Risk Analysis DOI: 10.1111/j.1539-6924.2010.01429.x

Assessment of Catastrophic Risk Using Bayesian Network Constructed from Domain Knowledge and Spatial Data

Lianfa Li,[1,2] Jinfeng Wang,[1] Hareton Leung,[2] and Chengsheng Jiang[1]

Prediction of natural disasters and their consequences is difficult due to the uncertainties and complexity of multiple related factors. This article explores the use of domain knowledge and spatial data to construct a Bayesian network (BN) that facilitates the integration of multiple factors and quantification of uncertainties within a consistent system for assessment of catastrophic risk. A BN is chosen due to its advantages such as merging multiple source data and domain knowledge in a consistent system, learning from the data set, inference with missing data, and support of decision making. A key advantage of our methodology is the combination of domain knowledge and learning from the data to construct a robust network. To improve the assessment, we employ spatial data analysis and data mining to extend the training data set, select risk factors, and fine-tune the network. Another major advantage of our methodology is the integration of an optimal discretizer, informative feature selector, learners, search strategies for local topologies, and Bayesian model averaging. These techniques all contribute to a robust prediction of risk probability of natural disasters. In the flood disaster's study, our methodology achieved a better probability of detection of high risk, a better precision, and a better ROC area compared with other methods, using both cross-validation and prediction of catastrophic risk based on historic data. Our results suggest that BN is a good alternative for risk assessment and as a decision tool in the management of catastrophic risk.

KEY WORDS: Bayesian network; domain knowledge; risk analysis; spatial data mining

1. INTRODUCTION

Emergency catastrophic events like natural disasters are affected by complex factors that are both diverse (natural, environmental, ecological, demographic, and socioeconomic) and may have a large measure of uncertainty.(1) The interactions of these factors are complex and affected by random fluctuations. Many scholars have done studies on natural disasters resulting in the construction of complex systematic theories thereof and suggestions of specialist methods for risk assessment.(1–4) Shi *et al.*(3) proposed a systematic theory of natural disasters by dividing the risk-related factors into three aspects: inducing factors, environmental factors, and vulnerability. Based on practical surveys, Guo and Chen(5) concluded that there is a monotonously decreasing relationship (the solid line in Fig. 1) between loss risk and predictability of natural disasters or a monotonously increasing relationship (the dash-dot line in Fig. 1) between loss risk and mitigation delay. As seen from Fig. 1, if predictability of natural disasters is improved (early warning) or timely mitigation actions are carried out, loss can be decreased considerably, thus lowering the risk and saving more lives.

[1]State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Resources Research, Chinese Academy of Sciences, Beijing, China.
[2]Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong.
*Address correspondence to Lianfa Li, LREIS, Institute of Geographical Sciences and Resources Research, Chinese Academy of Sciences, Rm. 1305, No. All, Road Datun, Anwai, District Chaoyang, Beijing, China 100101; lspatial@gmail.com.

0272-4332/10/0100-0001$22.00/1

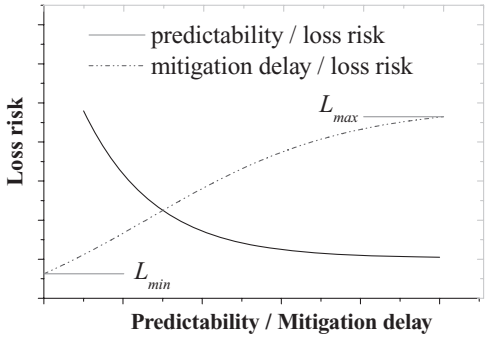


Fig. 1. Relationship between predictability of natural disasters and loss risk and between mitigation delay and loss risk.

Generally, the traditional approaches use the following form, or a derivation thereof, to model catastrophic risk:

$$R = \int \int C(V)P(V \mid A)P(A)\, dV\, dA, \qquad (1)$$

where $P(A)$ is the probability of the disaster event, A, $P(V \mid A)$ is the probability of vulnerability for a certain individual (V) given event A, and $C(V)$ is the damage potential of V. Estimation of risk using Equation (1) is dependent on multiple factors and these are subject to uncertainty.[6] Many previous approaches, however, ignore or are limited in quantifying the uncertainty of these factors and their interactions.[7] Building on top of the traditional approaches, we apply new techniques to enhance the predictability of natural disasters and thus to decrease the potential loss risk (Fig. 1).

Recently, with the development of geographical information science, data mining, and artificial intelligence, new techniques have been used in assessing catastrophic risk.[8–11] There is still the need for more exploration of new methods and applications. As pointed out in NASA's report,[12] some existing methods of data analysis lack consideration of domain knowledge, making it difficult to interpret the results; there are relatively few studies that merge multiple sources of information (some of which may be significant) although natural disasters are affected by multiple uncertainty factors.

As an exploration of pragmatic and efficient methods, this article proposes an uncertainty inference model, a Bayesian network, which is initially constructed according to domain knowledge and then fine-tuned by learning from historic data. Specifically, we explore how to fuse knowledge and spatial data from multiple sources to construct an adaptive network, how to combine components of kernel density analysis, exposure analysis, and vulnerability analysis into a consistent system, and how to use the network to make a robust prediction that overcomes overfitting.

Our model is based on Bayesian network (BN) that, as a directed acyclic graph, is able to represent uncertainty interdependences between factors that describe many real-world domains, such as public emergency catastrophic events. They have many advantages (over traditional probabilistic methods) including merging domain knowledge and multiple-source data within a consistent system, flexible network structure beneficial for searching a locally optimal solution, and inference under missing data conditions. Furthermore, by adding nodes of utilities and decision, a BN can easily become a tool for supporting reasonable decision making. With new information (evidence) at hand, the tool can instantly update the risk assessment and an adaptive decision can be made accordingly. Therefore, it is a good integration approach to assessing catastrophic risk.

Although BNs have been widely applied in many domains, for example, economy, public health, ecological risk assessment, and mineral exploitation, to the best of our knowledge, there are only a few reports of BNs being applied in catastrophic risk assessment that include debris flow,[13] earthquakes,[14] and avalanches.[7]

Compared with the small number of BN applications in catastrophic risk assessment, our study focuses on learning and robust prediction of risk using BN. We use data mining and learning methods to improve the BN's risk assessment. Hence, our modeling is different from those studies based mostly on domain knowledge. Through this study, we make the following contributions:

1. We use kernel density analysis (KDA) to preprocess the spatial data set. KDA is an approach to modeling the intensity of a certain event (e.g., how far a vulnerable individual is close to a flooded river) or a quantity (e.g., amount of loss) spread across the geospatial landscape.[15] In our study, this method is employed to preprocess the data to obtain the intensity buffer classification of the features (e.g., rivers, roads, or residential areas). This approach considers the influence of spatial distance on intensity classification of features, which is beneficial for making a robust prediction.
2. We design an optimal discretization algorithm. This is a supervised learning algorithm

that relates the discretization of quantitative factors to classification of the target variable. This algorithm is beneficial to improving the performance of the learned models in particular when we have little knowledge about the discretization and its influence on the risk prediction.

3. We design a generic modeling framework of a BN according to domain knowledge. Using data mining techniques such as informative feature selector, optimal discretizer, and search strategies, the framework facilitates integration of multiple quantitative factors and qualitative factors within a consistent system to make uncertainty inferences. It combines domain knowledge and historic data within an integration platform for multidisciplinary communication among experts in different fields (geographers, construction engineers, knowledge engineers, and economists). Aggregative use of multiple techniques is beneficial for improving the robustness of our method. Bayesian model averaging (BMA) can be used to enhance the prediction's robustness.

Our method was successfully applied in monitoring the flood disaster. Using both cross-validation and historic data to predict the new situations, our BMA prediction achieved a better probability of detection of high risk, a better precision, and ROC area than the other learning methods. Our encouraging result has implications for using BNs as an approach to assessment and decision making of catastrophic risk. If nodes of utilities and decision actions are added to the BN, a better decision-making functionality can easily be implemented on the basis of robust prediction by our methodology. Our methodology of risk assessment, although using the flood as the study case, is based on a generic modeling framework and it can be easily adjusted and extended to other types of catastrophic risk such as seismic and typhoon risk if relevant factors are selected and relevant domain knowledge is incorporated in the framework.

2. MODELING FRAMEWORK OF BAYESIAN NETWORK

This section briefly describes the BN modeling framework. Specifically, Section 2.1 introduces knowledge of the factors associated with catastrophic risk, Section 2.2 introduces the BN, and Section 2.3 proposes the BN framework.

2.1. Factors Associated with Catastrophic Risk

There are many factors associated with catastrophic risk. As described in Shi,[3] we can divide these factors into three aspects, namely, inducing factors, environmental factors, and vulnerability factors.

1. Disaster-inducing factors: As direct exposure-related factors, inducing factors are mainly responsible for occurrence of the hazards and are closely related to the occurrence of catastrophic loss. For instance, heavy rainfall may induce floods and landslides; extremely intense wind may cause typhoons or cyclones; and movements of the earth's crust may induce earthquakes.
2. Environmental factors for breeding disasters: Environmental factors are relevant to the environment that breeds the disasters. Such a factor can be either physical or artificial and is able to mitigate or aggregate the destructive power of a hazard. For instance, land with good water-soil conservation capabilities can prevent a mudslide or mitigate its destructive effect, while a flood has more destructive power on the infrastructure or residences close to a river floodplain than those further away from the floodplain.
3. Vulnerability: This is the degree to which a system or subsystem is likely to experience and adapt to harm due to exposure to a hazard. Different systems and individuals have different vulnerability due to their differences in adaptation to harm. For instance, young people are less vulnerable to a flood than seniors; and a house with a lightweight steel structure has a better ability to withstand earthquakes.

Table I gives a brief summary of the three types of factors for five natural disasters, namely, flood, typhoon, earthquake, tsunami, and landslide. Basically, these factors embody the domain knowledge about the mechanism of disasters, that is, their occurrence and effects, which is the source for the subsequent modeling of BN as described in the next section.

Table I. Division of Factors for Loss Risk in Five Disasters

Category	Flood	Typhoon	Earthquake	Tsunami	Landslide
Typical inducing factors	Heavy rainfall, dam collapse, etc.	Extreme climate events, for example, El Nino	Release of interstitial fluid pressure	Earthquake deep at sea	Heavy rainfall
Environmental factors	Water, soil, vegetation, elevation, slope, etc.	Location, close to sea, vegetation, elevation, etc.	Location, soil, close to fault, etc.	Location, close to sea, etc.	Water, soil, vegetation, elevation, slope, etc.
Vulnerability	Materials, structure and the number of stories of buildings; demographic and socioeconomic conditions; age, knowledge, and income of individuals				

2.2. Bayesian Network

A BN is a probabilistic graphical model that encodes a set of random variables and their probabilistic interdependencies through a directed acyclic graph (DAG) consisting of nodes and edges. It is a good method for modeling uncertainties and interactions between related factors inherent in monitoring and prediction of catastrophic risk. Given below is a brief introduction to BNs.

Definition 1: Given a set of random variables (rv), V, a BN is an ordered pair (B_S, B_P) such that

1. $B_S = G(V, E)$ is a directed acyclic graph, called the *network structure* of B, where $E \in V \times V$ is the set of directed edges, representing the probabilistic conditional dependency relationship between rv nodes that satisfies the *Markov property*, that is, there are no direct dependencies in B_S that are not already explicitly shown via edges, E, and
2. $B_P = \{\gamma_u : \Omega_u \times \Omega_{\pi_u} \rightarrow [0 \ldots 1] \mid u \in V\}$ is a set of *assessment functions*, where the state space Ω_u is the finite set of values of u; π_u is the set of parent nodes of u, indicated by B_S; if X is a set of variables, Ω_X is the Cartesian product of all state spaces of the variables in X; and γ_u uniquely defines the joint probability distribution $P(u \mid \pi_u)$ of u conditional on its parent set, π_u.

BNs are based on the Bayesian theorem, that is, inference of the posterior probability (a.k.a. *belief*) of a hypothesis according to some evidence. In assessment of catastrophic risk, evidence comes from inducing factors and environmental and vulnerability factors (Table I), while the hypothesis refers to the risk that is classified as several states of loss. Let r be such a hypothesis variable of loss risk and its state space is Ω_r. The risk can be classified as seven states, that is, $\Omega_r = \{none, slight, light, moderate, heavy, major, destroyed\}$ whose definitions of the damage states are given in Table II. In practice, the risk can be classified into two states, for example, $\Omega_r = \{low\ loss, high\ loss\}$, or $\{no\ loss, loss\}$ for convenience of surveying the loss when the training samples just involves two states of loss such as Wang's[16] survey of earthquake loss and the binary survey of flood loss in our flood study. In our binary classification, the threshold for the damage factor of "*high risk*" is 10% (damage factor), that is, if over 10% of the properties or people at a certain place are damaged, this place will be classified as "high loss."

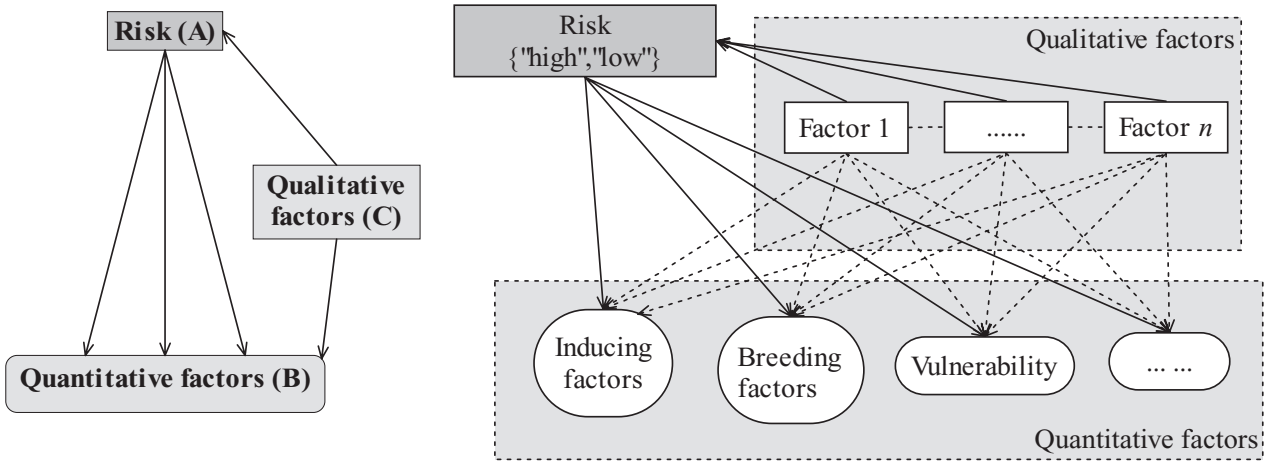
Table II. Different States of the Risk and Their Damage Definition

Damage State	Damage Factor Range (%)	Central Damage Factor (%)
None	0	0
Slight	0–1	0.5
Light	1–10	5
Moderate	10–30	20
Heavy	30–60	45
Major	60–100	80
Destroyed	100	100

In a specified BN, given some evidences at hand, we can estimate the posterior probability or belief of the target variable r as the risk probability by calculating the marginal probability:

$$Bel(r_k) = \sum_{u_i \in V, u_i \neq r} p(u_1, u_2, \ldots, r, \ldots, u_n), \quad (2)$$

where $p(u_1, \ldots, u_n) = \prod_{u_i \in V} p(u_i \mid \pi_{u_i})$ is the joint probability over V.



A. Disease-symptoms pattern and influence pattern from qualitative factors

B. Initial network framework for risk assessment (constructed from domain knowledge)

Fig. 2. Modeling framework of Bayesian network.

In practice, we often use an efficient algorithm of exact inference or approximate inference rather than the marginalization of the joint probability to compute *Bel* in Equation (2).

2.3. Initial Network Framework

We devise an initial network framework of BN according to domain knowledge. The domain knowledge comes from the generalization and classification of factors associated with natural disasters as described in Section 2.1. To simplify the implementation of the domain knowledge in the BN, we use two relationships to represent the knowledge: one is the relationship between quantitative independent factors such as rainfall, elevation, and slope, and the target variable, that is, loss risk, while the other is the influence of qualitative factors such as vegetation, landform, and soil type on the relationship between quantitative factors and the target variable.

1. When a quantitative factor is closely associated with the target variable (risk), abnormal values of the quantitative factor may indicate higher loss risk (probability). This relationship between quantitative factors and the loss risk is similar to that between a disease and the related symptoms of the patient catching the disease: a certain disease often causes the patient to have some abnormal test results or symptoms. Similarly, if a study region has a high loss risk, it often has a higher or abnormal measurement value of some quantitative factors. Based on empirical knowledge, such a relationship becomes a basic pattern for our Bayesian modeling framework. We call this relationship the *disease-test pattern* or *disease-symptom pattern*.
2. Another aspect of domain knowledge is the influence of qualitative factors such as soil type, geological type, and vegetation on the relationship between quantitative factors and loss risk. In this study, we regard these qualitative factors as contributing factors to the *disease-symptom pattern* relationship. The influence of qualitative factors on quantitative factors and loss risk naturally becomes our second basic pattern for BN modeling and we call it an *influence pattern*.

From the above two basic patterns (Fig. 2A), we can construct the BN modeling framework (Fig. 2B). We use a simple diverging connection[17] to model the "disease-symptom" pattern. First, we assume that the quantitative factors used are independent. Under this assumption, we can specify this connection using the loss risk ("disease") node as root and quantitative nodes ("symptoms") as leaves. In this pattern, each leaf node has an edge directed from the root node. If we do not temporarily consider the influence from qualitative factors, the diverging connection is a typical naïve Bayes that is often used in

medical diagnosis. From this connection, we get the likelihood of the target node, $r =$ "high loss":

$$L(m_1, \ldots, m_n \mid R) = P(m_1 \mid R) \cdot \cdots \cdot P(m_n \mid R), \quad (3)$$

where R is the target node of loss risk and m_i the ith node of the independent quantitative factors. Then using the normalization constant, μ, we get the posterior probability, or belief, of the risk variable: $Bel(R) = \mu P(R) \cdot L(m_1, \ldots, m_n \mid R)$.

To model the influence pattern, we use the fundamental rule of probability calculus:[18]

$$P(A, B \mid C) = P(B \mid A, C) P(A \mid C). \quad (4)$$

If we regard the risk factor as A, the independent quantitative factors as B, and the qualitative factors as C, according to the rule, it is natural to have the edges directed from the qualitative factor nodes (C) to the quantitative factor (B) and risk (A) nodes to obtain our framework (Fig. 2B). In BN, an edge represents a probabilistic dependence ($P(A \mid B)$) and the node (A) at the edge's arrow is statistically dependent on the one (B) at the edge's source. According to Equation (4), if C has a statistical influence on A and B (i.e., A and B are statistically conditional on C), this can be specified as the product of two probability dependencies, that is, $P(B \mid A,C)$ and $P(A \mid C)$, and thus we can direct an edge from C to A to represent the dependence $P(A \mid C)$, and direct two edges from A and C to B to represent the dependence $P(B \mid A,C)$. According to the BN principle, such a product of probabilities in Equation (4) can be implemented in the network structure as in Fig. 2A. When A represents different independent factors and B different qualitative factors, this interdependence between the three factors can be extended to multiple factors with the independence assumption among quantitative factors (C). This gives us the BN's initial framework (Fig. 2B). Furthermore, in order to decrease the computation burden, we can use domain knowledge to confirm the interdependency relationship between quantitative factors (B) and qualitative factors (C). Thus, domain knowledge is used to select the links in the final network by removing those noninterdependent relationships. This method of removing the noninterdependence conforms to the simplification principle of Occam's window[19] to select the model. Thus, in Fig. 2, we use dotted lines to denote such a relationship finally determined by domain knowledge.

On the other hand, there may be interdependencies between qualitative factors and these interdependencies can be learned from the data set using various search algorithms. Then, the learned interdependencies from the data are used to determine the local structure of the network framework, thereby constructing a complete BN.

In the framework, we must ensure independence between quantitative factors. Furthermore, since the predictive-related factors are quantitative (continuous) or qualitative (discrete or categorical), we need to develop methods to fuse such different types of variables in the BN. In our methodology, we use PCA and Quinlan's information measures to select independent quantitative factors, an optimal discretization algorithm to discretize the quantitative factors, thus enabling the BN to integrate both qualitative and quantitative factors within a consistent system, search algorithms to obtain the local structure, and BMA to obtain a robust prediction of the loss risk.

On the basis of the principles of naïve Bayes, Occam's window's simplification,[19] and probability calculus, our BN framework is constructed. Given its independence assumption and theoretical foundation, our framework, although artificial, is reasonable and adequately describes the probability relationships between the variables. This framework combines simple domain knowledge (relevant factors and their classification) and learning to obtain a robust assessment of catastrophic risk. The framework is practical and very useful, in particular when we lack domain knowledge of the probability interdependences among risk-related factors. But, if we have a clearer knowledge of the interdependence, we can directly construct the network.

3. BACKGROUND OF THE FLOOD CASE STUDY

3.1. Study Area

The study area is one part of the basin of the Heihe River that is located between east longitude 96°42′ and 102°04′ and between north latitude 36°45′ and 42°40′ in northwest China. Fig. 3 shows the study area, which includes 11 counties, for example, Jinta, Jia Yuguan, Jiuquan, Gaotai, Zhangye, etc. within Gansu Province.

The study area, unlike the southern regions of China, does not often experience heavy rainfalls. But due to worsening local ecological systems and poor water-soil conservation ability, a few heavy rainfalls can cause moderate flood disasters as recorded in history.[20] In the summer (July) of three successive

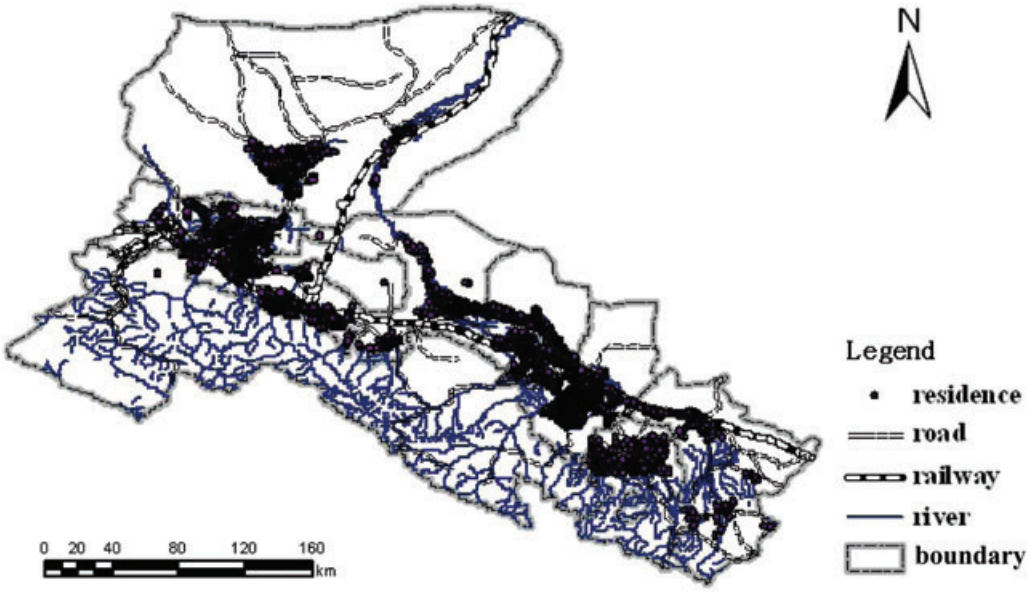


Fig. 3. The study area.

years, 2006, 2007, and 2008, three flood disasters happened in this region.

3.2. Study Goal

The study goal is to use the flood disaster data set and the related factors to construct a robust BN learner that can be used to predict the flood's loss risk. Our methods are compared with nine other probabilistic or nonprobabilistic models.

In the case study, we used three data sets of the same study area from three successive years, July 2006, July 2007, and July 2008.

For model validation, the data from 2006 and 2007 were, respectively, employed to validate the model using the usual 10×10 cross-validation. In this validation, the data set was randomly divided into 10 buckets of equal size. Nine buckets were used for training and the last bucket was used as the test. The procedure was iterated 10 times and the results were averaged. The various methods are compared using scalar measures, that is, the probability of detection (pd), probability of false alarms (pf), *precision,* and receiver operating characteristic (ROC) area.

Furthermore, the 2006 data set was used as the training data to teach the models to predict the 2007 and 2008 loss risks. Thus, through four comparisons (the 2006 and 2007 cross-validations, the 2007 and 2008 predictions), we validated our methodology.

3.3. Data Set

This study's data set is based on the grid format. In this format, the study region is subdivided into compartments or cells (pixels) on the basis of a spatial data set obtained from a rectangular grid. The grid's resolution is about 500 m. The data set includes data of three relevant factors and loss survey from 2006, 2007, and 2008, respectively. The involved factors are described as follows.

3.3.1. Predictive Factors

The predictive factors cover the three aspects as described in Section 2.1.

1. One inducing factor: heavy rainfalls (r_f) are the direct cause of the flood.
2. Six environmental factors: elevation (e), slope (s), daily mean wind velocity (d_{mn}), daily maximum wind velocity (d_{max}), normalized difference vegetation index (NDVI, n), and geology type (g). These factors correspond to the physical geographical environment in which the flood occurs and can aggregate or mitigate the flood's destructive power: a higher altitude (e) indicates less influence from the flood; a larger wind velocity (d_{mn} and d_{max}) implies more destruction indirectly related to the flood; geological conditions (g) have an influence on the indirect damages of the flood;

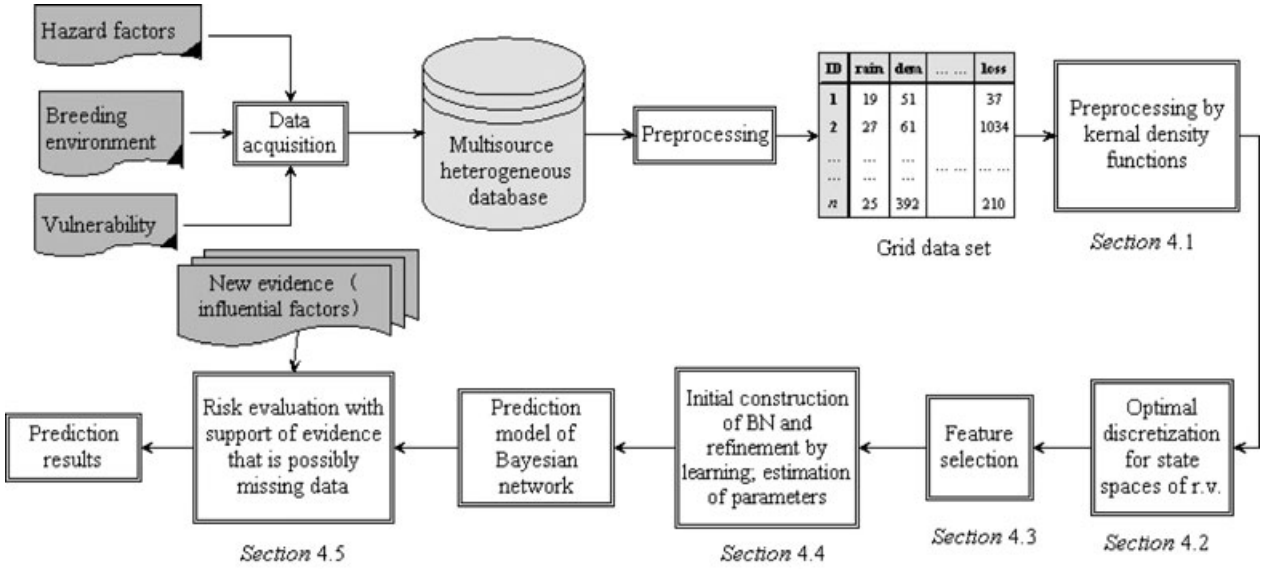


Fig. 4. Procedure for risk assessment in our methodology.

NDVI (n) is an indicator of the study area's vegetation and also has secondary effects on the flood disaster.

3. Three vulnerability-related factors: whether close to residents (r_e), whether close to roads (r_o), and whether close to rivers (r_i). These factors are associated with the location and surroundings of the vulnerable individuals (e.g., human beings, houses, or constructions). If individuals are closer to the flooded river (r_i), they are more vulnerable to the disaster's damage. If a flood is closer to a residential area (r_e), this area can be more vulnerable to the loss. Conversely, if individuals are closer to a road, they can have a greater chance provided by the road (r_o) for escaping the disaster and thus have less vulnerability.

Among these factors, the inducing factor, r_f, and the five environmental factors, e, s, n, d_{mn}, and d_{max}, are quantitative factors, while the remaining environmental factor, g, and the three vulnerability factors, r_i, r_o, and r_e, are qualitative factors.

3.3.2. The Target Variable

The target variable is the loss caused by the flood. We obtained the loss data using thematic mapping (TM) images in combination with yearbooks, statistics, and references from practical surveys. Using TM images, we obtained the submergence depth and duration of the flood and then we examined and confirmed the loss situation within the submergence range referring to other materials and statistics from practical surveys.[21] We used a binary categorization variable to indicate whether a unit (a cell in the grid data set) has a "high loss" or "low loss" risk ("1" representing "high loss" and "0" representing "low loss"). Basically, the areas having "high loss" each have a loss proportion of over 10%.

4. FLOOD RISK ASSESSMENT

The assessment of catastrophic risk is defined by the following steps:

1. Preprocess the data set using a kernel density function (Section 4.1).
2. Discretize quantitative factors using the optimal multisplits algorithm (Section 4.2).
3. Select independent quantitative factors using principal component analysis (PCA) and Quinlan's information measures (Section 4.3).
4. Build the BN with discrete or categorical predictive factors (Section 4.4).
5. Perform a robust prediction of the catastrophic risk (Section 4.5).

Figure 4 shows the procedure of the flood risk assessment. To obtain the grid data set from multiple heterogeneous sources, we apply various preprocessing steps, for example, converting the vector

data and resampling the grid data into the target grid data set at the standardized resolution and projection. We perform these steps in a GIS environment such as ARCGIS. The following sections describe major techniques of the procedure in Fig. 4.

4.1. Using Kernel Density Functions to Preprocess the Data Set

KDA is a nonparametric unsupervised learning procedure. The kernel, k, is a probability density function that is symmetric around the origin and decreases with an increasing distance from the origin. We can use the normal density function to simulate a kernel function, $K_\lambda(z, Z_i)$.[15] Then, we can summarize the kernel density values of any unit from the sample (observation) units to obtain the intensity value of any unit in the geographical area:

$$Density(z) = \frac{1}{n}\sum_{i=1}^{n} Z_i \cdot K_\lambda(z, Z_i), \qquad (5)$$

where n is the number of sample units. According to $Density(z)$, we obtain the classification of the predictive factors. Z_i is a count of a certain type of event or a quantity of the feature.

Kernel density estimation represents the concept of a spatial correlation, that is, closer spatial distance (d) between geospatial features means more correlation or more influence between them.[22] Consideration of spatial correlation in using KDA to quantify predictive factors is in particular useful for risk analysis of natural disasters given its reasonable assumption.

How to set the bandwidth or search radius, λ, is determined by empirical knowledge and the design goal. A big λ means more generalization over the entire study area while a small λ means overlocalization over the area. Since our goal is to reflect the influence of relevant indicators such as rivers on the damage to the individuals exposed to the flood, λ can be set according to the biggest influence range of the relevant indicators in practical disasters. In the flood disaster, three variables, that is, r_e, r_o, and r_i, are quantified and classified using KDA. Fig. 5 shows the classifications of the three variables via KDA (Fig. 5A for r_i; Fig. 5B for r_o, and Fig. 5C for r_e).

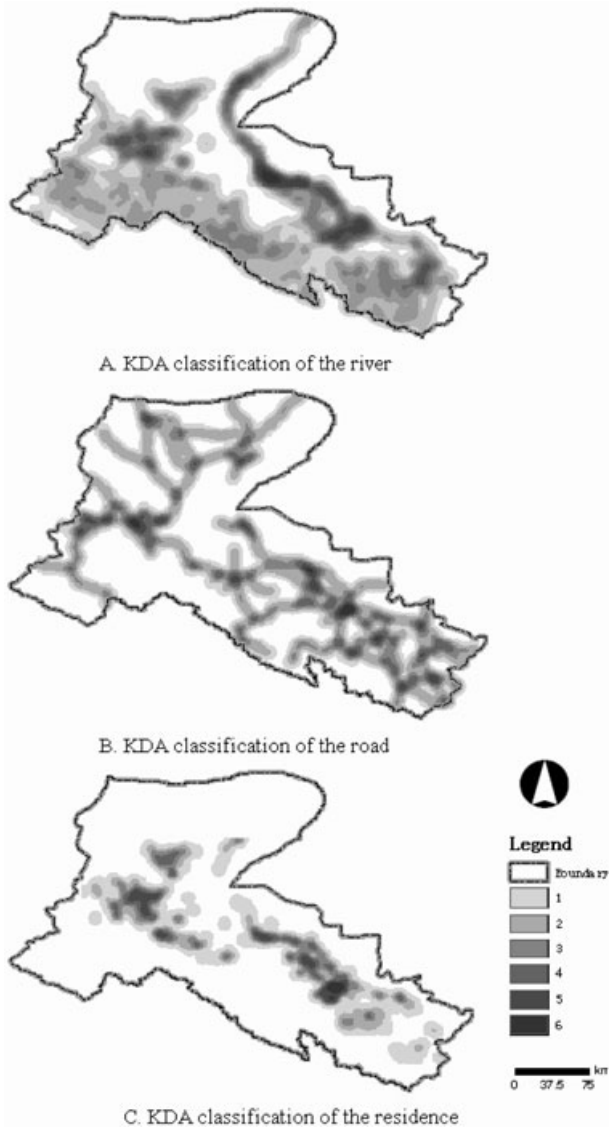


Fig. 5. KDA classification of r_i, r_o, and r_e.

4.2. Optimal Discretization of Quantitative Factors

This step involves discretizing the quantitative factors. The discretization will be used first in the selection of independent predictive factors described in Section 4.3, and then in BN modeling, as inputs of the discrete state space Ω_u. We use a supervised learning algorithm to find the optimal splits to discretize quantitative factors for the BN to achieve the data-adaptive prediction of the target variable. We describe this in two parts: Section 4.2.1 introduces the concept of Quinlan's measure that is used later in our algorithm, while Section 4.2.2 presents the discretization algorithm.

4.2.1. Quinlan's Information Measures

Quinlan's information gain ratio (GR)[23] is used to measure the contribution of the splits of each

Table III. Splits of Quantitative Factors Discretized

Quantitative Factors	Discretized Intervals (Splits)
Elevation (e)	[0, 117), [117, 1411.5), [1411.5, 2285), [2285, $+\infty$)
Slope (s)	[0, 1.095), [1.095, 4.1), [4.1, 6.495), [6.495, 17.705), [17.705, $+\infty$)
Rainfall (r_f)	[0, 48.0), [48.0, 56.5), [56.5, 57.5), [57.5, 58.5), [58.5, 2515.0), [2515.0, 2764.5), [2764.5, 2850.5), [2850.5, 3897.5), [3897.5, 4901.5), [4901.5, 4994.0), [4994.0, 7524.0), [7524.0, 7689.5), [7689.5, $+\infty$)
Daily mean wind velocity (d_{mn})	[0, 24.5), [24.5, 29.5), [29.5, 30.5), [30.5, $+\infty$)
Daily max. wind velocity ($d_{\max}$)	[0, 72.5), [72.5, 73.5), [73.5, 76.5), [76.5, 77.5), [77.5, 80.5), [80.5, 81.5), [81.5, $+\infty$)
NDVI (n)	[0, 0.015), [0.015, 0.075), [0.075, 0.115), [0.115, 0.225), [0.225, 0.305), [0.305, 0.675) [0.675, $+\infty$)

quantitative factor to risk prediction. GR measures the information GR given the discretization of the variable to be assessed. GR takes into account the information that the discretized variable contains. Qianlan's GR also measures the contribution of an indicator to the prediction and thus can be used as a means of feature selection (Section 4.3).

4.2.2. Optimal Multisplitting Discretization Algorithm

The algorithm is designed according to the "recursion" idea in the algorithm by Fulton *et al.*[24] and the minimal description length (MDL) stopping criteria in Fayyad and Irani's algorithm.[25] It recursively finds the optimal splits of a continuous predictor based on the discretization's contribution to the class prediction. Compared with other supervised methods, this algorithm can achieve the same or better splits with the number of intervals adjusted adaptively, although we need to set the maximum number of intervals.

This algorithm assumes that optimal cut points fall on the boundary points that are defined as the points between two successive attribute values of the sorted instances that have two different class labels. This assumption has been theoretically proven to be reasonable.[26] The algorithm uses GR as the goodness criterion for discretization. It recursively searches the boundary points from big to small until the optimal cut points with the maximum GR are obtained:

$$GR(k, 1, i) = \max_{1 \le j < i} (GR(k-1, 1, j) + GR(1, j+1, i)), \quad (6)$$

where $GR(k, j, i)$ denotes the maximum GR that results when the training instances j through i are partitioned into k intervals. The best k-split is the one that maximizes $GR(k,1,N)$, where N is the cardinality of the set of values of the continuous predictor.

Although we can set a maximum number of intervals, the algorithm makes use of Occam's MDL principle of information theory as the stopping criterion,[27] thus adaptively adjusting the number of discretization intervals. The algorithm can decide whether a candidate cut point is acceptable and new partitions are unnecessary according to the MDL criterion.

This algorithm considers the characteristics of the data such as the variance. If a factor has a big variance and a split can improve the discretization's contribution to classification, the split will be kept. The even discretization is simple and easy to use but some of its splits may be unnecessary (e.g., only the rainfall beyond a certain threshold can result in a flood disaster and the discretization below such a threshold is meaningless for the flood risk prediction). So we can use this algorithm to automatically detect and identify such threshold splits when we have little knowledge of the risk-related factors and experts cannot give precise splits. The splits identified by this algorithm can be adjusted according to domain knowledge if necessary.

In our study, six quantitative factors, namely, r_f, e, s, d_{mn}, $d_{\max}$, and n were discretized using this algorithm. Table III gives the splits of these factors.

Then, the splits were used to discretize the corresponding factors of the validation and test data sets to supply discrete versions of the continuous factors.

4.3. Feature Selection

To obtain independent quantitative factors, we employ PCA to detect the underlying independence among quantitative factors (Section 4.3.1) and then use Quinlan's GR to confirm the relationship of each

Table IV. Loading and GR of Quantitative Factors

Measures	E	s	r_f	d_{mn}	d_{max}	n
Loading	0.76	0.979	0.811	0.883	0.965	0.981
#Component	5	2	4	1	1	3
GR	0.077	0.69	0.079	0.039	0.052	0.131

Table V. GR of Qualitative Factors Used

Measures	Close to Residences? (r_e)	Geology Type (g)	Close to River? (r_i)	Close to Road? (r_o)
GR	0.066	0.056	0.132	0.0132

factor to the target variable (risk) to get the set of independent quantitative factors (Section 4.3.2).

4.3.1. PCA to Detect Underlying Factors

PCA[28] is a classic statistical method used to explain variability among observed variables in terms of fewer unobserved variables called principal components (PC). In this study, we used the commonly used varimax rotation strategy to make distinct the PC. If a PC's eigenvalue is greater than 1.0, it will be selected as a predictive factor.

4.3.2. Selection of Quantitative Factors According to Their Loadings and GR

From PCA, we obtain a subset of PCs with eigenvalues greater than 1.0. Next we select those quantitative factors within each PC whose loading is maximum and close to or above 0.8 according to empirical knowledge, and then from these responsible factors for each PC, we select one predictive factor whose GR is relatively large with the loading threshold 0.8 to avoid information loss while selecting independent PCs. A loading value of 0.8 or more makes the quantitative factor contain most information of the principal component. If a PC has several predictive factors whose loading coefficients are equal to or bigger than 0.8, we just select the one with the largest GR among them. Thus the features selected, while maintaining independence, are informative and beneficial for the prediction of risk.

Using the above selection criteria, we selected the predictive factors from the quantitative and qualitative factors in the flood's study. Table IV shows the loading and GR of the quantitative factors and Table V shows the GR scores of qualitative factors. In total, we selected nine predictive factors including five quantitative factors, namely, n, r_f, s, e, and d_{max}, as well as four qualitative factors, namely, g, r_e, r_o, and r_i.

4.4. Model Construction and Estimation of Parameters

This section describes learning of the BN's local structure (Section 4.4.1) and estimation of assessment parameters (Section 4.4.2).

4.4.1. Learning of Local Structure of Qualitative Factors

Once the independent quantitative factors have been selected, they are used to construct an initial network (Fig. 2). We then use the learning algorithms to learn the local structure of the qualitative factors from the training data set.

The learning uses a quality score to measure the network's quality. There are three kinds of score measures that bear a close resemblance: the Bayesian approach, the information criterion, and the minimum description length. In this study, we used the Bayesian approach, which uses the *a posteriori* probability of the learned structure given the training instances as a quality measure. The Bayesian approach can achieve a good effect as it is unaffected by the specific structure, unlike other measures.[29]

A search algorithm can be applied to the space of the network structures to find the locally optimal network with a high-quality score. Table VI shows various typical algorithms to obtain the topology of local network. In this table, the methods in bold font were used in our methodology.

4.4.2. Learning of Assessment Parameters

Once the BN's structure has been constructed, the CPT parameters for each node in the BN can be obtained in two ways.

1. If the training data sets are missing, we can elicit CPT from domain knowledge by consulting domain experts, modeling, or using various yearbooks, statistics, or references.
2. If we have enough data and we do not have clear knowledge about a disaster, we can learn

Table VI. Methods for Construction, Inference, and Prediction of BN

Steps	Type		Methods
Structure	Domain knowledge based		**Construct BN according to domain or empirical knowledge**
	Dependency analysis based		Conditional independence (CI)[17]
	Search scoring based[29]	Quality measure	**Bayesian approach,** information criterion approach, and minimum description length approach
		Learning methods	Heuristic search strategies: **K2, hill climbing (HC)**, and **TAN** etc. general-purpose search strategies: **Tabu, simulated annealing (SA), and genetic algorithm (GA),** etc.
Parameter learning	Domain knowledge based		**Reports, statistics, and experienced models**
	Distribution based		**Dirichlet-based parameter estimator**
	With missing data[17]		Expectation maximization, Gibbs sampling
Inference	Exact inference		**Joint probability,** naïve Bayesian, graph reduction, and polytree[30]
	Approximate inference		Forward simulation, random simulation[30]

CPT from the data set by using a learning algorithm (Table VI).

4.5. Robust Prediction of the Flood Disaster Risk

To mitigate the sampling bias and model uncertainties (also avoiding the overfitting problem), we use BMA and Occam's window[19,31] to produce a robust prediction of the flood risk.

Assume r to be the target variable of risk, D to be the training data set, and M_i to be the ith model of BN. Then we can get the averaged value of the probability of the target variable being a certain state using BMA:

$$pr(r \mid D) = \sum_{k=1}^{K} p(r \mid M_k, D)p(M_k \mid D), \qquad (7)$$

where K is the number of models selected and

$$p(M_k \mid D) = \frac{p(D \mid M_k)p(M_k)}{\sum_{k=1}^{K} p(D \mid M_l)p(M_l)} \qquad (8)$$

is the weight of Bayes factor that is ratios of marginal likelihoods or of posterior odds to prior odds for different models. We use the BN's inference algorithms (Table VI) to obtain $p(D \mid M_k)$ and assume that the prior probability of each model ($p(M_k)$) is the same.

While BMA can average the predictions of the models obtained using various learning algorithms (Table VI), we can also use Occam's window to select the qualified models and remove those poor models, thus improving the computation efficiency. Occam's window[19] has two principles: (1) if a model receives much less support (e.g., the ratio of 20:1) than the model with maximum posterior probability, then it should be dropped; (2) complex models that receive less support than their simpler counterparts should be dropped.

We use six search algorithms shown in Table VI (in bold font) to get the local structures of qualitative factors and use BMA and Occam's window to average the qualified models, thus decreasing model bias and improving the robustness.

5. EVALUATION

To evaluate our method, we compared it with other methods (Section 5.1) using scalar performance measures (Section 5.2).

5.1. Methods Compared

Our methods were compared with other prediction methods. This section provides a simple introduction of these methods.

The methods compared include both nonprobabilistic and probabilistic methods. Nonprobabilistic methods do not output the risk probability of each predicted instance, but instead output its class label directly and these methods include J48, RF, RT, SMO, and Winnow. Probabilistic methods predict the risk probability distribution of each test instance and classify the instance according to the distribution. These methods include LR, NB, RBF, MPer, and our seven BN methods (six search algorithms and the BMA averaged prediction). Table VII gives brief descriptions and references for these methods, and predictive factors used.

Table VII. Prediction Methods Compared

T	Method	Description and Reference	Predictive Factors Used
NP	J48	A C4.5 decision tree[23] recursively partitions the training data by means of attribute splits and generates a pruned or unpruned tree using the information-theoretical concept of entrophy.	All predictive factors (no discretization): quantitative: $r_f, e, s, d_{\text{mn}}, d_{\max}, n$; qualitative: r_i, r_o, r_e, g
	RF	A forest of random trees[32] is a meta learning scheme that embodies several base-classifiers (CART) that are built independently and participate in a voting procedure to obtain a final class prediction.	
	RT	A tree that considers K randomly chosen attributes at each node without pruning.[33]	
	SMO	Sequential minimal optimization algorithm for training a support vector classifier[34] globally transforms nominal attributes into binary ones and multiclass problems are solved using pairwise classification.	
	Win	Winnow and Balanced Winnow algorithm[35] updates a vector of parameters used to construct its weight vector that has an inner product with the vector of features as the prediction by repeated corrections.	
P	LR	Logistic regression[36] directly estimates the posterior probabilities by fitting data to a logistic curve.	Independent quantitative factors # (no discretization): $r_f, e, s, d_{\max}, n$
	NB	Naïve Bayes[37] assumes that the presence of a feature of a class is unrelated to the presence of any other feature.	
	RBF	Normalized Gaussian radial basis function (RBF) network[38] comprise a hidden layer of RBF nodes and an output layer with linear nodes and its output activity is normalized by the total input activity in the hidden layer.	
	MPer	Multilayer perceptron, a back-propagation classifier[38] whose network can be built manually or created by an algorithm and can also be monitored and modified during training time (the nodes in this network are all sigmoid).	
	BN	BNs are constructed based on our framework (Fig. 2) and six search strategies (Table VI) for the local structures among qualitative factors. Among the six search algorithms, K2 uses a hill climbing (HC) algorithm restricted by an order of the variables; HC searches locally optimal network by adding, deleting, and reversing arcs without any restriction of the variables' order; Tan determines the maximum weight spanning tree and returns a NB network augmented with a tree; Tabu uses tabu search for finding a well scoring and is similar to HC; simulated annealing (SA) and genetic algorithm (GA) are generic probabilistic metaheuristic algorithms (SA: physical mechanism; GA: biological mechanism). All the predictions from the six strategies are averaged to get a robust prediction using BMA.	Independent quantitative factors # (discretization) and qualitative factors: Quantitative: $r_f, e, s, d_{\max}, n$; qualitative: r_i, r_o, r_e, g

Note: T = type; P = probabilistic methods; NP = nonprobabilistic methods.
#Independent quantitative factors include $r_f, e, s, d_{\max}, n$; see Section 4.3 regarding feature selection.

5.2. Performance Measures

We use four scalar measures, that is, *pd*, *pf*, *precision*, and *ROC area*, for the comparison.

1. *Pd* refers to the probability of detection of "high loss" risk and it measures the proportion of correctly predicted positive instances among the actual positive ones. If a method achieves a higher *pd*, it can detect more positive instances (more cell units of "high loss" risk detected).
2. *Pf* refers to the probability of false alarms and a good method has a low *pf*.
3. *Precision* refers to the proportion of true positives among the instances predicted as positive, but it cannot measure how the method detects the actual positive instances. Good *precision* does not always mean a good *pd*. A method with high *precision* but a lower *pd* is less useful since it cannot detect significant positive instances (less units of "high loss" risk detected).
4. *ROC area* is the area between the horizontal axis and the ROC curve, and it is a comprehensive scalar value representing the model's expected performance. The ROC area is

between 0.5 and 1, where a value close to 0.5 is less precise, while a value close to 1.0 is more precise. A larger ROC area indicates better prediction performance.

In terms of risk assessment, security and warning are the major concerns and a good method should detect more positive ("high loss" risk) instances (a high *pd* and lower *pf*). Thus *pd* and *pf* are the main scalar measures of performance. *Precision* is a secondary measure used with *pd*. In other words, a good method should have a high *pd* and low *pf* with an acceptable *precision*.

6. RESULTS

This section presents the results that include the learned topologies of the BNs (Section 6.1) and the prediction comparisons of the different methods using both 10 × 10 cross-validation (Section 6.2) and using the historical data to predict the new situations (Section 6.3).

6.1. Network Topologies of Local Structure

We constructed the BNs according to the modeling framework (Fig. 2). Specifically, we first obtained an initial framework of the BN using the selected independent quantitative factors and qualitative factors. On the basis of this framework, we used six search strategies, namely, K2, hill climbing(HC), Tan, Tabu, simulated annealing (SA), and genetic algorithm (GA), to search the qualified local structure among the four qualitative factors, namely, r_i, g, r_e, and r_o. The searches help fine-tune the network structure. Fig. 6 shows partial topologies of the local structure constructed.

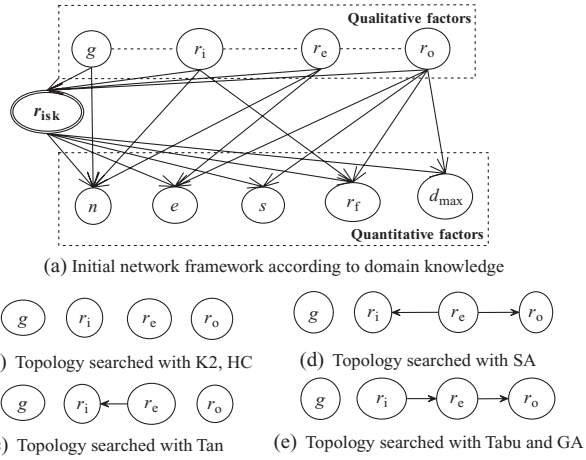


Fig. 6. Initial framework (a) and partial topologies of local structure (among qualitative factors) searched using the six algorithms.

Table VIII. Comparison of Prediction Models in the 2006 Cross-Validation

Model Type	Model	*pd*	*pf*	*Precision*	*ROC Area*
Probabilistic models	BN (BMA)	0.876	0.161	0.783	0.924
	BN (K2)	0.875	0.188	0.693	0.917
	BN (HC)	0.875	0.188	0.693	0.916
	BN (Tan)	0.836	0.143	0.739	0.922
	BN (Tabu)	0.837	0.175	0.699	0.911
	BN (AN)	0.766	0.129	0.742	0.904
	BN (GA)	0.874	0.188	0.693	0.917
	LR	0.638	0.107	0.743	0.879
	NB	0.783	0.144	0.725	0.891
	RBF	0.797	0.171	0.693	0.879
	MPer	0.733	0.105	0.742	0.828
Nonprobabilistic models	J48	0.801	0.133	0.745	0.88
	RF	0.747	0.093	0.746	0.901
	RT	0.709	0.125	0.734	0.798
	SMO	0.774	0.127	0.747	0.823
	Win	0.711	0.167	0.711	0.772

6.2. Performance Comparison Using Cross-Validation

This section presents the comparison of our methods (six search strategies and the BMA method) with nine other methods using 10 × 10 cross-validation. We respectively used the 2006 and 2007 data as the validation data set.

Tables VIII and IX, respectively, list the scalar measures, *pd*, *balance,* and *precision*, for all the 16 methods. As seen from these tables, totally, our methods have a relatively good *pd*, *precision,* and *ROC area* that indicates that our methods were able to capture many units (cells) of high risk in the flood disaster. On the other hand, although some of our methods such as BN(Tabu) in Table VIII and BN(HC) in Table IX do not have the highest *precision*, the difference in precision between them and the other non-BN models is small. In particular, we can see that the prediction of BMA has an improvement either in the 2006 or 2007 validation although such an improvement is slight or small in *pd* and *ROC area* but significant in *precision*. The averaged BN prediction using BMA has the best *pd* (0.876 for 2006; 0.873 for 2007), the best *precision* (0.783 for 2006; 0.835 for 2007), and the best *ROC area* (0.924 for 2006; 0.890 for 2007) compared with other

Table IX. Comparison of Prediction Models in the 2007 Cross-Validation

Model Type	Model	*pd*	*pf*	*Precision*	*ROC Area*
Probabilistic models	BN (BMA)	0.873	0.403	0.835	0.890
	BN (K2)	0.825	0.229	0.606	0.881
	BN (HC)	0.830	0.231	0.605	0.879
	BN (Tan)	0.771	0.161	0.670	0.889
	BN (Tabu)	0.792	0.214	0.613	0.871
	BN (AN)	0.596	0.126	0.668	0.869
	BN (GA)	0.805	0.205	0.626	0.869
	LR	0.635	0.117	0.629	0.850
	NB	0.715	0.188	0.619	0.822
	RBF	0.691	0.183	0.618	0.819
	MPer	0.619	0.122	0.681	0.845
Nonprobabilistic models	J48	0.634	0.115	0.702	0.801
	RF	0.651	0.102	0.731	0.810
	RT	0.627	0.125	0.682	0.801
	SMO	0.659	0.118	0.706	0.771
	Win	0.474	0.142	0.586	0.665

Table X. Comparison of Prediction Models Using the 2006 Data to Predict the Disaster Risk of the 2007 Flood

Model Type	Model	*pd*	*pf*	*Precision*	*ROC Area*
Probabilistic models	BN (BMA)	0.828	0.205	0.851	0.887
	BN (K2)	0.673	0.179	0.616	0.856
	BN (HC)	0.673	0.178	0.612	0.836
	BN (Tan)	0.687	0.21	0.62	0.823
	BN (Tabu)	0.741	0.223	0.62	0.823
	BN (AN)	0.532	0.211	0.454	0.754
	BN (GA)	0.673	0.179	0.6166	0.756
	LR	0.250	0.062	0.633	0.761
	NB	0.671	0.181	0.613	0.755
	RBF	0.705	0.199	0.602	0.842
	MPer	0.479	0.142	0.591	0.826
Nonprobabilistic models	J48	0.680	0.202	0.591	0.807
	RF	0.544	0.169	0.579	0.820
	RT	0.566	0.202	0.546	0.684
	SMO	0.548	0.135	0.634	0.706
	Win	0.331	0.077	0.640	0.627

non-BN methods. Across all the compared models, the 2006 and 2007 cross-validations demonstrated that BMA effectively decreased the model bias and improved the robustness of the risk prediction.

6.3. Performance Comparison of Prediction

This section presents the comparison of our BN-based methods with nine other methods using the 2006 data to train the learners used to predict the risk of the 2007 and 2008 flood disaster in the same area.

Tables X and XI, respectively, list the scalar measures, *pd*, *pf*, *precision,* and *ROC area* of the 2007 and 2008 risk prediction for all the 16 methods. These tables show that, totally, our methods have a relatively good *pd* that indicates that our methods are able to predict most units of high risk in the flood disaster. From these tables, we can see that our method also has a reasonable probability of false alarms (*pf*). In particular, the BN-based average prediction using BMA has moderately improved the probability of detection (*pd*: 0.828 vs. 0.250–0.740, an improvement of about 12–200% for 2007; 0.914 vs. 0.349–0.718, an improvement of about 27–162% for 2008) and *precision* (0.851 vs. 0.454–0.640, an improvement of about 32–88% for 2007; 0.805 vs. 0.554–0.763, an improvement of about 6–45% for 2007). The BMA prediction also has a slightly better ROC area in either 2007 or 2008 prediction (0.881–0.887). Across all the compared models, the 2007 and 2008 predictions

Table XI. Comparison of Prediction Models Using the 2006 Data to Predict the Disaster Risk of the 2008 Flood

Model Type	Model	*pd*	*pf*	*Precision*	*ROC Area*
Probabilistic models	BN (BMA)	0.914	0.123	0.805	0.881
	BN (K2)	0.642	0.122	0.718	0.877
	BN (HC)	0.629	0.126	0.707	0.851
	BN (Tan)	0.718	0.20	0.634	0.852
	BN (Tabu)	0.632	0.122	0.633	0.814
	BN (AN)	0.509	0.199	0.554	0.679
	BN (GA)	0.645	0.123	0.722	0.870
	LR	0.617	0.233	0.616	0.814
	NB	0.645	0.124	0.722	0.728
	RBF	0.641	0.125	0.714	0.810
	MPer	0.614	0.096	0.756	0.828
Nonprobabilistc models	J48	0.52	0.084	0.750	0.733
	RF	0.473	0.086	0.725	0.830
	RT	0.420	0.079	0.721	0.775
	SMO	0.559	0.112	0.708	0.724
	Win	0.349	0.053	0.763	0.648

demonstrated that BMA considerably decreased the model bias and improved the robustness of the risk prediction.

Figs. 7 and 8, respectively, show the maps of the 2007 and 2008 BMA prediction of risk probability. In these two figures, the degree of grayness represents the probability of high risk. We can see that the region of higher risk (darker region) is close to rivers and residential areas and this result is consistent with the practical situation. From Figs. 7 and 8, we can see

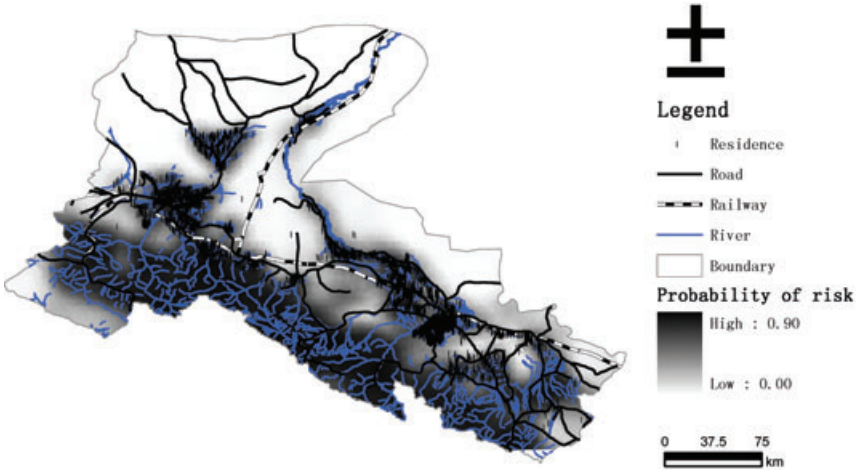


Fig. 7. Robust prediction map of risk probability using the 2006 data to train our BN to predict the risk probability of the 2007 flood disaster.

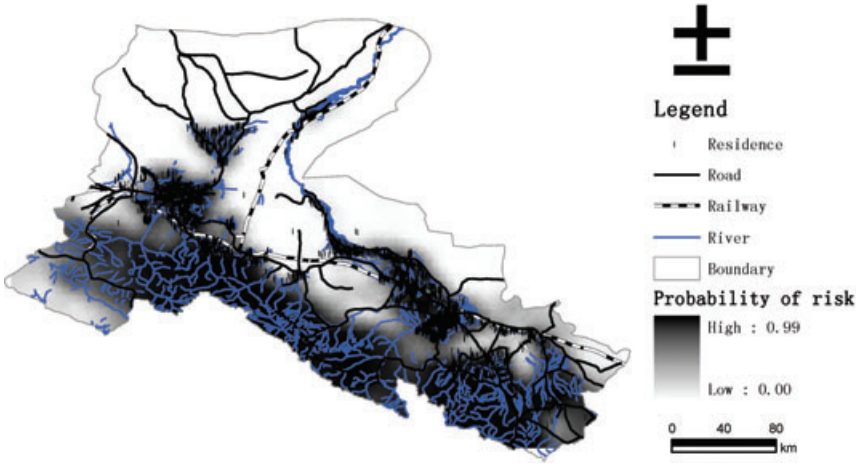


Fig. 8. Robust prediction map of risk probability using the 2006 data to train our BN to predict the risk probability of the 2008 flood disaster.

that the pattern of the 2007 predicted risk probability is somewhat different from that of the 2008 prediction. In the 2007 prediction, the upper half of the study area has more parts (cells) of high risk (higher probability of "high loss" risk) than that in the 2008 prediction while in the 2007 prediction the lower half of the area has fewer parts (units) of high risk than that in the 2008 prediction. The 2007 and 2008 predictions of high risk are consistent with the practical situation of the flood loss and their probability of detection (*pd*), respectively, reached 0.828 and 0.914 moderately greater than that of the other methods (0.250–0.741).

6.4. Sensitivity Analysis

Sensitivity analysis tries to detect how much impact the predictive factors each have on the uncertainty of the target variable, loss risk. We use Shannon's mutual information[30] to measure the sensitivity of the target variable. Using this analysis, we can find the indicator with the largest uncertainty that affects risk assessment.

Table XII presents the result of sensitivity analysis in the 2006 cross-validation, listing Shannon mutual information of nine variables with the loss risk as the target variable. From this table, we can see that r, n, r_i, and g have relatively larger values, which indicate their greater influence on the loss risk. The sensitivity analysis is reasonable given the domain knowledge: (1) rainfall (r) is the direct cause of the flood; (2) the vegetation (indirectly indicated by NDVI, n) and the geology condition (g) have a significant influence on the ability of the water-soil conservation (a poor water-soil conservation can result in serious loss under a flood); (3) being closer to a flooded river (r_i)

Table XII. Shannon Mutual Information with the Loss Risk as the Target Variable

Predictive Factors	Shannon Mutual Information
Rainfall (r)	0.287
NDVI (n)	0.260
Close to river? (r_i)	0.223
Geology type (g)	0.202
Elevation (e)	0.158
Slope(s)	0.122
Daily maximum wind speed ($d_{\max}$)	0.099
Close to residences? (r_e)	0.051
Close to road? (r_o)	0.031

indicates more vulnerability. According to the sensitivity analysis, we can focus our efforts on significant aspects (e.g., improving the water-soil conservation) that help mitigate the loss risk of natural disasters.

7. DISCUSSION

Only a few applications of BNs in catastrophic risk analysis have been reported.[13,39–41] Basically, the reported studies used domain knowledge to construct the BN and obtained the BN's assessment parameters from experts or by kind of modeling. Different from these previous studies, our study proposes a generic modeling framework of BN that integrates relevant quantitative and qualitative factors within a consistent system for assessment of catastrophic risk. Our method focuses on robustness of the model's prediction and this is also distinct from previous studies. Our methodology integrates an optimal discretizer, informative feature selector, and search strategies within the modeling system to obtain a robust prediction. In the study of flood disaster, the BMA prediction performed better than other methods based on the results from both cross-validations and prediction from historic data. With a better *precision* and *pd*, our method can detect more units (cells) of "high loss" risk than the other methods, in particular when using the historic data to predict the new situation. Note that detection of more units of "high loss" risk is significant for monitoring and management of catastrophic risk.

Although previous studies reported good applications of the BN in the assessment of catastrophic risk, we have no idea how well the BN performed in these applications when compared with other methods, since few of these studies provide such comparisons. In this study, we not only proposed a BN risk assessment methodology but also compared it with other methods. In the comparison, our methodology's robustness in *pd*, *precision,* and *ROC area* has been clearly tested. If a prediction method like ours is able to maintain a robust prediction, more high-risk units can be detected. This means an improvement in the risk's predictability and a corresponding decrease or avoidance of loss according to Fig. 1.

8. CONCLUSIONS AND FUTURE WORK

In this article, we develop a BN methodology of risk assessment of natural disasters. The methodology is based on a generic modeling framework of BN and can be generalized as an integration of techniques, namely, an optimal discretizer, informative feature selector, search strategies for local topologies, and robust predictor.

In the flood study, our methodology achieved a better *pd*, *precision,* and *ROC area* in cross-validations and predictions of new situations. This illustrates that our method is able to detect more units of "high loss" risk. This improved detection of high risk has implications for risk assessment and management in that the robust prediction can support more precise information for decision making. Our methodology is based on a general modeling framework and the techniques used are applicable for other natural disasters. Thus, our methodology can be easily extended to other natural disasters if relevant domain knowledge is incorporated in this framework and relevant data are available.

In the future, we will consider the following aspects as the extension of this work:

1. We will add nodes of utility and decision actions to the learned BN and test how different actions result in different loss risk probability and thus different expected utility of loss. The expected values of loss serve as important information for decision making and resource allocation.
2. We will extend the BN modeling by incorporating the exposure models, that is, to model the occurrence probability of a natural hazard. The incorporation of exposure information of natural hazards within the modeling system is beneficial for enhancing the predictability of natural disasters.

3. We will consider influence of the temporal indicator on relevant factors: to test how statistical variables (e.g., rainfall) of different periods (daily, weekly, monthly, seasonable) influence the prediction of the risk probability.
4. We will explore the underlying sparsity of the feature selection that may be beneficial for enhancing the robust prediction of catastrophic risk.

ACKNOWLEDGMENTS

The authors are thankful for the constructive comments from the area editor and three anonymous referees. We appreciate the support from the NSFC grant (40601077/D0120), MOST grants (2007AA12Z233 and O88RA204SA), and PolyU grant (H-ZG20).

REFERENCES

1. Kondratyev KY, Krapivin VF, Varotsos CA. Natural Disasters as Interactive Components of Global Ecodynamics. Chichester, UK: Springer, 2006.
2. Easterling DR, Meehl GA, Parmesan C, Changnon SA, Karl TR, Mearns LO. Climate extremes: Observations, modeling, and impacts. Science, 2000; 289(5487):2068–2074.
3. Shi P. Theory on disaster science and disaster dynamics. Natural Disasters (in Chinese), 2002; 11(3):1–9.
4. William JP, Arthur AA. Natural Hazard Risk Assessment and Public Policy. New York: Springer-Verlag New York, 1982.
5. Guo Z, Chen X. Strategies Against Earthquakes for Cities (in Chinese). Beijing, China: Earthquake Press, 1992.
6. Alexander D. Natural Disasters. New York: Chapman & Hall, 1993.
7. Gret-Regamey A, Straub D. Spatially explicit avalanche risk assessment linking Bayesian networks to a GIS. Natural Hazard and Earth System Sciences, 2006; 2006(6):911–926.
8. Huang C. Risk Analysis of Natural Disasters (in Chinese). Beijing, China: Beijing Normal University Press, 2001.
9. Jiang H, Eastman JR. Application of fuzzy measures in multicriteria evaluation in GIS. International Journal of Geographical Information Science, 2000; 14(2):173–184.
10. Li L, Wang J, Wang C. Typhoon insurance pricing with spatial decision support tools. International Journal of Geographical Information Science, 2005; 19(3):363–384.
11. Zhang P, Steinbach M, Kumar V, Shekhar S. Discovery of Patterns in Earth Science Data Using Data Mining. New Generation of Data Mining Applications. New York: John Wiley & Sons, 2005.
12. NASA. 2nd NASA Data Mining Workshop: Issues and Applications in Earth Science. 2nd NASA Data Mining Workshop, CA: Pasadena, 2006.
13. Antonucci A, Salvetti A, Zaffalon M. Hazard assessment of debris flows by credal networks. Pp. 98–103 in Pahl-Wostl C, Schmidt S, Rizzoli AE, Jakeman AJ (eds). iEMSs 2004: Complexity and Integrated Resources Management, Transactions of the 2nd Biennial Meeting of the International Environmental Modelling and Software Society. Osnabrück, Germany: iEMSs.
14. Bayraktarli YY, Yazgan U, Dazio A, Faber HM. Capabilities of the Bayesian probabilistic networks approach for earthquake risk management. First European Conference on Earthquake Engineering and Seismology 3–8, Sept.; Geneva, Switzerland 2006. 1458:1–10.
15. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag, 2001.
16. Wang Z, Yuan F, Sun Y. General introduction of engineering damage of Wenchuan Ms. 8.0 earthquake. Journal of Earthquake Engineering and Engineering Vibration (in Chinese), 2008; 28(Suppl):1–113.
17. Korb KB, Nicholson AE. Bayesian Artificial Intelligence. Boca Raton, FL: Chapman & Hall/CRC, 2004.
18. Jensen VF, Nielsen DT. Bayesian Network and Decision Graphs. New York: Springer, 2007.
19. Hoeting AJ, Madigan D, Raftery EA, Volinsky TC. Bayesian model averaging: A tutorial. Statistical Science, 1999; 14(4):382–417.
20. Dai C. Trend analysis of rainfalls evolution of Heihe Basin. Techniques of Gansu's Water and Electricity (in Chinese), 2008; 44(3):177–182.
21. Wang J, Meng J. Characteristics and tendencies of annual runoff variations in the Heihe river basin during the past 60 years. Scientia Geographica Sinica (in Chinese), 2008; 28(1):83–88.
22. Tobler WR. Cellular Geography, Philosophy in Geography. Dordrecht: Reidel, 1979.
23. Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kauffman, 1993.
24. Fulton T, Kasif S, Salzberg S (eds). Efficient Algorithms for Finding Multi-Way Splits for Decision Trees. Proc Twelfth International Conference on Machine Learning. San Francisco, CA: Kaufmann, 1995.
25. Fayyad U, Irani K (eds). Multiple-Interval Discretization of Continuous-Valued Attributes for Classification Learning. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence. Chambéry, France, August 28–September 3, 1993. San Mateo, CA: Kaufmann, 1993.
26. Elomaa T, Rousu J. Finding optimal multi-splits for numerical attributes in decision tree learning, 1996, Contract No.: NC-TR-96-041.
27. Dougherty J, Kohavi R, Sahami M (eds). Supervised and Unsupervised Discretization of Continuous Features Machine Learning: Proceedings of the Twelfth International Conference. San Francisco, CA: Morgan Kaufmann, 1995.
28. Dunteman G. Principal Component Analysis. Newbury Park, CA: Sage, 1999.
29. Bouckaert RR. Bayesian Belief Network: From Construction to Inference [Dissertation]. Utrecht: Universiteit Utrecht, 1995.
30. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco, CA: Morgan Kaufmann, 1988.
31. Cox AL. Risk Analysis: Foundations, Models and Methods. Norwell, MA: Kluwer Academic Publishers, 2001.
32. Breiman L. Random forests. Machine Learning, 2001; 45(1):5–32.
33. Geurts P. Contributions to Decision Tree Induction: Bias/Variance Tradeoff and Time Series Classification. Liège, Belgium: University of Liège, 2002.
34. Platt J. Machines using sequential minimal optimization. In Schoelkopf B, Burges C, Smola A (eds). Advances in Kernel Methods—Support Vector Learning. Cambridge, MA: MIT Press, 1998.
35. Littlestone N. Learning quickly when irrelevant attributes are abound: A new linear threshold algorithm. Machine Learning, 1988; 2(4):285–318.

36. Hosmer D, Lemeshow S. Applied Logistic Regression, 2nd ed. Hoboken, NJ: John Wiley and Sons, 2000.
37. John HG, Langley P (eds). Estimating Continuous Distributions in Bayesian Classifiers. Proc the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, Canada, 1995.
38. Mitchell T. Machine Learning. Columbus: McGraw Hill, 1997.
39. Bayraktarli YY, Ulfkjaer J, Yazgan U, Faber MH. On the Application of Bayesian Probabilistic Networks for Earthquake Risk Management, the 9th International Conference on Structural Safety and Reliability (ICOSSAR 05). Rome, Italy, 19–23, June, 2005.
40. Hincks T, Aspinall A, Woo G. An Evidence Science Approach to Volcano Hazard Forecasting. The International Association of Volcanology and Chemistry of the Earths Interior (AVCEI) 2004.
41. Straub D. Natural Hazards Risk Assessment Using Bayesian Networks. The 9th International Conference on Structural Safety and Reliability (ICOSSAR 05). Rome, Italy, 19–23, June, 2005.