GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models

Wei Chen, Xiaoshen Xie, Jianbing Peng, Jiale Wang, Zhao Duan and Haoyuan Hong

ABSTRACT
The main purpose of this paper is to explore some potential applications of sophisticated machine learning techniques such as the kernel logistic regression, Naïve-Bayes tree and alternating decision tree models for landslide susceptibility analysis at Taibai county (China). Initially, a landslide inventory map containing the information of 212 historical landslide locations was prepared. Seventy percentage (148) of landslides were randomly selected for training models and the remaining were used for validation. Additionally, 12 landslide conditioning factors were considered and the thematic layers were prepared in GIS. Subsequently, these three models were applied to build landslide susceptibility maps. The performances of the models were compared using the receive operating characteristic curves, kappa index, and statistical evaluation measures. The results show that the KLR model has the highest AUC values of 0.910 and 0.936 for training and validation datasets, respectively. The KLR model also has the highest degree of goodness-of-fits (84.5%) for the training dataset. The NBTree model has the highest goodness-of-fits (91.4%) for the validation dataset. However, the KLR model has the preferable balance performance for both the training and validation process. The results of this study demonstrate the benefit of selecting the optimal machine learning techniques in landslide susceptibility mapping.

ARTICLE HISTORY
Received 24 November 2016
Accepted 27 January 2017

KEYWORDS
Landslide susceptibility; kernel logistic regression; Naïve-Bayes tree; alternating decision tree; China

1. Introduction
Landslide is defined as the downslope movement of soil and rock affected by gravity (Malamud et al. 2004). It is one of the most frequent and catastrophic geologic hazards causing enormous casualties and severe economic losses (Mahalingam & Olsen 2016; Mahalingam et al. 2016; Myronidis et al. 2016; Pradhan et al. 2016). However, the complexity of soil and rock condition, topography, hydrology, and human activities makes accurate landslide forecast an extremely difficult task and presents a major challenge to the global change research community (Wu & Sidle 1995; Pielke et al. 2003; Gokceoglu & Sezer 2012).

A broad range of methods and techniques have been suggested by early researchers to improve the prediction accuracy of landslides. These methods and techniques are mainly physically based
or statistically based (Van Westen & Terlien 1996; Duc 2013; Thanh & De Smedt 2014). The physically based methods evaluate the safety factor of slopes based on the detailed geomorphic and geologic properties at site-specific locations and generally provide accurate results. However, it would be quite expensive and not practical in terms of regional landslide susceptibility assessments.

The statistically based methods assess the correlation between the landslide conditioning factors and the landslide occurrence. Various approaches, varying from simple expert knowledge to sophisticated mathematical procedures, have been used to produce landslide susceptibility maps assisted by the geographic information system (GIS) technology. Bivariate and multivariate statistical models, such as frequency ratio (Akgun et al. 2008; Yalcin et al. 2011a; Chen et al. 2016a), weights-of-evidence (Dahal et al. 2008; Regmi et al. 2014; Chen et al. 2016a), certainty factor (Pourghasemi et al. 2013; Dou et al. 2014; Hong et al. 2017), evidential belief function (Tien Bui et al. 2012b; Feizizadeh et al. 2014; Ding et al. 2016; Pourghasemi & Kerle 2016), and logistic regression (Tien Bui et al. 2011; Yalcin et al. 2011a; Süzen & Kaya 2012; Manzo et al. 2013; Shahabi et al. 2015; Chen et al. 2016d) models, have been widely used by early researchers. Expert knowledge-based models, such as the analytical hierarchy process model (Yalcin et al. 2011b; Pourghasemi et al. 2012; Shahabi et al. 2014; Youssef 2015; Althuwaynee et al. 2016; Chen et al. 2016c) and fuzzy approaches (Aksoy & Ercanoglu 2012; Zhu et al. 2014; Shahabi et al. 2015; Vakhshoori & Zare 2016), can also be seen in many literature works.

With the development of soft computing, many sophisticated mathematical procedures are increasingly adopted in landslide susceptibility modelling, such as artificial neural networks (Pradhan & Lee 2010; Yilmaz 2010; Xu et al. 2012b; Tien Bui et al. 2016b), support vector machines (Xu et al. 2012a; Peng et al. 2014; Colkesen et al. 2016; Pham et al. 2016; Tien Bui et al. 2016a), Naïve Bayes (Tien Bui et al. 2012a; Pham et al. 2015a; Tsangaratos & Ilia 2016), multivariate adaptive regression splines (Wang et al. 2015; Conoscenti et al. 2016; Pourghasemi & Rossi 2016), random forest (Catani et al. 2013; Youssef et al. 2016; Hong et al. 2016; Pourghasemi & Kerle 2016), and decision trees models (Pradhan 2013; Lombardo et al. 2015; Pham et al. 2016).

The kernel logistic regression (KLR) is a kernel version of logistic regression that constructs a linear logistic regression model in a high-dimensional space by using a kernel function (generally, Radial Basis Function). It is powerful to classify the data that is difficult to be distinguished in the current dimensional space. In case of the Naïve-Bayes tree (NBTREE), it is a hybrid algorithm of decision tree and Naïve Bayes. The NBTREE takes advantage of the simple structures decision tree and creates a naïve-bayes classifier at each leaf node to facilitate the classification. The alternating decision tree (ADTree) develops the structures of decision tree and then combines it with boosting algorithm. The main advantages of the NBTREE and the ADTREE are that they have easier-to-construct structures and easier-to-interpret classification rules.

The main purpose of this study is to compare and explore potential applications of three models, namely, KLR, the NBTREE, and the ADTREE in landslide susceptibility mapping at Taibai County (China). KLR, NBTREE, and ADTREE models have been used in few studies on landslide susceptibility mapping and showed satisfactory results (Hong et al. 2015; Pham et al. 2015b; Pham et al. 2016; Tien Bui et al. 2016b). Therefore, investigation of these new methods for landslide susceptibility is highly necessary.

2. Study area

The study area is located in the hinterland of Qinling Mountains which lies within longitudes 107°03'E to 107°46'E and latitudes 33°38'N to 34°10'N. It covers a surface area of about 2,780 km² (Figure 1). The altitudes of the study area range from 740 m to 3,767 m. In the study area the continental monsoon climate prevails, as well as alpine climate. The mean annual temperature is 7.8 °C, the mean annual sunshine duration is about 1,840 hours, and the mean annual precipitation is around 700 mm, most of which falls between July and September.
Figure 1. Landslide inventory and location map of study area.
The tectonic system of Taibai County belongs to the Shang-Dan ophiolite tectonic belt of eastern Qinling Mountains. This area crosses two secondary tectonic units: the north Qinling Mountains tectonic belt and the south Qinling Mountains tectonic belt. This tectonic system divides the study area into two tectonic units and controls the development of landslides in the study area.

3. Data preparation

3.1. Landslide inventory

It is well known that information about landslides occurred in the past and present will do great help to predict the landslides in the future. Hence, it is crucial to gather the data reliably when mapping the landslide inventory map. In the present study, in order to compile a reliable landslide inventory map, early reports, aerial photographs interpretation, coupled with comprehensive field surveys by GPS (Global Positioning System) to locate the landslide positions. A total of 212 landslides were marked in the study area (Figure 1). The landslide inventory consists of 197 slides, 10 falls, and 5 debris flows (Hungr et al. 2014). The smallest landslide is about 20 m², the largest is about $1.1 \times 10^5$ m², and the average is about 7,500 m². Several techniques have been adopted to deal with the landslide polygon to point representation, such as centroids (Xu et al. 2012a; Dou et al. 2015; Oliveira et al. 2015; Chen et al. 2016; Youssef et al. 2016; Chen et al. 2017), seed cells (Süzен & Doyuran 2004a, 2004b), and diagnostic area (Lombardo et al. 2014). As more than 85% of the landslides in the study area are less than 10,000 m²; therefore, the centroid technique was used to deal with the transformation of landslide polygon to point in the present study.

The landslide inventory also shows that most landslides occurred along the main valleys and road networks. This means that the road networks have a strong impact on landslide occurrences (Brenning et al. 2014; Tien Bui et al. 2014).

3.2. Landslide conditioning factors

In this study, 12 factors were chosen according to the literature review and the general geo-environment of the study area. These conditioning factors are slope aspect, slope angle, altitude, profile curvature, plan curvature, NDVI, landuse, lithological unit, distance to rivers, distance to roads, distance to faults, and mean annual precipitation. A digital elevation model (DEM) with the resolution of 30 m × 30m was extracted from the ASTER GDÉM data. The data-sets were collected from the National Aeronautics and Space Administration (NASA) website (http://reverb.echo.nasa.gov/reverb/). The DEM data was applied to extract slope aspect, slope angle, altitude, profile curvature, and plan curvature of the study area using ArcGIS 10.0 software (Figure 2(a–e)). The Landsat8 OLI images with the resolution of 30 m × 30 m were adopted to extract NDVI and land use using ENVI 5.1 software (Figure 2(f,g)). The data-sets were collected from the U.S. Geological Survey (USGS) website (http://landsat.usgs.gov/landsat8.php). The lithological unit map was compiled from the geological map in 1:500,000 scale, and was grouped into six classes according to their lithological characters and geological ages (Figure 2(h)). Distance to rivers, distance to roads, and distance to faults were produced from topographic map (1:50,000), and geological map (1:500,000), respectively (Figure 2(i–k)). The mean annual precipitation map was compiled from the precipitation data of the study area (Figure 2(l)) (http://www.sxmb.gov.cn/index.php). The detailed information of the classes for each landslide conditioning factors is shown in Table 1. Finally, all thematic maps were converted into the same resolution of 30 m × 30 m, and the study area covers 2,240 columns and 2,020 rows.

In this study, the correlation between landslides and conditioning factors were calculated using frequency ratio model (Akgun et al. 2008; Chen et al. 2016b), and the calculated frequency ratios of each landslide conditioning factor were normalized in the range of 0–1. Then, the normalized frequency ratios were used as inputs to build models and produce landslide susceptibility maps.
4. Methodology

4.1. KLR model

KLR is a kernel version of logistic regression. It aims to classify the data which is difficult to be distinguished in current dimensional space by constructing a linear logistic regression model in a high-dimensional space. The KLR can be carried as follows:

$$\text{logit}(p) = w \cdot \varphi(x) + b$$  \hspace{1cm} (1)
where \( w \) is a vector of input variables (landslide conditioning factors), \( \varphi(\cdot) \) does a nonlinear transformation to each input variable, and \( b \) is a vector of constant. A logit function can be written as follows:

\[
\text{logit}(p) = \log_e \left( \frac{p}{1-p} \right)
\]

Hence, we could write the transformation of Equation (2) as:

\[
p = \frac{1}{1 + \exp \{-w \cdot \varphi(x) - b\}}
\]
There is a predigestion trick in the calculation of $\varphi(x)$: as $\varphi(x)\varphi'(x)$ is inevitable outcome in the calculation procedure, we define the inner product between the images of vectors as the kernel function. That is:

$$K(x_i, x_j) = \varphi(x_i)\varphi'(x_j)$$  \hspace{1cm} (4)$$

The Mercer’s condition (Mercer 1909) must be obeyed thus the interpretation of the kernel function will be valid. Several kernel functions have been suggested such as the liner kernel, the polynomial kernel, radial basis function (RBF), and the sigmoid kernel (Lin & Lin 2003). In this study, the
radial basis function kernel is applied which is the most commonly used in practical applications (Cawley & Talbot 2005). The RBF kernel is as follows:

\[ K(x_i, x_j) = \exp\left( -\frac{||x_i - x_j||^2}{2\delta^2} \right) \]  

(5)

where \( \delta \) is the turning parameter that controls the sensitivity of the kernel. A optimal \( \mathbf{w} \) that can be found by minimizing a cost function according to the representer theorem (Kimeldorf & Wahba 1971; Schölkopf et al. 2001).

### 4.2. Naïve-Bayes tree model

The NBTree is a hybrid algorithm of decision tree and Naïve-Bayes (Kohavi 1996). Its learning procedure is exactly the same as C4.5 decision tree (Quinlan 2014) except that the leaf nodes created are Naïve-Bayes categorizers instead of nodes predicting a single class.

C4.5 is an extension of ID3. A tree growing follows the attribute selection measure based on the concept of entropy which is defined as the degree of disorder. Suppose that D is a set of cases and \( |D| \) represents the total number of cases. These cases can be classified into \( m \) classes: \( D_i \) \( (i = 1, 2, \ldots, m) \). \( |D_i| \) represents the number of the cases that belongs to the class \( D_i \). The expected entropy for classifying the set \( D \) can be calculated as:

\[ \text{Entropy}(D) = -\sum_{i=1}^{m} \left( \frac{|D_i|}{|D|}\right) \log_2 \left[ \left( \frac{|D_i|}{|D|}\right) \right] \]  

(6)

If we partition the set \( D \) on attribute \( A \) (having \( v \) distinct values), then the expected entropy is expressed as:

\[ \text{Entropy}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j) \]  

(7)

The differences between the \( \text{Entropy}(D) \) and the \( \text{Entropy}_A(D) \) is defined as the information gain, InfoGain, whose value determines the attribute to be split on.

\[ \text{InfoGain}(A) = \text{Entropy}(D) - \text{Entropy}_A(D) \]  

(8)
However, InfoGain biases toward attributes having many distinct values and the corresponding number of splits is unreasonable. To solve this problem, InfoGain is normalized by SplitInfo in C4.5 decision tree. The SplitInfo is defined as follows:

$$\text{SplitInfo} = - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \left( \frac{|D_j|}{|D|} \right)$$ (9)

SplitInfo is a kind of Entropy on the split point of an attribute. The information gain ratio in C4.5 decision tree is defined as follows:

$$\text{GainRatio}(A) = \frac{\text{InfoGain}(A)}{\text{SplitInfo}(A)}$$ (10)

Naive-Bayes is a popular learning algorithm in pattern recognition due to its simplicity and linear run-time. It assumes that predictive attributes $a_1, a_2, \ldots, a_m$ are conditionally independent. Given a class attribute $c_j$, a class attribute of class set $C$, the joint distribution is then given compactly by:

$$P(a_1, a_2, \cdots, a_m | c_j) = \prod_{i=1}^{m} P(a_i | c_j)$$ (11)

Naive-Bayes aims to decide and choose the class that maximizes the posteriori probability, and then the classification rule of NB is:

$$c^* = \operatorname{argmax}_{c_j \in C} P(c_j | a_1, a_2, \cdots, a_m) = \frac{P(c_j) \prod_{i=1}^{m} P(a_i | c_j)}{\sum_{j=1}^{k} P(c_j) \prod_{i=1}^{m} P(a_i | c_j)}$$ (12)

$$= \propto \operatorname{argmax}_{c_j \in C} P(c_j) \prod_{i=1}^{m} P(a_i | c_j)$$

where $k$ is the total number of classes.

4.3. Alternating decision tree model

Alternating decision tree is a classifier combining boosting algorithm and decision tree (Freund & Mason 1999). Its representation is a direct generalization of decision tree where each decision node is replaced by two nodes: a splitter node and a prediction node.

A base ruler mapping from instances to real numbers consists of a precondition $c_1$, a base condition $c_2$, and two real numbers $a$ and $b$. The prediction is $a$ if $c_1 \land c_2$ or $b$ if $c_1 \land \neg c_2$. $\neg$ denotes negation (NOT). The values of $a$ and $b$ are calculated as:

$$a = \frac{1}{2} \ln \frac{W_+(c_1 \land c_2)}{W_-(c_1 \land c_2)}, \quad b = \frac{1}{2} \ln \frac{W_+(c_1 \land \neg c_2)}{W_-(c_1 \land \neg c_2)}$$ (13)

where $W(p)$ represents the total weight of the training instances that satisfy the predicate $p$. The best $c_1$ and $c_2$ are selected by minimizing the $Z_t(c_1, c_2)$ which is defined as:

$$Z_t(c_1, c_2) = 2 \sqrt{W_+(c_1 \land c_2)W_-(c_1 \land c_2)} + \sqrt{W_+(c_1 \land \neg c_2)W_-(c_1 \land \neg c_2)} + W(\neg c_2)$$ (14)
If symbol $R$ was defined as a set base rules, a new rule could be defined as $R_{t+1} = R_t + r_t$. $r_t(x)$ represents the two prediction values ($a$ and $b$) at each layer of the tree (with $T$ prediction layers). $x$ represents a set of instances. The classification of an instance is the sign of the sum of all the prediction values in $R_{t+1}$:

$$
\text{Class}(x) = \text{sign} \left( \sum_{i=1}^{T} r_t(x) \right)
$$

The algorithm starts by finding the best constant prediction for the entire data-set (Freund 7 Mason 1999). A cross validation is usually applied to make the selection (Dietterich 2000).

### 4.4. Preparation of training and validation data-set

In landslide susceptibility analysis, the dependent variable is a binary variable representing landslides or non-landslide. Models introduced above should be constructed using the training data-set that contains landslide pixels and non-landslide pixels, and validated by a validation data-set that contains landslide pixels and non-landslide pixels. It is generally recommended to use equal landslide and no-landslide pixels (Yesilnacar & Topal 2005; Nefeslioglu et al. 2008; Tien Bui et al. 2016b). With respect to replicate sampling, it is reported that single-sample studies run the risk of accidentally yielding a poor model, and the calculation of multiple models based on independent random samples should be advocated (Heckmann et al. 2014). Brenning (2005) took 50 samples to compare error rates across different sample sizes and statistical methods. Lombardo et al. (2016) also produced a set of 50 randomly generated replicates to assess landslide triggering thickness susceptibility. Cama et al. (2016) evaluated the robustness of predictive results on a total number of 80 repetitions. However, Heckmann et al. (2014) pointed out that these may not be enough and extracted 1,000 replicates. Therefore, in this paper, 70% of landslide locations (centroid, and a landslide location is shown as a pixel) were selected randomly to build landslide susceptibility models and the rest (30%) were used as validation purpose. Then, 212 non-landslide pixels were randomly selected outside the landslide polygons, and were also split into two parts (70/30) (Figure 1). Therefore, the final training data-set contained 70% (148) of the landslide pixels (points) and 148 non-landslide pixels. The validation data-set contained the remaining 30% (64) of the landslide pixels (points) and 64 non-landslide pixels. This generating and splitting process was repeated more than 10 times and the goodness of fit and prediction ability of each replicate have been assessed through area under the receiver operating characteristic (ROC) curves (AUC) and the statistical evaluation measures to find the best combination. Besides, during the running process of the models, 10-fold cross-validation method was used to reduce variability of the models results.

### 4.5. Multicolinearity analysis of landslide conditioning factors

In landslide susceptibility mapping, it is important to check the multicollinearities of landslide conditioning factors. The tolerance (TOL) and variance inflation factor (VIF) are two widely used indexes for multicollinearity checking. A TOL value less than 0.1 or a VIF value larger than 10 indicates a serious multicollinearity (Tien Bui et al. 2011). In this study, the TOL and VIF values were calculated with the training dataset using IBM SPSS Statistics 21 software, and the results are shown in Table 2. The results show that there are no multicollinearities among the 12 landslide conditioning factors.

### 4.6. Model evaluation and assessment

#### 4.6.1. Statistical evaluation measures

By using different cut-off values, the predicted probabilities were classified into one of the response levels (i.e. landslide or non-landslide). For each cut-off value, there always exists misclassification
situation. TP (true positive) and TN (true negative) represent the number of landslides and non-landslides that were correctly classified to the corresponding classes. Whereas FP (false positive) represents the number of non-landslides that were misclassified to the landslides class and FN (false negative) represents the number of landslide that were misclassified to the non-landslides class. Sensitivity, Specificity, and Accuracy were calculated as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (16)
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \quad (17)
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \quad (18)
\]

Accuracy is the proportion of landslide and non-landslide pixels that were correctly classified; Sensitivity is the proportion of landslide pixels that are correctly classified as landslide occurrences; Specificity is the proportion of the non-landslide pixels that are correctly classified as non-landslide.

4.6.2. The ROC curve and kappa index

Undoubtedly, if a susceptibility model leads to high percentage of correct classification and a low percentage of false positive and false negative classification, it could be considered as a valid model (Gobin et al. 2001). The model validity can be graphically represented on an ROC curve. The ROC curve was created by plotting ‘sensitivity’ versus ‘1 – specificity’ as the cut-off value varies. As the ROC curve cannot clearly explain the performance of a model, the area under the ROC curve (AUC) is usually applied for a quantitative comparison. An AUC value of 1 indicates a perfect model that correctly classified all landslide and non-landslide pixels, while an AUC value of 0 indicates a non-informative model (Walter 2002).

The reliability of the landslide models can be assessed using the kappa index (Hoehler 2000). For the interpretation of the magnitude of the kappa index: almost perfect (0.81–1.0), substantial (0.61–0.80), moderate (0.41–0.60), fair (0.21–0.40), slight (0–0.2), and poor (<0) (Landis & Koch 1977).

5. Results

5.1. Importance of conditioning factors by different models

The result of importance of conditioning factors by different models is shown in Table 3. It can be seen that all the conditioning factors have contributed to the models. Distance to roads, with an average merit (AM) value of 0.309, is the most important factor for three models, followed by
altitude and distance to rivers, which also have similar importance to the three models, respectively. The other conditioning factors have different contributions depending on the models used. Besides, all the three models yield small contributions by profile curvature, plan curvature, and slope aspect.

As all 12 conditioning factors have positive contributions to the three models, all 12 conditioning factors were used in the analysis to build landslide susceptibility maps.

5.2. Model results and analysis

In the present study, the KLR model was applied to extract the spatial relationship between landslides and conditioning factors. A forward stepwise process was carried out to incorporate the
conditioning factor that has an important contribution to the landslide. The calculated probability values were classified into five relative susceptibility classes using geometrical interval method: very low (0.001-0.017) (11.20%), low (0.017-0.058) (28.06%), moderate (0.058-0.157) (23.53%), high (0.157-0.400) (15.49%), and very high (0.400-0.998) (21.72%) (Figures 3 and 6).

NBTree model was applied to classify the pixel using the training data-set. Each pixel contains the information of conditioning factors and landslide (or non-landslide). At each decision node, a Naive-Bayes classifier will be constructed based on the sample set that filters into the node. The NBTree continues at data subset which includes instances of different class and stops at ‘pure’ data subset in which the instances are of the same class. At each leaf node (prediction node), an actual

Figure 3. Landslide susceptibility map derived from the KLR model.
The ADTree model was used to estimate the susceptibility index for each pixel using the training data-set. Pixels go down along the paths defined by the ADTree model. When reaching a decision value will be given representing the contribution to the landslides (or non-landslide). All the contributions encountered in the NBTree were summed to give the final prediction which is used to generate the landslides susceptibility map. Five relative susceptibility classes were divided according to the prediction value using geometrical interval method: very low (0.009-0.013) (15.68%), low (0.013-0.028) (29.92%), moderate (0.028-0.180) (21.53%), high (0.180-0.464) (14.11%), and very high (0.464-0.912) (18.84%) (Figures 4 and 6).

Figure 4. Landslide susceptibility map derived from the NBTree model.
node, a path will continue with the child node corresponding to the outcome of decision node. When reaching a prediction node, a path will continue with all of the children of the node. Susceptibility index was calculated by summing up all the values of prediction nodes which are encountered in the ADTree. In this study, geometrical interval method was applied to divide the landslide susceptibility index into five relative susceptibility classes, i.e. very low (0.056-0.196) (22.09%), low (0.196-0.356) (31.23%), moderate (0.356-0.540) (21.95%), high (0.540-0.749) (12.32%), and very high (0.749-0.990) (12.41%) (Figures 5 and 6).

Figure 5. Landslide susceptibility map derived from the ADTree model.
5.3. Model performance and validation

KLR, NBTree, and ADTree models were constructed using the training data-set, and 10-fold cross-validation method was used during the training and validation process to reduce variability of the result. The result is shown in Table 6. The ROC curves and AUC values for KLR, NBTree, and ADTree are shown in Figure 7.

The KLR has the highest accuracy with a value of 0.845, followed by the ADTree (0.841), and the NBTree (0.814). For the sensitivity value, the ADTree shows a value of 0.892, indicating the probability to correctly classify landslide pixels to the landslide class is 89.2%, followed by the KLR (0.865) and the NBTree (0.858). The KLR has the highest value of specificity (0.824) explaining that 82.4% non-landslide pixels are correctly classified to the non-landslide class. It is followed by the ADTree (0.791) and the NBTree (0.770).

The AUC values are 0.910, 0.880, and 0.903 for the KLR, the NBTree, and the ADTree models, respectively. It indicates a high degree of fit to the training data-set and the highest one is for the KLR model. Kappa indexes are 0.689, 0.628, and 0.682 for the KLR, the NBTree, and the ADTree models, respectively, which indicate a substantial agreement between prediction and observation.

The prediction capability and confirm accuracy of the trained susceptibility models was validated using the validation dataset. The AUC value, the kappa index, and the three statistical indexes introduced above were used in the validation process. The results are shown in Table 7. The ROC curves and AUC values for these three models are shown in Figure 8.

The results show that the NBTree has the highest classification accuracy (0.914) for the validation data-set, followed by the KLR model (0.898) and the ADTree (0.820). For the sensitivity, the highest is for the NBTree with a quite high value of 0.984, indicating that 98.4% of the landslide pixels are correctly classified into the landslide class. The KLR has the second highest sensitivity value of

| Table 6. Model performance. |
|-----------------------------|
| Models | KLR | NBTree | ADTree |
| Parameter                  |     |       |        |
| True positive              | 128 | 127   | 132    |
| True negative              | 122 | 114   | 117    |
| False positive             | 26  | 34    | 31     |
| False negative             | 20  | 21    | 16     |
| Sensitivity                | 0.865 | 0.858 | 0.892 |
| Specificity                | 0.824 | 0.770 | 0.791 |
| Accuracy                   | 0.845 | 0.814 | 0.841 |
| Kappa index                | 0.689 | 0.628 | 0.682 |
| AUC                        | 0.910 | 0.880 | 0.903 |
0.922. Whereas, the ADTree can only correctly classify 78.1% of landslides pixels. The highest specificity is 0.875 for the KLR, indicating that 87.5% of non-landslide pixels are correctly classified. It is followed by the ADTree (0.859), and the NBTree (0.844) models.

The ROC curves show that all the models have high AUC values, which indicate good prediction capabilities. The KLR and the NBTree have close AUC values (0.936 and 0.935, respectively),

Table 7. Model validation.

| Parameter      | KLR  | NBTree | ADTree |
|---------------|------|--------|--------|
| True positive | 59   | 63     | 50     |
| True negative | 56   | 54     | 55     |
| False positive| 8    | 10     | 9      |
| False negative| 5    | 1      | 14     |
| Sensitivity   | 0.922| 0.984  | 0.781  |
| Specificity   | 0.875| 0.844  | 0.859  |
| Accuracy      | 0.898| 0.914  | 0.820  |
| Kappa index   | 0.797| 0.828  | 0.641  |
| AUC           | 0.936| 0.935  | 0.917  |
followed by the ADTree model (0.917). The kappa indexes are 0.797, 0.828, and 0.641 for KLR, NBTree, and ADTree models, respectively. The results also indicate a substantial agreement between predicted and observed landslides.

6. Discussion and conclusion

Spatial prediction of landslides is considered to be useful for land-use planning and the first important step in landslide hazard and risk assessment (Fell et al. 2008). Therefore, it is necessary to select a susceptibility model with high prediction capability which depends on the methods used. Various methods for landslide susceptibility modelling have been suggested by the previous researchers, while the prediction accuracy of these methods is still debated (Akgun 2012). Some advanced machine learning methods such as KLR, NBTree, and ADTree have been applied in biomedical field (Liu et al. 2005; Lee et al. 2006), acoustic field (Katz et al. 2006; Karsmakers et al. 2007), Internet field (Cai et al. 2000), etc., and showed satisfactory results. Therefore, the investigation of new methods is highly necessary and it will help reach reasonable conclusions. As the application of these models for landslide susceptibility mapping has seldom been used. Thus, this study carried out an

Figure 8. ROC curves and AUC analysis for the three models with validation data-set.
investigation and comparison study of KLR, NBTree, and ADTree model for landslide susceptibility mapping.

The prediction capability of a susceptibility model depends on the method used. The goodness-of-fits and the validations of three susceptibility models are good and there is no significant difference between the AUC values of these three models (Figures 7 and 8). For the statistical indexes, they differed significantly between the models. In terms of classification accuracy, the KLR and the ADTree models are really close for the training data-set and are higher than that of the NBTree model by 3%. For the validation data-set, the NBT model has the highest classification accuracy of 91.4%. Therefore, the NBTree model showed a notable prediction power for the validation data-set and gave a poor account of goodness-of-fit for the training data-set. On the contrary, the ADTree has a high degree of goodness-of-fit for the training data-set and made a poor prediction for the validation data-set. In general, the KLR model has more balance for the training and validation data-sets in terms of the statistical index (Tables 6 and 7), while the NBTree and the ADTree models show significantly variance.

As not all the selected landslide conditioning factors have equal predictive capabilities, and sometimes some of them may create noises that reduce the prediction quality (Tien Bui et al. 2016b). In the present study, no multicollinearity was found among the 12 landslide conditioning factors (Table 2), and all the conditioning factors have contributions to the models. However, the results also showed that different conditioning factors had different contributions to models (Tables 3–5). In general, distance to roads, altitude, distance to rivers, and land-use have the highest importance to the three models, whereas profile curvature, plan curvature, and slope aspect reveal low predictive capabilities. In order illustrate the importance of landslide conditioning factors to the three used models visually, the normalized predictive capabilities of conditioning factors for the three models were calculated to indicate the relative importance of conditioning factors (Figure 9). It could be observed that the contribution of distance to roads occupies the highest percentages of 17.667%, 19.846%, and 18.945% for KLR, NBTree, and ADTree models, respectively. The highest contribution of distance to roads is related to the fact that 80.4% of landslide locations occurred along the main valleys and road networks (less than 500 m) (Figure 1). This high contribution of distance to roads also give rise to the phenomenon that the very high susceptibility classes follow similar spatial pattern of the roads. The slope aspect occupies the lowest percentages of 2.916%, 1.092%, and 2.391% for the three models, respectively. The NDVI produces higher contribution than lithological unit for the NBTree and ADTree models, while it is on the contrary for the KLR model. Therefore, it can be concluded that landslide conditioning factors tend to have different contributions depending on the types of models used (Tien Bui et al. 2016b).

Figure 9. Relative importance of conditioning factors to landslide susceptibility models.
Considering the overall performance and the model construction, the KLR model is considered better in this paper, because it shows a stable classification ability on both training and validation dataset. The landslide susceptibility maps produced in this study could be helpful for land-use planner to choose suitable construction sites.

Acknowledgments

The authors would like to express their gratitude to the Editor-in-Chief (Prof. Ramesh Singh) and two anonymous reviewers for their helpful comments on the manuscript. This research was supported by the Key National Basic Research Program of China (973 Program) (Grant no. 2014CB744702) and the Research Cultivation Fund of Xi’an University of Science and Technology (Grant No. 201722).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Key National Basic Research Program of China (973 Program) [grant number 2014CB744702]; and the Research Cultivation Fund of Xi’an University of Science and Technology [grant number 201722].

ORCID

Wei Chen http://orcid.org/0000-0002-5825-1422
Haoyuan Hong http://orcid.org/0000-0001-6224-069X

References

Akgun A. 2012. A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at Izmir, Turkey. Landslides. 9:93–106.
Akgun A, Dag S, Bulut F. 2008. Landslide susceptibility mapping for a landslide-prone area (Findikli, NE of Turkey) by likelihood-frequency ratio and weighted linear combination models. Environ Geol. 54:1127–1143.
Aksoy B, Ercanoglu M. 2012. Landslide identification and classification by object-based image analysis and fuzzy logic: an example from the Azdavay region (Kastamonu, Turkey). Compu Geosci. 38:87–98.
Althuwaynee OF, Pradhan B, Lee S. 2016. A novel integrated model for assessing landslide susceptibility mapping using CHAID and AHP pair-wise comparison. Int J Remote Sens. 37:1190–1209.
Brenning A. 2005. Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat Haz Earth Sys. Sci. 5:853–862.
Brenning A, Schwinn M, Ruiz-Páez A, Muenchow J. 2014. Landslide susceptibility near highways is increased by one order of magnitude in the Andes of southern Ecuador, Loja province. Nat Haz Earth Sys Discussions. 2:1945–1975.
Cai D, Delcher A, Kao B, Kasif S. 2000. Modeling splice sites with Bayes networks. Bioinformatics. 16:152–158.
Cama M, Conoscenti C, Lombardo L, Rotigliano E. 2016. Exploring relationships between grid cell size and accuracy for debris-flow susceptibility models: a test in the Giampilieri catchment (Sicily, Italy). Environ Earth Sci. 75:1–21.
Catani F, Lagomarsino D, Segoni S, Tofani V. 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat Haz Earth Sys Sci. 13:2815–2831.
Cawley GC, Talbot NL. 2005. The evidence framework applied to sparse kernel logistic regression. Neurocomputing. 64:119–135.
Chen W, Chai H, Sun X, Wang Q, Ding X, Hong H. 2016a. A GIS-based comparative study of frequency ratio, statistical index and weights-of-evidence models in landslide susceptibility mapping. Arabian J Geosci. 9:1–16.
Chen W, Ding X, Zhao R, Shi S. 2016b. Application of frequency ratio and weights of evidence models in landslide susceptibility mapping for the Shangzhou District of Shangluo City, China. Environ Earth Sci. 75:1–10.
Chen W, Li W, Chai H, Hou E, Li X, Ding X. 2016c. GIS-based landslide susceptibility mapping using analytical hierarchy process (AHP) and certainty factor (CF) models for the Baozhong region of Baoji City, China. Environ Earth Sci. 75:1–14.
Chen W, Pourghasemi HR, Zhao Z. 2016d. A GIS-based comparative study of Dempster-Shafer, logistic regression, and artificial neural network models for landslide susceptibility mapping. Geocarto Int. DOI: 10.1080/10106049.2016.1140824

Chen W, Wang J, Xie X, Hong H, Trung NV, Bui DT, Wang G, Li X. 2016. Spatial prediction of landslide susceptibility using integrated frequency ratio with entropy and support vector machines by different kernel functions. Environ Earth Sci. 75:1–15.

Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Duan Z, Ma J. 2017. A comparative study of logistic model tree, random forest, and classification tree regression models for spatial prediction of landslide susceptibility. CATENA. 151:147–160.

Colkesen I, Sahin EK, Kavzoglu T. 2016. Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. J Afr Earth Sci. 118:53–64.

Conoscenti C, Rotigliano E, Cama M, Caraballos-Arias NA, Lombardo L, Agnesi V. 2016. Exploring the effect of absence selection on landslide susceptibility models: A case study in Sicily, Italy. Geomorphology. 261:222–235.

Dahal RK, Hasegawa S, Nonomura A, Yamanaka M, Masuda T, Nishino K. 2008. GIS-based weights-of-evidence modelling of rainfall-induced landslides in small catchments for landslide susceptibility mapping. Environ Geol. 54:311–324.

Dietterich TG. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning. 40:139–157.

Ding Q, Chen W, Hong H. 2016. Application of frequency ratio, weights of evidence and evidential belief function models in landslide susceptibility mapping. Geocarto Int. 1–21.

Dou J, Oguchi T, Hayakawa YS, Uchiyama S, Saito H, Paudel U. 2014. GIS-based landslide susceptibility mapping using a certainty factor model and its validation in the Chuetsu Area, Central Japan. In: Sassa K, Canuti P, Yin Y, editor. Landslide science for a safer geoenvironment. Cham: Springer; p. 419–424.

Dou J, Yamagishi H, Pourghasemi HR, Yunus AP, Song X, Xu Y, Zhu Z. 2015. An integrated artificial neural network model for the landslide susceptibility assessment of Osado Island, Japan. Nat Hazards. 78:1749–1776.

Duc DM. 2013. Rainfall-triggered large landslides on 15 December 2005 in Van Canh district, Binh Dinh province, Vietnam. Landslides. 10:219–230.

Feizizadeh B, Blaschke T, Nazmfar H. 2014. GIS-based ordered weighted averaging and Dempster–Shafer methods for landslide susceptibility mapping in the Urmia Lake Basin, Iran. Int J Digital Earth. 7:688–708.

Fell R, Corominas J, Bonnard C, Cascini L, Leroi E, Savage WZ. 2008. Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning. Eng Geol. 102:99–111.

Freund Y, Mason L. The alternating decision tree learning algorithm. Proceedings of the ICML; 1999; Bled, Slovenia.

Gobin A, Campling P, Feyen J. 2001. Logistic modelling to identify and monitor local land management systems. Agricultural Systems. 67:1–20.

Gokceoglu C, Sezer E. 2012. Soft computing modeling in landslide susceptibility assessment. In: Pradhan B, Buchroithner M, editor. Terrigenous mass movements. Berlin, Heidelberg: Springer; p. 51–90.

Heckmann T, Gegg K, Gegg A, Becht M. 2014. Sample size matters: investigating the effect of sample size on a logistic regression susceptibility model for debris flows. Nat Haz Earth Syst Sci. 14:259–278.

Hoehler FK. 2000. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J Clin Epidemiol. 53:499–503.

Hong H, Pourghasemi HR, Pourtaghi ZS. 2016. Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models. Geomorphology. 259:105–118.

Hong H, Chen W, Xu C, Youssef AM, Pradhan B, Tien Bui D. 2017. Rainfall-induced landslide susceptibility assessment at the Chongren area (China) using frequency ratio, certainty factor, and index of entropy. Geocarto Int. 32:139–154.

Hong H, Pradhan B, Xu C, Bui DT. 2015. Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. CATENA. 133:266–281.

Hung Or, Leroueil S, Picarelli L. 2014. The Varnes classification of landslide types, an update. Landslides. 11:167–194.

Karsmakers P, Pelkmans K, Suykens JA, Hamme HV. 2007. Fixed-size kernel logistic regression for phoneme classification. Proceedings of the INTERSPEECH; Antwerp, Belgium.

Katz M, Schaffoner M, Andelic E, Kruger SE, Wendemuth A. 2006. Sparse kernel logistic regression using incremental feature selection for text-independent speaker identification. Proceedings of the 2006 IEEE Odyssey-The Speaker and Language Recognition Workshop; IEEE.

Kimeldorf G, Wahba G. 1971. Some results on Tchebycheffian spline functions. J Math Anal Appl. 33:82–95.

Kohavi R. 1996. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. Proceedings of the KDD; Citeseer.

Landis JR, Koch GG. 1977. The measurement of observer agreement for categorical data. Biometrics. 33:159–174.

Lee H, Tu Z, Deng M, Sun F, Chen T. 2006. Diffusion kernel-based logistic regression models for protein function prediction. OMICS: J Integr Biol. 10:40–55.
Lin H-T, Lin C-J. 2003. A study on sigmoid kernel for SVM and the training of non-PSD kernels by SMO-type methods. Taipei: National Taiwan University, Department of Computer Science and Information Engineering. 
Liu K-Y, Lin J, Zhou X, Wong ST. 2005. Boosting alternating decision trees modeling of disease trait information. BMC Genet. 6:1. 
Lombardo L, Cama M, Conoscenti C, Marker M, Rotigliano E. 2015. Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). Nat Hazards. 79:1621–1648. 
Lombardo L, Cama M, Maerker M, Rotigliano E. 2014. A test of transferability for landslides susceptibility models under extreme climatic events: application to the Messina 2009 disaster. Nat Hazards. 74:1951–1989. 
Lombardo L, Fubelli G, Amato G, Bonasera M. 2016. Presence-only approach to assess landslide triggering-thickness susceptibility: a test for the Mili catchment (north-eastern Sicily, Italy). Nat Hazards. 84:565–588. 
Malahangim R, Olsen MJ. 2016. Evaluation of the influence of source and spatial resolution of DEMs on derivative products used in landslide mapping. Geomat Nat Haz Risk. 7:1835–1855. 
Malahangim R, Olsen MJ, O’Banion MS. 2016. Evaluation of landslide susceptibility mapping techniques using lidar-derived conditioning factors (Oregon case study). Geomat Nat Haz Risk. 7:1884–1907. 

Malamud BD, Turcotte DL, Guzzetti F, Reichenbach P. 2004. Landslide inventories and their statistical properties. Earth Surf Proc Landforms. 29:687–711. 
Manzo G, Tofani V, Segoni S, Battistini A, Catani F. 2013. GIS techniques for regional-scale landslide susceptibility assessment: the Sicily (Italy) case study. Int J Geogr Inf Sci. 27:1433–1452. 
Merce J. 1909. Functions of positive and negative type, and their connection with the theory of integral equations. Phil Trans R Soc London A, containing papers of a mathematical or physical character. 209:415–446. 
Myronidis D, Papageorgiou C, Theophanous S. 2016. Landslide susceptibility mapping based on landslide history and analytic hierarchy process (AHP). Nat Hazards. 81:245–263. 

Nefeslioglu HA, Duman TY, Durmaz S. 2008. Landslide susceptibility mapping for a part of tectonic Kelkit Valley (Eastern Black Sea region of Turkey). Geomorphology. 94:401–418. 
Oliveira SC, Zézere JL, Garcia RA. 2015. Structure and characteristics of landslide input data and consequences on landslide susceptibility assessment and prediction capability. In: Lollino G, Giordan D, Crosta GB, Corominas J, Azzam R, Wasowski J, Sciarra N, editors. Engineering Geology for Society and Territory-Volume 2. Cham: Springer: p. 189–192. 
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 BT, Bui DT, Dholakia M, Prakash I, Pham HV. 2016. A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area. Geotech Geol Eng. 1–18. 

Pham BT, Tien Bui D, Pourghasemi HR, Indra P, Dholakia MB. 2015a. Landslide susceptibility assessment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. Theor. Appl. Climatol. DOI: 10.1007/s00704-015-1702-9 
Pham BT, Tien Bui D, Prakash I, Dholakia MB. 2015b. A comparison study of predictive ability of support vector machine and Naive Bayes tree methods in landslide susceptibility assessment at an area between Tehri Garhwal and Pauri Garhwal, Uttarakhand State (India) using GIS. Proceedings of the National Symposium on Geomatics for Digital India Conventions of ISG & ISRS; Jaipur, India. 
Pielke R, Schellnhuber H, Sahagian D. 2003. Non-linearities in the Earth System. Global Change Newsletter. 55:11–15. Pourghasemi HR, Kerle N. 2016. Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran. Environ Earth Sci. 75:1–18. 
Pourghasemi HR, Pradhan B, Gokceoglu C. 2012. Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran. Nat Hazards. 63:965–996. 
Pourghasemi HR, Pradhan B, Gokceoglu C, Mohammad M, Moradi HR. 2013. Application of weights-of-evidence and certainty factor models and their comparison in landslide susceptibility mapping at Haraz watershed, Iran. Arabian J Geosci. 6:2351–2365. 

Pourghasemi HR, Rossi M. 2016. Landslide susceptibility modeling in a landslide prone area in Mazandaran Province, north of Iran: a comparison between GLM, GAM, MARS, and M-AHP methods. Theor Appl Climatol. DOI: 10.1007/s00704-016-1919-2 
Pradhan AMS, Tarolli P, Kang HS, Lee JS, Kim YT. 2016. Shallow landslide susceptibility modeling incorporating rainfall statistics: a case study from the Deokjeok-ri Watershed, South Korea. Int J Erosion Control Eng. 9:18–24. 
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. Compu Geosci. 51:350–365. 
Pradhan B, Lee S. 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environ Model Softw. 25:747–759. 
Quinlan JR. 2014. C4.5: programs for machine learning. London: Elsevier.
Yilmaz I. 2010. The effect of the sampling strategies on the landslide susceptibility mapping by conditional probability and artificial neural networks. Environ Earth Sci. 60:505–519.

Youssef AM. 2015. Landslide susceptibility delineation in the Ar-Rayth area, Jizan, Kingdom of Saudi Arabia, using analytical hierarchy process, frequency ratio, and logistic regression models. Environ Earth Sci. 73:8499–8518.

Youssef AM, Pourghasemi HR, Pourtaghi ZS, Al-Katheeri MM. 2016. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides. 13:839–856.

Zhu AX, Wang R, Qiao J, Qin CZ, Chen Y, Liu J, Du F, Yang L, Zhu T. 2014. An expert knowledge-based approach to landslide susceptibility mapping using GIS and fuzzy logic. Geomorphology. 214:128–138.