Geoscience Frontiers 12 (2021) 101249

HOSTED BY

Contents lists available at ScienceDirect

Geoscience Frontiers

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

Research Paper

Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management

Zizheng Guo [a], Yu Shi [b], Faming Huang [b,*], Xuanmei Fan [c], Jinsong Huang [d]

[a] Faculty of Engineering, China University of Geosciences, Wuhan 430074, China
[b] School of Civil Engineering and Architecture of Engineering, Nanchang University, Nanchang 330031, China
[c] State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu 610059, China
[d] ARC Centre of Excellence for Geotechnical Science and Engineering, University of Newcastle, NSW 2287, Australia

ARTICLE INFO

Article history:
Received 3 February 2021
Revised 1 June 2021
Accepted 4 June 2021
Available online 06 June 2021
Handling Editor: E. Shaji

Keywords:
Landslide susceptibility
Frequency ratio
C5.0 decision tree
K-means cluster
Classification
Risk management

ABSTRACT

Machine learning algorithms are an important measure with which to perform landslide susceptibility assessments, but most studies use GIS-based classification methods to conduct susceptibility zonation. This study presents a machine learning approach based on the C5.0 decision tree (DT) model and the K-means cluster algorithm to produce a regional landslide susceptibility map. Yanchang County, a typical landslide-prone area located in northwestern China, was taken as the area of interest to introduce the proposed application procedure. A landslide inventory containing 82 landslides was prepared and subsequently randomly partitioned into two subsets: training data (70% landslide pixels) and validation data (30% landslide pixels). Fourteen landslide influencing factors were considered in the input dataset and were used to calculate the landslide occurrence probability based on the C5.0 decision tree model. Susceptibility zonation was implemented according to the cut-off values calculated by the K-means cluster algorithm. The validation results of the model performance analysis showed that the AUC (area under the receiver operating characteristic (ROC) curve) of the proposed model was the highest, reaching 0.88, compared with traditional models (support vector machine (SVM) = 0.85, Bayesian network (BN) = 0.81, frequency ratio (FR) = 0.75, weight of evidence (WOE) = 0.76). The landslide frequency ratio and frequency density of the high susceptibility zones were 6.76/km^2 and 0.88/km^2, respectively, which were much higher than those of the low susceptibility zones. The top 20% interval of landslide occurrence probability contained 89% of the historical landslides but only accounted for 10.3% of the total area. Our results indicate that the distribution of high susceptibility zones was more focused without containing more "stable" pixels. Therefore, the obtained susceptibility map is suitable for application to landslide risk management practices.

1. Introduction

Landslides are one of the most frequent geomorphologic processes occurring in China, especially in the Loess Plateau area (Zhuang et al., 2018; Peng et al., 2019; Tang et al., 2020). Located in the transition zone between the first and second units of China's topography, this area contains various complicated geological environments. The majority of the deposit area is covered by loess characterized by a heightened sensitivity to suction stress and collapsibility, which are prone to catastrophic landslides under rainfall (Zhang and Liu, 2010; Lian et al., 2020; Shu et al., 2020). For instance, the Xiangning landslide in Shanxi Province of the Loess Plateau, which occurred on March 15, 2019, caused 20 deaths and 13 injuries (Zhao and Zhao, 2020). Because these events are commonly characterized by large volumes and fast movement, it is difficult for local governments to perform successful emergency response campaigns. To avoid the potential loss of lives or damage to the environment and social economy, it is necessary to develop techniques to assess landslide hazards and risks (Corominas et al., 2014; Crawford et al., 2018).

* Corresponding author.
E-mail address: faminghuang@ncu.edu.cn (F. Huang).

https://doi.org/10.1016/j.gsf.2021.101249

Landslide susceptibility modelling allows us to portray the spatial distribution of landsliding at a regional scale (Guzzetti et al., 2006) and can be subsequently used for risk management and land use planning (Fell et al., 2008; Chen et al., 2019). There are three main types of approaches that have been developed for landslide susceptibility modelling: heuristic models, physically based models and data-driven models (Zêzere et al., 2017; Shu et al., 2019; Medina et al., 2021). Professional knowledge and experiences play an important role in heuristic models because these models estimate landslide potential by mainly considering investigator opinions. Hence, subjective judgement from experts on environmental variables is more important than on-site data and evidence (Barredo et al., 2000; Van Den Eckhaut et al., 2010). Physically based models can better reveal the mechanism of landslide occurrence, but the data acquisition about input parameters over large areas is generally an operational challenge (Sorbino et al., 2010; Bueechi et al., 2019). Meanwhile, most physical models are time-consuming because of the application of complex hydrological assumptions (e.g., Rossi et al., 2013). Hence, data-driven models have become more widely used in the recent past and can be divided into two categories, statistically based models and machine learning models (Pradhan, 2013; Reichenbach et al., 2018; Youssef and Pourghasemi, 2021). The former approach is mainly based on the process of weighting landslide-related factors using probabilistic or statistical techniques. Specifically, numerous attempts have been made by researchers to develop multicriteria methods (Feizizadeh et al., 2014; Kouli et al., 2014), bivariate or multivariate statistical methods (Yalcin et al., 2011; Felicísimo et al., 2013; Schlögel et al., 2018). However, these approaches generally require environmental factors that exhibit a normal distribution, which are seldom available in some cases. To remedy this, various machine learning methods have been employed in landslide susceptibility assessments, such as the extreme learning machine (Huang et al., 2017), artificial neural network (Lee et al., 2003; Yilmaz, 2009; Wang et al., 2019), and tree-based models (e.g., random forest) (Tsangaratos and Ilia, 2016; Wu et al., 2020). Additionally, several hybrid methods have also been proposed by integrating machine learning with statistical approaches, such as the fuzzy logic relation (Pradhan, 2011) and rough set-SVM (Peng et al., 2014). Generally, these models not only have higher accuracies but also allow the use of different types of environmental factors as input variables because these models are inherently nonlinear. Thus, unlike statistical models, machine learning models require less a priori historical landslide data (Pham et al., 2019), which helps such methods to be implemented in the determination of novel relationships in datasets.

Nevertheless, the occurrence of landslides is a geology-based engineering problem. Although many machine learning models are better choices considering their prediction accuracy, they fail to describe the mechanism of landslide occurrence. Additionally, spurious correlations and overfitted points are drawbacks of these "black box" models. In contrast, a decision tree (DT) algorithm is a "white box" model, which means that it offers an accurate depiction of the relationships between the input and output data (Ma et al., 2017). Moreover, it can present the importance of input factors, which is important for result analysis and factor selection. Its basic concept is that the complex group is split, including its independent variables, into several simpler groups by conditional methods, which may lead to an easier solution (Quinlan, 1993). Several algorithms have been proposed in the literature to implement the DT model and to perform landslide susceptibility evaluations, such as classification and regression tree, chi-square automatic interaction detector decision tree, and ID3 (Ture et al., 2009; Pradhan, 2013; Tsangaratos and Ilia, 2016). However, very few studies have employed the C5.0 algorithm, which is a relatively novel approach. Hence, landslide susceptibility mapping was conducted using the C5.0 decision tree model in this study. Meanwhile, to test the reliability of the resulting landslide susceptibility map, two machine learning algorithms and two statistically based models were utilized as a comparison, namely, support vector machine (SVM), Bayesian network (BN), frequency ratio (FR) and weight of evidence (WOE) models.

In addition to the calculation of landslide occurrence probability, the classification of landslide susceptibility is another challenge in seeking a suitable susceptibility map. Some studies adopted the expert-based approach, which divided the histogram of the probability map into different categories (Guzzetti et al., 1999; Dai and Lee, 2002). However, this type of continuously changing data into two or more categories was associated with great uncertainty. Currently, statistically based classification methods are commonly available, including natural breaks, quantiles, equal intervals and standard deviations (Baeza et al., 2016; Zhao and Chen, 2020). However, such methods can only be performed on a GIS platform and are neither statistically tested nor fully automated (Ayalew and Yamagishi, 2005). In fact, a few studies have shown the disadvantages of this type method. For example, the natural break is useful only when there are large jumps in the dataset. The equal interval approach is a cut-off-dependent approach in which the results may vary with the breakpoints of reclassification (Zhou et al., 2018). To overcome these drawbacks, a novel classification method based on the K-means cluster algorithm was proposed in this study to rapidly classify landslide susceptibility.

The present study aims to generate a reliable landslide susceptibility map at a regional scale. Yanchang County on the Loess Plateau (China) was selected as the area of interest for this case study. This region has experienced various landslide hazards throughout the past few decades, so this study is of great importance to update the study of model application, especially with the purpose of improving the accuracy of the results. Specifically, our objective mainly includes: (i) the determination of the probability of landslide occurrence by using a decision tree approach, (ii) the achievement of landslide susceptibility zonation by the K-means cluster algorithm, and (iii) the analysis and validation of model effectiveness.

2. Materials

2.1. Study area

This study was conducted in Yanchang County, which is located in the Yan'an city of the Loess Plateau area, China, and includes 13 towns (Fig. 1). It lies between longitudes ranging from 109°33′E to 110°30′E and latitudes ranging from 36°14′N and 36°46′N, with a total area of approximately 2368 km^2. The population of this area is approximately 1.4×10^5, so it is considered a populated region with a density of more than 50 people/km^2. The elevation of the area ranges from 473 m above sea level (a.s.l.) at the river valley to 1369 m a.s.l. at the highest peak, characterized by a low southeast and high northwest trending terrain (Fig. 1b).

From the perspective of geology, a total of 5 sedimentary rock units exist in the area, which are only associated with Triassic and Quaternary deposits. The strata contain the Yongping (T_3y), Wayaobao (T_2w), Hujiacun (T_3h), and Tongchuan (T_2t) formations from the Triassic period, and Quaternary loess (Li, 2018; Guo et al., 2019a). Among them, $Qp_{1\text{-}3}$ and T_2t, which are composed of alternating layers of sandstone and mudstone, are the most common outcrops. The climate regime is the mainland monsoon,

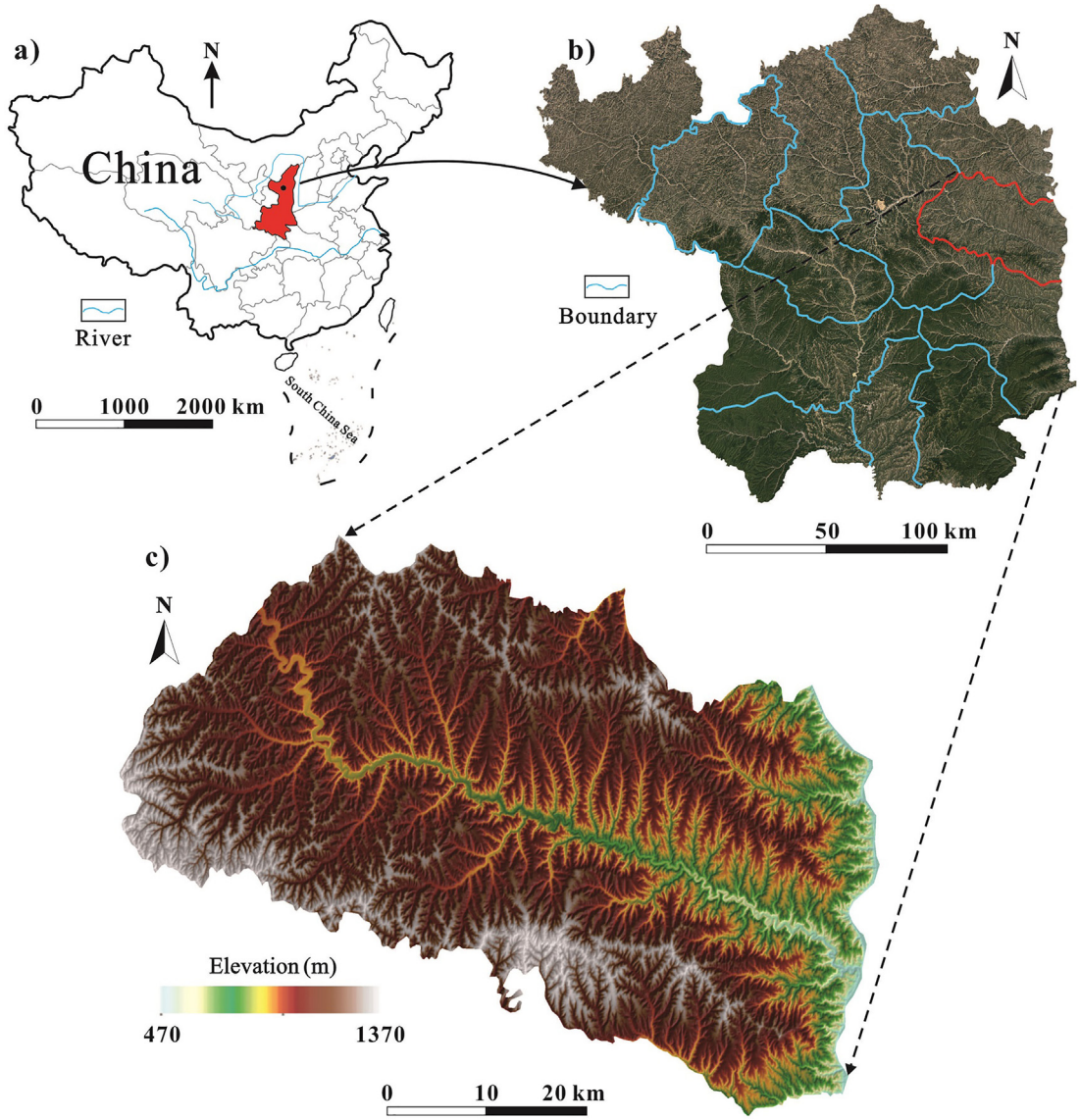

Fig. 1. Location of the study area (the coordinate system used is WGS84). (a) Location of Shaanxi Province in China, (b) remote sensing image of the Yan'an City from Google Earth, and (c) the topography map of the Yanchang County showed by the digital elevation model.

which is dominated by two factors, i.e., the orographic effects of the Qinghai-Tibet Plateau and the prevailing northwest wind from Mongolia (Li, 2018). Since it is a mountainous area, the climate can vary locally due to elevation differences: in the area with elevations below 800 m a.s.l., the annual rainfall mostly ranges from 450 mm to 500 mm, whereas in the western part with elevations above 1100 m, the average annual rainfall reaches approximately 550 mm (Guo et al., 2019a). Approximately 60%–70% of the total annual rainfall is concentrated in the rainy season (July–September). The rivers and valleys in Yanchang County are well developed, and dozens of large and small streams compose a complex stream network, among which the Yan River is the largest one. It flows through the area from west to east, and the settlements are mainly distributed along the banks of the river.

2.2. Data sources

The data used in this study mainly include: (i) the digital elevation model (DEM), (ii) remote sensing (RS) images, (iii) the geological map, (iv) detailed geological survey reports, (v) photos of landslides, and (vi) aerial images. Detailed information on the data is listed in Table 1.

Table 1
The sources and characteristics of the data used in the paper.

No.	Data	Scale	Resolution (m)	Source	Purpose
i	DEM	/	30	http://www.gscloud.cn/	Preparing 9 factor maps: Fig. 3a, b, c, d, e, f, g, h, j
ii	RS images	/	30	Landsat 8 satellite	Preparing 3 factor maps: Fig. 3k, l, m
iii	Geological map	1:100,000	/	Local monitoring institute, archived documents, literature	Preparing lithology map
iv	Geological survey reports	/	/		Landslide inventory, model validation

2.3. Landslide mapping

Landslide inventory is a basic but essential tool for landslide hazard management, representing a fundamental base of knowledge on the spatial distribution of existing landslides (Tian et al., 2019). In the present study, the landslide inventory was obtained based on different data as follows: (i) the geological survey report provided by Xi'an Center of China Geological Survey and (ii) visual interpretation of remote sensing images taken from Google Earth (https://www.google.com/earth/). Then the detailed locations of the landslides were compared with previous literature (Li, 2018; Guo et al., 2019a) because it can help to calibrate the landslide inventory (Guzzetti et al., 2012; Taylor et al., 2015). The survey reports contained the detailed characteristics of landslides (e.g., location, area, volume, etc.), so a specific GIS-based database was created to store and process all the collected data, which were linked to the landslide distribution in ArcGIS.

The landslide inventory reveals 82 landslide hazards in the area (Fig. 2a), with a total area of 3.09 km^2, accounting for approximately 0.1% of the total area of Yanchang County. The cumulative volume of the landslides is approximately 5.2×10^7 m^3, among which one landslide has a volume greater than 10×10^6 m^3, and 37 landslides have volumes ranging from 10×10^5 m^3 to 10×10^6 m^3 (Guo et al., 2019a). Considering the Varnes classification system (Varnes, 1978; Hungr et al., 2014), these landslides can be roughly divided into three categories (Fig. 2b–d, modified from Li, 2018): (i) small and shallow earth slides, most of which involve the ground surface soil layer; (ii) shallow debris flows; and (iii) composite earth slide–debris flows. However, the total number of landslides classified as types (ii) and (iii) are small (less than 10), and their depths are similar to that of type (i). Given that our main objective is not to include landslide typology in the landslide susceptibility assessment, we took all the landslides as being a single group in the current analysis. A comparison with some recent studies also showed that it was acceptable for use in regional assessments (Shu et al., 2019; Medina et al., 2021). From the perspective of occurrence time, most of the landslides occurred during the rainy season, whereas only a few landslides occurred in the dry season period, thus indicating that rainfall is an important factor triggering landslides in the area. Meanwhile, since loess, which is widely distributed, has a high collapsibility and a great water sensitivity, the region is characterized by many gullies and evident soil erosion. Many loess landslides are activated under severe rainfall events, engineering activities and agriculture irrigation practices.

2.4. Preparation of influencing factors

According to the field survey and available data, fourteen influencing variables associated with topographical, hydrological, geological and environmental factors were prepared for the preliminary analysis of landslide susceptibility mapping. Specifically, these factors included the elevation, slope, aspect, plan curvature, profile curvature, surface roughness (SDS), surface cutting

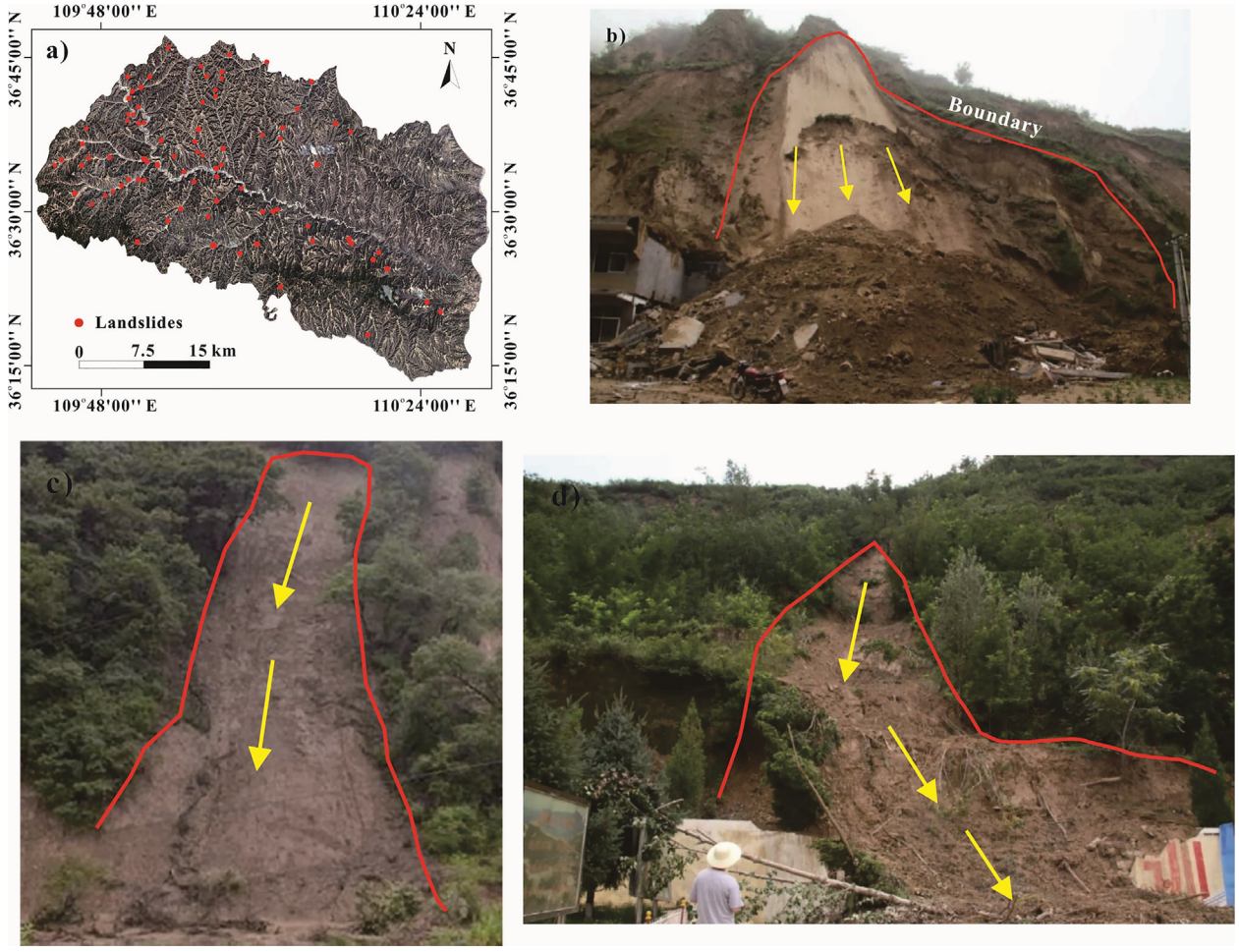

Fig. 2. Preparation of the landslide inventory in the Yanchang County. (a) Spatial distribution of the landslides, (b) a shallow landslide, (c) a small-scale debris flow, and (d) a composite earth slide–debris flow.

depth, relief degree of land surface (RDLS), distance to rivers, topographic wetness index (TWI), modified normalized difference water index (MNDWI), normalized difference vegetation index (NDVI), normalized difference barren index (NDBI) and lithology. Then, the corresponding thematic factor layers were obtained in raster form based on the GIS platform. Among these factors, aspect and lithology are discrete variables that have fixed categories, and the other twelve factors are continuous variables. To achieve reasonable classification of these variables, every variable was discretized into several small classes using the same intervals first. Then, the categories with similar frequency ratio values (see methodology section for the principles of FR method) were placed into the same group (Fig. 3). Generally, the division of factors is rough if the number of classes is small, while the model complexity is large if the number of classes is high (Huang et al., 2020a). Some studies (Aditian et al., 2018; Chang et al., 2020; Huang et al., 2020a) have shown that class numbers between 4 and 12 are conductive to landslide susceptibility analysis, and the classification of all factors in this study fits this point well. The preparation of these maps and their impacts on landslide occurrence are described below.

Elevation (Fig. 3a): Environmental settings on slopes normally vary with elevation; thus, elevation is often considered an important factor driving landslide occurrence. The digital elevation model (DEM) with a 30-m resolution of Yanchang County was downloaded from the website (http://www.gscloud.cn/home).

Slope (Fig. 3b): The slope angle can directly affect slope stability and has been widely used in landslide susceptibility analysis (e.g., Catani et al., 2013). The slope map of the study area was created from the digital elevation model (DEM) with a resolution of 30 m. It was divided into five classes.

Aspect (Fig. 3c): Aspect was also derived from the DEM and was first divided into nine classes (i.e., flat area, north, northeast, east, southeast, south, southwest, west, northwest) according to the geographic orientations of the topography. Then, the classes with similar FR values were combined into one class, and the resulting map contained seven classes.

Plan curvature (Fig. 3d): Plan curvature can be described as the curvature of a hypothetical contour that passes through a specific pixel. It reflects the rate of change of aspect along a contour and thus can affect the flow of water (direction and amount, etc.) across a surface. The plan curvature map was generated from the DEM and divided into seven classes.

Profile curvature (Fig. 3e): Profile curvature influences the acceleration and deceleration of flow through slopes, thus some valuable information about erosion and deposition is provided (Wu et al., 2020). Its values ranged from 0 to 30 and were divided into five classes.

Surface roughness (SDS) (Fig. 3f): Surface roughness is an index that can reflect the fluctuation degree and erosion intensity of the land surface. It can be defined as the variability of the slope angle in a specific area and is calculated as the standard deviation of the slope (i.e., SDS) (Atkinson and Massari, 1998). Mathematically, it can be estimated as follows:

$$\mathrm{SDS} = 1/\cos(S) \tag{1}$$

where S is the slope. The SDS values of Yanchang County vary from 1 to 1.57, and are divided into six classes.

Surface cutting depth (Fig. 3g): The surface cutting depth is defined as the difference between the lowest elevation and the average elevation in a certain area around a given point (Huang et al., 2020b). It can be used to represent the degrees of erosion of a surface. The neighbourhood statistics tool in ArcGIS was used to calculate the shape of the area around a specific cell. The statistic type tool was then used to obtain the mean and minimum values of elevation within the neighbourhood, which had a size of 3 m × 3 m cells. The difference in the mean and minimum values obtained from the raster calculator tool was the map of the surface cutting depth. This map was divided into six classes.

Relief degree of land surface (RDLS) (Fig. 3h): Defined as the maximum difference in height per unit area, RDLS can reveal terrain characteristics, so it was considered as an important influencing factor in some previous studies (Tien Bui et al., 2012; Huang et al., 2018). The procedure to generate the RDLS map was similar to that of the surface cutting depth. The maximum and minimum values of elevation within the neighbourhood were calculated in ArcGIS, and their difference was the RDLS value. It ranged from 0 to 85 and was divided into nine classes.

Distance to rivers (Fig. 3i): Rivers impact slope stability because they can cut and erode riverbanks, and these behaviours reshape and carve geomorphology. Moreover, fluctuations of the water level affect the water table of the slopes to a great extent. We used the distance to rivers as the index to reflect the effect of rivers on landslides, and the index had eight classes.

Topographic wetness index (TWI) (Fig. 3j): As a hydrological parameter, the TWI describes the topographic attributes of hydrological processes, because both slope and local upslope contributing areas are considered (Moore et al., 1991). The equation to calculate TWI is expressed:

$$\mathrm{TWI} = \ln(a/\tan\beta) \tag{2}$$

where a is the upslope area draining from a specific cell and $\tan\beta$ is the slope angle of this cell. To generate the TWI map, the hydrographic statistics tools in ArcGIS were used to calculate the flow direction and accumulation at each cell. Normally, areas with smaller TWI values are recognized as landslide-prone area (Achour and Pourghasemi, 2020). The results of Yanchang County fitted well this point: The TWI ranging from 2.925 to 26.369 was divided into five classes.

Modified normalized difference water index (MNDWI) (Fig. 3k): MNDWI is a hydrological factor that can reflect the water information of the ground surface (McFeeters, 1996; Xu, 2006). Hence, the effect of hydrological conditions on landslide occurrences can be recognized by this factor to a certain degree. It was obtained from the Landsat 8 images by using the equation as follows:

$$\mathrm{MNDWI} = \frac{P(Green) - P(MIR)}{P(Green) + P(MIR)} \tag{3}$$

where P ($Green$) and P (MIR) are measurements of the spectral reflectance from remote sensing images, and they are the green band and middle infrared band, respectively. In this study, MNDWI was divided into four classes and it can be found that landslides were mainly distributed in the area with large MNDWI values.

Normalized difference vegetation index (NDVI) (Fig. 3l): NDVI reflects the greenness degree of an area and may change the distribution of soil and hydrological processes on slopes. Therefore, it can be used as a proxy for the land use map (Arabameri et al., 2020). It is normally derived from remote sensing images and can be calculated using the equation:

$$\mathrm{NDVI} = \frac{P(NIR) - P(Red)}{P(NIR) + P(Red)} \tag{4}$$

where P (NIR) and P (Red) represent the spectral reflectance of the infrared band and red band, respectively. In this study, the Landsat 8 images were selected as data sources to create NDVI maps. NDVI values in the area had a range of 0.054–0.879 that was divided into five classes.

Normalized difference building index (NDBI) (Fig. 3m): NDBI reflects the density of building distribution, thus it can be consid-

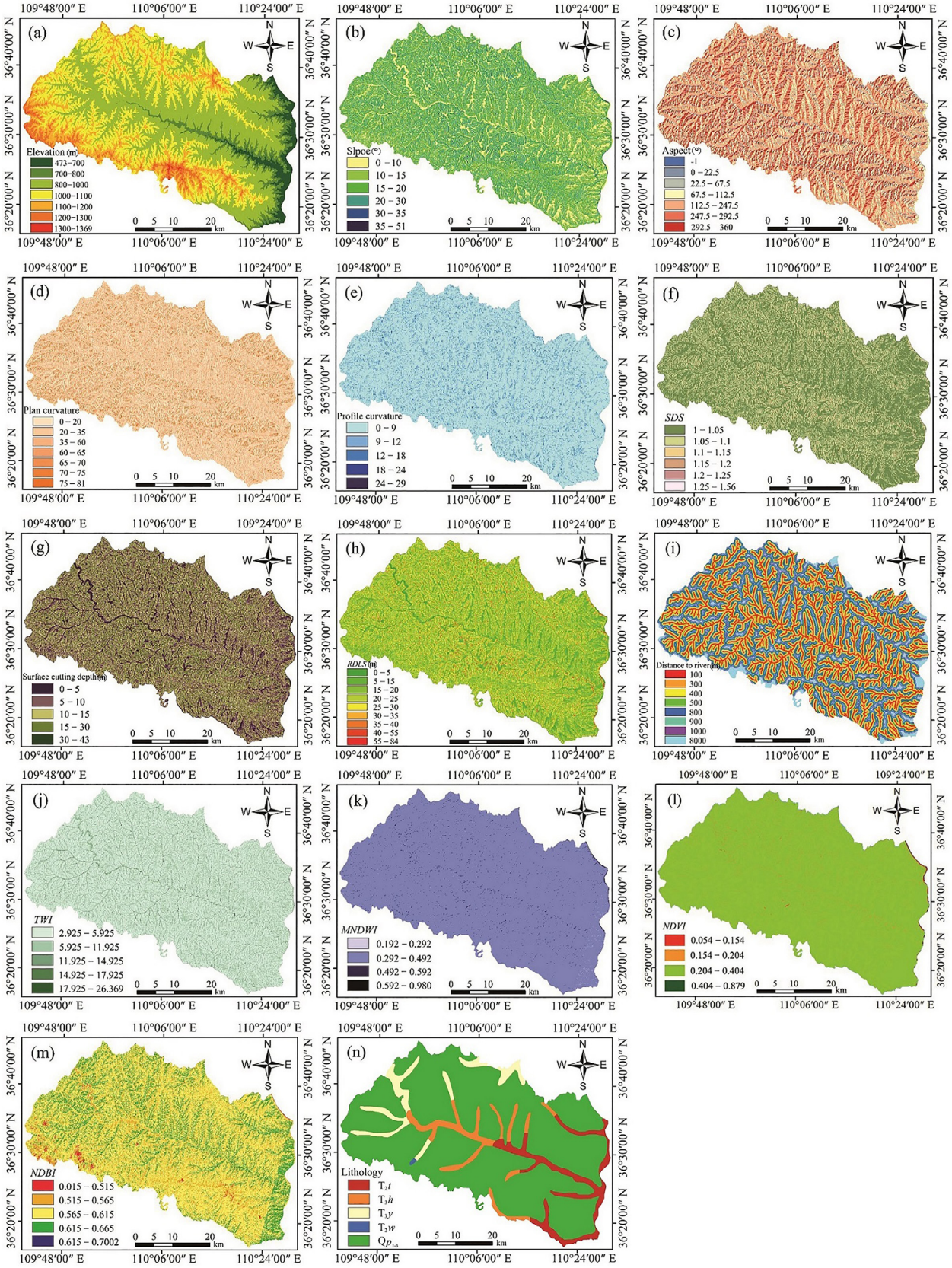

Fig. 3. Influencing factor maps used for landslide susceptibility modelling. (a) Elevation, (b) slope, (c) aspect, (d) plan curvature, (e) profile curvature, (f) SDS, (g) surface cutting depth, (h) RDLS, (i) distance to rivers, (j) TWI, (k) MNDWI, (l) NDVI, (m) NDBI and (n) lithology.

ered an index to represent the population and engineering activities. The principle to generate the NDBI map is similar to that of NDVI (Huang et al., 2018). The NDBI can be obtained as follows:?>

$$\mathrm{NDBI} = \frac{P(MIR) - P(NIR)}{P(MIR) + P(NIR)} \tag{5}$$

where $P(MIR)$ and $P(NIR)$ are the spectral reflectance of the middle infrared band and infrared band of Landsat 8 images, respectively.

Lithology (Fig. 3n): The mechanical and hydrological properties (e.g., permeability and friction angle) of rock masses vary between lithological units, so this factor may greatly influence slope stability. The lithology map was obtained from the geological map at a scale of 1:100,000. We find that the lithology units outcropping in the study area are limited, including only Triassic (T) and Quaternary (Q) units. According to chronological order, the Triassic group was divided into middle (T_2) and upper (T_3) units.

3. Methodology

3.1. Frequency ratio model

One basic assumption in landslide susceptibility assessment based on data-driven models is that future landslides are more likely to occur under the same/similar external conditions that led to past landslides (Zêzere et al., 2017). Hence, it is normally an important step to analyze the correlation between historical landslides and environmental conditions. Frequency ratio (FR) is a commonly-used model on this issue which can expose the statistical associations between landslide distributions and each influencing factor. The principle of it has been reported by some previous studies (e.g., Yilmaz, 2009; Yalcin et al., 2011). If one landslide-related factor is divided into several categories, the FR of one given category can calculated as follows:

$$\mathrm{FR} = \frac{N_i/TN}{A_i/TA} \tag{6}$$

where N_i is the area of landslides in the ith category, TN is the total area of landslides in the study area, A_i is the area of ith category, and TA is the total area of the study area. In general, the FR is an index to reflect the density of landslide distribution in a certain range of one influencing factor. If FR is more than 1, the category can be considered as positive to landsliding. On the contrary, the value of less than 1 represents a negative condition for the landslide occurrence.

3.2. C5.0 decision tree model

The DT model is essentially a tree involving a set of decision nodes, among which the root and each internal node are labelled with a question (Pradhan, 2013). The arcs descend from each root node to leaf nodes, where a solution to the associated question is offered. A split is created at each node by making a binary decision, which separates one class or several classes from the global dataset. The C5.0 is a kind of algorithm which calculates the best splits based on information gain ratio (IGR). The IGR is considered as a probability-based measure used to calculate the level of uncertainty reduction. Generally, the decision tree grows down by calculating the split with the biggest IGR until the best solution is available. The IGR is calculated as follows:

$$\mathrm{GainRatio} = \frac{Gains(N,T)}{Ent(T)} \tag{7}$$

where GainRatio is the IGR, N is the global dataset, T is the predictor variable, and $Gains(N,T)$ is the entropy difference between the original and new nodes, which is calculated as follows:

$$Gains(N,T) = \left[\sum_{i}^{t} P(C_i|N)\log_2 P(C_i|N)\right] \times \left(\sum_{j}^{k} \frac{|T_j|}{|N|} - 1\right) \tag{8}$$

where C is a set of target variable, t is the category number of C, K is the category number of T, i.e, C_i (i = 1,2,. . .,t), T_j (j = 1,2,. . .,k).

In the modelling process, the boosting algorithm was adopted to improve the robustness of the C5.0 DT model, which can control both the bias and variance (Dou et al., 2020). Its basic procedure can be divided into the following steps: (i) initialize the weights of the training sample; (ii) obtain training subsets sequentially; (iii) calculate the error of the subsets and update the weights; and (iv) end the training and evaluate the classification results. Additionally, a cross-validation method was utilized to investigate the verification accuracy of the model. According to Yao et al. (2008), the dataset is divided into n folds, among which one fold is utilized for validation, and the other folds ($n-1$ folds) are the training data. When every fold is selected as the validation set, n model accuracies are obtained by iterating the same step. Finally, the average of these obtained accuracies, which is referred to as the cross-validation accuracy (CAV), is considered the model accuracy. Compared to the traditional approach, this method is helpful for overfitting problems and improving the generalization capability of the DT model.

3.3. K-means cluster algorithm

Clustering is a useful unsupervised learning technique because it achieves the division of unknow objects into several groups. The members in each group have similar properties and characteristics. The K-means cluster algorithm is a relatively simple way to implement clustering analysis. Detailed information about the algorithm has been presented by Hartigan and Wong (1978) and Melchiorre et al. (2008). The principle for it is as follows:

(i) A certain number K initial centroids of the given input data are determined randomly, then the data are divided into several groups. The Euclidean distance (d) between the data and centroids are calculated as follows:

$$d(X_t, X_\xi) = \sqrt{\sum_{u=1}^{n} (X_{ut} - X_{u\xi})^2} \tag{9}$$

where X_t and X_ξ are input data and given centroids, respectively, u is the data property, l is the number of the properties. When modelling the landslide susceptibility, the occurrence probability of the landslide is the data property, so n is set as 1 in this study.

(ii) After the first time calculation of d, the obtained result is recorded as $D^{(0)}$={$D_1^{(0)}$, $D_2^{(0)}$, . . ., $D_k^{(0)}$}, and the new centroids are updated using the following equation:

$$X_\xi^{(m)} = \frac{1}{h_\xi^{(m-1)}} \sum_{X_t = D_{t\xi}^{(m-1)}} X_t \tag{10}$$

where X_ξ^m is the new centroid, m is the number of iterations, $h_\xi^{(m-1)}$ is the number of data in the new group based on new centroids.

(iii) iterating the above step, and ending the calculation process when $_\xi^{(m)} = h_\xi^{(m-1)}$ and $D^{(m)} = D^{(m-1)}$. The obtained centroids are the clustering centres of the input data.

3.4. Contribution of influencing factors

It is essential to evaluate the relative importance of influencing factors because it can help to establish appropriate factor systems for landslide susceptibility analysis. Many techniques have been employed to quantify the contribution of factors, such as random forest, learning vector quantization and principle component analysis (Pavel et al., 2011; Tang et al., 2020; Youssef and Pourghasemi,

2021). In this study, the C5.0 package (Kuhn et al., 2015) based on the R 4.0.3 environment was used to calculate the contribution of factors. The package provides a function named C5.0imp to achieve this goal. By default, the function determines the factor importance by calculating the percentage of training datasets that fall into the terminal nodes.

3.5. Accuracy measures

Some studies evaluate the result accuracy by accounting for the number or frequency of landslides located in different susceptibility zones (e.g., Ahmed, 2015; Guo et al., 2019b), but it is not logical because the areas of different landslide susceptibility zones vary. Hence, our validation task was conducted mainly using two indicators: landslide frequency ratio (FR) and landslide frequency density (FD). These indicators can be calculated as follows:

$$\mathrm{FR} = \frac{S_i/S}{A_i/A} \tag{11}$$

$$\mathrm{FD} = N_i/A_i \tag{12}$$

where S_i is the landslide area in each susceptibility zone, S is the total area of landslides, A_i is the area of a specific landslide susceptibility zone, A is the total area of the study area, and N_i is the number of landslides in each susceptibility zone. Considering the landslide area and susceptibility area at the same time, the indicators reveal the landslide distribution more appropriately. Additionally, the receiver operating characteristic (ROC) curve that has been widely accepted, was also used in this study. The area under the curve (AUC) is an important index to reflect the model accuracy. The details on its principle have been introduced by previous literature (e.g., Frattini et al., 2010; Wu et al., 2020). All these methods are done with the help of historical landslide locations.

4. Modelling procedure

After determining the landslide inventory map and influencing factor maps, the results from the FR analysis were used as the input, and the C5.0 decision tree and K-means cluster algorithm were integrated to generate the final landslide susceptibility map. The study area contains a total of 2,622,482 cells, while 3432 cells involve landslides, which were divided into two parts: 70% of the cells (2402 cells) were randomly selected as the training dataset, and the remaining 30% (1030 cells) were used for validating the model. Meanwhile, the same number of non-landslide cells were also selected because they offer the necessary information on unfavorable conditions for landslide occurrence. As a result, during the training process, the attribute matrix $F^{(t)}_{4804\times14}$ presenting the influencing factors of these cells was set as the input data, while the output data was the occurrence probability matrix of the landslide events ($P^{(t)}_{4804\times1}$), which was presented as binary response data, i.e., 0 and 1. Similar settings were established during the validation stage: the influencing factor matrix ($F^{(v)}_{30,000\times8}$) and landslide occurrence probability matrix ($P^{(v)}_{30,000\times1}$) of validation cells were considered as input data and output data, respectively.

The modelling process adopted SPSS Modeler software, which mainly contained the following procedures:

(i) The training dataset was first input into the software to generate the decision tree based on the C5.0 algorithm. The boosting and cross-validation techniques were used to improve the robustness of the training results;

(ii) The influencing factor values on every cell were extracted using GIS and were subsequently input into the constructed C5.0 DT model. The occurrence probabilities of landslides in these cells were obtained, and all the values were expressed in nondimensional terms ranging from 0 to 1.

(iii) The factor values of all landslide cells and the same number of non-landslide cells and their landslide state (0: non-landslide; 1: landslide) were combined into a matrix. The matrix was then input to R 4.0.3 software, and the C5.0 package was loaded to calculate the contribution of such factors. In this process, the number of boosting iterations was set to 1, and the confidence factor was 0.95.

(iv) The occurrence probability matrix obtained from (ii) was introduced into SPSS software, and the K-means cluster algorithm was used to obtain five centroids in the dataset.

(v) The data near one centroid were reclassified into the same group, and this centroid was considered the center of this group. The average of two adjacent centroids was considered the cut-off value between different susceptibility zones, because this value distinguished two groups of data with different properties. Based on this, the landslide susceptibility map was created, where the study area was classified into five susceptibility zones ranging from very low to very high.

(vi) Finally, the model accuracy was validated by analyzing the distribution of landslide inventory points and random points. Moreover, the model's performance was compared with other models.

The overall flow chart of this study was showed in Fig. 4.

5. Results

5.1. Factor importance and landslide susceptibility mapping

The final resulting classification scheme of each factor and the FR of each category are presented in Table 2. We can see the influences of different categories within one factor on landslide occurrence. For instance, most landslides distributed in areas with middle elevations (800–1100 m), and areas with lower and higher elevations only report a small number of landslides. The FRs of all classes were less than 1 except for the plan curvature of 0–20, thus indicating that this class had a positive effect on landslides. FRs roughly increased with increasing surface cutting depth. This is mainly because a large cutting depth provides a longer distance and larger space for the movement of slopes. Most landslides are located within 500 m of rivers, which reveals that rivers in the area are a positive factor for landslide occurrences. The number of landslides in the Quaternary and upper Triassic units was the largest, where the outcropped lithologies were mainly loess, sandstone and mudstone.

The obtained importance of the influencing factors is shown in Fig. 5. Among the 14 factors, five factors had relatively high contributions to landslide susceptibility, namely, lithology (importance measurement (IM) = 1), elevation (IM = 0.97), slope (IM = 0.97), distance to rivers (IM = 0.93) and aspect (IM = 0.91). Seven factors had a relatively low importance, including surface cutting depth (IM = 0.19), TWI (IM = 0.13), NDBI (IM = 0.10), plan curvature (IM = 0.06), SDS (IM = 0.04), NDVI (IM = 0.01) and MNDWI (IM = 0). The two factors (profile curvature and RDLS) with IM values between 0.4 and 0.5 had moderate contributions to landslide susceptibility. Overall, this probably indicates that geological (lithology) and topographical factors (e.g., elevation, slope and aspect) are more important for landsliding in the region than the evaluated hydrological (e.g., TWI and MNDWI) and environmental (e.g., NDBI and NDVI) factors. Additionally, none of them had a neg-

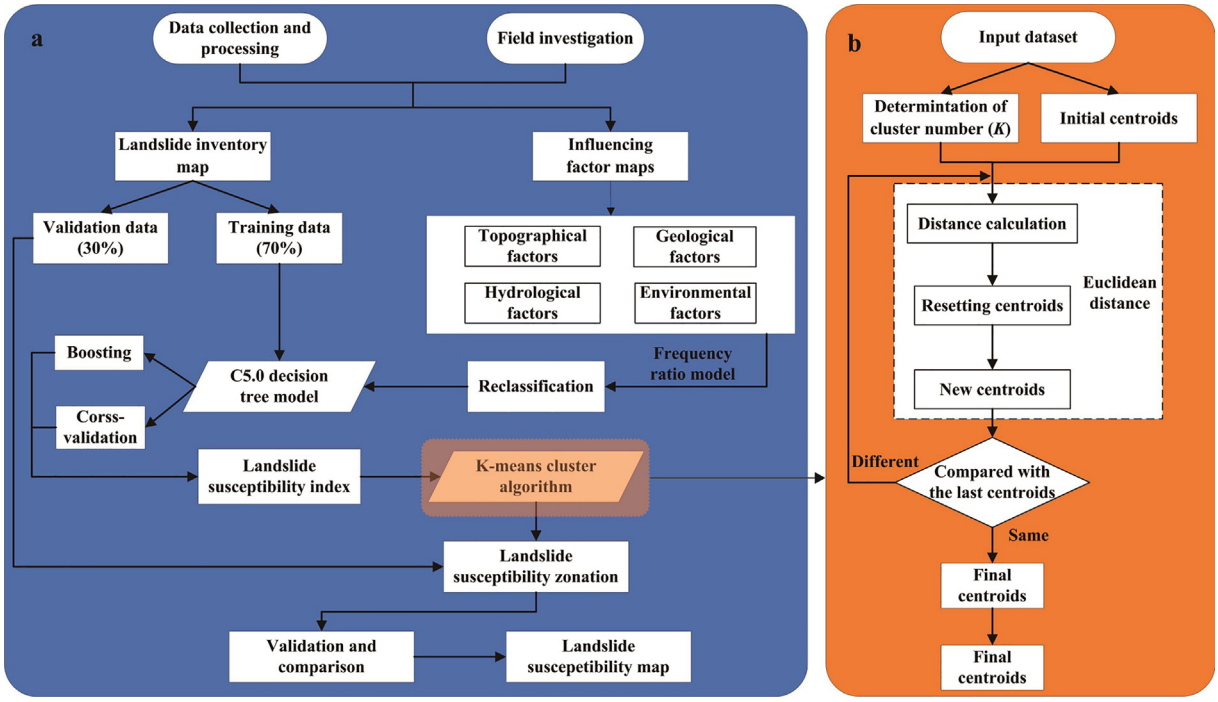

Fig. 4. The flow chart of the study. (a) The modelling process for landslide susceptibility mapping, and (b) the process of K-means cluster algorithm.

Table 2
The frequency ratio of each category within the influencing factors.

Factor	Category	FR	Factor	Category	FR	Factor	Category	FR
Elevation (m)	400–700	0	Profile curvature	9–12	0.900	Distance to river (m)	400–500	0.748
	700–800	0.026		12–18	0.772		500–800	0.391
	800–1000	1.663		18–24	0.566		800–900	0.705
	1000–1100	0.823		24–30	0		900–1000	0.937
	1100–1200	0.381	SDS	1–1.05	0.567		>1000	0.211
	1200–1300	0.552		1.05–1.1	1.692	TWI	2.925–5.925	1.198
	1300–1400	0		1.1–1.15	2.121		5.925–11.925	0.723
Slope	0°–10°	0.198		1.15–1.2	1.638		11.925–14.925	0.130
	10°–15°	0.605		1.2–1.25	0.666		14.925–17.925	0.054
	15°–20°	1.253		1.25–1.57	0		17.925–26.369	0
	20°–30°	1.923	Surface cutting depth	0–5	0.227	MNDWI	0.192–0.292	0.467
	30°–35°	1.616		5–10	0.593		0.292–0.492	0.933
	35°–51°	0		10–15	1.122		0.492–0.592	1.858
Aspect	−1°	0		15–25	2.078		0.592–0.980	0
	0°–22.5°	0.640		25–30	2.446	NDVI	0.054–0.154	0
	22.5°–67.5°	1.459		30–44	0		0.154–0.204	1.441
	67.5°–112.5°	0.765	RDLS	0–5	0.037		0.204–0.404	0.988
	112.5°–247.5°	1.252		5–15	0.298		0.404–0.879	0
	247.5°–292.5°	0.486		15–20	0.672	NDBI	0.015–0.515	0
	292.5°–360°	0.939		20–30	1.147		0.515–0.565	1.179
Plan curvature	0–20	1.390		30–35	1.875		0.565–0.615	0.867
	20–35	0.994		35–40	2.843		0.615–0.665	1.399
	35–60	0.646		40–45	2.272		0.665–0.7	0
	60–65	0.440		45–55	1.861	Lithology	T_2w	0
	65–70	0.210		55–85	0		T_3h	1.956
	70–75	0.090	Distance to river	0–100 m	1.036		T_3y	2.423
	75–82	0.426		100–300 m	1.873		T_2t	0.084
Profile curvature	0–9	1.035		300–400 m	1.034		Qp_{1-3}	0.933

ative value of importance, so it is reasonable to consider them all as influencing factors for the landslide susceptibility analysis in Yanchang County. However, such results do not mean that every factor can improve the model performance. As Glade and Crozier (2005) reported, adding data into the input dataset can improve the predictive capability of the model with a given complexity, but the model performance perhaps decreases if the data availability continuously increases.

As shown in Fig. 6a, a landslide susceptibility map created adopting the method proposed in this study was presented, where the K-means cluster was used to obtain the landslide susceptibility zonation. To make a comparison, the natural break, which is a commonly used classifier in ArcGIS, was also applied to generate the landslide susceptibility map (Fig. 6b). We can see some differences between these two maps. The map from the K-means cluster method had fewer very high (high) susceptibility areas than the

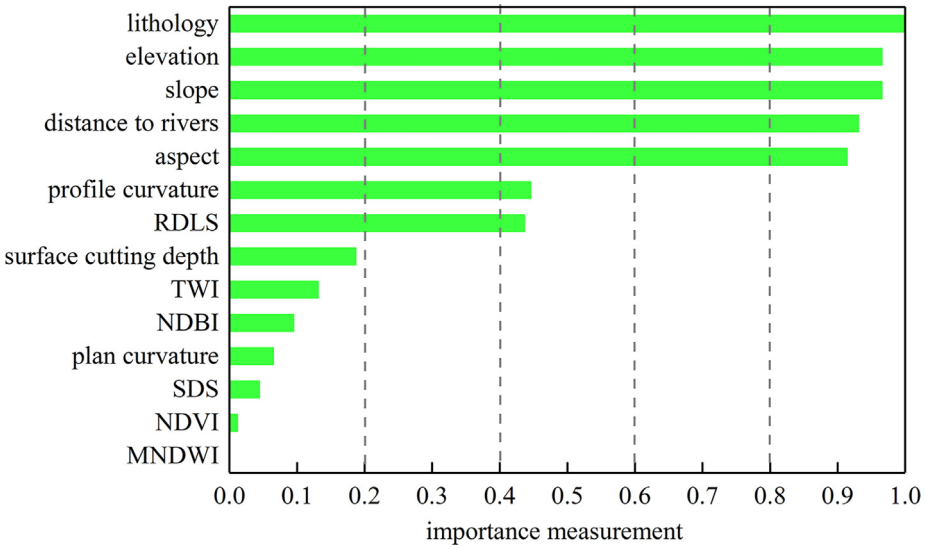

Fig. 5. Relative importance calculated from C5.0 package in *R* software.

map obtained employing natural breaks. In contrast, Fig. 6b had fewer very low (low) susceptibility areas than Fig. 6a. This is mainly because the two classifications led to different thresholds of probability of landslides calculated by the C5.0 DT model.

Regarding the spatial distribution, the high- and very high-susceptibility zones were mostly distributed in areas near the small river networks, especially in the central and northern parts of the area. However, the eastern part where the largest river flows through is mainly low susceptibility areas. This is mainly because this area is characterized by low elevations and flat slope angles, which make it difficult to induce landslides. The summary regarding the landslide characteristics shows that nearly all landslides in the study area have a shallow depth of less than 10 m. Because the amount of annual rainfall in the area is generally small, the most frequent triggering factor is not heavy rainfall events, but engineering activities. This leads us to the conclusion that the distribution of settlement may have a large impact on landslides. Actually, the centre of Yanchang County was mainly constructed along rivers, where fast urban development has been observed over the past decade. Hence, shallow landslides have been induced by human engineering activities. This phenomenon explains why the distance to rivers has a relatively large weight on landslide susceptibility. Another spatially distributed zone associated with high/very high landslide susceptibility is located in the southern part of the study area. Given that it covers many areas with moderate elevations and relatively high slope angles (20°–35°), the special properties of loess deposits may be the reason for slope instability in this zone. This can be explained by a similar conclusion has been obtained on the Loess Plateau by previous works (e.g., Zhang and Liu, 2010; Shu et al., 2020). Additionally, an important shared characteristic of zones characterized by high landslide susceptibility is that they are mostly covered by the strata of $Qp_{1\text{-}3}$ (Quaternary loess) and T_3y (mainly composed of siltstone and mudstone). This can also be inferred from the results listed in Table 2: the frequency ratios of these strata are 3.225 ($Qp_{1\text{-}3}$) and 3.3 (T_3y), which are much larger than those of other lithologies in the area.

Overall, the resulting landslide susceptibility maps have similar patterns as the landslide spatial distribution, especially regarding the high susceptibility zones. From the perspective of qualitative analysis, the well-matched relationship between susceptibility zonation and landslide locations represents the appropriate identification of exiting landslides employing these models.

5.2. Model validation and accuracy analysis

To better quantitatively explain the rationality and performance of the model, two machine learning approaches (support vector machine and Bayesian network models) and two statistically-based approaches (frequency ratio and weight of evidence models) were used for comparison. The K-means cluster was selected as the classification method. The resulting landslide susceptibility maps are shown in Fig. 7. The differences among the four maps here were more evident than in Fig. 6. Given that the models used to calculate landslide susceptibility were inherently different (statistical models and machine learning models), such differences were reasonable.

According to Eqs. (11) and (12), two statistical indicators, namely, FR and FD were calculated (Fig. 8). Generally, the histograms of the indicator for different susceptibility levels show the regularity: the higher the susceptibility level is, the larger the indicator is (Tang et al., 2020; Zhao and Chen, 2020). It can be seen that the results fit well with this regularity. For the very low and low susceptibility levels, the FR index (Fig. 8a) was below 1 for all models. However, from the moderate level on, the FR values evidently increased, and the very high susceptibility level had the largest values. The FR values representing the very high level of the five models used were 6.76 (C5.0 DT), 3.59 (SVM), 3.95 (BN), 3.69 (FR) and 2.69 (WOE), which were approximately four times the values representing the moderate and low susceptibility levels. Similar results were also observed in the FD curves (Fig. 8b). This revealed that the landslide pixel distribution became gradually denser with increasing landslide susceptibility levels. Among these models, the C5.0 DT model had the best performance. On the one hand, it had the largest FR and FD values at the very high and high susceptibility levels compared with the other models. For example, the FR values for the very high and high levels were 6.76 and 2.93, respectively, which were significantly greater than those of the other models. On the other hand, the FR and FD values of the C5.0 DT model for the low susceptibility levels were small. Its FR and FD at the very low level were 0.076/km^2 and 0.01/km^2, respectively, which were smaller than those of the other models. Such results exposed that the landslide susceptibility map generated by the C5.0 DT model reported the highest concentration of landslides in the area with a high susceptibility level, and there were few landslides erroneously classified into the low susceptibility zones.

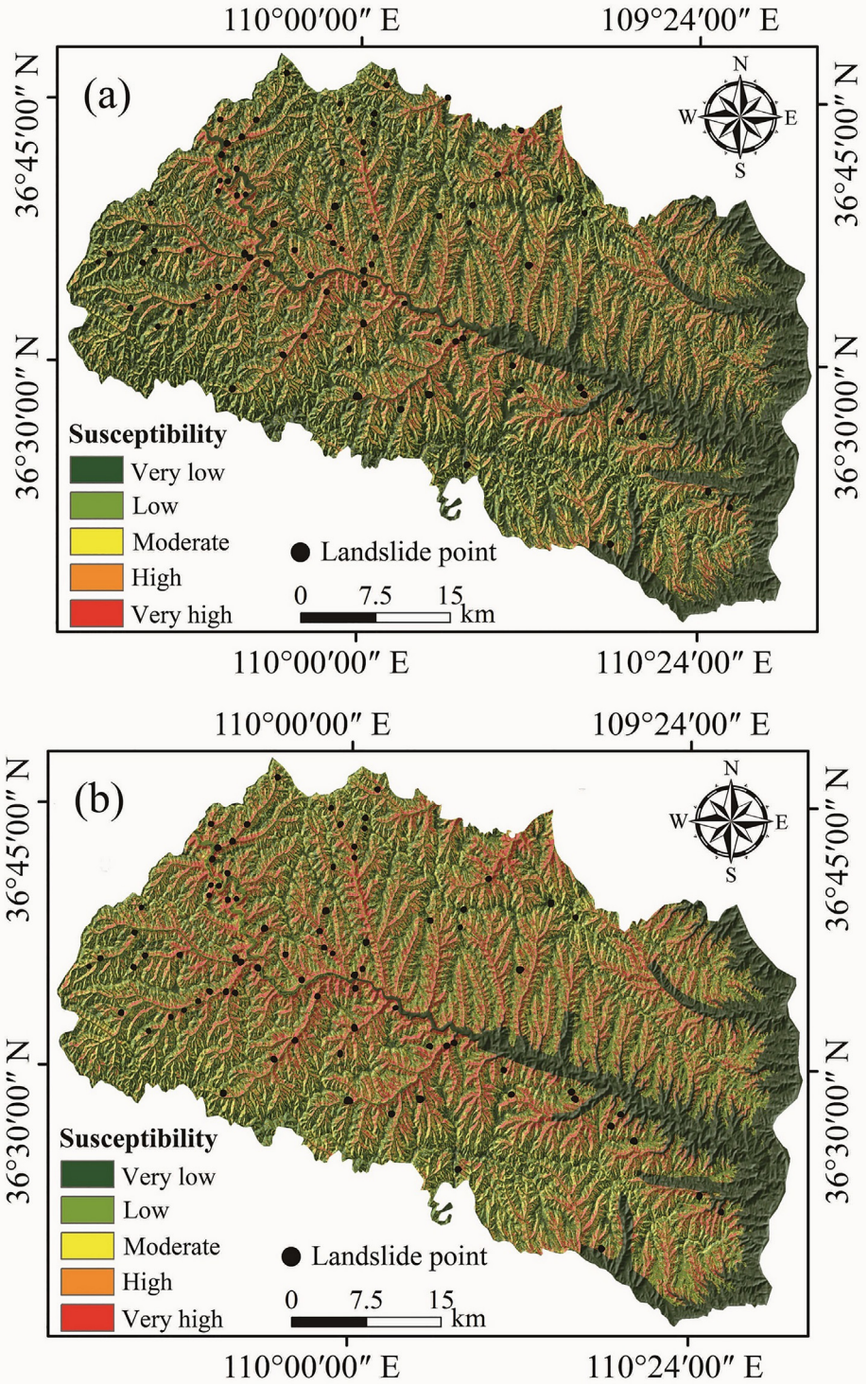

Fig. 6. Landslide susceptibility maps generated by using the C5.0 decision tree algorithm. (a) K-means cluster classification method, (b) Natural breaks classification method.

Second, the occurrence probability of all 82 landslides was analysed. Differing from the first step, which was pixel-based, this assessment was based on landslide events. The pixel with the highest occurrence probability within each landslide event was selected to characterize the susceptibility of the landslide. This step was conducted by using the "zonal statistics as table" tool in ArcGIS. The distribution of the occurrence probability of these landslides was compared with that of the randomly selected points contained in the training dataset. It is evident that highest number of landslide points is distributed in zones with very high and high susceptibility levels (Fig. 9a). For the five models, the percentages of landslides in the range of the top 20% interval of the occurrence probability were 89.0% (C5.0 DT), 96.3% (SVM), 85.4% (BN), 47.6% (FR) and 81.7% (WOE). In contrast, the random points (i.e., non-landslide pixels) were mostly located in the very low and low susceptibility zones (Fig. 9b). The percentages of points corresponding to the top 20% interval of the occurrence probability were 10.1% (C5.0 DT), 18.6% (SVM), 9.6% (BN), 1.6% (FR) and 10.0% (WOE). These results also confirm the satisfactory performance of the model used for identifying landslides.

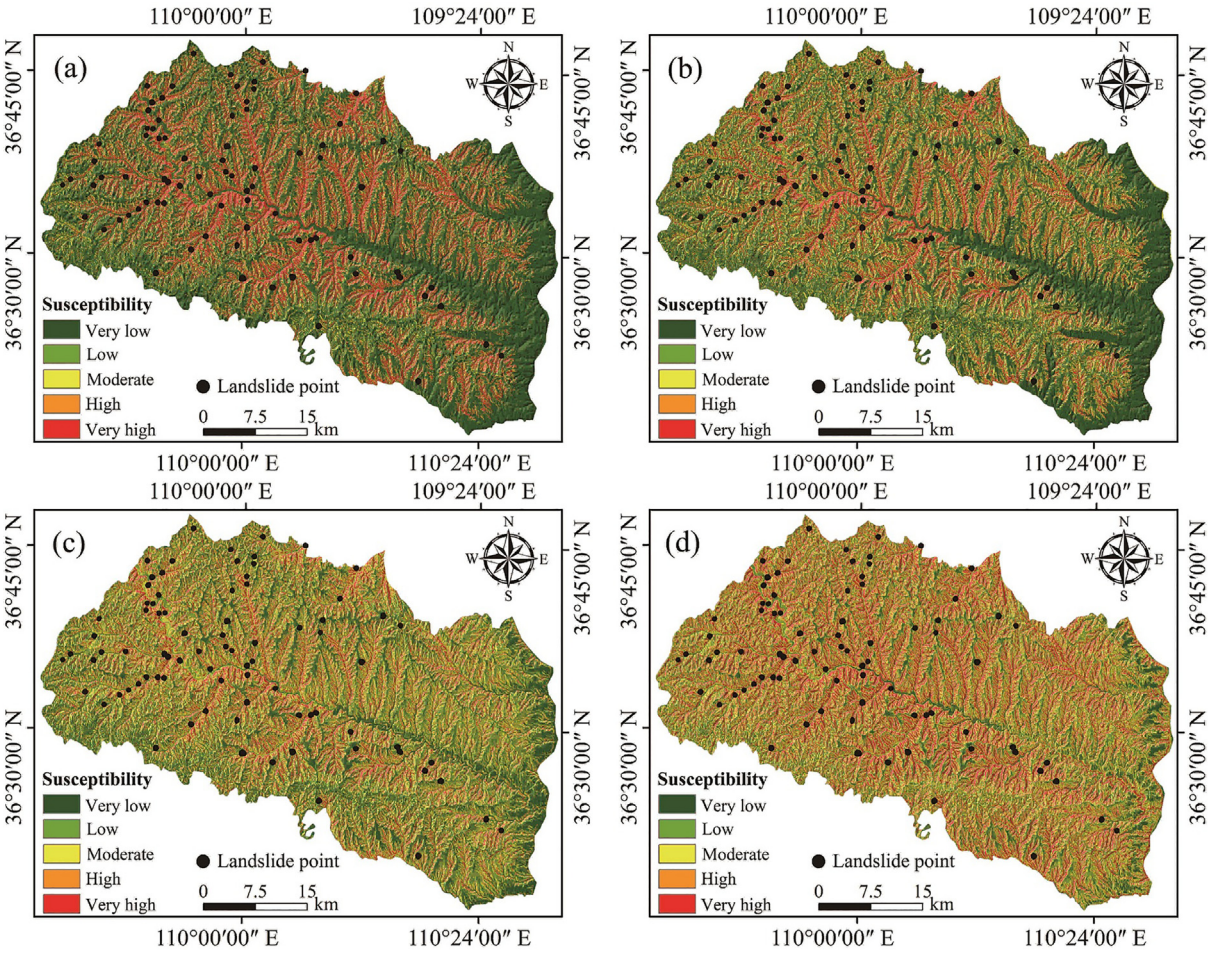

Fig. 7. Landslide susceptibility maps using different models and K-means cluster classification method. (a) SVM, (b) BN, (c) FR, and (d) WOE models.

Finally, two types of curves were computed to quantify the accuracy of the maps by adopting two different datasets. One is the receiver operating characteristic (ROC) curve obtained by using the training dataset (i.e., the landslide pixels and the same number of non-landslide pixels), whereas the other curve uses all the pixels to calculate the ROC curve to show the prediction performance of the obtained landslide occurrence probability. As seen in Fig. 10a, the accuracy of the C5.0 DT model had the highest area under the curve (AUC) of 0.883, followed by the SVM (0.850) and BN (0.813) models. The AUC values of the other two models were relatively small, less than 77%, among which the accuracy of the FR model was the worst. From the perspective of prediction performances (Fig. 10b), the AUCs of these models were 0.855 (C5.0 DT), 0.825 (SVM), 0.799 (BN), 0.719 (WOE), and 0.620 (FR). The C5.0 DT model and the FR model were still the best and worst models, respectively. Moreover, the prediction performance of all AUCs of the curves presented slightly smaller but similar trends compared to those curves that used the training dataset. This is mainly because the latter was a supervised classification, while the modelling process for most datasets contained in the former was unsupervised. On the other hand, the similar increase mode between the two curves showed that the models are robust when addressing the different datasets.

Overall, according to the statistics and ROC evaluation, the susceptibility maps resulted from all the models showed decent fitting/predictive performance towards recorded landslides. In particular, the accuracies of machine learning methods, namely the C5.0 DT, SVM and BN models, increased by at least 5% compared with two statistically-based models. Among all the models, the performance of the C5.0 DT model was better than the other models by 3.0%–13.5%, thus indicating it can be a powerful tool to mapping landslide susceptibility at a regional scale.

5.3. Comparison of susceptibility zonation methods from the viewpoint of risk management

To clarify the differences between classification methods, the natural breaks method implemented in ArcGIS was applied to generate landslide susceptibility maps (Fig. 11). The distribution of susceptibility level areas and the corresponding percentage of landslides in each level were computed and are shown in Fig. 12. Generally, an ideal susceptibility model should have a less area containing the high susceptibility class, while this area contains more landslides. In this regard, the machine learning models reported better performances than the statistically based models. Compared with the natural beak classification method (Fig. 12b), the C5.0 DT model using the K-means cluster classification (Fig. 12a) had smaller high susceptibility zones, but fewer landslides were contained. Similar results can also be observed among the different models. When using K-means cluster classification, the SVM and BN models placed more landslides in high susceptibility zones than the C5.0 DT model (Fig. 12a). This should be understood from two perspectives. On the one hand, the proposed model in this study had fewer landslides occurring in high susceptibility areas because the total area of this susceptibility level was smaller. When using the K-means cluster classification,

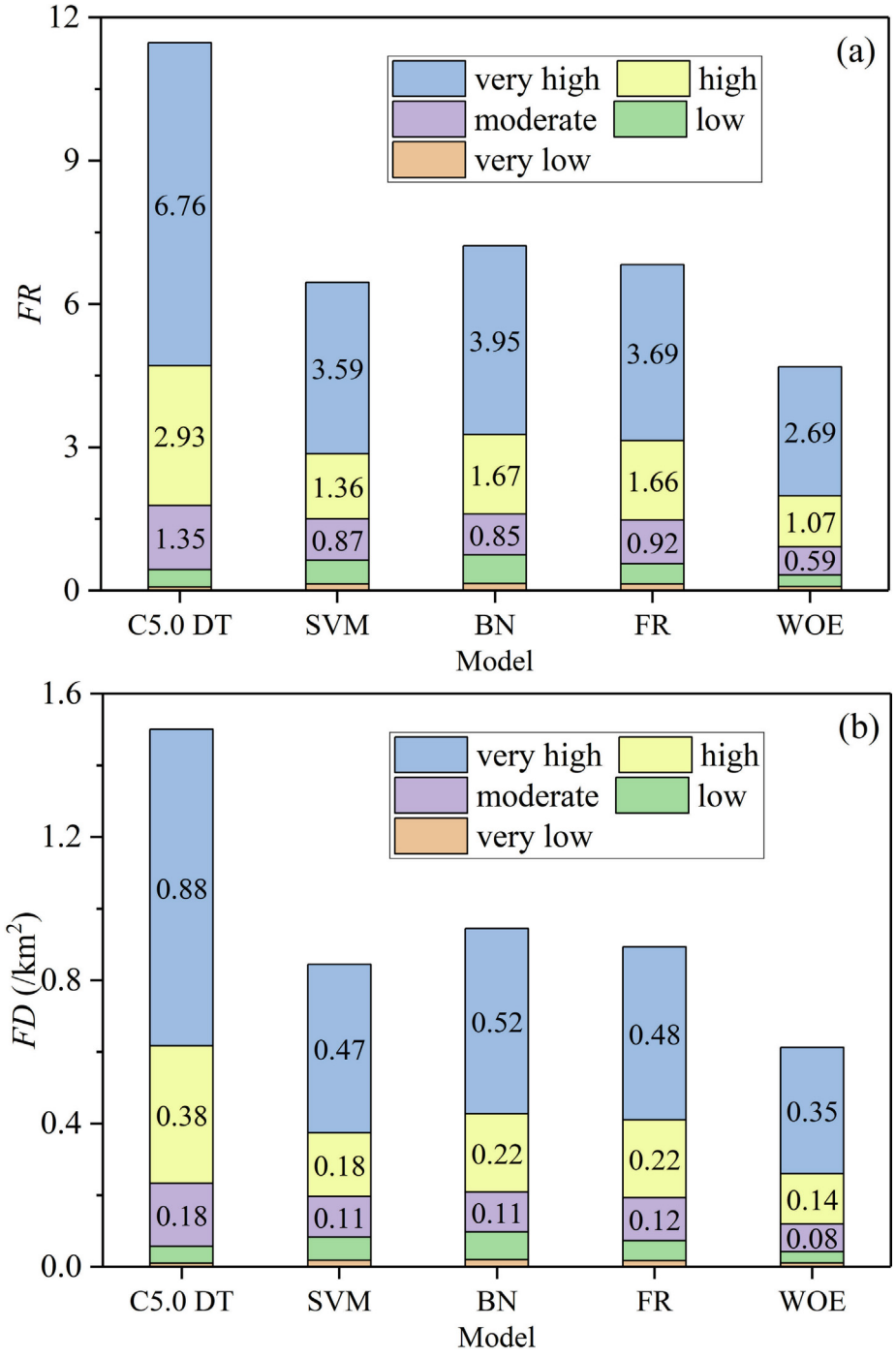

Fig. 8. Model accuracies presented by the statistical indicis. (a) FR, and (b) FD. The number in the bar shows the value for each landslide susceptibility level.

the cumulative percentage of landslides occurring under high and very high susceptibility levels was 67.1% (C5.0 DT), 78.3% (SVM) and 69.8% (BN). This percentage changed to 79.9% (C5.0 DT), 78.2% (SVM) and 70.2% (BN) when the natural breaks classification was applied. However, the increase in the total area with higher susceptibility levels can also trigger false alerts in many stable pixels. Taking the extreme situation as an instance, if the whole study area is classified into a high susceptibility area, all landslides are identified correctly, but it is evident that the model is not reasonable. In other words, a good model depends on its ability to identify landslide initiation points without classifying large areas as unstable (Goetz et al., 2011). From this viewpoint, the proposed procedure identified approximately 70% of landslides in high susceptibility areas, which only covered 16.6% of the study area. In contrast, in the map obtained from the C5.0 DT model and the natural breaks methods, high susceptibility zones covered 23.2% of the total area, which was evidently higher. Such differences indicated the higher efficiency of the proposed model than other models. This can also be confirmed by the statistical index FR (Table 3).

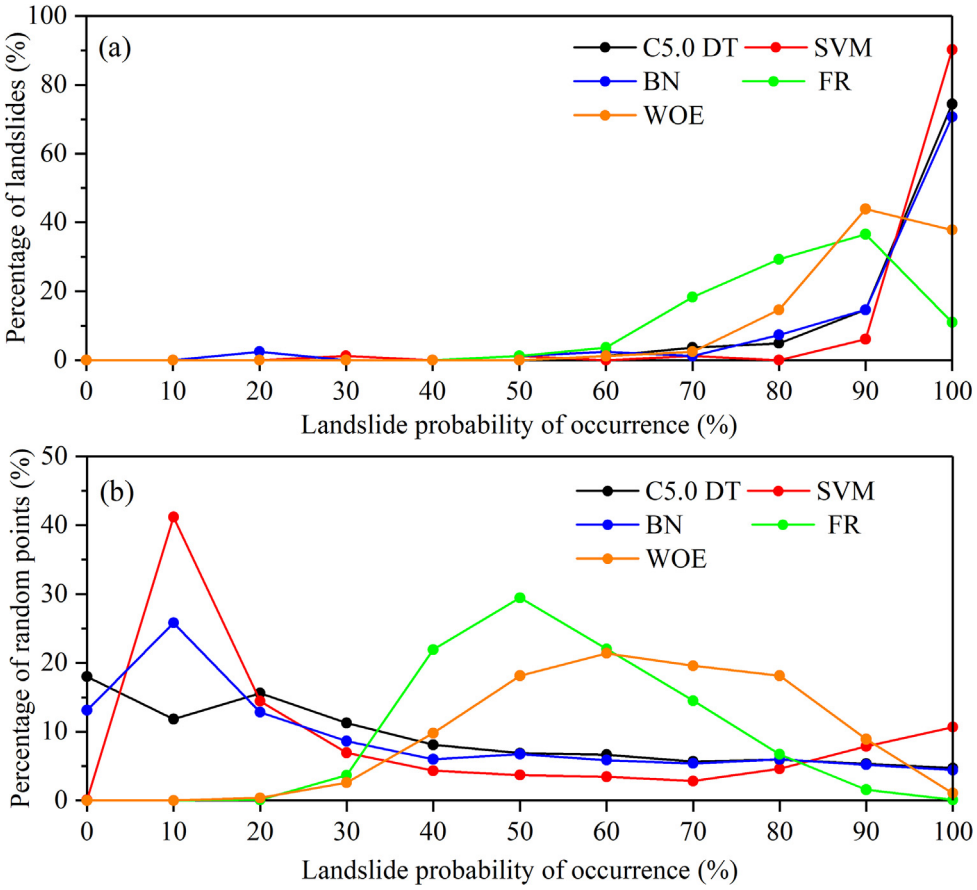

Fig. 9. The distribution of points versus the landslide probability of occurrence. (a) Landslide points, and (b) random points.

6. Discussion

6.1. Understanding the accuracy of landslide susceptibility maps

The accuracy of the proposed model in this study is approximately 88%, which is not superior to some machine learning models used by other studies (e.g., Achour and Pourghasemi, 2020). However, the evident difference in the statistical index between each susceptibility level showed that most landslide inventory points were identified in the final map. Moreover, as Goetz et al. (2015) reported, a "correct" model in a set of competing models does not exist in reality when various techniques and methods are available for the purpose of model selection. Hence, it is more important to select a model according to the specific scientific goals of the study, not only using the accuracies expressed by "boring" numbers.

In the case of landslide risk management, one of the main objectives for the civil protection department is the detection of more landslides in potential higher risk areas. However, misclassification costs (Cantarino et al., 2019) or time costs should also be considered because they have a close relationship with the final decisions (Carrara et al., 1991). A comparison of the multiple indicators mentioned above indicated that in the map created by the C5.0 DT and K-means cluster model, the distribution of very high and high susceptibility zones was exceptionally focused, meaning that they covered most landslide initiation areas, but few places without existing landslides were identified as landslide-prone zones. This is a user-friendly feature because this model allows researchers to focus on more specific but fewer areas with a high degree of landslide susceptibility, which may reduce the time cost and subsequently improve the efficiency of landslide risk management. We do not claim that the proposed procedure has the best accuracy relative to other methods; However, it does provide a balance between accuracy and efficiency, and this balance may be rather important for landslide risk practices in some areas. Hence, we are not expecting perfect results from any model because all of them have limitations. Users and decision makers need to select a model depending on their management objectives. Last but not least, different findings from multiple comparisons demonstrate that a single metric may only provide limited insight into the full model performances because each metric may fail to totally capture the local conditions of certain geomorphological characteristics (Reichenbach et al., 2018).

6.2. Insights on the evaluation of factor importance and model performance

The relative importance of each influencing factor was evaluated based on the standardized approach of C5.0 in the R environment. However, this might be only a rough estimate because the sampling times of the training data in the algorithm were limited. In fact, the k-fold cross-validation technique has been applied when performing susceptibility modelling, so it should be expected that the importance of factors would not have evident differences when only using SPSS Modeler. This is also the reason why we computed the factor importance using a separate procedure, not SPSS Modeler. On the other hand, this means that a certain degree of uncertainty was contained in the factor evaluation, but its impact on the final susceptibility mapping may have been minor. In particular, given that the relationship between landslides and

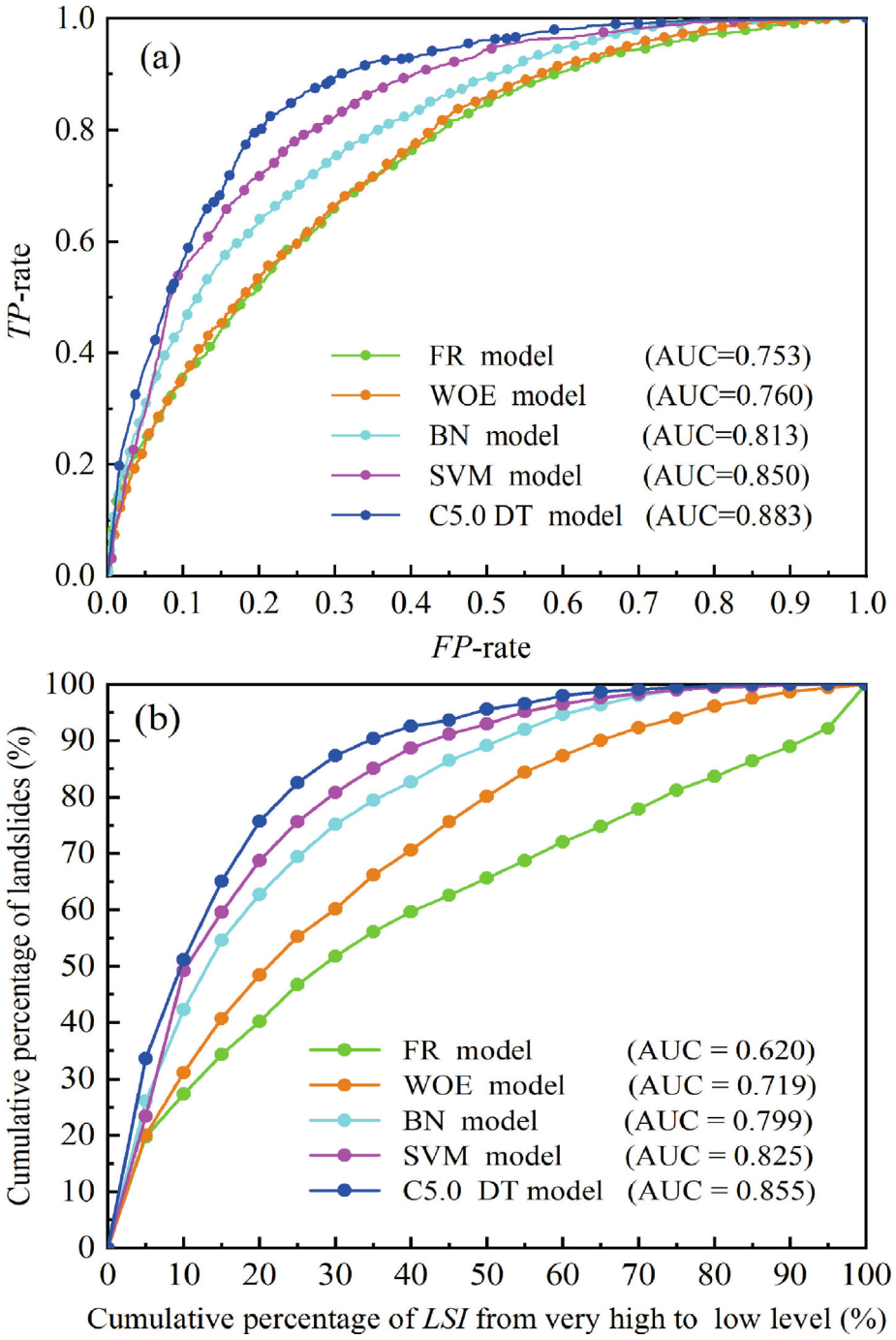

Fig. 10. Model accuracies presented by (a) the ROC curve using the training dataset (i.e., landslide pixels and random selected no-landslide pixels), and (b) the curve showing the predictive performance by using all the pixels in the study area.

influencing factors at the regional scale is a nonlinear high-dimensional system, the predicted contribution of each factor may vary among many conditions, including geomorphic conditions, input data resolutions and even modelling techniques (Catani et al., 2013; Goetz et al., 2015). Hence, the present results on the factor importance are considered preliminary and data-based, and deeper analysis on geomorphological and geological conditions is necessary to better understand and explain the contribution of such factors. Relevant experiences from similar studies can be observed and considered (Yesilnacar and Topal, 2005; Segoni et al., 2020). For instance, Yesilnacar and Topal (2005) tested the importance of factors for landslides in the Hendek region (Turkey). Although the results obtained from the forward stepwise logistic regression and neural network models were different, they finally determined the most important factors through field visits and analyses of the geology.

In the current analysis, all factors had positive importance, excluding MNDWI (importance = 0). However, its importance is not negative, which means it did not decrease the model's performance. An additional calculation also supported this point: the landslide susceptibility map without considering MNDWI had an ROC value of 0.87, which was less than (almost the same) that of the map considering this factor. Hence, from the perspective of model complexity, it is also necessary to investigate the contribu-

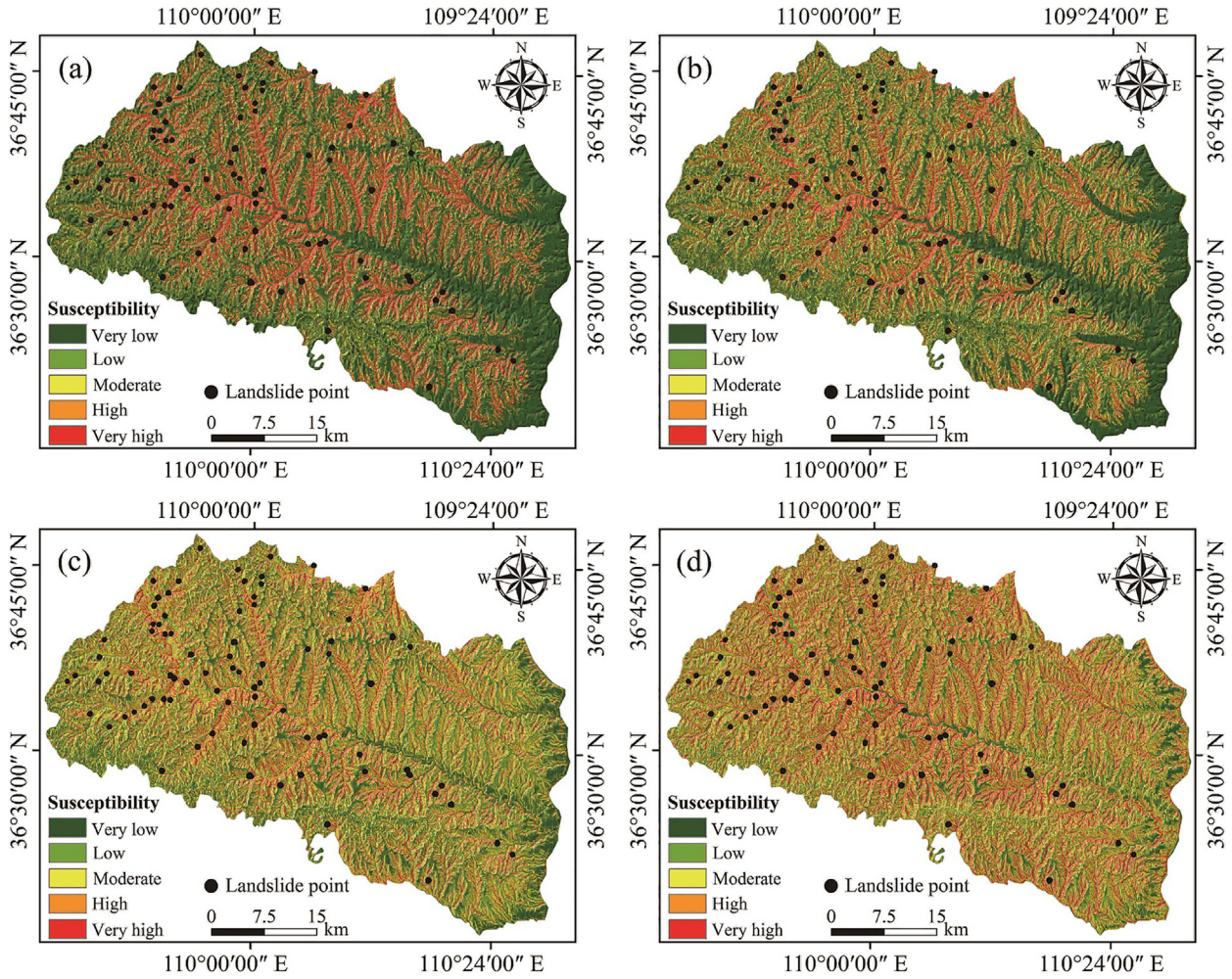

Fig. 11. Landslide susceptibility maps using different models and natural breaks classification method. (a) SVM, (b) BN, (c) FR, and (d) WOE models.

tion of one factor to landslide susceptibility through geomorphological analyses. It should be noted that we are not denying the validity of data-based techniques. On the contrary, we highlight the importance of analysis on geomorphic conditions, which may benefit subsequent assessment and make the results have more geomorphological significance.

In addition to the methods employed in this study, many criteria have been applied in the literature to evaluate model performance. For example, Goetz et al. (2015) used the true positive rate value measured at a 10% false positive rate to compare models. Guzzetti et al. (2006) proposed criteria to rank the quality of a landslide susceptibility assessment. The principles behind these criteria are easy to understand, i.e., they are used to translate the ROC curves into a cut-off measure that can be used to assess specific classification or prediction requirements. Another potential result of different evaluation criteria is the generation of different predictive surfaces (Cantarino et al., 2019). Under this condition, users must be aware of the backgrounds and final objectives of model's applications, because even a small difference in the evaluation measure may have practical consequences (Beguería, 2006). With this respect, statistical indices may be helpful but they were absent in the present study. Another challenge related to landslide susceptibility assessment is how to define strategies for an "optimal" combination of multiple forecasts, and their associated terrain zonation (Rossi et al., 2010). However, this study focused on the introduction of a modelling procedure and a model comparison method and failed to address this issue. As we can see, the SVM model had a relatively better performance in the higher risk zone, whereas the C5.0 DT and statistically based models were better at identifying stable pixels (Fig. 9). Hence, their combination in the future may help overcome the uncertainties inherent in dealing with the differences among landslide susceptibility classes.

7. Conclusions

Reliable landslide susceptibility mapping is the integral step for risk assessment and mitigation. Compared with previous studies, the contributions of this study mainly include two aspects: (i) adopting a novel machine learning model, the C5.0 decision tree, to calculate the landslide occurrence probability, and (ii) employing the K-means cluster algorithm to perform susceptibility zonation. Based on the experiences obtained from Yanchang County located in northwestern China, we stated that although the model required more time to achieve zonation, it presented a superior performance. On the one hand, the ROC analysis indicated that both the model accuracy (AUC = 0.883) expressed by the training data and the predictive performance (AUC = 0.855) expressed by all the pixels were the best, compared with two commonly used machine learning algorithms (SVM (AUC = 0.850, 0.825), BN (AUC = 0.813, 0.799)) and two statistically based models (FR (AUC = 0.753, 0.620), WOE (AUC = 0.760, 0.719)). On the other hand, the comparison of statistical indicators revealed that in the map created by the proposed model, high susceptibility zones had higher landslide frequency ratios and landslide density values

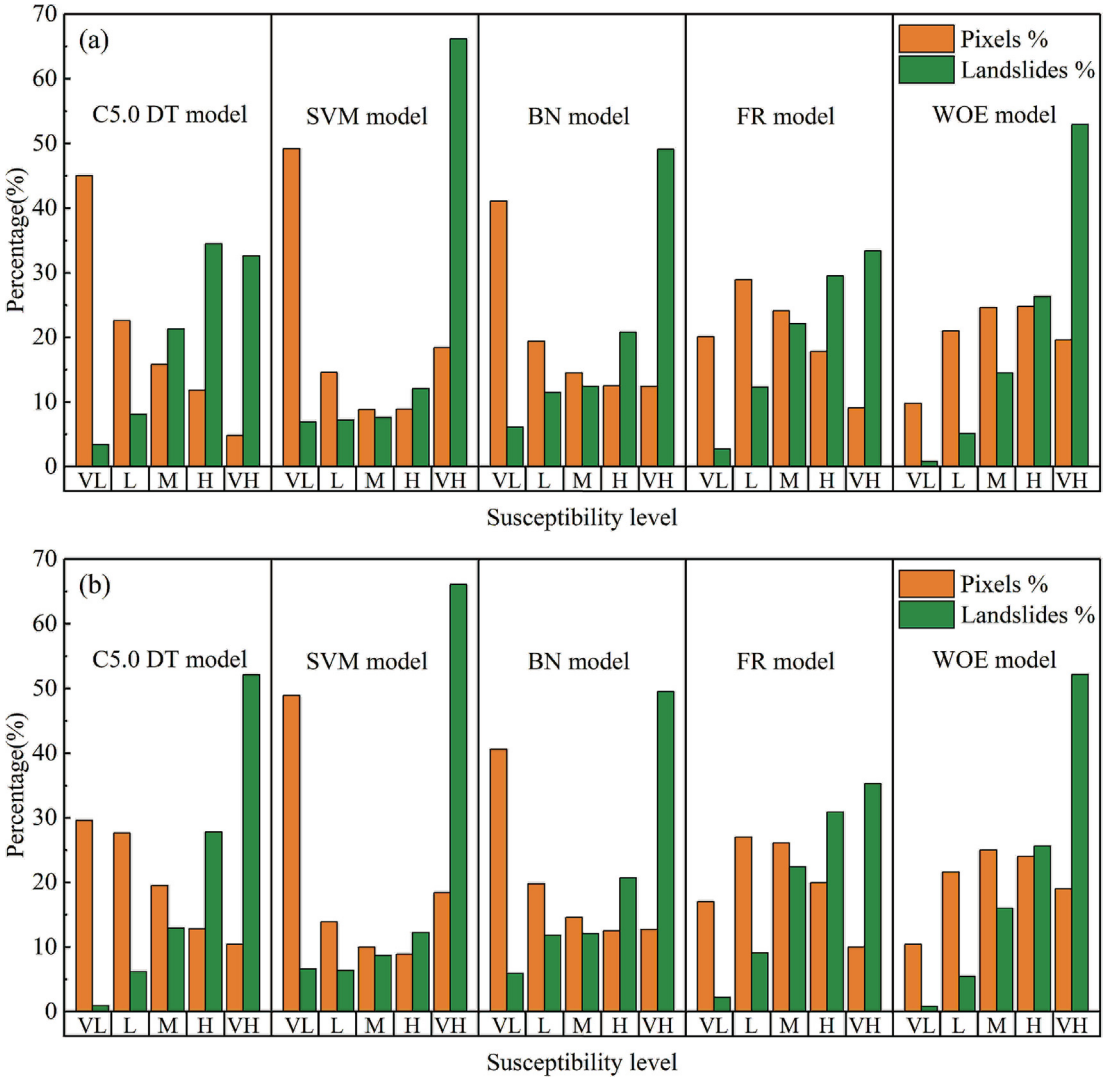

Fig. 12. The classification results considering different classification methods. (a) K-means cluster, (b) natural breaks. VL: very low, L: low, M: moderate, H: high, VH: very high.

Table 3
FR values in each susceptibility level.

Model	Susceptibility level				
	Very low	Low	Moderate	High	Very high
C5.0 DT + K-means cluster	0.076	0.358	1.350	2.927	6.755
C5.0 DT + natural breaks	0.031	0.225	0.663	2.171	4.997
SVM + K-means cluster	0.140	0.490	0.869	1.361	3.593
SVM + natural breaks	0.135	0.463	0.869	1.368	3.598
BN + K-means cluster	0.149	0.594	0.855	1.665	3.953
BN + natural breaks	0.144	0.597	0.838	1.665	3.904

than other models but had smaller total areas. This indicated that the distribution of landslides in high susceptibility zones was more focused without containing more "stable" pixels. Moreover, only 3.4% of known landslide pixels were located in very low susceptibility zones, and only one landslide event was determined to have an occurrence probability less than 0.5. Such results indicated that there were few false classifications of historical landslides. Hence, the obtained map is more readily available for decision makers because it may reduce the associated time costs and subsequently improve the efficiency of landslide risk management.

In this study, the model performances were quantified and evaluated based on multiple metrics, but some of them exposed different trends and findings. Hence, it is recommended that the assessment of model performance should be linked with the specific backgrounds and objectives of landslide risk management. The proposed model performed well with respect to the pure predictive performance, but some uncertainties exist in the current analysis; in particular the accuracy of higher risk zones identified by the K-means cluster classification should be improved. Nevertheless, the outcomes seem to be coherent and indicate that if the landslide inventory and the input data are well prepared, the C5.0 decision tree combined with the K-means cluster model can identify the landslide locations effectively. Although the modelling procedure was conducted in a mountainous area located on the Loess Plateau (China), it can be replicated in other similar settings. The resulting map can be used as a basic tool for engineers and decision makers in land use planning and can assist in the reduction of future landslide risks by implementing various measures for prevention and mitigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research is funded by the National Natural Science Foundation of China (Grant Nos. 41807285 and 51679117), Key Project of the State Key Laboratory of Geohazard Prevention and Geoenvironment Protection (SKLGP2019Z002), the National Science Foundation of Jiangxi Province, China (20192BAB216034), the China Postdoctoral Science Foundation (2019M652287 and 2020T130274), the Jiangxi Provincial Postdoctoral Science Foundation (2019KY08), and Fundamental Research Funds for National Universities, China University of Geosciences (Wuhan).

References

Achour, Y., Pourghasemi, H.R., 2020. How do machine learning techniques help in increasing accuracy of landslides susceptibility maps?. Geosci. Front. 11, 871–883.
Aditian, A., Kubota, T., Shinohara, Y., 2018. Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia. Geomorphology 318, 101–111.
Ahmed, B., 2015. Landslide susceptibility mapping using multi-criteria evaluation techniques in Chittagong Metropolitan Area, Bangladesh. Landslides 12, 1077–1095.
Atkinson, P.M., Massari, R., 1998. Generalised linear modelling of susceptibility to landsliding in the central apennines, Italy. Comput Geosci. 24, 373–385.
Arabameri, A., Chen, W., Loche, M., Zhao, X., Li, Y., Lombardo, L., Cerda, A., Pradhan, B., Bui, D.T., 2020. Comparison of machine learning models for gully erosion susceptibility mapping. Geosci. Front. 11, 1609–1620.
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.
Baeza, C., Lantada, N., Amorim, S., 2016. Statistical and spatial analysis of landslide susceptibility maps with different classification systems. Environ. Earth Sci. 75, 1318.
Barredo, J., Benavides, A., Hervás, J., van Westen, C.J., 2000. Comparing heuristic landslide hazard assessment techniques using GIS in the Tirajana basin, Gran Canaria Island, Spain. Int. J. Appl. Earth Obs. Geoinf. 2, 9–23.
Beguería, S., 2006. Validation and evaluation of predictive models in hazard assessment and risk management. Nat. Hazards 37, 315–329.
Bueechi, E., Klimeš, J., Frey, H., Huggel, C., Strozzi, T., Cochachin, A., 2019. Regional-scale landslide susceptibility modelling in the Cordillera Blanca, Peru—a comparison of different approaches. Landslides 16, 395–407.
Cantarino, I., Carrion, M.A., Goerlich, F., Martinez Ibañez, V., 2019. A ROC analysis-based classification method for landslide susceptibility maps. Landslides 16, 265–282.
Carrara, A., Cardinali, M., Detti, R., Guzzetti, F., Pasqui, V., Reichenbach, P., 1991. GIS techniques and statistical models in evaluating landslide hazard. Earth Surf. Processes Landf. 16, 427–445.
Catani, F., Lagomarsino, D., Segoni, S., Tofani, V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat. Hazards Earth Syst. Sci. 13, 2815–2831.
Chang, Z., Du, Z., Zhang, F., Huang, F., Chen, J., Li, W., Guo, Z., 2020. Landslide susceptibility prediction based on remote sensing images and GIS: comparisons of supervised and unsupervised machine learning models. Remote Sens. 12, 502.
Chen, L., Guo, Z., Yin, K., Shrestha, D.P., Jin, S., 2019. The influence of land use and land cover change on landslide susceptibility: a case study in Zhushan Town, Xuan'en County (Hubei, China). Nat. Hazards Earth Syst. Sci. 19, 2207–2228.
Corominas, J., van Western, C.J., Frattini, P., Cascini, L., Malet, J.P., Fotopoulou, S., Catani, F., Van Den Eeckhaut, M., Mavrouli, O., Agliardi, F., Pitilakis, K., Winter, M.G., Pastor, M., Ferlisi, S., Tofani, V., Hervás, J., Smith, J.T., 2014. Recommendations for the quantitative analysis of landslide risk. Bull. Eng. Geol. Environ. 73, 209–263.
Crawford, M.H., Crowley, K., Potter, S.H., Saunders, W.S.A., Johnston, D.M., 2018. Risk modelling as a tool to support natural hazard risk management in New Zealand government. Int. J. Disast. Risk Re. 28, 610–619.
Dai, F.C., Lee, C.F., 2002. Landslide characteristics and slope instability modeling using GIS Lantau Island, Hong Kong. Geomorphology 42, 213–228.
Dou, J., Yunus, A.P., Bui, D.T., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.-W., Han, Z., Pham, B.T., 2020. Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan. Landslides 17, 641–658.
Feizizadeh, B., Jankowski, P., Blaschke, T., 2014. A GIS based spatially-explicit sensitivity and uncertainty analysis approach for multi-criteria decision analysis. Comput. Geosci. 64, 81–95.
Felicísimo, Á.M., Cuartero, A., Remondo, J., Quirós, E., 2013. Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. Landslides 10, 175–189.
Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., Savage, W.Z., 2008. Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. Eng. Geol. 102, 85–98.
Frattini, P., Crosta, G., Carrara, A., 2010. Techniques for evaluating the performance of landslide susceptibility models. Eng. Geol. 111, 62–72.
Glade, T., Crozier, M.J., 2005. A review of scale dependency in landslide hazard and risk analysis. In: Glade, T., Anderson, M.G., Crozier, M.J. (Eds.), Landslide Hazard and Risk. Wiley, Chichester, pp. 75–138.
Goetz, J.N., Guthrie, R.H., Brenning, A., 2011. Integrating physical and empirical landslide susceptibility models using generalized additive models. Geomorphology 129, 376–386.
Goetz, J.N., Brenning, A., Petschko, H., Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput. Geosci. 81, 1–11.
Guo, T., Zhang, J., Han, Y., Zhong, Y., Tan, J., Wei, J., 2019a. Evaluation of landslide susceptibility in Yanchang County based on particle swarm optimization—based support vector machine. Geol. Sci. Tech. Infor. 38, 236–243 (in Chinese).
Guo, Z., Yin, K., Huang, F., Fu, S., Zhang, W., 2019b. Evaluation of landslide susceptibility based on landslide classification and weighted frequency ratio model. Chin. J. Rock Mech. Eng. 38, 287–300 (in Chinese with English abstract).
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.
Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., Galli, M., 2006. Estimating the quality of landslide susceptibility models. Geomorphology 81, 166–184.
Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M., Chang, K.-T., 2012. Landslide inventory maps: New tools for an old problem. Earth-Sci. Rev. 112, 42–66.
Hartigan, A., Wong, M.A., 1978. A k-means clustering algorithm. Appl. Stat. 28, 100–108.
Huang, F., Yin, K., Huang, J., Gui, L., Wang, P., 2017. Landslide susceptibility mapping based on self-organizing-map network and extreme learning machine. Eng. Geol. 223, 11–22.
Huang, F., Yao, C., Liu, W., Li, Y., Liu, X., 2018. Landslide susceptibility assessment in the Nantian area of China: a comparison of frequency ratio model and support vector machine. Geomatics. Nat. Hazards Risk 9, 919–938.
Huang, F., Cao, Z., Guo, J., Jiang, S.-H., Li, S., Guo, Z., 2020a. Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping. Catena 191, 104580.
Huang, F., Zhang, J., Zhou, C., Wang, Y., Huang, J., Zhu, L., 2020b. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. Landslides 17, 217–229.
Hungr, O., Leroueil, S., Picarelli, L., 2014. The Varnes classification of landslide types, an update. Landslides 11, 167–194.
Kouli, M., Loupasakis, C., Soupios, P., Rozos, D., Vallianatos, F., 2014. Landslide susceptibility mapping by comparing the WLC and WofE multi-criteria methods in the West Crete Island. Greece. Environ. Earth Sci. 72, 5197–5219.
Kuhn, M., Weston, S., Coulter, N., Culp, M., Quinlan, J. R., 2015. Package 'C50'. https://cran.r-project.org/web/packages/C50/C50.pdf (accessed 5 December 2020).
Li, Y., 2018. Study on Formation and Motion Mechanism of Shallow Loess Landslide-taking Yanchang County as the Study Area. M.S. Thesis, Chang'an University, 55 pp (in Chinese).
Lian, B., Peng, J., Zhan, H., Huang, Q., Wang, X., Hu, S., 2020. Formation mechanism analysis of irrigation-induced retrogressive loess landslides. Catena 195, 104441.
Lee, S., Ryu, J.-H., Min, K., Won, J.-S., 2003. Landslide susceptibility analysis using GIS and artificial neural network. Earth Surf. Processes Landf. 28, 1361–1376.
Ma, J., Tang, H., Liu, X., Hu, X., Sun, M., Song, Y., 2017. Establishment of a deformation forecasting model for a step-like landslide based on decision tree C5.0 and two-step cluster algorithms: a case study in the Three Gorges Reservoir area, China. Landslides 14, 1275–1281.
McFeeters, S.K., 1996. The use of the normalized difference water index (NDWI) in the delineation of open water features. Int. J. Remote Sens. 17, 1425–1432.
Medina, V., Hürlimann, M., Guo, Z., Lloret, A., Vaunat, J., 2021. Fast physically-based model for rainfall-induced landslide susceptibility assessment at regional scale. Catena 201, 105213. https://doi.org/10.1016/j.catena.2021.105213.
Melchiorre, C., Matteucci, M., Azzoni, A., Zanchi, A., 2008. Artificial neural networks and cluster analysis in landslide susceptibility zonation. Geomorphology 94, 379–400.
Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrol. Process. 5, 3–30.
Pavel, M., Nelson, J.D., Jonathan Fannin, R., 2011. An analysis of landslide susceptibility zonation using a subjective geomorphic mapping and existing landslides. Comput. Geosci. 37, 554–566.

Peng, J., Wang, S., Wang, Q., Zhuang, J., Huang, W., Zhu, X., Leng, Y., Ma, P., 2019. Distribution and genetic types of loess landslides in China. J. Asian Earth Sci. 170, 329–350.
Peng, L., Niu, R., Huang, B., Wu, X., Zhao, Y., Ye, R., 2014. Landslide susceptibility mapping based on rough set theory and support vector machines: A case of the Three Gorges area, China. Geomorphology 204, 287–301.
Pham, B.T., Shirzadi, A., Shahabi, H., Omidvar, E., Singh, S.K., Sahana, M., Asl, D.T., Ahmad, B.B., Quoc, N.K., Lee, S., 2019. Landslide susceptibility assessment by novel hybrid machine learning algorithms. Sustainability 11, 4386.
Pradhan, B., 2011. Use of GIS-based fuzzy logic relations and its cross application to produce landslide susceptibility maps in three test areas in Malaysia. Environ. Earth Sci. 63, 329–349.
Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350–365.
Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Mateo, CA, USA, p. 307 p..
Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M., Guzzetti, F., 2018. A review of statistically-based landslide susceptibility models. Earth-Sci. Rev. 180, 60–91.
Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A.C., Peruccacci, S., 2010. Optimal landslide susceptibility zonation based on multiple forecasts. Geomorphology 114, 129–142.
Rossi, G., Catani, F., Leoni, L., Segoni, S., Tofani, V., 2013. HIRESSS: a physically based slope stability simulator for HPC applications. Nat. Hazards Earth Syst. Sci. 13, 151–166.
Schlögel, R., Marchesini, I., Alvioli, M., Reichenbach, P., Rossi, M., Malet, J.-P., 2018. Optimizing landslide susceptibility zonation: Effects of DEM spatial resolution and slope unit delineation on logistic regression models. Geomorphology 301, 10–20.
Segoni, S., Pappafico, G., Luti, T., Catani, F., 2020. Landslide susceptibility assessment in complex geological settings: sensitivity to geological information and insights on its parameterization. Landslides 17, 2443–2453.
Shu, H., Hürlimann, M., Molowny-Horas, R., González, M., Pinyol, J., Abancó, C., Ma, J., 2019. Relation between land cover and landslide susceptibility in Val d'Aran, Pyrenees (Spain): Historical aspects, present situation and forward prediction. Sci. Total Environ. 693, 133557.
Shu, H., Ma, J., Guo, J., Qi, S., Guo, Z., Zhang, P., 2020. Effects of rainfall on surface environment and morphological characteristics in the Loess Plateau. Environ. Sci. Pollut. Res. 27, 37455–37467.
Sorbino, G., Sica, C., Cascini, L., 2010. Susceptibility analysis of shallow landslides source areas using physically based models. Nat. Hazards. 53, 313–332.
Tang, Y., Feng, F., Guo, Z., Feng, W., Li, Z., Wang, J., Sun, Q., Ma, H., Li, Y., 2020. Integrating principal component analysis with statistically-based models for analysis of causal factors and landslide susceptibility mapping: A comparative study from the Loess Plateau area in Shanxi (China). J. Clean. Prod. 277, 124159.
Taylor, F.E., Malamud, B.D., Freeborough, K., Demeritt, D., 2015. Enriching Great Britain's National Landslide Database by searching newspaper archives. Geomorphology 249, 52–68.
Tian, Y., Xu, C., Ma, S., Xu, X., Wang, S., Zhang, H., 2019. Inventory and spatial distribution of landslides triggered by the 8th August 2017 M_W 6.5 Jiuzhaigou earthquake, China. J. Earth Sci. 30, 206–217.
Tien Bui, D., Pradhan, B., Lofman, O., Revhaug, I., Dick, O.B., 2012. Landslide susceptibility assessment in the Hoa Binh province of Vietnam: A comparison of the Levenberg–Marquardt and Bayesian regularized neural networks. Geomorphology 171-172, 12–29.
Tsangaratos, P., Ilia, I., 2016. Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece. Landslides 13, 305–320.
Ture, M., Tokatli, F., Kurt, I., 2009. Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Syst. Appl. 36, 2017–2026.
Varnes, D.J., 1978. Slope movement types and processes. In: Schuster, R.L., Krizek, R. J. (Eds.), Landslides, Analysis and Control, Special Report 176: Transportation Research Board. National Academy of Sciences, Washington, DC., pp. 11–33.
Van Den Ecekhaut, M., Marre, A., Poesen, J., 2010. Comparison of two landslide susceptibility assessment in the Chamopagne-Ardenne region (France). Geomorphology 115, 141–155.
Wang, Y.i., Fang, Z., Hong, H., 2019. Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China. Sci. Total Environ. 666, 975–993.
Wu, Y., Ke, Y., Chen, Z., Liang, S., Zhao, H., Hong, H., 2020. Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. Catena 187, 104396.
Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Inter. J. Remote Sens. 27, 3025–3033.
Yalcin, A., Reis, S., Aydinoglu, A.C., Yomralioglu, T.A., 2011. GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods for landslide susceptibility mapping in Trabzon, NE Turkey. Catena 85, 274–287.
Yao, X., Tham, L.G., Dai, F.C., 2008. Landslide susceptibility mapping based on Support Vector Machine: A case study on natural slopes of Hong Kong, China. Geomorphology 101, 572–582.
Yesilnacar, E., Topal, T., 2005. Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). Eng. Geol. 79, 251–266.
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). Comput. Geosci. 35, 1125–1138.
Youssef, A.M., Pourghasemi, H.R., 2021. Landslide susceptibility mapping using machine learning algorithms and comparison of their performance as Abha Basin, Asir Region, Saudi Arabia. Geosci. Front. 12, 639–655.
Zêzere, J.L., Pereira, S., Melo, R., Oliveira, S.C., Garcia, R.A.C., 2017. Mapping landslide susceptibility using data-driven methods. Sci. Total Environ. 589, 250–267.
Zhang, M., Liu, J., 2010. Controlling factors of loess landslides in western China. Environ. Earth Sci. 59, 1671–1680.
Zhao, B., Zhao, Y.Q., 2020. Investigation and analysis of the Xiangning landslide in Shanxi Province, China. Nat. Hazards 103, 3837–3845.
Zhao, X., Chen, W., 2020. Optimization of computational intelligence models for landslides susceptibility evaluation. Remote Sen. 12, 2180.
Zhou, C., Yin, K., Cao, Y., Ahmed, B., Li, Y., Catani, F., Pourghasemi, H.R., 2018. Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China. Comput. Geosci. 112, 23–37.
Zhuang, J., Peng, J., Wang, G., Javed, I., Wang, Y., Li, W., 2018. Distribution and characteristics of landslide in Loess Plateau: A case study in Shaanxi province. Eng. Geol. 236, 89–96.