ORIGINAL ARTICLE

A comparative study of sequential minimal optimization-based support vector machines, vote feature intervals, and logistic regression in landslide susceptibility assessment using GIS

Binh Thai Pham[1,2] · Dieu Tien Bui[3] · Indra Prakash[4] · Long Hoang Nguyen[5] · M. B. Dholakia[6]

Received: 22 August 2015 / Accepted: 3 May 2017 / Published online: 17 May 2017

Abstract Landslide susceptibility assessment using GIS has been done for part of Uttarakhand region of Himalaya (India) with the objective of comparing the predictive capability of three different machine learning methods, namely sequential minimal optimization-based support vector machines (SMOSVM), vote feature intervals (VFI), and logistic regression (LR) for spatial prediction of landslide occurrence. Out of these three methods, the SMOSVM and VFI are state-of-the-art methods for binary classification problems but have not been applied for landslide prediction, whereas the LR is known as a popular method for landslide susceptibility assessment. In the study, a total of 430 historical landslide polygons and 11 landslide affecting factors such as slope angle, slope aspect, elevation, curvature, lithology, soil, land cover, distance to roads, distance to rivers, distance to lineaments, and rainfall were selected for landslide analysis. For validation and comparison, statistical index-based methods and the receiver operating characteristic curve have been used. Analysis results show that all these models have good performance for landslide spatial prediction but the SMOSVM model has the highest predictive capability, followed by the VFI model, and the LR model, respectively. Thus, SMOSVM is a better model for landslide prediction and can be used for landslide susceptibility mapping of landslide-prone areas.

Keywords Landslides · GIS · Sequential minimal optimization (SMO) · Support vector machines (SVM) · Vote feature intervals (VFI) · India

Introduction

A landslide is one of the most widespread and devastating natural hazards causing heavy loss to property, infrastructure, and a lot of casualties annually all over the world (Cevik and Topal 2003; Liu et al. 2009; Yin et al. 2010). According to the Centre for Research on the Epidemiology of Disasters, landslides are responsible for at least 17% casualties among the deadly natural hazards throughout the world (Lacasse and Nadim 2009). India is one of the top Asian countries affected by landslides (Pham et al. 2015). Landslides in India mainly occur in the Himalayan range (Onagh et al. 2012). The Defense Terrain Research Laboratory reported that Himalayan landslides kill at least one person per 100 km^2 with over 220 fatalities every year (Mukane 2014). Many efforts have been made to minimize the damages caused by landslides in this Himalayan area during recent decades (Das et al. 2010). One of the

✉ Binh Thai Pham
phambinhgtvt@gmail.com; binhpt@utt.edu.vn

[1] Department of Civil Engineering, Gujarat Technological University, Nr.Visat Three Roads, Visat - Gandhinagar Highway, Chandkheda, Ahmedabad, Gujarat 382424, India

[2] Department of Geotechnical Engineering, University of Transport Technology, 54 Trieu Khuc, Thanh Xuan, Hanoi, Vietnam

[3] Geographic Information System Group, Department of Business Administration and Computer Science, Telemark University College, Hallvard Eikas Plass 1, 3800 Bø, Telemark, Norway

[4] Department of Science and Technology, Bhaskarcharya Institute for Space Applications and Geo-Informatics (BISAG), Government of Gujarat, Gandhinagar, India

[5] University of Transport Technology, 54 Trieu Khuc, Thanh Xuan, Hanoi, Vietnam

[6] Department of Civil Engineering, LDCE, Gujarat Technological University, Ahmedabad, Gujarat 380015, India

effective solutions is to produce landslide susceptibility maps of landslide-prone areas (Mathew et al. 2009).

Landslide susceptibility map can be used to minimize human loss and property through proper land use planning by decision makers (Dai et al. 2002). Landslide susceptibility can be expressed as the spatial probability of landslide occurrences (Varnes 1984). Assessment of landslide susceptibility is based on the assumption that future landslides would be more likely to occur under similar conditions to those of the previous landslides (Varnes 1984). Therefore, the spatial relationship between past landslide occurrences and a set of affecting factors is usually carried out using different statistical methods.

More recently, many statistical methods have been developed and applied successfully to produce landslide susceptibility maps for many regions in the world. Common methods are frequency ratio (Poudyal et al. 2010; Yalcin et al. 2011), weight of evidence (Dahal et al. 2008; Neuhäuser and Terhorst 2007), evidential belief function (Althuwaynee et al. 2012; Lee et al. 2013), artificial neural networks (Ermini et al. 2005; Yilmaz 2009), decision trees (Hwang et al. 2009; Yeon et al. 2010), and support vector machines (Yao et al. 2008; Yilmaz 2010). Even though these methods have performed relatively well, their performance is different in different areas due to local geo-environment factors. Thus, making comparisons between various modeling techniques is felt necessary to select a suitable method to produce a reliable landslide susceptibility map which may be applicable in wider areas (Akgun 2012). Therefore, the main objective of the present study is to apply and compare the predictive capability of three different machine learning methods, namely sequential minimal optimization-based support vector machines (SMOSVM), vote feature intervals (VFI), and logistic regression (LR) for spatial prediction of landslide occurrences. Out of these methods, the SMOSVM and VFI are the state-of-the-art methods for binary classification problems but have not been applied so far for landslide prediction, whereas another method of the LR is known as a popular method for landslide susceptibility assessment.

As a case study, a part of Uttarakhand State (India), which is one of the landslide-prone areas of Himalaya, has been selected for landslide susceptibility assessment. For validation and comparison of results, statistical index-based methods and the receiver operating characteristic (ROC) curve have been used. Data processing and modeling have been done using Weka 3.7.12 and ArcMap 10.2 software.

Description of study area

The study area is located in the middle of the Tehri Garhwal and Pauri Garhwal districts in the Uttarakhand State (India) which is a landslide-prone area of Himalaya, between latitudes 29°56′38″N to 30°09′37″N and longitudes 78°29′01″E to 78°37′06″E, covering an area of 323.815 km^2 (Fig. 1). Elevation in the area varies from 380 m to 2180 m above sea level, with mean elevation of 1081 m. Slope angles in this area are very steep up to 70°. About 85.45% of the hill slopes are having average slope (15°–45°).

Broadly, four types of land covers have been classified in the area which are non-forest (39.02%), dense forest (31.96%), open forest (22.36%), and scrub land (6.67%). Soil in this area is mainly of two types: silty and loamy. Silty soil is classified as fine and occupies 26.27% of the study area. Loamy soil is classified into four categories, namely skeletal loamy, coarse loamy, fine loamy, and mixed loamy. Skeletal loamy occupies major area (42.02%), followed by coarse loamy (20.1%), fine loamy (8.02%), and mixed loamy (3.6%), respectively.

The study area is situated in a subtropical monsoon region having three separate seasons: summer (March to June), monsoon (June to September), and winter (October to February). Heavy rainfall usually occurs in the monsoon season. The annual precipitation varies from 770 to 1684 mm. Temperature in the study area varies from 1.3 °C in winter to 45 °C in summer. General relative humidity varies between 54 and 63%, and the highest is about 85% (http://pauri.nic.in/pages/display/55-the-land).

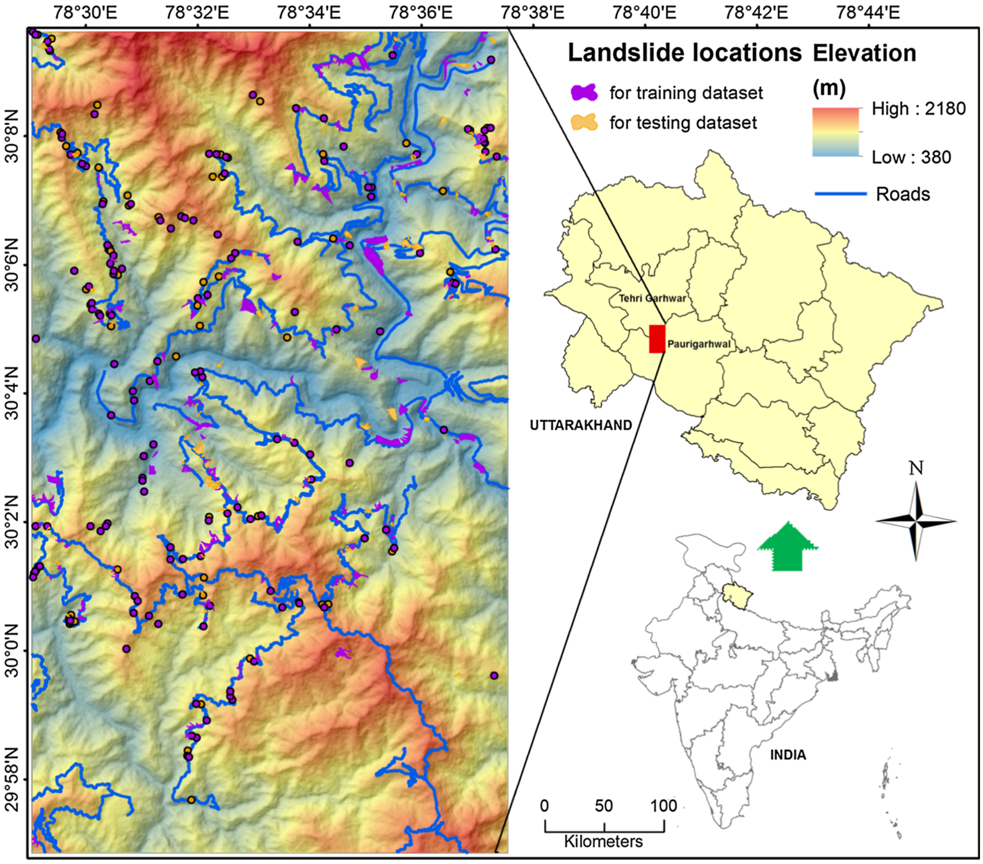
Methodology

The methodology in the present study involves five steps: (1) data collection and interpretation, (2) dataset preparation, (3) building landslide models using different methods (SMOSVM, VFI, and LR), (4) validation and comparison of the predictive capability of these landslide models, and (5) delineation of landslide susceptibility maps.

Data collection and interpretation

A landslide inventory map was constructed with 430 landslide locations which were identified by interpretation of Google Earth images using Google Earth pro 7.0 software, and LANDSAT-8 satellite images. Out of these landslides, a total of 236 landslides with area larger than 400 m^2 (equal to a pixel size of DEM 20 m) were represented as polygons, and 194 landslides with areas smaller than 400 m^2 were represented as points (Fig. 1). The largest landslide area is about 199,574 m^2. Newspaper records, historical landslide reports, and extensive field investigation were then employed to validate these landslide locations. Most of these landslides are translation type (325 locations), and the remaining landslides are rotational type (105 locations). It is shown in Fig. 1 that landslides in

Fig. 1 Location of landslides in the study area



the study area usually occur along roads and highways. Islam et al. (2014) stated that landslides in this study area annually occur mainly during monsoon season. Examples of landslide photographs in the study area are shown in Fig. 2.

In addition, the selection of landslide affecting factors in the modeling is very important (Tsangaratos and Ilia 2016). In the present study, a total of 11 landslide conditioning factors (slope angle, slope aspect, elevation, curvature, lithology, soil, land cover, distance to roads, distance to rivers, distance to lineaments, and rainfall) were selected based on the analysis of the geo-environmental characteristics and mechanism of landslide occurrences in the study area. Thematic maps considering these conditioning factors were generated and constructed as the raster data with grid size of 20 × 20 m for analysis.

A digital elevation model (DEM) in the study area with a spatial resolution of 20 × 20 m was generated from the state topographic map available on the published literature (http://www.ahec.org.in/wfw/maps.htm). Using the DEM data, four geomorphologic factors were extracted including slope angle, slope aspect, elevation, and curvature. Slope angle map (Fig. 3a) was constructed with six classes (0°–8°, 8°–15°, 15°–25°, 25°–35°, 35°–45°, and >45°). These classes are based on the analysis of frequency and the natural mechanism of landslide occurrences in the study area as landslide is more susceptible in the average slopes (15°–45°), less susceptible in very gentle slopes (smaller than 8°), and very high slopes (larger than 45°) (Pham et al. 2015; Varnes 1984). A slope aspect map (Fig. 3b) was generated with nine classes, namely flat (−1), north (0–22.5 and 337.5–360), northeast (22.5–67.5), east (67.5–112.5), southeast (112.5–157.5), south (157.5–202.5), southwest (202.5–247.5), west (247.5–292.5), and northwest (292.5–337.5). The classification of these aspect classes is based on the fact that different slope facing directions have different impaction of solar radiation and rainfall on the slopes which controls the moisture of terrain affecting landslide occurrences (Varnes 1984). Different classes have been selected for the elevation map (Fig. 3c) including <600, 600–750, 750–900, 900–1050, 1050–1200, 1200–1350, 1350–1500, 1500–1650, 1650–1800, and >1800 m which is based on the analysis of topographic characteristics in conjunction with frequency analysis of landslide occurrences in the study area (Pham et al. 2015). A curvature map (Fig. 3d) was constructed with three classes as concave (<−0.05), flat (−0.05 to 0.05), and convex (>0.05) which is based on the fact that frequency of landslide is more in concave and convex areas than flat areas (Varnes 1984).

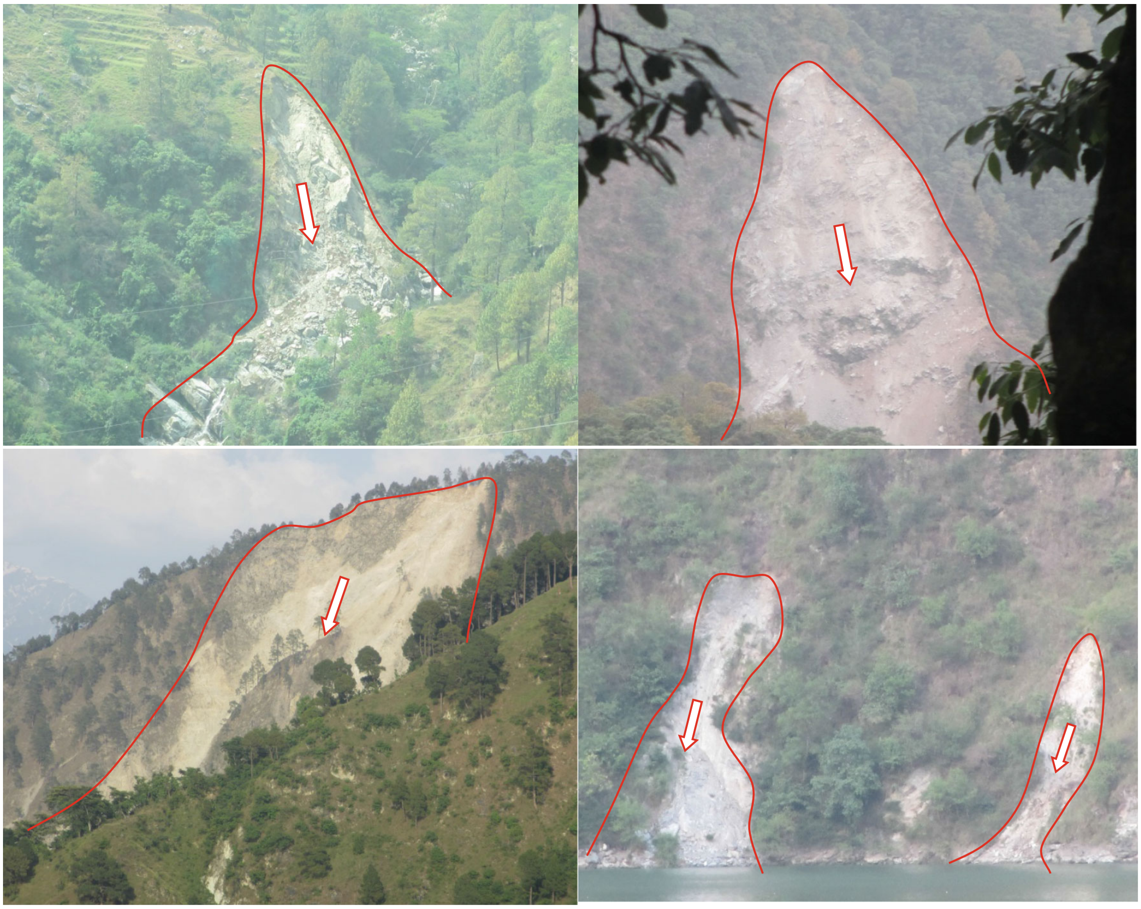


Fig. 2 Examples of landslides in study area

A lithological map of the study area (Fig. 3e) was extracted from the state geological map. Lithology has been classified into six groups, namely Amri group (quartzite, phyllite), Blaini and Krol group (boulder bed and limestone), Jaunsar group (phyllite and quartzite), Bijni group (quartzite, phyllite), Tal group (sandstone, shale, quartzite, phyllite, and limestone), and Manikot shell limestone (limestone). A land cover map (Fig. 3f) was extracted from the state land cover map with four classes including non-forest, dense forest, open forest, and scrub land. A soil map (Fig. 3g) was also extracted from the state soil map, and it includes five classes: coarse loamy, fine loamy, fine silt, skeletal loamy, and mixed loamy. Rainfall data were extracted from meteorological data which were compiled for 30 years from 1984 to 2014 from the climate forecast system reanalysis (CFSR) in global weather data for SWAT (NCEP 2014). A rainfall map (Fig. 3h) was then constructed based on spline interpolation method (Kawamura et al. 1992) with different classes (<900, 900–1000, 1000–1100, 1100–1200, 1200–1300, 1300–1400, 1400–1500, and >1500 mm) based on the frequency analysis in the study and adjacent area (Pham et al. 2016f).

Road and river networks were obtained from Google Earth images and drainage analysis in GIS. A distance to roads map (Fig. 3i) was constructed by buffering the road sections on slope angles larger than 15° in the study area, and six classes of distance to roads (0–40, 40–80, 80–120, 120–160, 160–200, and >200 m) were selected based on the frequency analysis in the study area and adjacent area (Pham et al. 2016f). The distance to rivers map (Fig. 3j) was also constructed by buffering rivers sections on slope angles larger than 15° in the study area, and the distance classes were classified into six intervals (0–40, 40–80, 80–120, 120–160, 160–200, and >200 m) based on the frequency analysis in the study area and adjacent area (Pham et al. 2016f). Lineaments were extracted from LANDSAT-8 satellite images using Geomatica 2015 software. A distance to lineaments map (Fig. 3k) was built by buffering the lineaments in the study area. Distance to lineaments map shows various classes (0–50, 50–100, 100–150, 150–200, 200–250, 250–300, 300–350, 350–400, 400–450, 450–500, and >500 m) which is based on the frequency analysis in the study area and adjacent area (Pham et al. 2016f).

Fig. 3 Landslide affecting factor maps: **a** slope angle map, **b** slope aspect map, **c** elevation map, **d** curvature map, **e** lithological map, **f** land cover map, **g** soil map, **h** rainfall map, **i** distance to roads map, **j** distance to rivers map, **k** distance to faults map

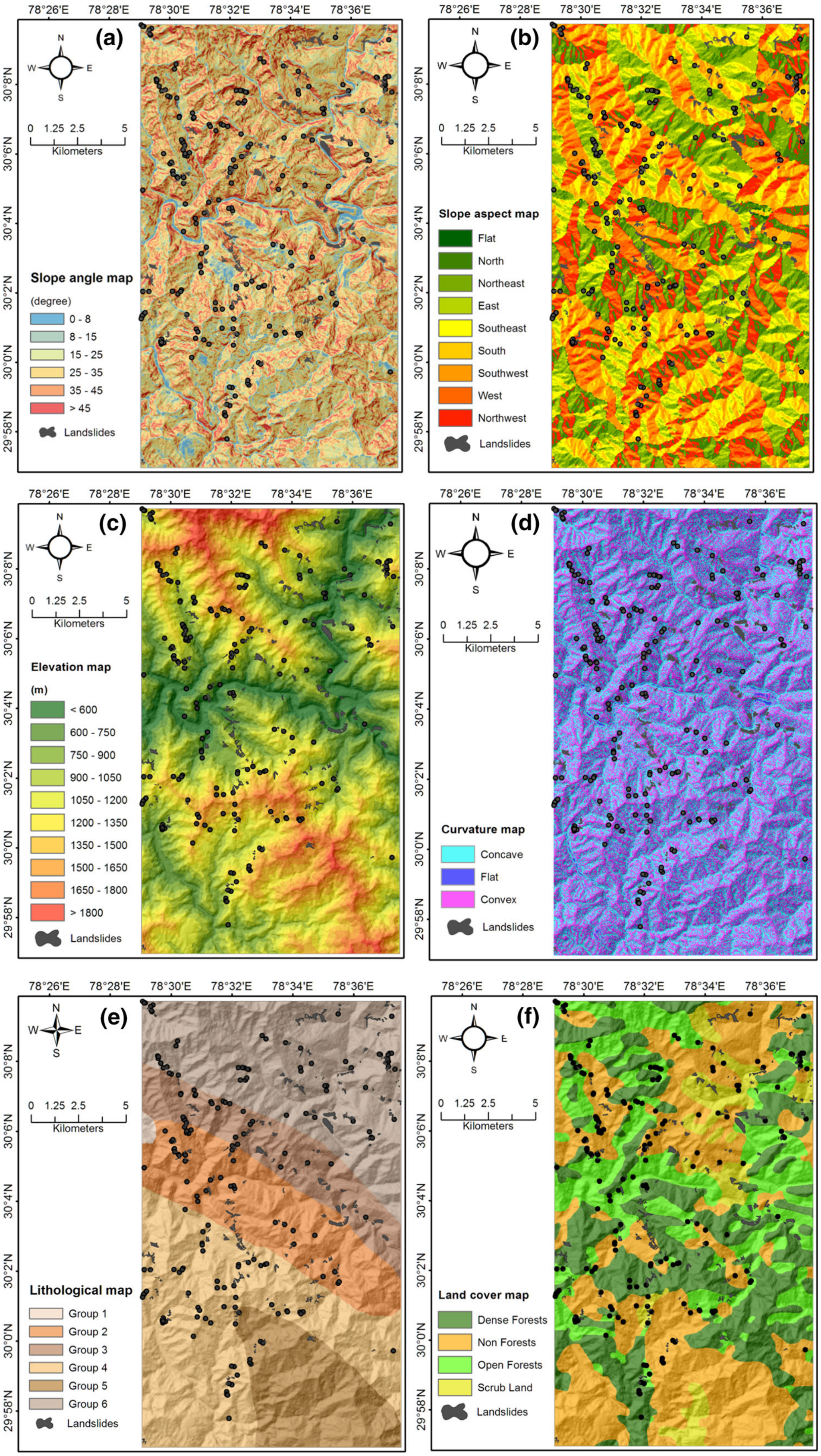
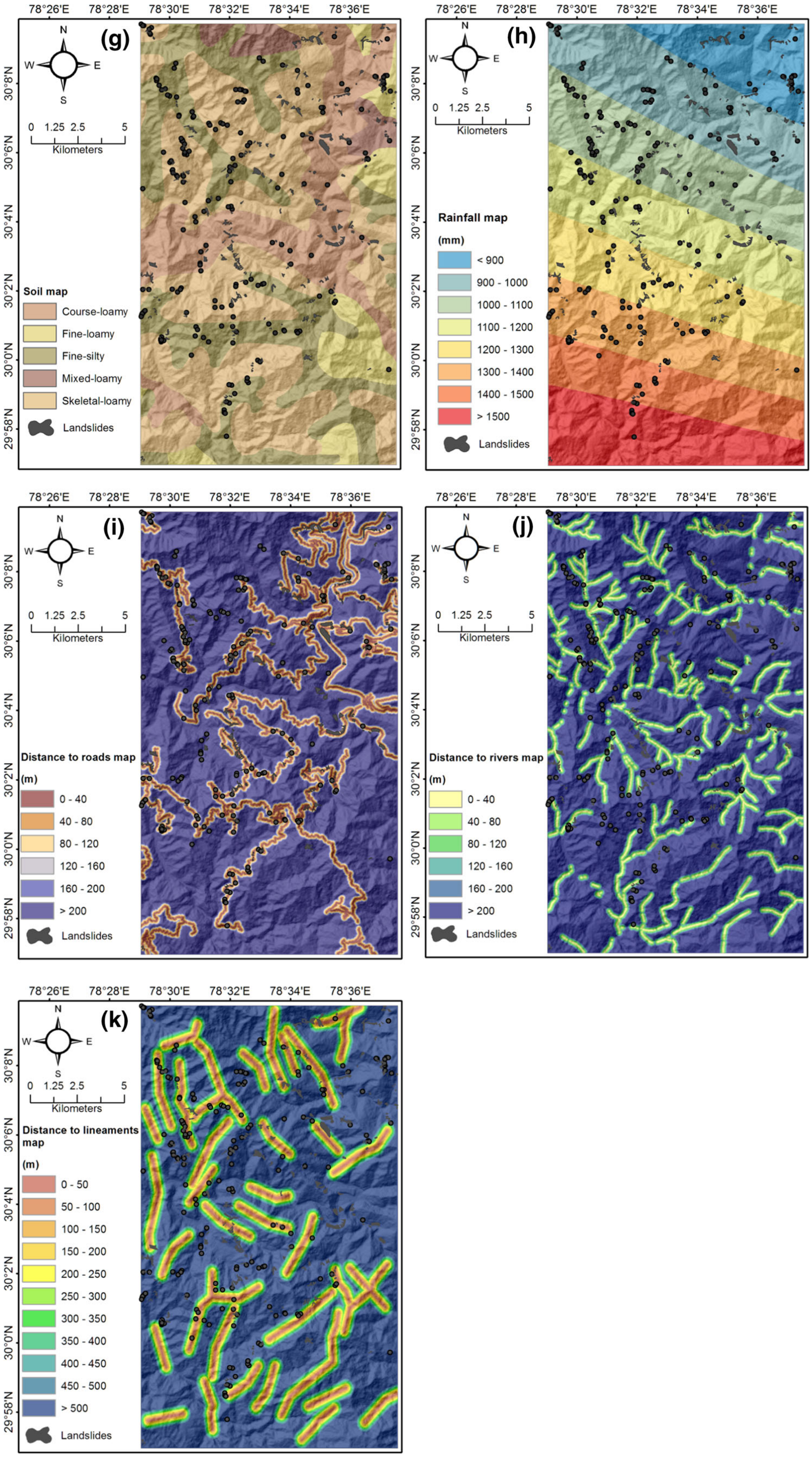


Fig. 3 continued



Dataset preparation

According to Tien Bui et al. (2016b), landslide susceptibility maps are viewed as a binary classification. Therefore, both landslides and non-landslides have been considered to construct classification inputs for landslide models. For landslide susceptibility modeling, the dataset is to be split into two subsets including a training dataset and a testing dataset (Chung and Fabbri 2003).

In the present study, for generating the training dataset, 70% of the landslide locations (301 landslides) were selected randomly from landslide inventory map. These landslides were then converted into pixels of 20 × 20 m size. A total number of landslide pixels in the training dataset are 6133. These landslide pixels were then combined with 6133 non-landslide pixels which were randomly extracted from non-landslide areas. Finally, the training dataset was obtained by sampling these landslide and non-landslide pixels with the 11 landslide conditioning factors.

For generating the testing dataset, 30% of the remaining landslide locations (129 landslides) were also converted into pixels with a size of 20 × 20 m with 1614 landslide pixels. A total of 1614 non-landslide pixels were also extracted randomly from non-landslide areas. These landslide and non-landslide pixels were sampled with the 11 landslide conditioning factors to generate the testing dataset.

The training dataset was then used for building the landslide models, while the testing dataset was employed for validating and comparing the performance of the landslide models.

Landslide susceptibility classifiers

Sequential minimal optimization-based support vector machines

Sequential minimal optimization-based support vector machines (SMOSVM) is a hybrid approach of support vector machines (SVM) and sequential minimal optimization (SMO). The SVM is one of the most effective methods for classification with high accuracy (Kavzoglu et al. 2014; Peng et al. 2014; Pourghasemi et al. 2013; Yilmaz 2010). Despite the merits, the SVM also has a limitation in sophisticated studies with large input data (Lai et al. 2006) because the SVM uses inequality constraints in solving large-scale quadratic programming problems leading to great computational complexity (Lai et al. 2006). Therefore, the SMOSVM was introduced by Platt (1999) to handle this problem of the SVM (Platt 1999). The SMOSVM has been utilized successfully for brain tumor classification (Deepa and Aruna 2011), involving designing of very large-scale integration systems (Kuan et al. 2012). So far, the SMOSVM has not been explored for landslide spatial prediction.

The SMOSVM method is based on the theorem that the large quadratic programming problem generated in the SVM (Vapnik 2000) can be broken into a series of the smallest possible quadratic programming problems (Platt 1999). These small quadratic programming problems are tackled analytically using two Lagrangian multipliers per step instead of using a time-consuming numerical quadratic programming optimization with an inner loop (Flake and Lawrence 2002). Therefore, the SMOSVM is faster than the SVM. Different kernel functions define the feature space for classifying the training set examples (Luo and Cheng 2012) used in the SMOSVM. It is very important to select a suitable kernel function for classification in the SMOSVM because different kernel functions will give different results (Luo and Cheng 2012). In this study, the SMOSVM was evaluated for the predictive capability in landslide susceptibility assessment and the radial basis function (RBF) kernel was chosen as it is the most suitable kernel function for landslide model (Pham et al. 2016c).

Giving a training dataset (x, y) in which $x = x_i$, $i = 1, 2, \ldots, 11$ is the vector of the 11 landslide conditioning factors, and $y = (y_1, y_2)$ is the vector of classified variables including landslide and non-landslide classes. The SMO is utilized to optimize the quadratic programming problem through two main steps: (1) identifying and solving analytically the two Lagrange multipliers (Platt 1999) and (2) choosing suitable Lagrange multipliers to optimize the quadratic programming problem using heuristics (Platt 1999).

The quadratic programming problem arisen during training process of the SVM is shown as following expression:

$$\text{Maxi min } e: R(\beta_i) = \sum_{i=1}^{11} \beta_i - \frac{1}{2}\sum_{i=1}^{11}\sum_{j=1}^{11} \beta_i \beta_j y_i y_j k(x_i, x_j)$$

$$\text{Subject to: } \sum_{i=1}^{11} \beta_i y_i = 0 \text{ vs } 0 \le \beta_i \le a,\ i = 1, 2, \ldots, 11 \tag{1}$$

where β_i are positive real constants, a is the complexity parameter (Vapnik 2000), and $k(x_i, x_j)$ is the RBF kernel that is defined as an infinite dimensional feature space (Vapnik and Vapnik 1998). The RBF kernel function is given by Eq. (2) as follows:

$$k(x_i, x_j) = \exp\left\{-\|x_i - x_j\|_2^2/\sigma^2\right\},\ \sigma^2 \text{ is the squared bandwidth} \tag{2}$$

Vote feature intervals

Vote feature intervals (VFI) is a classification algorithm which is based on attribute discretization (Demiröz and Güvenir 1997). The VFI is a non-incremental approach using a set of feature intervals in representing a range of feature values (Demiröz and Güvenir 1997). Features in the VFI method are considered as independent variables rather than dependent ones (Marsolo et al. 2007). The VFI method has been employed successfully in classification such as in computer sciences for coping with highly imbalanced datasets (Del Gaudio et al. 2014) and in medical sciences for diagnosis of erythema-to-squamous diseases (Nanni 2006). This method has been utilized for the first time in landslide susceptibility assessment in the present study.

The VFI is carried out in two main phases: (1) training phase and (2) classification phase. In training phase, feature intervals are first constructed by calculating the lowest and highest feature value around each class for each feature. Next, in the classification phase, a feature vote is calculated for each class based on each interval of each feature, and then the vote of each feature interval is integrated to produce outputs (Malviya and Umrao 2014). The advantage of the VFI is that it ignores the missing feature values occurring in both training and classification phases; therefore, it provides classification accuracy (Demiröz and Güvenir 1997).

Let an instance $t = (t_1, t_2, \ldots, t_{11}, k_j)$ where $t_1, t_2, \ldots, t_{11}$ is the feature values of the 11 landslide conditioning factors, k_j, $j = 1, 2$, is the classified classes which represent landslide or non-landslide, t_f is the feature value of the test sample t. The VFI algorithm is presented below.

If t_f is unknown (missing), the factors with missing values are simply ignored.

If t_f is known, the feature interval of each factor is calculated, and then for each class, the vote of each factor is calculated as below:

$$\begin{aligned}&\text{factor\_vote}[t,k] = \text{interval\_class\_vote}[t,i,k]\\&\text{interval\_class\_vote}[t,i,k] \text{ is the vote of factor t given to class } k\end{aligned} \tag{3}$$

These vote vectors are summed up to obtain a total vote vector (vote[k_1], vote[k_2]). Finally, the class corresponding to the highest total vote is selected as the predicted class (Demiröz and Güvenir 1997).

Logistic regression

Logistic regression (LR) is a multivariate analysis method which was proposed in late 1960s and early 1970s (Cabrera 1994; Lee 2005). The LR is well known as an efficient method for binary classification problems including landslide spatial prediction (Lee 2005; Ohlmacher and Davis 2003). The LR has been proven more efficient than other methods such as certainty factor, likelihood ratio, artificial neural networks, and multi-criteria decision analysis for landslide susceptibility assessment (Akgun 2012; Devkota et al. 2013; Lee et al. 2007). In general, the LR is known as a promising method which should be used for landslide prediction and assessment (Das et al. 2010).

For landslide spatial prediction, the main principle of the LR is to use the mathematical concept of the logit–natural logarithm to analyze the spatial relationship between a set of landslide affecting factors and the obscene and presence of a landslide event (Akgun 2012). In the present study, the LR is used as a benchmark model to compare with the SMOSVM and VFI models which have been applied first time in the landslide assessment.

Suppose $z = z_i$, $i = 1, 2, \ldots, 11$ represents the vector of 11 landslide affecting factors, and $f = (f_1, f_2)$ represents outcome variables of landslide or non-landslide. The LR is trained using the logit–natural logarithm as following equation:

$$f = f(P) = \ln\left(\frac{P}{1-P}\right) = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_n z_n \tag{4}$$

Based on the above logit–natural logarithm, the probability of a landslide event can be obtained as following equation:

$$P = P(f|z) = \frac{e^{\alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_n z_n}}{1 + e^{\alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_n z_n}} \tag{5}$$

where α_0 is the intercept condition, $\alpha_1, \alpha_2, \ldots, \alpha_n$ are the regression coefficients (Cabrera 1994).

Delineation of landslide susceptibility classes

Landslide susceptibility classes were classified by reclassification of landslide susceptibility indexes (LSI) which were generated from training process of three landslide models. The LSI indicates how susceptible an area is to landslide occurrences. The LSI was first calculated for all the pixels in the study area. Thereafter, it was sorted in descending order. The reclassification of the LSI can be done using mathematical methods such as quantiles, natural breaks, standard deviation, (equal intervals (Ayalew et al. 2004), and equal area percentage (Pradhan and Lee 2010). These methods are described briefly below.

The quantiles-based technique takes into account different values in the same susceptible class. The natural breaks method builds the boundaries in big jumps existing in the LSI values. The equal intervals method considers the relative relationship among susceptible classes. The standard deviation technique uses the average value of the LSI to create the susceptible class breaks (Akgun et al. 2008). The equal area percentage technique is carried out on the base dividing the LSI values according to area percentage from small LSI values to high ones (Pradhan and Lee 2010).

Among the above methods, the equal area percentage technique is the most widely used (Pradhan and Lee 2010; Tien Bui et al. 2016b). Therefore, in this study, the equal area percentage technique was selected to classify the LSI values. Landslide susceptibility maps were then constructed into five classes: very low (40%), low (20%), moderate (20%), high (10%), and very high (10%).

Model performance validation

The performance of three landslide models (SMOSVM, VFI, and LR) was validated using statistical index-based methods and receiver operating characteristic curve.

Statistical index-based methods

In the present study, statistical indexes (sensitivity, specificity, and accuracy) were selected to evaluate the performance of landslide models. These indexes were calculated based on the values from the confusion matrix which is a table indicating a visualization of the performance of an algorithm (Alizadehsani et al. 2013). For two classes of landslide and non-landslide, the confusion matrix has two rows and two columns that show four values such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN) (Table 1). The TP infers the number of pixels that were correctly predicted as landslide; the FP is the number of pixels that were incorrectly predicted as landslides; the TN means the number of pixels that were correctly predicted as non-landslide; the FN is the number of pixels that were incorrectly predicted as non-landslide (Bennett et al. 2013).

Sensitivity is defined as the proportion of landslide pixels which are correctly classified as landslide. *Sensitivity* can only be calculated from the pixel being defined as landslide (Pham et al. 2016b). This means that *sensitivity* indicates how good the prediction of the model is for identifying landslide pixels when only looking at the pixels being defined as landslide.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

Specificity is defined as the proportion of non-landslide pixels which are correctly classified as non-landslide. It means that *specificity* can only be calculated from the pixels being defined as non-landslide (Pham et al. 2016d). *Specificity* indicates how good the prediction of the model is for identifying non-landslide pixels when only looking at the pixels being defined as non-landslide.

$$\text{Specificity} \ = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{7}$$

Accuracy is defined as the proportion of landslide and non-landslide pixels that are correctly classified. The *accuracy* is equal to 1 (100%) indicating the optimal model. Higher *accuracy* value indicates better predictive models.

$$\text{Accuracy} \ = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

Receiver operating characteristic curve

Receiver operating characteristic (ROC) curve is a useful method to determine the quality of the probabilistic model by characterizing its ability to reliably predict the occurrence or non-occurrence of landslide events (Feizizadeh et al. 2014). The ROC curve shows the trade-off between the two values including "*sensitivity*" on the X-axis and "100-*specificity*" on the Y-axis (Dou et al. 2014). Area under the curve (AUC) indicates how good landslide model is (Pham et al. 2016e). The AUC value obtained using the training dataset indicates how good the relationship between the inputs and the outputs, and the AUC value using the testing dataset shows how good is the model predictive capability (Pham et al. 2017). The model has a perfect performance if the AUC value equals to 1 (Pradhan 2013; Pradhan and Lee 2010). Higher AUC value indicates better performance of landslide model (Pham et al. 2016a).

Results and analysis

Landslide susceptibility maps using the SMOSVM, VFI and LR models

The landslide susceptibility maps constructed using the SMOSVM, VFI and LR models are shown in Figs. 4, 5, and 6, respectively. To evaluate the performance of these maps, the landslide inventory map has been used in conjunction with these susceptibility maps. Landslide density (LD) is then calculated for each susceptible class and is shown in Table 2. The LD is a ratio between the percentage of landslide pixels and the percentage of class pixels in each class on landslide susceptibility map (Pham et al. 2016f).

Table 1 Confusion matrix

	Actual landslide class	Actual non-landslide class
Predicted landslide class	True positive (TP)	False positive (FP)
Predicted non-landslide class	False negative (FN)	True negative (TN)

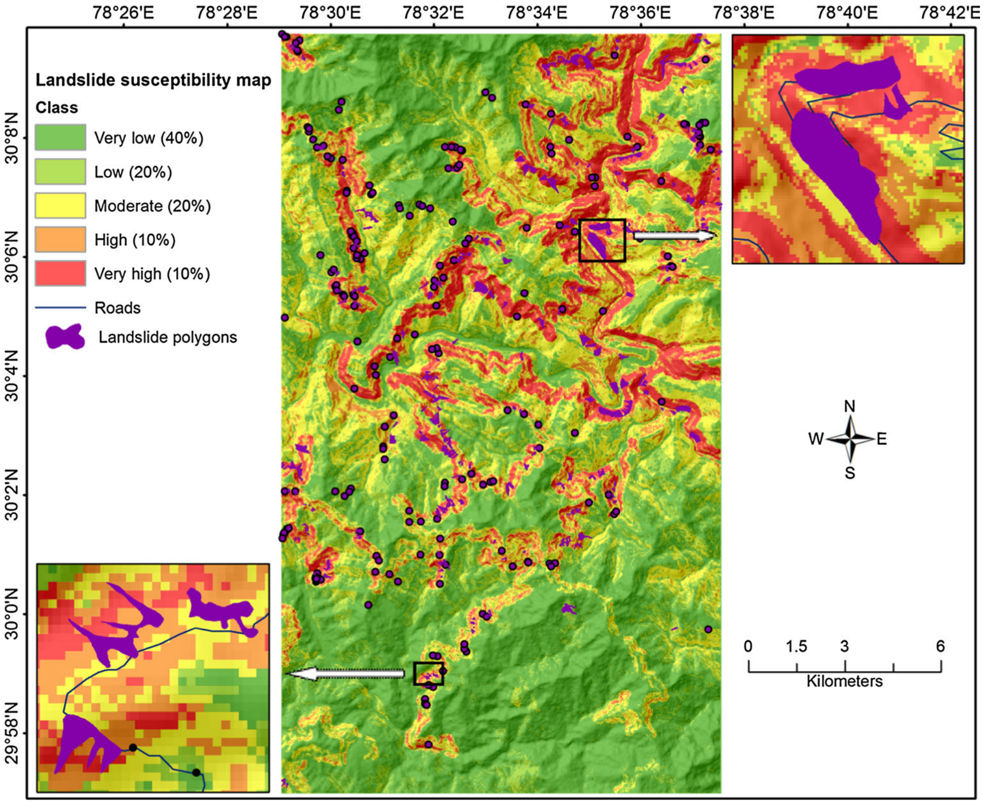


Fig. 4 Landslide susceptibility map using the SMOSVM model

Landslide density analysis results (Table 2) show that landslide pixels were observed mainly in the very high class (LD = 5.42 for the SMOSVM model, LD = 4.68 for the VFI model, and LD = 4.01 for the LR model) and high class (LD = 2.4 for the SMO model, LD = 2.29 for the VFI model, and LD = 2.34 for the LR model). Landslide pixels were observed very few in moderate class (LD = 0.7 for the SMOSVM model, LD = 0.98 for the VFI model, and LD = 1.07 for the LR model), low class (LD = 0.22 for the SMOSVM model, LD = 0.2 for the VFI model, and LD = 0.44 for the LR model), and very low class (LD = 0.05 for the SMOSVM model, LD = 0.13 for the VFI model, and LD = 0.16 for the LR model). Result analysis shows that three susceptibility maps produced from three landslide models have a good performance but the susceptibility map produced by the SMOSVM model is better than those produced from other models (VFI and LR) as LD in very high class of the SMOSVM model (5.42) is higher than those of the VFI model (4.68) and the LR model (4.01).

Performance of models and their comparison

Predictive capability of three landslide models (SMOSVM, VFI, and LR) has been validated using statistical index-based methods. The values of the confusion matrix were first extracted (Table 3), and then the values of statistical indexes were calculated as shown in Table 4.

For the training dataset, the SMOSVM model has the highest value of sensitivity (82.14%), followed by the VFI model (76.74%), and the LR model (73.66%), respectively. As for the specificity, the VFI model has the highest value (86.63%), followed by the SMOSVM model (82.26%), and the LR model (74.48%), respectively. As per the accuracy, the SMOSVM model has the highest value (82.20%), followed by the VFI model (80.91%), and the LR model (74.06%), respectively.

For the testing dataset, the SMOSVM model has the highest value of sensitivity (81.19%), followed by the VFI model (75.27%), and the LR model (73.11%), respectively. Regarding the specificity, the VFI model has the highest value (81.02%), followed by the SMOSVM model (76.87%), and the LR model (74.19%), respectively. As for the accuracy, the SMOSVM model has the highest value (78.87%), followed by the VFI model (77.85%), and the LR model (73.64%), respectively.

Furthermore, the performance of three landslide models (SMOSVM, VFI, and LR) has been validated using the ROC curve, as shown in Figs. 7 and 8. As for the training dataset, the analysis of ROC curve shows that the SMOSVM model has the highest value of AUC (0.891), followed by the VFI model (0.862), and the LR model

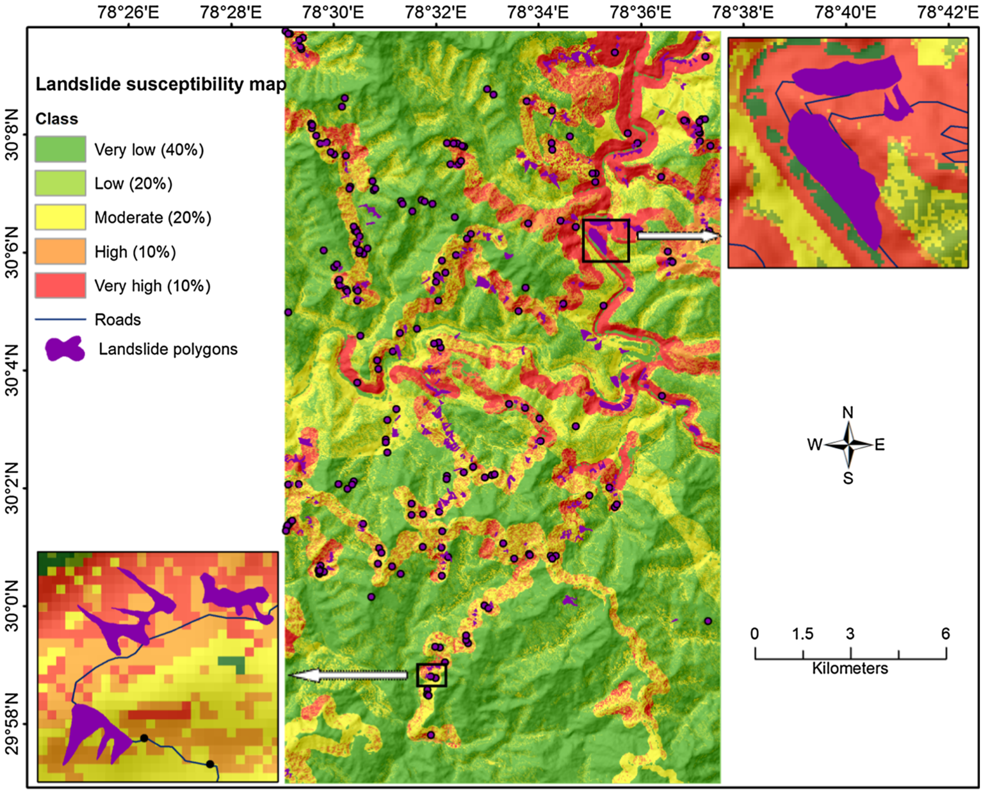


Fig. 5 Landslide susceptibility map using the VFI model

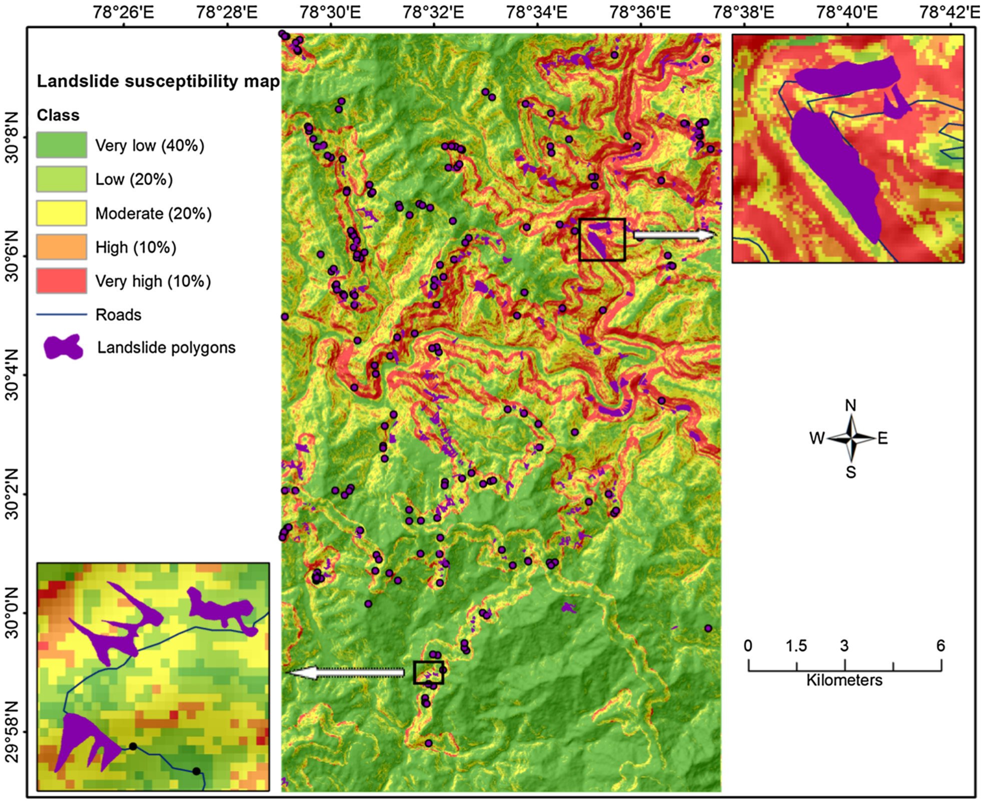


Fig. 6 Landslide susceptibility map using the LR model

Table 2 Landslide density on landslide susceptibility maps of different landslide models

No	Class	% Class pixels	% Landslide pixels			LD		
			SMOSVM	VFI	LR	SMOSVM	VFI	LR
1	Very low	40	1.83	5.23	6.35	0.05	0.13	0.16
2	Low	20	4.45	4.05	8.86	0.22	0.2	0.44
3	Moderate	20	13.97	19.67	21.32	0.7	0.98	1.07
4	High	10	24.51	23.33	23.36	2.4	2.29	2.34
5	Very high	10	55.23	47.72	40.11	5.42	4.68	4.01

Table 3 Confusion matrix for different landslide models

Models	Dataset		True landslide	True non-landslide
SMOSVM	Training	Predictive landslide	5046	1086
		Predictive non-landslide	1097	5036
	Testing	Predictive landslide	1213	401
		Predictive non-landslide	281	1333
VFI	Training	Predictive landslide	5440	692
		Predictive non-landslide	1649	4484
	Testing	Predictive landslide	1339	275
		Predictive non-landslide	440	1174
LR	Training	Predictive landslide	4593	1539
		Predictive non-landslide	1642	4491
	Testing	Predictive landslide	1207	407
		Predictive non-landslide	444	1170

Table 4 Performance of landslide models

Models	Dataset	Sensitivity (%)	Specificity (%)	Accuracy (%)
SMOSVM	Training	82.14	82.26	82.20
	Testing	81.19	76.87	78.87
VFI	Training	76.74	86.63	80.91
	Testing	75.27	81.02	77.85
LR	Training	73.66	74.48	74.06
	Testing	73.11	74.19	73.64

(0.806), respectively. Similarly, the analysis of ROC curve for the testing dataset also shows that the SMOSVM model has the highest value of AUC (0.856), followed by the VFI model (0.826), and the LR model (0.806), respectively.

Discussion and conclusions

Landslide susceptibility assessment has been done for producing the landslide susceptibility maps of part of landslide-prone area of Uttarakhand region of Himalaya using three different machine learning methods, namely SMOSVM, VFI and LR. Out of these methods, the SMOSVM and VFI are state-of-the-art methods for binary classification problems but have not been applied for landslide prediction, whereas the LR is another known popular method for landslide susceptibility assessment.

Regarding validation and comparison of landslide models, the ROC curve is well known as a standard method; however, the ROC curve only validates the general performance of models and it does not show the classification accuracy of landslide and non-landslide classes. Moreover, the ROC curve used for landslide prediction is affected by some factors such as (i) geo-environmental characteristics of the study area, (ii) landslide affecting factors and landslide inventory map, (iii) the analyzing methods used. In addition, Bennett et al. (2013) have also suggested to use multiple evaluation criteria for the validation of models. Therefore, in the present study, statistical index-based methods, which can fill the gap of the ROC curve method, have been also used for validation of landslide models.

Analysis of the results shows that all three landslide models (SMOSVM, VFI, and LR) have good performance

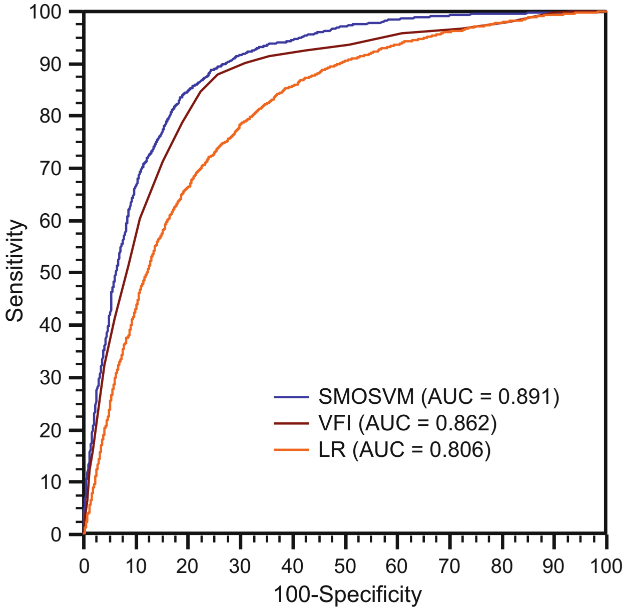


Fig. 7 Analysis of the ROC curve of three landslide models (SMOSVM, VFI, and LR) using training dataset

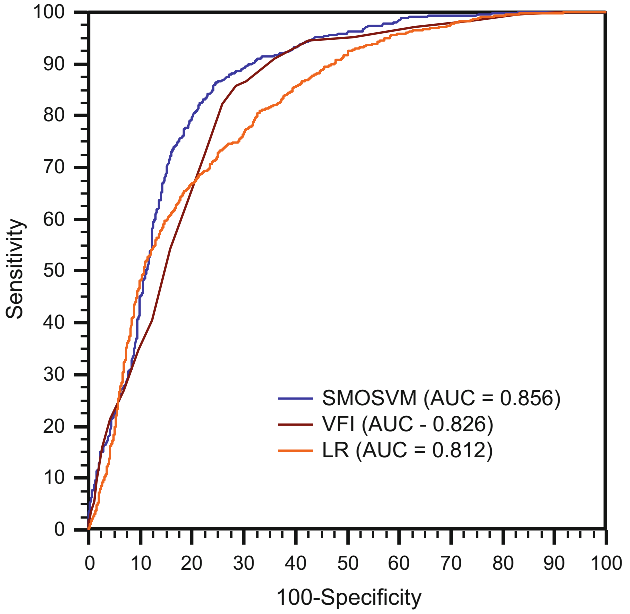


Fig. 8 Analysis of the ROC curve of three landslide models (SMOSVM, VFI, and LR) using testing dataset

for landslide susceptibility assessment in the present study but the SMOSVM model (AUC = 0.856) has the highest predictive capability, followed by the VFI model (AUC = 0.826), and the LR model (AUC = 0.806), respectively. Analysis results are reasonable because the SMOSVM used the SMO technique which might improve not only the processing speed but also improve the performance of the SVM classifier. The optimization techniques can also generally improve the performance of a single landslide model (Tien Bui et al. 2016a). Moreover, the SVM classifier used in the SMOSVM is considered as one of the best methods for spatial prediction of landslides (Pham et al. 2016c).

As for the VFI, it is known as an efficient classifier for binary classification problems. The VFI uses a set of feature intervals for representing a range of affecting factor values which can enhance its predictive capability of landslide occurrences. However, its performance might be affected by the independent assumption of variables (Demiröz and Güvenir 1997). Thus, the performance of VFI model observed in the present study is better than the LR model, but it is lower than the SMOSVM model.

The LR model is already well-known good landslide model (Marsolo et al. 2007) as it uses a sequence of convergence criterions to maximize the likelihood function for predicting landslide occurrences (Pham et al. 2016c). In the present study though the predictive capability of the LR model is relatively good (AUC 0.806), its performance is not better than the SMOSVM and VFI models which have been applied first time in the landslide study.

In conclusion, the SMOSVM has the highest predictive capability compared to other two methods of the VFI and LR even though all the three models have performed well in the present study for landslide susceptibility assessment. Thus, the SMOSVM is a more promising method which can be used as a better alternative for landslide spatial prediction and development of landslide susceptibility maps for land use planning and hazard management.

Acknowledgements Authors are thankful to the Director, Bhaskarcharya Institute for Space Applications and Geo-Informatics (BISAG), Department of Science & Technology, Government of Gujarat, Gandhinagar, Gujarat, India, for providing facilities to carry out this research work.

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 İzmir, 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

Alizadehsani R et al (2013) A data mining approach for diagnosis of coronary artery disease. Comput Methods Programs Biomed 111:52–61

Althuwaynee OF, Pradhan B, Lee S (2012) Application of an evidential belief function model in landslide susceptibility mapping. Comput Geosci 44:120–135

Ayalew L, Yamagishi H, Ugawa N (2004) Landslide susceptibility mapping using GIS-based weighted linear combination, the case in Tsugawa area of Agano River, Niigata Prefecture, Japan. Landslides 1:73–81

Bennett ND et al (2013) Characterising performance of environmental models. Environ Model Softw 40:1–20

Cabrera AF (1994) Logistic regression analysis in higher education: An applied perspective. In: Smart JC (ed) Higher education: handbook of theory and research, vol 10. Agathon press, New york, pp 225–256

Cevik E, Topal T (2003) GIS-based landslide susceptibility mapping for a problematic segment of the natural gas pipeline, Hendek (Turkey). Environ Geol 44:949–962. doi:10.1007/s00254-003-0838-6

Chung C-JF, Fabbri AG (2003) Validation of spatial prediction models for landslide hazard mapping. Nat Hazards 30:451–472

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

Dai F, Lee C, Ngai YY (2002) Landslide risk assessment and management: an overview. Eng Geol 64:65–87

Das I, Sahoo S, van Westen C, Stein A, Hack R (2010) Landslide susceptibility assessment using logistic regression and its comparison with a rock mass classification system, along a road section in the northern Himalayas (India). Geomorphology 114:627–637

Deepa S, Aruna D (2011) Second order sequential minimal optimization for brain tumour classification. Eur J Sci Res 64:377–386

Del Gaudio R, Batista G, Branco A (2014) Coping with highly imbalanced datasets: a case study with definition extraction in a multilingual setting. Nat Lang Eng 20:327–359

Demiröz G, Güvenir H (1997) Classification by voting feature intervals. In: Machine Learning: ECML-97, pp 85–92

Devkota KC et al (2013) Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling-Narayanghat road section in Nepal Himalaya. Nat Hazards 65:135–165

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 (eds) Landslide science for a safer geoenvironment. Springer, pp 419–424

Ermini L, Catani F, Casagli N (2005) Artificial neural networks applied to landslide susceptibility assessment. Geomorphology 66:327–343

Feizizadeh B, Roodposhti MS, Jankowski P, Blaschke T (2014) A GIS-based extended fuzzy multi-criteria evaluation for landslide susceptibility mapping. Comput Geosci 73:208–221

Flake GW, Lawrence S (2002) Efficient SVM regression training with SMO. Mach Learn 46:271–290

Hwang S, Guevarra IF, Yu B (2009) Slope failure prediction using a decision tree: a case of engineered slopes in South Korea. Eng Geol 104:126–134

Islam M, Chattoraj S, Ray CP (2014) Ukhimath landslide 2012 at Uttarakhand, India: causes and consequences. Int J Geomat Geosci 4:544

Kavzoglu T, Sahin EK, Colkesen I (2014) Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. Landslides 11:425–439

Kawamura H, Sasaki T, Otsuki T (1992) Spline interpolation method. Google Patents

Kuan T-W, Wang J-F, Wang J-C, Lin P-C, Gu G-H (2012) VLSI design of an SVM learning core on sequential minimal optimization algorithm very large scale integration (VLSI) systems. IEEE Trans Very Large Scale Integr (VLSI) Syst 20:673–683

Lacasse S, Nadim F (2009) Landslide risk assessment and mitigation strategy. In: En Sassa K, Canuti P (eds) Landslides-disaster risk reduction. Springer, Berlin, pp 31–61

Lai KK, Yu L, Zhou L, Wang S (2006) Credit risk evaluation with least square support vector machine. In: Wang GY, Peters JF, Skowron A, Yao Y (eds) Rough sets and knowledge technology. Springer, pp 490–495

Lee S (2005) Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. Int J Remote Sens 26:1477–1491

Lee S, Ryu J-H, Kim I-S (2007) Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: case study of Youngin, Korea. Landslides 4:327–338

Lee S, Hwang J, Park I (2013) Application of data-driven evidential belief functions to landslide susceptibility mapping in Jinbu, Korea. Catena 100:15–30

Liu C, Liu Y, Wen M, Li T, Lian J, Qin S (2009) Geo-hazard initiation and assessment in the Three Gorges Reservoir. In: Wang F, Li T (eds) Landslide disaster mitigation in Three Gorges Reservoir, China. Springer, pp 3–40

Luo S-T, Cheng B-W (2012) Diagnosing breast masses in digital mammography using feature selection and ensemble methods. J Med Syst 36:569–577

Malviya R, Umrao BK (2014) Comparison of NBTree and VFI machine learning algorithms for network intrusion detection using feature selection. Int J Comput Appl 108:35–38

Marsolo K, Twa M, Bullimore M, Parthasarathy S (2007) Spatial modeling and classification of corneal shape information technology in biomedicine. IEEE Trans on Inf Technol Biomed 11:203–212

Mathew J, Jha V, Rawat G (2009) Landslide susceptibility zonation mapping and its validation in part of Garhwal Lesser Himalaya, India, using binary logistic regression analysis and receiver operating characteristic curve method. Landslides 6:17–26

Mukane P (2014) India's worst landslides, and why these might not be the last ones. http://www.dnaindia.com/india/

Nanni L (2006) An ensemble of classifiers for the diagnosis of erythemato-squamous diseases. Neurocomputing 69:842–845

NCEP (2014) Global weather data for SWAT. http://globalweather.tamu.edu/home

Neuhäuser B, Terhorst B (2007) Landslide susceptibility assessment using "weights-of-evidence" applied to a study area at the Jurassic escarpment (SW-Germany). Geomorphology 86:12–24

Ohlmacher GC, Davis JC (2003) Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. Eng Geol 69:331–343

Onagh M, Kumra V, Rai PK (2012) Landslide susceptibility mapping in a part of Uttarkashi district (India) by multiple linear regression method. Int J Geol Earth Environ Sci 2:102–120

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, Tien Bui D, Pourghasemi HR, Indra P, Dholakia MB (2015) Landslide susceptibility assesssment 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 122:1–19. doi:10.1007/s00704-015-1702-9

Pham BT, Bui DT, Dholakia MB, Prakash I, Pham HV, Mehmood K, Le HQ (2016a) A novel ensemble classifier of rotation forest and Naïve Bayer for landslide susceptibility assessment at the Luc Yen district, Yen Bai Province (Viet Nam) using GIS. Geomat Nat Hazards Risk. doi:10.1080/19475705.2016.1255667

Pham BT, Bui DT, Prakash I, Dholakia M (2016b) Evaluation of predictive ability of support vector machines and naive Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS. J Geomat 10:71–79

Pham BT, Pradhan B, Tien Bui D, Prakash I, Dholakia MB (2016c) A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). Environ Model Softw 84:240–250

Pham BT, Tien Bui D, Dholakia MB, Prakash I, Pham HV (2016d) 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 34:1–18. doi:10.1007/s10706-016-9990-0

Pham BT, Tien Bui D, Prakash I, Dholakia MB (2016e) Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS. Nat Hazards 83:1–31. doi:10.1007/s11069-016-2304-2

Pham BT, Tien Bui D, Pham HV, Le HQ, Prakash I, Dholakia MB (2016f) Landslide hazard assessment using random subspace fuzzy rules based classifier ensemble and probability analysis of rainfall data: a case study at Mu Cang Chai District, Yen Bai Province (Viet Nam). J Indian Soc Remote Sens. doi:10.1007/s12524-016-0620-3

Pham BT, Tien Bui D, Prakash I, Dholakia MB (2017) Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. CATENA 149(1):52–63

Platt JC (1999) Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B, Burges C, Smola A (eds) Advances in kernel methods, chap 12. MIT press, pp 185–208

Poudyal CP, Chang C, Oh H-J, Lee S (2010) Landslide susceptibility maps comparing frequency ratio and artificial neural networks: a case study from the Nepal Himalaya. Environ Earth Sci 61:1049–1064

Pourghasemi HR, Jirandeh AG, Pradhan B, Xu C, Gokceoglu C (2013) Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province, Iran. J Earth Syst Sci 2:349–369

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

Pradhan B, Lee S (2010) Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models. Environ Earth Sci 60:1037–1054

Tien Bui D, Nguyen QP, Hoang N-D, Klempe H (2016a) A novel fuzzy k-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS. Landslides 14:1–17

Tien Bui D, Pham BT, Nguyen QP, Hoang N-D (2016b) Spatial prediction of rainfall-induced shallow landslides using hybrid integration approach of least-squares support vector machines and differential evolution optimization: a case study in Central Vietnam. Int J Digit Earth 9:1–21. doi:10.1080/17538947.2016.1169561

Tsangaratos P, Ilia I (2016) Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece. Landslides 13:305–320

Vapnik VN (2000) The nature of statistical learning theory, ser. Statistics for engineering and information science, vol 21. Springer, New York, pp 1003–1008

Vapnik VN, Vapnik V (1998) Statistical learning theory, vol 1. Wiley, New York

Varnes DJ (1984) Landslide hazard zonation: a review of principles and practice, vol 3. UNESCO, Paris

Yalcin A, Reis S, Aydinoglu A, Yomralioglu T (2011) A 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, Dai F (2008) Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China. Geomorphology 101:572–582

Yeon Y-K, Han J-G, Ryu KH (2010) Landslide susceptibility mapping in Injae, Korea, using a decision tree. Eng Geol 116:274–283. doi:10.1016/j.enggeo.2010.09.009

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

Yilmaz I (2010) Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. Environ Earth Sci 61:821–836

Yin Y, Wang H, Gao Y, Li X (2010) Real-time monitoring and early warning of landslides at relocated Wushan Town, the Three Gorges Reservoir, China. Landslides 7:339–349