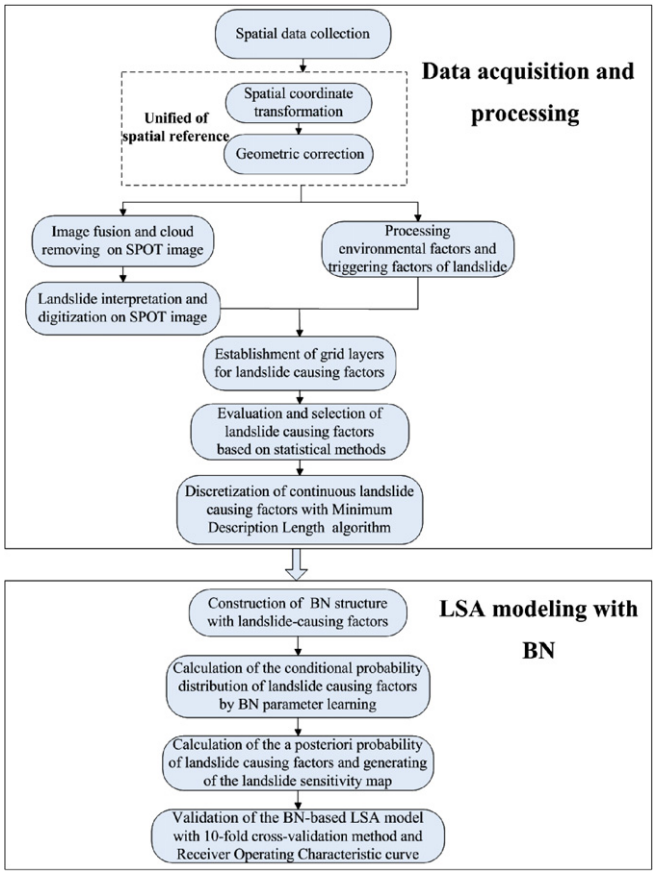

Fig. 2. Working procedure for BN-based landslide susceptibility analysis.

3.2. Data acquisition and processing

There are a variety of interrelated factors that affect landslides. The data required to prepare a map of landslide susceptibility can be divided into four groups: landslide inventory data, environmental factors, triggering factors, and elements at risk (van Westen et al., 2006).

3.2.1. Event-based landslide inventory

Postearthquake SPOT images were used as the main source of landslide detection, and preearthquake RS images and historical data were used as auxiliary information in landslide recognition (Table 1). Spectral, texture, shape, and additional feature characteristics were combined with expert prior knowledge and utilized in the visual interpretation of landslides. Terrain, roads, rivers, and settlements were also used in landslide recognition and to establish the event-based landslide inventory from May 12 to 18, 2008.

The dominant weather after the Wenchuan earthquake was rainy and cloudy. Therefore, SPOT images that covered the study area were partly clouded. FORMOSAT-2 and CBERS2B RS images were referenced to fill in the gaps of SPOT data due to clouds. Both multispectral (MS) and panchromatic (PAN) images were used in landslide interpretation. Image fusion, which is based on the multiplicative resolution merge method and the bilinear interpolation resampling technique in ERDAS IMAGINE, was utilized to produce a high resolution, true-color composite image to facilitate landslide recognition. Thin clouds in RS images were also reduced with ERDAS to improve the visualization. The complete preearthquake and postearthquake landslide inventory was then obtained by digitizing the data in ArcMap. Then preearthquake landslides were removed from the event-based landslide inventory.

Subsequently, an event-based landslide inventory map between May 12, 2008 and May 19, 2008 was generated (Fig. 3), resulting in the identification of 563 earthquake-triggered landslides in the study area and a total landslide area of 35.6 km^2. Most of the landslides are located within a zone that stretches along the Beichuan-Yingxiu Fault. In some locations, multiple landslides blend together, forming landslide complexes with no clear distinction between individual landslides. The most common types of landslides in the study area are falls (including rock

Table 1
RS images used in the event-based landslide inventory.

Image description	Time	RS Image type	Image resolution
Postearthquake RS image	05/18/2008	SPOT	2.5 m PAN and 10 m MS
	05/19/2008	FORMOSAT-2	2.0 m PAN and 8 m MS
	06/27/2008	CBERS2B	2.36 m HR (high resolution) and 19.5 m CCD images
Preearthquake RS image	11/10/2006	SPOT	2.5 m PAN and 10 m MS

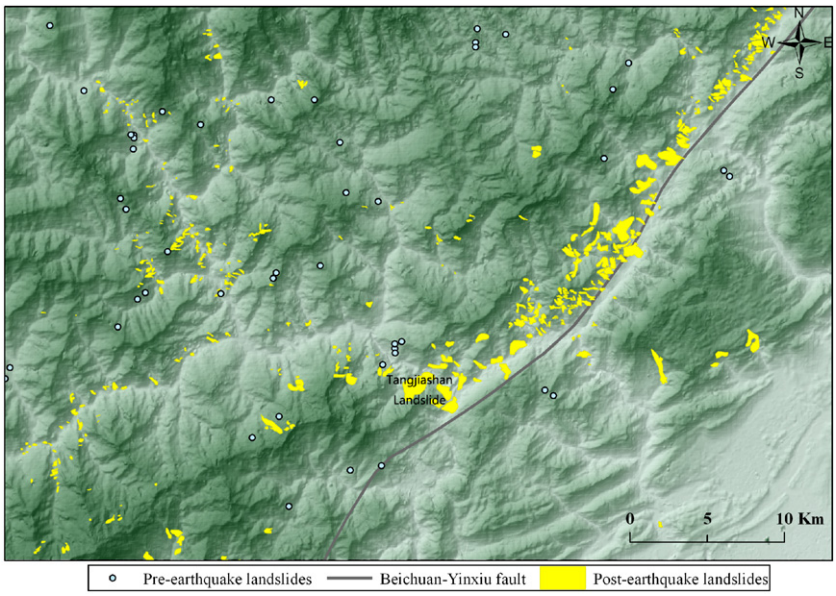

Fig. 3. Landslides triggered by the Wenchuan earthquake within the study area. The yellow regions are the postearthquake landslide areas, and the blue points are the locations of preearthquake landslides. The background image is a DEM. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

falls and debris falls), slides (rock slides, debris slides, and earth slides), and flows (debris flows). The dominant types of coseismic landslides are rock falls and rock slides, whereas earth slides were much less frequent. Most of the landslides in the study area occurred in Silurian slates and phyllites, Cambrian sandstones, and Devonian limestones. The largest landslide in the study area, the Tangjiashan landslide, occurred 5.5 km upstream from the city of Beichuan, along the Jian River. This landslide occurred on a dip slope of shale and slate and formed a river-blockage mass across the valley of the Jian River.

Finally, the postearthquake landslide data were rasterized onto a raster grid layer with a cell size of 25 m × 25 m using the cubic spline interpolation in ArcMap. A value was assigned to each pixel: 1 to indicate the presence of landslide and 0 to indicate the absence of landslide. The same grid cell size (25 m × 25 m) was used in the landslide-causing factor rasterizing process.

3.2.2. *Processing for landslide environmental factors*

Dozens of different landslide-causing factors can be used for LSA, e.g., geomorphic data, land cover and land-use data, geologic data, and hydrology data (van Westen et al., 2008). Eleven environmental factors were selected as landslide environmental factors: slope, aspect, relief amplitude, curvature, terrain roughness, lithology, distance from a fault, distance from a road, distance from a river, the Normalized Difference Vegetation Index (NDVI), and land cover (Table 2).

Landslides are closely related to terrain parameters. Five terrain factors were utilized in the study (Table 2). The digital elevation model (DEM) data used are the advanced spaceborne thermal emission and reflection Radiometer (ASTER) Global digital elevation model data. The preproduction accuracy of the DEM data was 20 m for vertical data and 30 m for horizontal data, both at a 95% confidence level (Hirano et al., 2003). The terrain parameters used were processed using the 3D extension tool in ArcMap.

Geological maps play an important role in LSA. With the impact of rainfall or an earthquake, different geological units display remarkable differences in landslide susceptibility. The geological maps used in this study were provided by the Geological Survey of China and the scale was 1:50,000. Two data layers were used: lithology and fault. The lithology in the study area is primarily composed of phyllite, slate, sandstone, limestone, and sandstone–shale. The Beichuan-Yingxiu Fault, the major causative fault in the Wenchuan earthquake, runs through the study area.

Three distance factors were included in the LSA: distance from a fault, distance from a road, and distance from a river. The distances were calculated using ArcMap spatial analysis tools. Near river channels or faults, the probability of slope failures is higher. Therefore, the landslide hazard level is closely related to the distance from a river or a fault. Furthermore, road construction cuts slopes and affects the stability of slopes and road drainage systems, which increases the chances of landslides nearby. As smaller roads have little influence on slope failures, only expressways, national highways, and provincial highways were considered in this study.

Land cover is also a key factor associated with the occurrence of landslides. The land cover data used were provided by the Data Sharing Infrastructure for Earth Science Systems (www.geodata.cn).

Table 2
Environmental data used in the landslide analysis.

Data classification		Data format	Scale or resolution
Terrain parameters	Slope	Geotiff	30 m
	Aspect	Geotiff	30 m
	Terrain roughness	Geotiff	30 m
	Relief amplitude	Geotiff	30 m
	Curvature	Geotiff	30 m
Geological map	Lithology	.Shp polyline	1:50,000
	Fault	.Shp line	1:50,000
Environmental elements	River	.Shp line	1:10,000
	Road	.Shp line	1:10,000
Land cover		.Shp polyline	1:25,000
NDVI		Geotiff	250 m

The standard classification of land cover consists of six types (forest and shrub, grassland, farmland, residential areas, wetlands, and water bodies) and 25 subtypes. The study area includes 10 subtypes: deciduous forest, deciduous broadleaf forest, coniferous and broadleaf mixed forest, brush, shrubland, irrigated land, dry land, urban, village, and inland water; the main subtype is coniferous and broadleaf mixed forest.

The NDVI data are from the MODIS AQUA data set. The data are 16-day, Level-3 250 m composite data (MYD13Q1). The processing tool we used is the MODIS Reprojection Tool. The NDVI is given by

$$NDVI = (\rho_{nir} - \rho_{red})/(\rho_{nir} + \rho_{red}), \quad (1)$$

where ρ_{nir} and ρ_{red} are the second- and the first-band reflectivities of the MODIS image, respectively. The NDVI is closely related to the vegetation cover. The effect of vegetation on the LSI is complex, and it is determined by the interaction of four different factors: mechanical stabilization due to the presence of roots, soil moisture depletion as a result of transpiration, surcharge from the weight of trees, and wind-breaking (Nilaweera and Nutalaya, 1999). Overall, the large area of forest cover has a relatively low probability of landslides.

3.2.3. Processing for landslide triggering factors

Earthquake-induced landslides are affected by the ground motion amplitude, spectrum, and duration of the earthquake, and the impact assessment can be represented by the peak ground acceleration (PGA), peak ground velocity (PGV), or Arias intensity (AI). The probability of earthquake-induced landslides is proportional to the Newmark displacement (Jibson et al., 2000). Among these three parameters, AI is directly related to the Newmark displacement (Harp and Wilson, 1995) and is the most relevant to the study of earthquake-induced landslides (Wang et al., 2010). Thus, we adopted AI as the triggering factor for landslides. AI is given by

$$I_\alpha = \frac{\pi}{2g}\int_0^{T_d} (\alpha(t))^2 dt, \quad (2)$$

where I_α is the value of AI, $\alpha(t)$ is the ground acceleration as a function of time, T_d is the total duration of the strong-motion record, and g is the gravitational acceleration (Arias, 1970). AI is measured in units of velocity (m/s).

The landslide inventory does not including precise timing information for each landslide, and the relationship between distance and AI is approximately linear. We used a simplified alternative method to determine the AI information for each pixel in the study area. Based on the seismic waveform data for each earthquake, the AI value of each observation station point was calculated. In this step, the earthquakes with Ms greater than 5.5 from May 12 to May 19, 2008 were selected (Table 3). Ordinary kriging, which is an interpolation algorithm based on the theory of regionalized variables, was used to derive the AI maps corresponding to each earthquake listed in Table 3, covering the study area. The spatial overlay was conducted with the AI maps obtained in the previous step, and the AI value used in the LSA is the result of the spatial overlay. The kriging interpolation and spatial overlay operations were performed in ArcMap.

3.2.4. Evaluating and selecting landslide-causing factors

Different factors contribute differently to landslides, and only the key factors are required in the LSA model (Carrara, 1983; Fell, 1994). To determine the key factors causing landslides, a hybrid method combining correlation analysis, gain ratio (GR) (Quinlan, 1993), and principal component analysis (PCA) was used. The GR and PCA were calculated using the "Select Attributes" tool in Weka, which is a popular open-source machine learning software written in Java (Witten and Frank, 2011).

Table 3
Parameters of the main shock and six major aftershocks of the Wenchuan earthquake.

Earthquake time (GMT+8)		Epicenter location		Depth (km)	Magnitude (Ms)
Data	Time	Latitude (°N)	Longitude (°E)		
05/12/2008	14:28:04.0	31.0	103.4	14	8.0
05/12/2008	14:43:15.0	31.0	103.5	33	6.0
05/12/2008	19:10:58.4	31.4	103.6	33	6.0
05/13/2008	04:08:50.1	31.4	104.0	33	5.7
05/13/2008	15:07:11.0	30.9	103.4	33	6.1
05/14/2008	10:54:36.5	31.3	103.4	33	5.6
05/16/2008	13:25:49.0	31.4	103.2	33	5.9
05/18/2008	01:08:23.4	32.1	105.0	33	6.0

The first step was to detect the correlation coefficient between each pair of landslide-causing factors and calculate the GR value of the landslide-causing factor with respect to the landslide. The correlation coefficient qualitatively measures the strength and direction of a linear relationship between two factors; it can be used to avoid multicollinear factors. The correlation coefficients range from -1 to $+1$, varying from negative correlations to positive ones. GR, introduced by Quinlan in 1993, is an important function to efficiently and effectively assess the correlation of an attribute with a feature class (Quinlan, 1993). A larger GR value indicates a stronger relationship between a landslide-causing factor and a landslide. The GR value of a landslide-causing factor f_a with respect to a landslide L can be calculated as

$$GR(L, f_a) = (H(L) - H(L|f_a))/H(f_a), \quad (3)$$

where the function $H()$ is the entropy of the data set (Witten and Frank, 2011). If the absolute value of correlation coefficients between the two landslide-causing factors is greater than 0.8, we assume that it indicates a strong correlation between factors, and only factors with large GR values were selected for the LSA.

In the second step, based on PCA, a further selection was done for the remaining landslide-causing factors of the first step. PCA is a typical statistical method that has been widely used in data analysis and dimension-reduction processing to transform a large number of possibly correlated variables into a smaller number of uncorrelated variables. The results of PCA can be represented by a number of principal components, and each of them can be explored in terms of component scores and loadings. In this study, if the absolute value of factor loading was greater than 0.4 for any principal component, the factor was selected for LSA modeling.

Based on the two steps above, relief amplitude, lithology, distance from a fault, distance from a road, distance from a river, NDVI, land cover, and AI were selected for the LSA. The GR and PCA values of the selected factors are shown in Tables 4 and 5, respectively.

3.2.5. Discretizing the continuous landslide-causing factors

The ability of BN to treat continuous data is limited (Jensen, 2001). The processing of continuous data in BN can be divided into two categories: discrete or given a probability distribution. Due to the limits of the study area, the probability distributions of the data cannot usually be accurately defined. Discretization is a better solution to handle continuous landslide-causing factors in BN (Dougherty et al., 1995).

Among the eight selected landslide-causing factors, the attribute values of lithology and land cover were defined as discrete

Table 4
GR values of the selected landslide-causing factors.

Factor	NDVI	Distance from a fault	Lithology	Distance from a river	Distance from a road	Relief amplitude	Land cover	AI
GR	0.02623	0.01901	0.03979	0.01105	0.01803	0.00454	0.00774	0.01426

Table 5
PCA values of the selected landslide-causing factors.

Principal components index	Factor loading (greater than 0.4)							
	NDVI	Distance from a fault	Lithology	Distance from a river	Distance from a road	Relief amplitude	Land cover	AI
1	–	0.548	–	–	–	–	–	0.493
2	–	–	–	0.518	0.42	–	−0.499	–
3	–	–	–	–	–	−0.926	–	–
4	0.765	–	–	–	–	–	–	–
5	–	–	−0.846	–	–	–	–	–
6	0.443	–	–	–	–	–	0.764	–

Table 6
Results of the discretization of the six continuous landslide-causing factors.

Factor	Cut points	Number of intervals
Relief amplitude	23.5, 36.5, 56.5, 116.5	5
AI	109.1, 109.3, 109.5, 111.4, 112.1, 112.3, 113.3, 113.8, 114.2, 114.7, 115.1, 115.4, 115.9, 116.4, 116.7, 116.9, 117.3, 117.9, 118.3, 120.7, 122.1	22
Distance from a fault	192.8, 315.7, 399.1, 1325.1, 1769, 1990.5, 2756.6, 3022.6, 3258.9, 3405.0, 3683.6, 4481.2, 4606, 4710.4, 5082.7, 5373.7, 7480.4, 7810.6	19
Distance from a river	860.5, 1325.1, 1947.2, 2219, 2276.8, 2496.5, 2852.9, 3076, 3283.6	10
Distance from a road	72.8, 77, 131.0, 147.8, 155, 609.5, 1051.3, 1219.5, 1343.1, 1598.2, 2049.0, 2121.7, 2229.9, 2355.5	15
NDVI	6009.5, 6166.5, 6246.5, 6385.5, 6477.5, 6557.5, 6637.5, 6720.5, 6836.5, 6917.5, 6999.5, 7099.5, 7188.5, 7278.5, 7360.5, 7468.5, 7549.5, 7630, 7711, 7847, 7950, 8036.5, 8146, 8231, 8327, 8508	27

and were classified into predefined categories according to expert knowledge, whereas the values of relief amplitude, distance from a fault, distance from a road, distance from a river, and NDVI were defined as continuous. Supervised discretization methods and a Weka discretization filter (Witten and Frank, 2011) based on a minimum description length (MDL) algorithm (Fayyad, 1993) were used to discretize the six continuous landslide-causing factors into categories with appropriate states (Table 6).

Given a landslide-causing factor f_a, by searching the cutoff point from the maximum attribute value to the minimum using the MDL algorithm, f_a can be divided into two adjacent categories, f_a^1 and f_a^2, when the value of information gain is minimized. Information gain can be calculated as

$$E = \frac{|f_a^1|}{|f_a|} Ent(f_a^1) + \frac{|f_a^2|}{|f_a|} Ent(f_a^2), \tag{4}$$

where $|f_a'|$ is the size of the expressed sequence and $Ent()$ is the information entropy. When the value of E is at a minimum, a cut point can be obtained forf_a. After a recursive search of the two adjacent categories f_a^1 and f_a^2, the cut points for f_a are obtained when the resulting information gain is smaller than a specified value (Table 6).

3.3. LSA modeling with BN

3.3.1. Overview of BN

The BN is a graphical model for probabilistic relationships among a set of variables, and the representation of a BN consists of two components: (G,P). The first component $G=(X,L)$ is a directed acyclic graph (DAG). The nodes $X=(X_1,...,X_v)$ of the graph G correspond to the variables (landslide-causing factors) used in the LSA. The links (arcs) $L \in X_v \times X_v$ of the graph G connect a set of directed edges among the nodes, which represent the direct causal influences of the linked variables, and the strengths of these influences are expressed by conditional probabilities (Pearl, 2000). The second component,P, describes a conditional distribution for each node given its parent nodes in G. The conditional distribution is typically specified by a conditional probability table (CPT). $BN=(G,P)$ is a Bayesian network with respect to G if it satisfies the local Markov property. If X_v represents a node in the BN, then the joint probability distribution of Xcan be given by

$$P(X) = \prod_{X \in V} P(X_v | X_{p\alpha(v)}), \tag{5}$$

where $p\alpha(v)$ is a set of parents of v (Russell and Norvig, 2003).

BN modeling consists of two main steps. The first step is constructing the structure of the BN, which can be used to specify the conditional dependence relationships between the variables. The BN structure may be defined with a priori expert knowledge or calculated using a structure learning algorithm and a training data set, or it may be determined using a combined approach. The next step is calculating the conditional distribution of each node in the BN structure. Here, BN parameter learning was the main method used to specify the CPT of each node. Once the BN models were constructed, the BN was used to calculate the posterior marginal distribution of the variables with the inference engine.

BNT (Bayes Net Toolbox for MATLAB) (Murphy, 2001) is an open-source MATLAB package for BN that includes all parts of BN

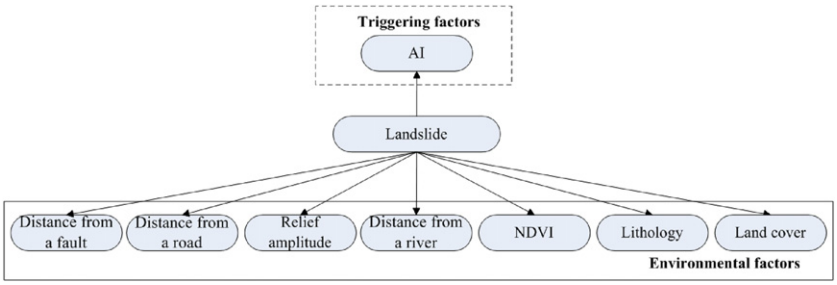

Fig. 4. Initial BN network used for the LSA. The initial network is a naive BN, the landslide is the root node, and each landslide-causing factor is a child node of the landslide.

modeling, such as structure learning, parameter learning, and inference engine. BNT was used in this study to develop the BN model for the LSA.

3.3.2. Constructing the BN structure used for the LSA

The BN structure was constructed under the conditions that the structure and probability contain uncertainty. The goal was to obtain conditional independence between landslide-causing factors and landslides. In this study, the tree-augmented naive Bayes (TAN) BN structure was chosen because it provides better prediction accuracy and a better representation of the correlation among variables than other classic models such as the naive Bayes model.

The initial network used for structure learning is a naive Bayes model, which contains an arrow from the landslide node to each of the landslide-causing factor nodes (see Fig. 4). The K2 algorithm (Cooper and Herskovits, 1992) is a classic search and score-based algorithm in the BNT structure learning algorithm and was adopted in this study. With the initial naive BN structure, the K2 algorithm adds a parent node to each landslide-causing factor at random with the hill climbing search algorithm (Cooper and Herskovits, 1992), allowing all the possible structures to be described. The Bayesian information criterion (BIC) (Schwarz, 1978) is then used to choose a better structure.

3.3.3. Calculating the conditional probability distribution of landslide-causing factors in the BN model

The conditional probability distribution of landslide-causing factors was determined by BN parameter learning. Using the previously constructed structure, the goal of BN parameter learning is to calculate the CPTs and update the prior probability distribution of landslide-causing factors in the existing BN model. The expectation maximization (EM) algorithm was adopted in BN parameter learning (Dempster et al., 1977).

The EM algorithm is a maximum-likelihood estimation method used to calculate parameter values from incomplete data (Dempster et al., 1977). Beginning with an arbitrary initial parameter assignment θ_0 and then applying the Expectation Step (E-step) and the Maximization Step (M-step) repeatedly, the EM algorithm seeks to find the maximum-likelihood estimation of the marginal likelihood. Given a likelihood function $L(\theta;X,Z)$, where θ is the parameter vector, X is the observed data, and Z represents the unobserved latent data or missing values, the E-step estimates the conditional distribution of Z given X, using the $\theta(t)$ values from the last M-step. The M-step finds the parameter that maximizes the value of the likelihood function.

3.3.4. Calculating the posterior probability of landslide-causing factors and generating the landslide sensitivity map

The links between nodes in a BN have a strong conditional independence relationship. When a node's parent is determined, the parent is independent of nodes other than the child node. This feature can be used to calculate the a posteriori probability of a variable when other evidence variables are observed. Thus, the BN model can be used to predict the impact of introducing evidence for certain variables by a posteriori probability. The posteriori probability is computed using the BN inference engine.

The junction tree (JT) (Lauritzen and Spiegelhalter, 1990) is one of the more popular inference engine algorithms and has been adapted for computing the marginal distributions of landslides. A JT consists of four steps: clustering nodes into cliques, connecting the cliques to form a junction tree, propagating information through the network, and answering a query. After the four steps are completed, the JT algorithm translates the BN into a junction tree, and using the message propagation through the tree, the probability of each node in BN can be calculated.

Suppose that only one landslide-causing factor F was observed, and the other factors were not observed. If the evidence was a 1D array and only one value F_i was assigned such that $F_i \in F$, then the a posteriori value of F could be obtained with a JT. By repeating this process, the probability distributions of all the landslide-causing factors were obtained.

Assuming that all the landslide-causing factors were observed, and the joint distribution values of each pixel in the study area could be calculated using a JT with the corresponding factor values, then the joint distribution value is the LSI. The value of the LSI ranges from 0 to 1, representing the susceptibility index for a landslide. The landslide susceptibility increases with increasing LSI value. With a coordinate transformation, the LSI values in the study area can be converted into a GIS grid format image. By classifying the LSI values, a landslide sensitivity map can be obtained.

3.3.5. Validating the BN model used for the LSA

The aim of the validation in this study was to verify the accuracy of the landslide inference results. The LSI value and the landslide value (i.e., the presence of a landslide (1) or the absence of a landslide (0)) were compared to evaluate the performance of the BN model used for the LSA. The models in the study were validated using the 10-fold cross-validation method (Stone, 1974) and an ROC curve (Hanley and McNeil, 1982).

With the 10-fold cross-validation method, the initial training data used for the LSA were divided into 10 subsets. One subset was used for the BN model validation, and the other 9 subsets were used to train the BN model. By repeating the cross-validation 10 times, each subset was tested once, and a total of 10 results were obtained.

By setting a threshold value, the LSI value can be interpreted as a variable with only two values: the presence of a landslide (1) or the absence of landslide (0). The threshold value has large influence on the classification results. When the threshold value is reduced, more landslide areas can be recognized, which

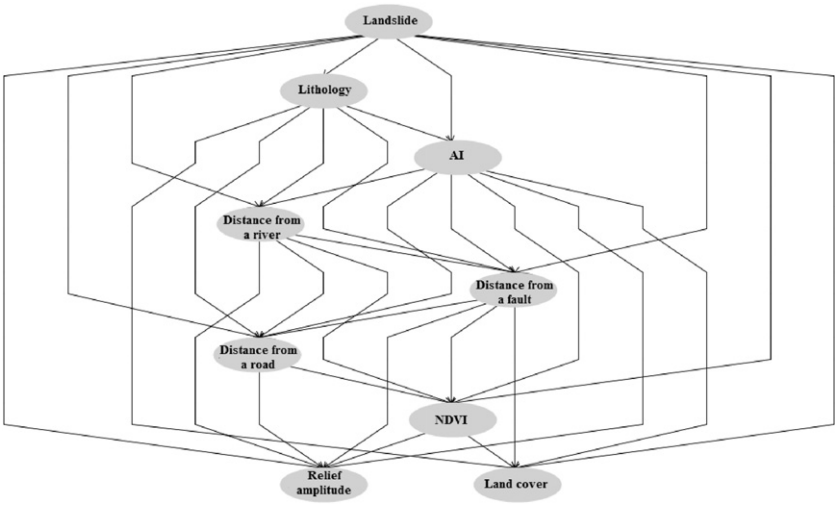

Fig. 5. Resulting structure of Bayesian networks for the landslide-susceptibility assessment.

Table 7
Conditional probability of the node lithology.

The value of the parent node (landslide)	Lithology				
	Phyllite	Slate	Sandstone	Limestone	Sandstone–shale
Absence of landslide (0)	0.85	0.90	0.88	0.71	0.66
Presence of landslide (1)	0.15	0.10	0.12	0.29	0.34

improves the identification of landslides and increases the true positive rate. However, some stable slopes will be mistaken as landslides, which mean that the false positive rate also increases.

To minimize the effect of the threshold value, an ROC was adopted to analyze the results of the cross-validation. The accuracy of the BN model is reported as the average of the 10 results. In essence, the ROC is a sensitivity or specificity curve that reflects the dynamic changes in the classification results with varying threshold values. The interpretation of the ROC curve is made by calculating the area under the curve (AUC), i.e., the area between the horizontal axis and the ROC curve; this area can be used to evaluate the performance of models, and its value ranges between 0.5 and 1.

4. Results and discussion

4.1. The BN model used for the LSA and validation

The BN model used for the LSA contains two main components: a qualitative component and a quantitative component. The qualitative component of the BN model is a DAG (Fig. 5) with nodes representing the landslide-causing factors used in the LSA and links representing the direct causal influences between the linked nodes. The BN structure reveals the complexity of the relationships between the landslide-causing factors that form the basis of the BN model. In Fig. 5, the landslide is the root node of the BN structure, and each landslide-causing factor is directly linked to it. Most of the links between the landslide-causing factors contain two variables connected in the following order: lithology and AI. This tie indicates that both the lithology and the AI have a large effect on landslides in the study area. The quantitative component of the BN model consists of the CPTs of each landslide-causing factor, which were calculated using BN parameter learning. The CPT of each child node was specified by assigning each combination a value selected from the parent nodes. The CPT of the node lithology is presented as an example in Table 7.

The left column is the value of the parent node of the lithology (landslide), and the five lithology columns contain the corresponding probability values of the five lithology categories relative to the parent node.

Two columns of data were obtained from the 10-fold cross-validation: the landslide value and the landslide probability calculated using the BN inference engine. And the ROC curves are shown in Fig. 6. In this study, the area under the ROC curve is 0.93, which means that the equivalent accuracy rate is 93%.

4.2. Probabilities associated with landslide-causing factors

Fig. 7 shows the probability distributions of the landslide-causing factors, allowing the quantification of relationships between landslide-causing factors and the probability of a landslide. For the discrete landslide-causing factors (lithology and land cover), the values shown in Fig. 7 are integer indexes. The index values of the lithology categories are phyllite (1), slate (2), sandstone (3), limestone (4), and sandstone–shale (5). The index values of the land cover categories are deciduous forest (1), deciduous broadleaf forest (2), coniferous and broadleaf mixed forest (3), brush (4), shrubland (5), irrigated land (6), dry land (7), urban (8), village (9), and inland water (10). For the six continuous landslide-causing factors, the categories are divided by the cut points listed in Table 6. The category index is calculated starting at one, and the value is increased by one if it enters the next category.

Based on Fig. 7, the occurrence of landslides is highly sensitive to relief amplitude when the relief amplitude is above 116.5 m (category value is 4), with an increasing belief (36%) for relief amplitude between 56.5 m (category value is 3), and 116.5 m. AI also has a large, nonlinear influence on landslides. As the AI value increases, the landslide probability fluctuates, but the overall trend is downward. This is consistent with the findings of Wang et al. (2010), who studied the relationship between AI and landslides induced by the Wenchuan earthquake. Based on Fig. 7, the highest landslide probability (34%) occurs when the AI is between 109.3 and 109.5 m/s. The landslide probability also varies as a function of the force of the earthquake and the

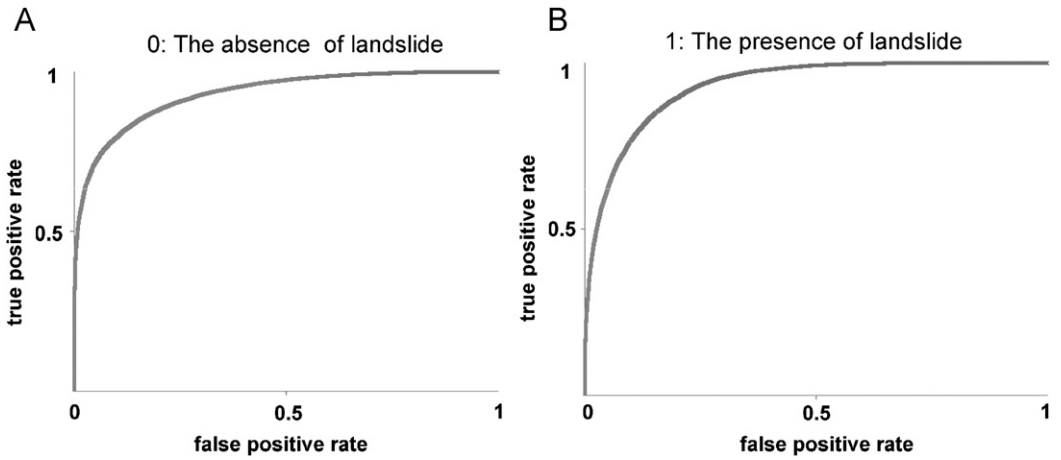

Fig. 6. ROC curve of the BN model used to LSA. (a) The absence of landslides; (b) the presence of landslides.

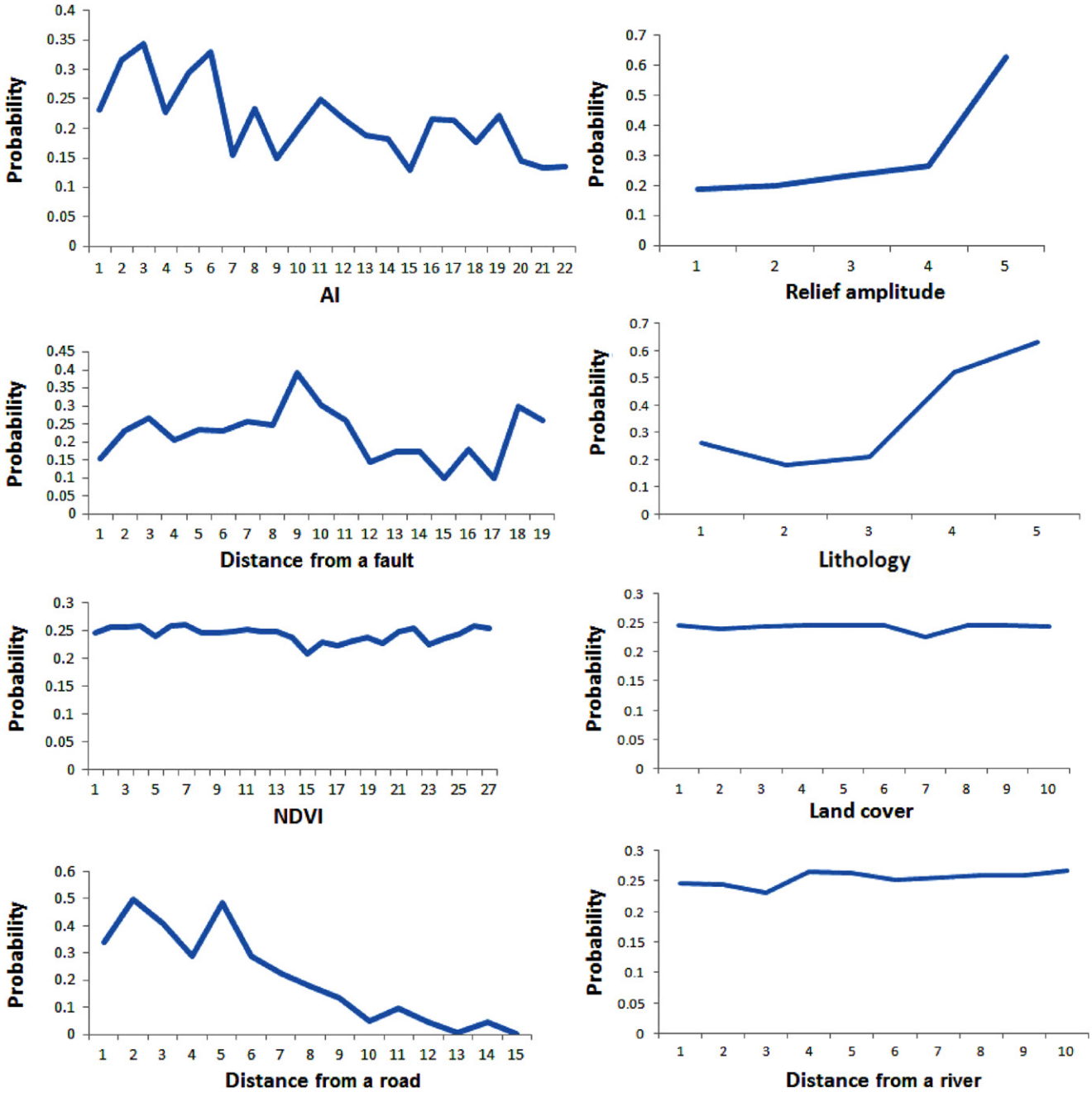

Fig. 7. Probability distributions of landslide-causing factors. The *x*-axis shows the value of each landslide-causing factor, and the *y*-axis shows the landslide probabilities corresponding to the factors.

fractures through different lithologies. Based on the trends shown in Fig. 7, the occurrence of landslides in the study area is most influenced by sandstone–shales (63%) and limestones (52%). In the study area, landslides mostly occurred in the area surrounding the fault zone, with the chance of landslides decreasing with an increasing distance from the causative fault. The landslide probability is highest at a distance of 3 km from the fault. The landslide probability increases as the distance from a road increases, which illustrates that the landslides in the study area have destroyed a large number of roads. The landslides were mainly caused by the earthquake, whereas land cover and distance from a river had only small impacts on landslides.

4.3. Spatial distribution of landslide susceptibility index

Fig. 8 presents the spatial distribution of LSI in the study area. Based the susceptibility index, the grid cells were grouped into five categories. Most regions in the western part of the study area had low landslide susceptibility. The regions with the very high susceptibility area in Fig. 8 are represented in red color, and most are located in the northeast of the study area. Very high and high landslide susceptibility index values are found in most areas close to the Beichuan-Yinxiu Fault, indicating that rock movement in the future will increases the probability of landslides.

Table 8 shows the mathematical statistics for the susceptibility index categories. With the Table 8, we can see that the analysis of this study resulted in a very high susceptibility area of 20.2 km^2 (i.e., 1.0% of the total study area), a high susceptibility level area of 21.4 km^2 (i.e., 1.1%), a moderate susceptibility area of 27.7 km^2 (i.e., 1.4%), and a low susceptibility area of 91.2 km^2 (i.e., 4.8%)

5. Conclusions

In this study, a BN approach was adopted for modeling the occurrence of earthquake-induced landslides. The landslide inventory was extracted from the high-resolution SPOT images, and 563 earthquake-triggered landslides were identified in the study area, with a total landslide area of 35.6 km^2. Eight landslide-causing factors (relief amplitude, lithology, distance from a fault, distance from a road, distance from a river, NDVI, land cover, and AI) were chosen as the independent variables for BN modeling. A BN model was developed through data training to determine the TAN structure, and the dependent relations between variables were quantified. Lithology and AI were confirmed as important factors contributing to landslides in the study area. Based on the a posteriori probability distribution, the occurrence of a landslide is highly sensitive to relief amplitudes above 116.5 m. AI values between 109.3 and 109.5 m/s and the sandstone–shale and limestone rock types also have significant influences on landslides in the study area. Model validation showed that LSA modeling with the BN used in this study is feasible.

Several key technologies will be required to use the BN model for more detailed LSA research in the future. For example, landslide-causing factors are often collected at different times and with different spatial resolutions. The primary problem in LSA data acquisition and processing lies in reducing the effects of temporal and spatial sampling variation and integrating the factor data into a unified space–time reference system. Other future objectives are improving the methods for evaluating and selecting landslide-causing factors using a smarter and more accurate algorithm, modeling the BN with continuous variables, and improving the performance of the BN model using larger data sets.

Table 8
Statistics for the susceptibility index categories.

Categories of susceptibility index	LSI value	Total number of pixels	Percentage value
Very low	0–0.05	2,806,191	0.917
Low	0.06–0.19	145,863	0.048
Moderate	0.20–0.40	44,357	0.014
High	0.41–0.65	34,259	0.011
Very high	0.66–1.0	32,354	0.010

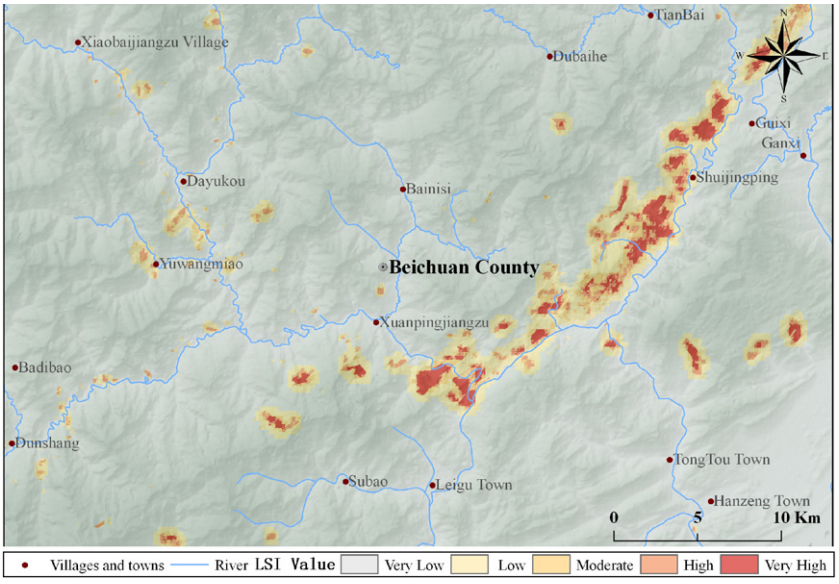

Fig. 8. Landslide susceptibility map of the study area using LSA modeling with BN. The LSI values were grouped into five categories through natural breakpoints in ArcMap. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

Acknowledgments

This research is supported by the Key Knowledge Innovative Project of Chinese Academy of Sciences (KZCX2-EW-318), the National Natural Science Foundation of China (40871181), National Basic Research Program of China, 973 Program (2007CB714402), and National Natural Science Foundation of China (41101363).

References

Aguilera, P., Fernández, A., Reche, F., Rumí, R., 2010. Hybrid Bayesian network classifiers: application to species distribution models. Environmental Modelling & Software 25, 1630–1639.

Aguilera, P.A., Fernández, A., Fernández, R., Rumí, R., Salmerón, A. Bayesian networks in environmental modelling. Environmental Modelling & Software, doi:10.1016/j.envsoft.2011.06.004. In press.

Alpaydin, E., 2010. Introduction to Machine Learning, 2nd ed. The MIT Press.

Arias, A., 1970. A measure of earthquake intensity. In: Hansen, R.J.E. (Ed.), Seismic Design of Nuclear Power Plants, MIT Press, Cambridge, pp. 438–489.

Ayalew, L., Yamagishi, H., 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology 65, 15–31.

Carrara, A., 1983. Multivariate models for landslide hazard evaluation. Mathematical Geology 15, 403–426.

Chung, C.-J., 2006. Using likelihood ratio functions for modeling the conditional probability of occurrence of future landslides for risk assessment. Computers & Geosciences 32, 1052–1068.

Cooper, G.F., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9, 309–347.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39, 1–38.

Dougherty, J., Kohavi, R., Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In: Prieditis, A., Russell, S. (Eds.), Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, pp. 194–202.

Fayyad, Irani, 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the International Joint Conference on Uncertainty in AI, pp. 1022–1027.

Fell, R., 1994. Landslide risk assessment and acceptable risk. Canadian Geotechnical Journal 31, 261–272.

Guzzetti, F., Carrara, A., Cardinali, M., Reichenbach, P., 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study. Central Italy. Geomorphology 31, 181–216.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

Harp, E.L., Wilson, R.C., 1995. Shaking intensity thresholds for rock falls and slides: evidence from 1987 Whittier Narrows and superstition hills earthquake strong-motion records. Bulletin of the Seismological Society of America 85, 1739–1757.

Hirano, A., Welch, R., Lang, H., 2003. Mapping from ASTER stereo image data: DEM validation and accuracy assessment. ISPRS Journal of Photogrammetry and Remote Sensing 57, 356–370.

Jensen, F.V., 2001. Bayesian Networks and Decision GraphsSpringer-Verlag Inc., New York.

Jibson, R.W., 1993. Predicting earthquake-induced landslide displacements using Newmark's sliding block analysis. Transportation Research Record, 9–17.

Jibson, R.W., Harp, E.L., Michael, J.A., 2000. A method for producing digital probabilistic seismic landslide hazard maps. Engineering Geology 58, 271–289.

Keefer, D., 1984. Landslides caused by earthquakes. Geological Society of America Bulletin 95, 406–421.

Lauritzen, S.L., Spiegelhalter, D.J., 1990. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems, Readings in Uncertain Reasoning. Morgan Kaufmann Publishers Inc. (pp. 415–448).

Lee, C., Huang, C., Lee, J., Pan, K., Lin, M., Dong, J., 2008. Statistical approach to earthquake-induced landslide susceptibility. Engineering Geology 100, 43–58.

Marcot, B.G., Steventon, J.D., Sutherland, G.D., McCann, R.K., 2006. Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. Canadian Journal of Forest Research 36, 3063–3074.

Murphy, K., 2001. The Bayes Net Toolbox for MATLAB. Computing Science and Statistics 33, 1024–1034.

Newmark, N.M., 1965. Effects of earthquakes on dams and embankments. Geotechnique 15, 139–159.

Nilaweera, N.S., Nutalaya, P., 1999. Role of tree roots in slope stabilisation. Bulletin of Engineering Geology and the Environment 57, 337–342.

Pearl, J., 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc.

Regmi, N.R., Giardino, J.R., Vitek, J.D., 2010a. Assessing susceptibility to landslides: using models to understand observed changes in slopes. Geomorphology 122, 25–38.

Regmi, N.R., Giardino, J.R., Vitek, J.D., 2010b. Modeling susceptibility to landslides using the weight of evidence approach: Western Colorado, USA. Geomorphology 115, 172–187.

Russell, S.J., Norvig, P., 2003. Artificial Intelligence: A Modern Approach, 2nd ed. Pearson Education.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 111–147.

Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecological Modelling 203, 312–318.

van Westen, C.J., Castellanos, E., Kuriakose, S.L., 2008. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview. Engineering Geology 102, 112–131.

van Westen, C.J., Rengers, N., Terlien, M.T.J., Soeters, R., 1997. Prediction of the occurrence of slope instability phenomenal through GIS-based hazard zonation. Geologische Rundschau 86, 404–414.

van Westen, C.J., van Asch, T.W.J., Soeters, R., 2006. Landslide hazard and risk zonation—why is it still so difficult? Bulletin of Engineering Geology and the Environment 65, 167–184.

Wang, X., Nie, G., Wang, D., 2010. Relationships between ground motion parameters and landslides induced by Wenchuan earthquake. Earthquake Science 23, 233–242.

Witten, I.H., Frank, E., 2011. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc.

Yilmaz, I., 2009. Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat–Turkey). Computers & Geosciences 35, 1125–1138.

Yin, Y., Wang, F., Sun, P., 2009. Landslide hazards triggered by the 2008 Wenchuan earthquake, Sichuan, China. Landslides 6, 139–152.