

RESEARCH ARTICLE

Open Access

Support vector machine modeling of earthquake-induced landslides susceptibility in central part of Sichuan province, China

Suhua Zhou¹ and Ligang Fang^{2*}

Abstract

Background: Support vector machine (SVM) modeling is a machine-learning-based method. It involves a training phase with associated input and a predicting phase with target output decision values. In recent years, the method has become increasingly popular. The aim of this study is to carry out prediction of earthquake-induced landslides distribution in the area affected by the April 20 2013 Lushan earthquake based on GIS and the SVM model. The current study was undertaken to investigate the prevalence of Impaired Fasting Glucose (IFG)/Type 2 Diabetes (T2D) and its risk factors in the adult population in Biyem-Assi-Yaoundé, Cameroon.

Results: A detailed inventory map containing 1289 landslides triggered by this earthquake was produced through interpretation of colored aerial photographs and extensive field surveys. Elevation, slope angle, slope aspect, land cover, distance from co-seismic faults, peak ground acceleration and geology unit were selected as the controlling parameters. Cross validation with grid search method were used to search the best modeling parameters. A grid cell size of 60 × 60 m was adopted to produce the landslide susceptibility maps. The study area was divided into 186175 grid cells and each grid consisted of seven layers representing the controlling parameters. 70% of the total landslides (1782 grid cells) were used as positive training samples and 1782 randomly selected points on the stable slopes were treated as negative training samples in concert with four kernel functions: linear, polynomial, radial basis function and sigmoid. These results were further validated using area-under-curve (AUC) analysis of success-rate curves and prediction-rate curves. Comparative analyses of landslide-susceptibility and area relation curves show that both the polynomial and radial basis function suitably classified the input data of both training dataset and validating dataset, though the radial basis function was a bit more successful in success rate curves. Four cases of landslide susceptibility were mapped. The generated landslide-susceptibility maps were compared with known landslide. About 20%-30% of the study area 26 (Linear 34.78%, Polynomial 30.49%, and radial basic 23.83%) was categorized into high and very high susceptible zones during the Lushan earthquake, containing more than 70% occurrence of landslides triggered by the earthquake (Linear 74.16%, Polynomial 85.32%, and radial basic 86.71%). However, in maps with sigmoid function, 62.27% of the area was found to be highly susceptible to landslides during the earthquake with almost the entire landslides occurrence.

Conclusion: Most of the high susceptible and very high susceptible area was concentrated along the seismogenic faults with a PGA of more than 0.52 g. This paper provides an example for selecting appropriate types of kernel functions for prediction mapping of seismic landslides using support vector machine modeling. The susceptibility maps for earthquake-induced landslides can be useful in landslide hazard mitigation by helping planners understand the probability of landslides in different regions.

Keywords: Earthquake-induced landslide; Support vector machine; Susceptibility; Geographic information system

* Correspondence: fangligang_csu@163.com

²Department of Civil Engineering, Central South University, Changsha, Hunan, China

Full list of author information is available at the end of the article

Background

Landslide is one of the most severe natural hazards in the world, causing thousands of death and great property loss per year. Earthquake is a dominant trigger of landslides in mountainous and tectonic-active areas. Landslides induced by an earthquake are usually large in number, huge in scale and wide in distribution. Earthquake-induced landslides can bring great damages to property and infrastructures in developed areas, leading to economic losses and fatalities sometimes. For example, more than 20,000 people were killed by landslides induced by the 2008 Wenchuan earthquake with a magnitude of Ms 8.0 and 34 large barrier lakes were produced, which threatened the residents who lived downstream of these dams [Yin et al. 2009]. In the 2010 Yushu earthquake (Ms 7.1) about 60 million in damages and 8 deaths were directly caused by earthquake-induced landslides [YP Yin et al., 2010].

Earthquake-induced landslides were hard to predict, but could be evaluated. Identifying a region's susceptibility to landslides during an earthquake was an effective and most economical way to provide planners with foreknowledge of dangerous regions thereby helping with land management and infrastructure planning. For earthquake-induced landslide, landslide susceptibility assessment was to evaluate location of landslide susceptibility zones where landslides could be induced in future earthquake shaking. Many different methods and techniques for assessing landslide susceptibility have been proposed and tested. These have already been systematically compared and their advantages and limitations outlined in Carrara et al. 1999, Huabin et al. 2005 and van Westen et al. 2008. Both deterministic and statistical methods have been used in earthquake-induced landslide susceptibility. For deterministic methods, assessment of earthquake-induced landslide susceptibility on a regional scale commonly required the employment of an analytical slope-stability method and the infinite-slope model [Jibson and Keefer, 1993; Jibson et al., 2000; Refice and Capolongo, 2002]. The deterministic method required calculation to determine the limit-equilibrium of the slope stability given the strength parameters of mass, failure depth, and groundwater conditions for every calculation point in the study area. This requirement caused immense problems in terms of data acquisition and control of spatial variability of the variables ([Carrara et al., 1999]. For statistical method, it was most common to use a statistical approach where landslide inventories and causative factors are utilized to build a susceptibility model for the prediction of future landslides. For instance, Kamp et al. 2008 carried out spatial prediction of landslides related to 2005 Kashmir earthquake-induced by use of a multi-criterion method. Lee et al. 2008 applied multivariate statistical methods in a study of shallow earthquake-induced landslides in central western Taiwan.

The results showed that landslide distribution can be predicted. Landslides induced by Wenchuan earthquake were assessed and predicted by Su et al. 2010 using logistic regression models, and were compared with the bivariate statistics, artificial neural networks, and support vector machine models by Xu et al. 2012a.

Among these approaches, the support vector machine (SVM) model has become increasingly popular. The SVM was originally developed by [Vapnik, 1995] as a new machine learning algorithm for pattern classification and non-linear regression. The main procedure involved in SVM modeling is a training phase with associated input and target output values. Recently, several authors have applied the SVM model successfully on landslide susceptibility mapping. [Gallus et al., 2008] compared several classification approaches of SVM, Gaussian process, and LR modeling, with SVM having the best results. [Xu et al., 2012b] examined the use of SVM model for landslide susceptibility mapping in an earthquake zone with combination of 4 kernel functions and 3 different training sets and found that radial-basis and polynomial kernel functions were suitable for modeling with any input training data. [Xu et al., 2012b] applied 6 different models in susceptibility mapping of landslides induced by the 2008 Wenchuan earthquake with SVM having a second best results outranked only by logistic regression. [Kavzoglu et al., 2013] also made a comparison of susceptibility results from multi-criteria decision analysis, SVM, and logistic regression and showed that multi-criteria decision analysis and SVM methods were better than logistic regression in shallow landslides susceptibility mapping. These applications proved that when used properly, SVM model in landslide susceptibility mapping might produce a good result. Two outstanding advantages of the SVM are: (a) Based on the principle of minimization structural risk; (b) Guarantee its performance by solving constrained quadratic form. Theoretically, it can achieve the optimal prediction result by using the SVM model. Its detailed mathematical formulas are introduced in [Vapnik and Cortes 1995].

In this study, we propose the application of SVM model to produce a landslide susceptibility map of the area hit by the April 20, 2013 Lushan earthquake on the ArcGIS platform. The goal of the study is to produce a relatively accurate landslide susceptibility map with optimal kernel functions. The 4 resultant cases are compared using AUC (area under curve) analysis to verify the susceptibility mapping results. This is done by comparisons with known landslide locations to establish the model's success rate, and its predictive accuracy.

Study area

On April, 2013, an Ms 7.0 earthquake, with a maximum source intensity of up to 9.0 on Chinese seismic scale,

struck Lushan county, Sichuan province, west China. The epicenter of the main shock was located in 30.3°N, 103.0°E, about 100 km southwest of Chengdu (Figure 1). The earthquake occurred on the southern segment of the Longmenshan fault zone. This area was celebrated for steep mountain landscapes and heavy tectonics. The April 20 earthquake was an strong aftershock of 2008 Wenchuan earthquake and were the most devastative earthquake in China since the 2008 Wenchuan earthquake [Wei-Min et al., 2013].

The study area had experienced serious shallow landslides during this earthquake, since the steep slopes and jagged ridges were susceptible to landslide while suffering heavy ground shaking. Topography of the study area ranges from river valley to mountainous. Elevation of the study area ranges from 596 m to 2872 m. Land use includes mainly cropland distributed on the ridges, slope wasteland on side-slopes and gullies and town in flat river valleys. Due to long-term human activity, many parts of the natural vegetation have been destroyed. Because it was right time for vegetation, earthquake-induced landslides were easy to be recognized according to landslide scars on aerial photos. Landslides triggered by the Lushan earthquake can be mainly classified into following types (Figure 2): (1) shallow-disrupted slope failures; (2) rock avalanches; and (3) deep-seated rocky or soil slides. Shallow-disrupted slope failures with a

composition of weathered and fractured superficial soils and rocks were widespread throughout the whole study area, because they could be triggered easily by weak shaking. Rock avalanches are usually originated on a high place along the river and road banks, with large potential energy, resulting in a high speed and a long run-out distance during sliding process. Deep-seated landslides are mainly distributed within a short distance from the main co-seismic ruptures, since only a strong ground shaking could trigger them.

This area was very tectonic-active with many folds and active faults trending NW–SE (Figure 1). The bedrock exposure in the area was dominated by Mesozoic volcanic rocks and Mesozoic group. The volcanic rocks, which comprised tuffs and lavas with intercalated sedimentary rocks. Intrusive rocks consisted mainly of granites, sandstone and dykes of various compositions. As a result of the abundant supply of rainfall and the local rich groundwater, almost all rocks in the study area had undergone a certain degree of weathering. In many slopes, weathering had penetrated deep into rock masses through joints and bedding planes (Figure 2).

Methods

Support vector machine (SVM) modeling

Support vector machine (SVM), as the representative’s kernel-based techniques, is a major development in

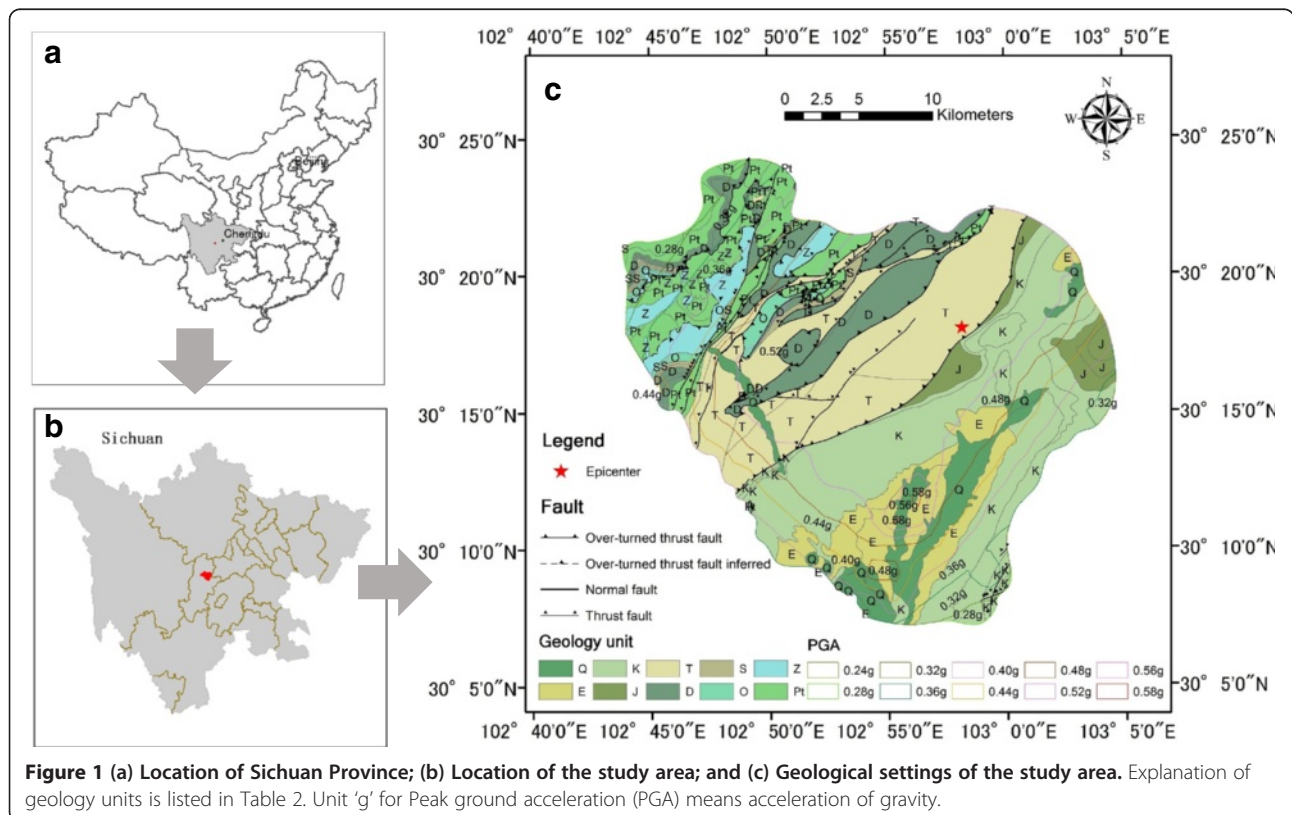


Figure 1 (a) Location of Sichuan Province; (b) Location of the study area; and (c) Geological settings of the study area. Explanation of geology units is listed in Table 2. Unit 'g' for Peak ground acceleration (PGA) means acceleration of gravity.



Figure 2 Typical types of landslides. (a) Rock avalanches, (b) translational slides and (c) shallow-disrupted slope failures.

machine learning algorithms. SVM is a group of supervised learning methods based on the statistical learning theory and the Vapnik-Chervonenkis (VC) dimension introduced by [V Vapnik and Cortes, 1995] and [Chervonenkis, 2013] that can be applied to pattern classification or non-linear regression.

For the linear separable condition, consider a set of training vectors with two classes as follows:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \quad (1)$$

where $x_i \in X \subset R^m$, $y_i \in \{1, -1\}$, $i = 1, 2, \dots, n$, that can be separated the two classes $[1, -1]$ by a hyper-plane (Figure 3):

$$(w \cdot x) + b = 0, w \in R^N, b \in R \quad (2)$$

where w is the normal of the hyper-plane, b is a scalar base, and (\cdot) denotes the scalar product operation.

After normalization, the geometrical margin between the two groups can be expressed as $\frac{2}{\|w\|}$. The operation of the SVM algorithm is to find the hyper-plane that gives the largest geometrical margin to the training examples. The maximum $\frac{2}{\|w\|}$ can be expressed as:

$$\underset{w, b}{\text{Minimize}} \frac{1}{2} \|w\|^2 \quad (3)$$

Subjecting to constrains:

$$y_i(w^T x_i + b) \geq 1 \quad i = 1, 2, \dots, n \quad (4)$$

Introducing the Lagrangian multiplier, the cost function can be defined as:

$$\phi(w, b; \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i (y_i [w \cdot x_i + b] - 1) \quad (5)$$

where $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)^T \in R_+^n$ is the Lagrangian multiplier, and the problem can be solved by dual minimization of Equation (5) with respect to w and b through standard procedures Equation (6). More detail of SVM was discussed in [Vapnik 1995].

$$\begin{cases} \nabla_b \phi(w, b; \alpha) = 0 \\ \nabla_w \phi(w, b; \alpha) = 0 \end{cases} \quad (6)$$

Mostly, however, the training vectors are non-separable, [Vapnik 1995] introduced an slack variables ξ_i modified the constraints as follows:

$$y_i((w \cdot x_i) + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, n, \quad \xi_i \geq 0 \quad (7)$$

To avoid a high value of ξ_i , some kind of penalty term C was introduced into the original optimization Equation (3), which can be modified as:

$$\underset{w, b}{\text{Minimize}} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad (8)$$

where $C > 0$ is the penalty factor to control the trade-off between the maximum margin and the minimum error. Additionally, a kernel function $k(x_i, x_j)$ is introduced by [Vapnik 1995] to transform the originally non-linear data pattern to a linear one in higher dimensional feature space (Figure 3).

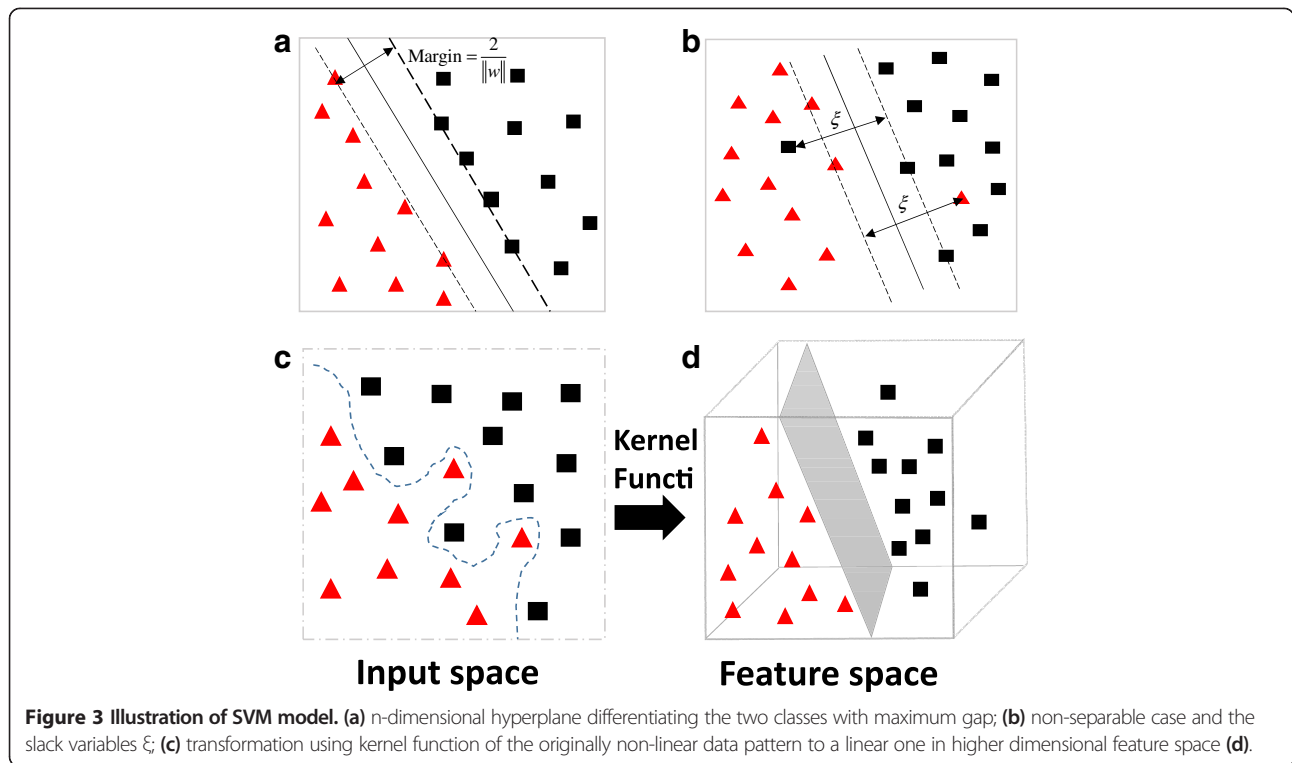


Figure 3 Illustration of SVM model. (a) n-dimensional hyperplane differentiating the two classes with maximum gap; (b) non-separable case and the slack variables ξ ; (c) transformation using kernel function of the originally non-linear data pattern to a linear one in higher dimensional feature space (d).

Selection of the kernel function is the main issue in SVM modelling. Theoretically, any function that satisfy the Mercer criteria can be used as kernel function, however, some of them work well in a wide variety of applications. The mathematical representation of some kernel functions are listed below:

$$\text{Linear} : K(x_i, x_j) = x_i^T x_j \tag{9}$$

$$\text{Polynomial} : K(x_i, x_j) = (\gamma x_i^T x_j + r)^p, \gamma > 0 \tag{10}$$

$$\text{Radial basis function} : K(x_i, x_j) = e^{(-\gamma \|x_i - x_j\|^2)}, \gamma > 0 \tag{11}$$

$$\text{Sigmoid} : K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \tag{12}$$

where γ is the gamma term in all the kernel function except linear; p is the polynomial order term in the kernel function for the polynomial kernel; r is the bias term in the kernel function for the polynomial and sigmoid kernels. Proper parameters, such as the order of polynomials and width of radial basis function, play a key role in governing the accuracy of the SVM modeling. Of these functions, polynomial and radial basis function (RBF) are the most-used kernels and are utilized in our research due to its good generalizing properties.

In reality, the unstable slope cases (with landslides) are recognized as positive pattern, while stable slope cases (without landslides) are recognized as negative pattern. Note that we often commonly have only a one-class dataset without negative data. One-class SVM models

also have been developed, but their theories are not reach perfection and they produce poor prediction efficiency than two-class SVM [Guo et al., 2005; Yao et al., 2008]. Hence, a two-class SVM modeling is utilized in this study.

To carry out the two-class SVM modeling, we established a spatial database containing all the landslides triggered by the earthquake and their controlling parameters. Then all the data layers were classified and rasterized in Arcgis and coded in Matlab7.01. The landslides as well as the same amount of selected stable slopes were randomly divided into two groups for training and validation purpose, respectively. We use the training dataset as input to train the SVM model, then the testing dataset were used to examine the model. Both the training and validation phase were completed in Matlab 7.01. Finally, all the cells in the study area were input into the established model for possibility prediction of landslide occurrence.

Data

Two kinds of data were indispensable in the two-class SVM modeling: (1) Units with landslides and units supposed to be considered as stable and conditions of these units and (2) Conditions of units that needed to be predicted. The former were samples used to train the two-class SVM model, while the latter were used as the input of the trained model to predict risk of region including them. All of the data representing a categorical attribute should be converted into numeric code before entering the SVM model.

Landslide inventory

Institute of Remote Sensing and Digital Earth (RADI) of the Chinese Academy of Sciences (CAS) took airborne images with a high-resolution of 0.6 m covering the earthquake-affected area on the morning right after the earthquake. Except a little part was masked by cloud, most of earthquake damage in this area was shown clearly on these image. All of these high-resolution images as well as a preliminary interpretation of earthquake-induced geo-hazards were proposed on Geo-Information Platform of Lushan Earthquake [Institute of Mountain Hazard and Environment, C. A. o. S., and Geomatics Center of Sichuan Province 2013] based on Tianditu online map service [Chen et al., 2013]. Due to a critical use for rescue after earthquake, this preliminary landslide inventory was incomplete, only location of suspected landslides were available.

The accurate detection of landslides is vital for landslide susceptibility analysis, so an inventory of landslides triggered by April 20 Earthquake was made with the help of Arcgis server. Firstly, the high resolution images provided by Tianditu map service were invoked into Arcgis through Arcgis server. These images were specified as the based map. Then, an empty vector layer with the same coordinate system as the base map was created for the storage of landslides. After that, experts in earthquakes and geo-hazards were called upon to visually interpret the base map according to their experiences, knowledge as well as previously identified landslide points. High-resolution pre-event satellite images of RADARSAT-2 and SPOT-4 (see Table 1) of the study area were geometrically rectified and matched to be taken into consideration as a contrast. The boundaries of landslides were interpreted on the base map and transformed into vector format and stored in ArcGIS system. A filed survey was finally conducted to check the accuracy of the interpretation, following which the interpreted images were modified. The resultant landslide inventory map is shown in Figure 4. Landslide–area ratio(LAR), defined as the percentage of the area affected by landslide activity, and landslide number density (LND) gives the number of landslides per square kilometer. In this study area, $LAR = (4.26 \text{ km}^2 / 674.45 \text{ km}^2) \times 100\% = 0.63\%$ and $LND = 1289 \text{ landslides} / 674.45 \text{ km}^2 = 1.91 \text{ km}^{-2}$.

Table 1 Images used in ArcGIS for interpretation

No.	Type	Date	Mode/resolution
1 ¹	RADARSAT-2	2012-03-04	WIDE/30 m
2 ¹	SPOT-4	2011-04-09	PAN/6.25 m
3 ¹	SPOT-4	2011-04-09	MS/12.5 m
4	Airborne images	2013-04-20	0.6 m

¹Data source can be found on <http://www.radi.ac.cn/yaan/yaanphoto/>.

Controlling parameters of landslides

Seven environmental variables were used to train the model and to predict the potential distribution for landslides (Figure 5 and Table 2). These variables included: (1) slope gradient, (2) slope aspect, (3) land cover, (4) distance to fault, (5) peak ground acceleration (PGA) distribution of April 20 Lushan earthquake, (6) elevation, (7) geology unit. The above selections were made based on the authors’ knowledge of the physical environment and landslides in the study area. Slope gradient, slope aspect, elevation were derived from digital elevation model (DEM). Land cover layer was derived from 1:20,000 scale digital vegetation cover maps. For ease of analysis, the 1:20,000 scale superficial and solid geology map covering the study area was divided into 10 groups based on chronostratigraphic unit. Other environmental variables were also divided into seven or eight classes manually, slope gradient was first broken at 15° because very fewer landslides were probable on shallower slopes. The PGA map is extracted from the United States Geology Survey (U.S.G.S) Shakemap (<http://earthquake.usgs.gov/earthquakes/shakemap>) (see Table 2).

Results

There are many free programs for SVM modelling [Chang and Lin, 2011; Joachims, 1999], which can be downloaded from the internet, providing all kinds of interfaces to other software. In this study, LibSVM [Chang and Lin, 2011] was employed to finish the computation of the SVM model on Matlab 7.01. The environmental parameters were derived and rasterized in ArcGIS 9.3. A grid cell size of 60 × 60 m was adopted to produce the landslide susceptibility maps. The study area was divided into 186,230 grid cells and each grid consisted of seven layers representing the environmental parameters.

Training and validation dataset

The two-class SVM requires both positive and negative data to train the model. The landslide inventory were randomly divided into two groups: 70% of the total (902 landslides with 1782 grid cells) were used as positive training samples. As mentioned before, negative training data was also needed. 1782 negative training points were generated within 120 m interval in both north and south direction of the positive points. A validation dataset contains 30% of the total landslides (387 landslides with 738 grid cells) and 738 negative points generated using the same way as negative training data. A total of 2520 landslide points were assigned the value of 1, while the same amount of negative points were assigned the value of 0.

In SVM modelling, the input of controlling factors should be as a vector of real numbers. For categorical attributes, a simple 1 of k coding is recommended to represent a k-category attribute. For instance, suppose a

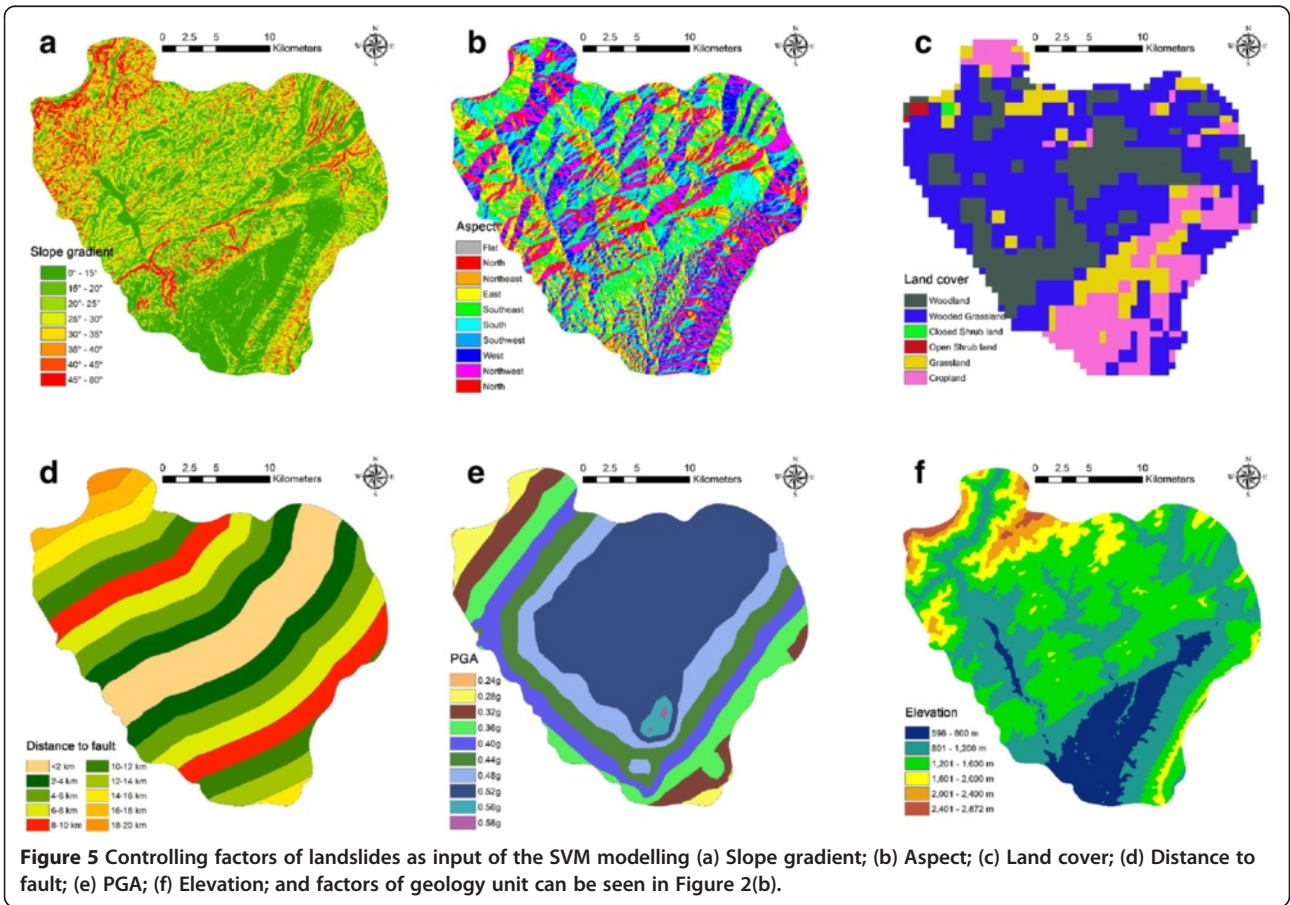
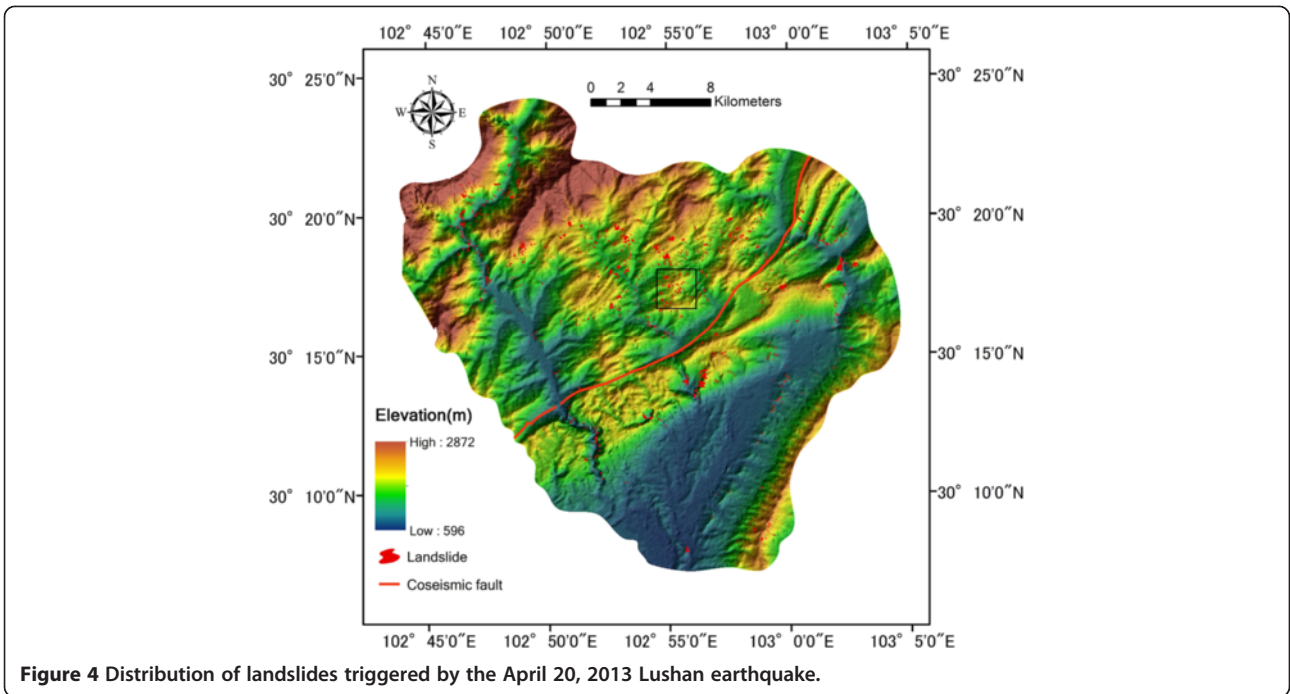


Table 2 Controlling parameters and their classes for this study

Controlling parameters	Classes
Elevation (m)	(1)(min)596-800;(2)800-1200(3)1200-1600;(4)1600-2000;(5)2000-2400;(6)24002872(max)
Slope gradient (°)	(1)< 15;(2)15-20;(3)20-25;(4)25-30;(5)30-35;(6)35-40;(7)40-45;(8)> 45
Aspect	(1)F;(2)N;(3)NE;(4)E;(5)SE;(6)S;(7)SW;(8)W;(9)NW;
Distance to co-seismic fault (km)	(1)< 2;(2)2-4;(3)4-6;(4)6-8;(5)8-10;(6)10-12;(7)12-14;(8)14-16;(9)16-18;(10)18-20
PGA	(1)0.24;(2)0.28;(3)0.32;(4)0.36;(5)0.40;(6)0.44;(7)0.48;(8)0.52;(9)0.56;(10)0.58
Land cover	(1)Woodland;(2)Wooded Grassland;(3)Closed Shrub land;(4)Open Shrub land;(5)Grassland;(6)Cropland
Geology unit	(1)Quaternary(2) Paleogene(3)Cretaceous(4)Jurassic(5)Triassic(6)Devonian(7)Silurian (8)Ordovician(9)Sinian(10)Proterozoic

1-dimensional a three-category attribute taking value {a, b, c}, Just turn it into 3-dimensional numbers such that a = (1,0,0), b = (0,1,0), c = (0,0,1). If the number of values in an attribute is not too large, such coding is more stable than using a single number to represent a categorical attribute ([Hsu et al., 2003]). Therefore, the seven environmental parameters were converted into a vector with 59 bits. Finally, a training dataset containing 3564 grids with 7 input variables were built through extracting the value of landslide conditioning factors in every grid.

Cross validation and grid search for SVM parameter optimization

The performance of the SVM model is depended on the choice of kernel functions and their parameters especially the penalty factor C and γ terms. In this study, a grid-search method with 5-folder cross-validation was used to locate the optimal values of C and γ [Hsu et al., 2003] as

follows: (1) Set a pair of (C, γ) values for SVM model; (2) Randomly divided the training dataset into 5 equal sized subsets; (3) Use Four subsets of them to train the SVM model; (4) Validate the trained model using the one remaining subset; (5) Repeat step three and four for five times for each of the subset; (6) Calculate the overall accuracy defined as the percentage of data which are correctly predicted.

Pairs of (C, γ) were generated through a grid search with $C = 2^{-8}, 2^{-7}, 2^{-6} \dots 2^6, 2^7, 2^8$ and $\gamma = 2^{-8}, 2^{-7}, 2^{-6} \dots 2^6, 2^7, 2^8$. For every pair of (C, γ), we can get an overall accuracy and the optimal C and γ corresponded to the highest overall accuracy.

The best value of C for linear was 2 with the overall accuracy 85.5%. The best C and γ for polynomial were found 4 and 1 with the overall accuracy 89.6%. In the case of RBF, the best C and γ were 16 and 1 respectively, with the overall accuracy 92% while sigmoid used 16 and 8 as the best C and γ .

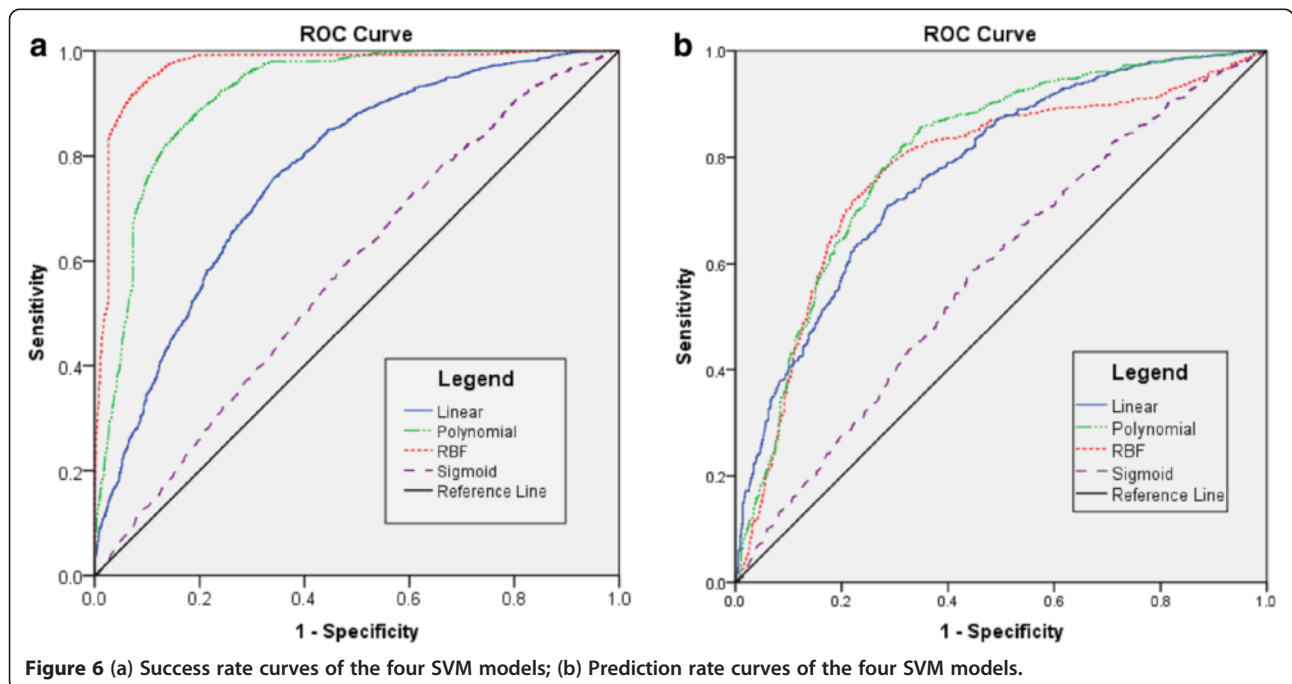


Figure 6 (a) Success rate curves of the four SVM models; (b) Prediction rate curves of the four SVM models.

Comparison of landslide susceptibility maps

ROC curve is one of the useful methods for representing the quality of deterministic and probabilistic detection, especially for landslide susceptibility assessment. The characterizes the quality of a forecast system by describing the system's ability to anticipate correctly the occurrence or non-occurrence of predefined event (Yesilnacar and Topal 2005). A true positive (TP) means prediction of a landslide for a point where a landslide does occur, while a false positive (FP) is a prediction of a landslide for a stable point. On the conversely, we can get the true negative (TN) and false negative (FN). A ROC space is defined by the false positive rate (FPR, defines as $FP/(FP + TN)$)

and true positive rate (TPR, defined as $TP/(TP + FN)$) as x and y axes respectively. The true-positive rate is also known as sensitivity in biomedicine, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as $1 - \text{specificity}$.

The area under the ROC curve (AUC) is an important measure of the accuracy of the binary classification. AUC values are typically between 0.5 and 1.0. If this area is equal to 1.0 then the roc curve consists of two straight lines, one vertical from (0, 0) to (0, 1) and the next horizontal from (0, 1) to (1, 1) this test is 100% accurate because both the sensitivity and specificity are 1.0 and there was no false positives and no false negatives. On

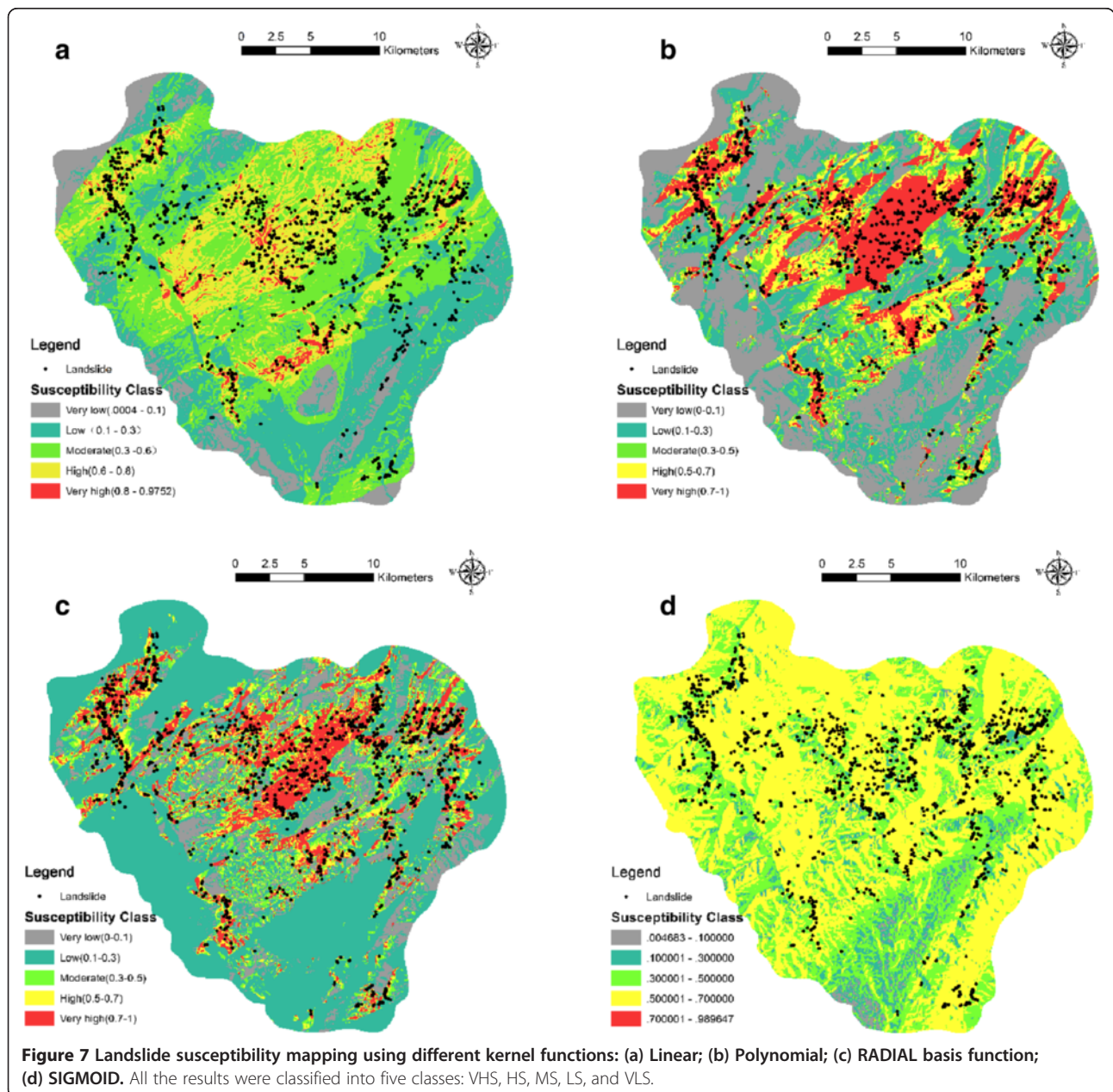


Figure 7 Landslide susceptibility mapping using different kernel functions: (a) Linear; (b) Polynomial; (c) RADIAL basis function; (d) SIGMOID. All the results were classified into five classes: VHS, HS, MS, LS, and VLS.

the other hand a test that can't discriminate between positive and negative corresponds to an ROC curve that is the diagonal line from (0, 0) to (1, 0). The AUC for this line is 0.5.

To evaluate the four landslide susceptibility maps, success-rate curves and prediction-rate curves were established, and values of area under curves (AUC) were also calculated [Hasegawa et al., 2009]. Higher AUC value indicated a higher capacity of correctly classifying the data with existing landslides. The success-rate curve was a measure of goodness of fit for SVM model and training data. The curve was obtained by comparing the four landslide susceptibility maps with the training dataset, (Figure 6a). Results indicated that RBF and polynomial had the highest AUC values 0.97 and 0.91 respectively, followed by linear (0.77), while model using sigmoid kernel function had the lowest AUC values of 0.58.

Nevertheless, the success-rate was not a suitable measure for the prediction capability of the landslide models because it was based on the landslide pixels that had already been used for building the model. To overcome this, prediction-rate curve and corresponding AUC values were obtained by comparing the four susceptibility maps with the validation dataset (Figure 6b). The results showed that model using polynomial kernel functions had the highest capacity of prediction with the AUC of 0.86, slightly better than RBF (0.82) and Linear (0.78). Same with success-rate curve, sigmoid had the lowest AUC values.

Discussion

Once the landslide susceptibility models were successfully trained in the training phase, they were then used to calculate the landslide susceptibility indexes (LSI) for all the pixels. The SVM classification output or result

was the decision values of each pixel. The results were then converted into raster data. Figure 7 showed the mapping results for the landslide susceptibility index (LSI) ranging from 0 to 1. 0 indicates no chance and 1 indicates 100% chance for occurrence of landslides.

The LSI values of each grid cell predicted using SVM with the linear, polynomial, radial basis, and sigmoid kernel functions were 0.0004-0.9752, 0.0001-0.9999, 0.0007-0.9948, and 0.0047-0.9896 respectively.

A few classification methods, such as natural breaks, equal intervals and defined interval, were used to distinguish the susceptibility classes for trial. Equal intervals classification was found not to be useful for its emphasis on the amount of one class value relative to other classes. Natural breaks are identified that best group similar values and that maximize the differences between classes and not useful for comparing multiple maps built from different underlying information. A series of specified interval sizes can be used to define the classes with different ranges in defined interval methods based on a comprehensive consideration of the data distribution. Moreover, the define interval classification allow comparison of different maps with similar ranges of attribute value.

The maps with continuous LSI values were then reclassified into five landslide susceptibility categories using the method of define intervals, i.e. very low susceptibility (VLS: less than 0.1), low susceptibility (LS: 0.1-0.3), moderate susceptibility (MS: 0.3-0.5), high susceptibility (HS: 0.5-0.7), and very high susceptibility (VHS: more than 0.7) (Figure 7).

The resultant landslide susceptibility maps were also compared with the landslide inventory. The coverage percentages of 5 susceptibility classes and the corresponding landslide occurrence are shown in Table 3. The results showed that the landslide frequency ratio (defined as the

Table 3 Landslide statistical results by different SVM kernel functions

Models		Susceptibility class					Success rate	Prediction rate
		VLS	LS	MS	HS	VHS		
Linear	%area	8.39	29.83	27.00	23.87	10.91	0.77	0.78
	%landslide	0.44	6.67	18.73	38.41	35.75		
	LND	0.05	0.22	0.69	1.61	3.28		
Polynomial	%area	28.01	27.95	13.56	12.66	17.83	0.91	0.86
	%landslide	1.27	4.52	8.89	16.71	68.61		
	LND	0.05	0.16	0.66	1.32	3.85		
Radial basic	%area	11.16	54.29	10.09	9.01	14.82	0.97	0.82
	%landslide	1.83	5.79	5.67	11.63	75.08		
	LND	0.16	0.11	0.56	1.29	5.07		
Sigmoid	%area	1.12	7.03	29.56	62.27	0.01	0.58	0.58
	%landslide	0.12	0.71	2.82	96.27	0.08		
	LND	0.11	0.10	0.10	1.55	7.78		

ratio of percentage of landslide occurrence in each class on that of area) gradually increased from the very low to the high susceptibility class and then suddenly jumped in very high susceptibility class.

According to maps, about 20%-30% of the study area (Linear 34.78%, Polynomial 30.49%, and RBF 23.83%) were categorized into high and very high susceptible zones during the Lushan earthquake, with 70%-80% occurrence of landslides triggered by the earthquake (Linear 74.16%, Polynomial 85.12%, and RBF 86.71%). However, in maps with sigmoid function, 62.27% of the area were found to be highly susceptible to landslides during the earthquake with almost all of the landslides occurrence.

Most of areas that classified as very high, high and moderate were concentrated along the seismic faults, suffering a high PGA of more than 0.52 g. This may be because earthquake is the trigger of the landslides used for training model to produce the landslide susceptibility map.

Conclusion

Based on the statistical learning theory, GIS technology, SVM model, and four types of kernel functions, including linear function, polynomial function, RBF function, and sigmoid function, this work has studied the prediction for spatial distribution of landslides triggered by the April 20, 2013 Lushan earthquake in Sichuan province of China. From the results of this study, the following conclusions can be drawn:

- (1) Cross validation and grid search was an efficient tool for parameters optimization. This method avoided the subjectivity in parameter selection for the SVM model.
- (2) The validation results by ROC method showed that RBF and polynomial function is the better than linear and sigmoid for the Lushan earthquake area. AUC of RBF shows a high accuracy of 97% (0.97) in case of success rate curves and 82% (0.82) in case of prediction rate curves, and that of polynomial are 91% (0.91) and 86% (0.86) respectively.
- (3) According to the landslide susceptibility index of each grid, the study area was divided into 5 classes of landslide susceptibility, namely very low, low, moderate, high and very high and 4 landslide susceptibility maps were generated Comparing with all 1289 landslides (2520 grid cells), The results show that the landslide frequency ratio gradually increases from the no to the high susceptibility class.
- (4) Most of landslide triggered by the earthquake happened in high and very high susceptible zones, which were concentrated along the seismic faults with a high PGA;
- (5) The SVM modelling related to the Lushan earthquake landslides can be applied to landslide

disaster prediction in other regions with potential seismic risks given appropriate kernel functions and model parameters

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SZ carried out the susceptibility modeling and drafted the manuscript. LF arranged the structure of the manuscript and participated in the discussion and conclusion of the study. Both authors read and approved the final manuscript.

Acknowledgement

This research is supported by State Key Development Program of Basic Research of China (Grant 2011CB710601) The data used in this paper was provided by the Department of Geotechnical Engineering, Central South University, China. We wish to express our sincere appreciation for the generous support.

Author details

¹Department of Civil and Structural Engineering, Kyushu University, Fukuoka, Japan. ²Department of Civil Engineering, Central South University, Changsha, Hunan, China.

Received: 7 August 2014 Accepted: 14 October 2014

Published online: 06 February 2015

References

- Carrara A, Guzzetti F, Cardinali M, Reichenbach P (1999) Use of GIS technology in the prediction and monitoring of landslide hazard. *Nat Hazards* 20(2-3):117-135
- Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27
- Chen Y-W, Yap K-H, Lee JY (2013) Tianditu: China's first official online mapping service. *Media. Culture & Society* 35(2):234-249
- Chervonenkis AY (2013) Early History of Support Vector Machines. *Festschrift in Honor of Vladimir N. Vapnik, Empirical Inference*, pp 13-20
- Gallus D, Abecker A, Richter D (2008) Classification of landslide susceptibility in the development of early warning systems. In: *Symposium on Headway in Spatial Data Handling*. Springer: Montpellier, France pp 55-75
- Guo Q, Kelly M, Graham CH (2005) Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecol Model* 182(1):75-90
- Hasegawa S, Dahal RK, Nishimura T, Nonomura A, Yamanaka M (2009) DEM-based analysis of earthquake-induced shallow landslide susceptibility. *Geotech Geol Eng* 27(3):419-430
- Hsu C-W, Chang C-C, Lin C-J (2003) A Practical Guide to Support Vector Classification. Technical report, Department of Computer Science, National Taiwan University.
- Huabin W, Gangjun L, Weiya X, Gonghui W (2005) GIS-based landslide hazard assessment: an overview. *Prog Phys Geogr* 29(4):548-567
- Institute of Mountain Hazard and Environment, C. A. o. S., and Geomatics Center of Sichuan Province (2013) Geo-Information Platform of Lushan Earthquake. <http://scgis.net/LSXEarthquake/>
- Jibson RW, Keefer DK (1993) Analysis of the seismic origin of landslides: examples from the New Madrid seismic zone. *Geol Soc Am Bull* 105(4):521-536
- Jibson RW, Harp EL, Michael JA (2000) A method for producing digital probabilistic seismic landslide hazard maps. *Eng Geol* 58(3):271-289
- Joachims T (1999) *SvmLight: Support Vector Machine, SVM-Light Support Vector Machine* <http://svmlight.joachims.org/>. University of Dortmund, 19(4)
- Kamp U, Growley BJ, Khattak GA, Owen LA (2008) GIS-based landslide susceptibility mapping for the 2005 Kashmir earthquake region. *Geomorphology* 101(4):631-642
- Kavzoglu T, Sahin E, Colkesen I (2014) Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides* 11(3):425-439
- Lee C-T, Huang C-C, Lee J-F, Pan K-L, Lin M-L, Dong J-J (2008) Statistical approach to earthquake-induced landslide susceptibility. *Eng Geol* 100(1):43-58
- Refice A, Capolongo D (2002) Probabilistic modeling of uncertainties in earthquake-induced landslide hazard assessment. *Comput Geosci* 28(6):735-749
- Su F, Cui P, Zhang J, Xiang L (2010) Susceptibility assessment of landslides caused by the wenchuan earthquake using a logistic regression model. *J Mt Sci* 7(3):234-245

- van Westen CJ, Castellanos E, Kuriakose SL (2008) Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview. *Eng Geol* 102(3):112–131
- Vapnik VN (1995) The nature of statistical learning theory. Springer-Verlag New York, Inc, p 188
- Vapnik V, Cortes C (1995) Support-vector networks. *Mach Learn* 20(3):273–297
- Wei-Min W, Jin-Lai H, Zhen-Xing Y (2013) Preliminary result for rupture process of Apr. 20, 2013, Lushan Earthquake, Sichuan, China. *CHINESE JOURNAL OF GEOPHYSICS-CHINESE EDITION* 56(4):1412–1417
- Xu C, Xu X, Dai F, Saraf AK (2012a) Comparison of different models for susceptibility mapping of earthquake triggered landslides related with the 2008 Wenchuan earthquake in China. *Comput Geosci* 46:317–329
- Xu C, Dai F, Xu X, Lee YH (2012b) GIS-based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China. *Geomorphology* 145:70–80
- 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(4):572–582
- Yesilnacar E, Topal T, (2005) Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). *Engineering Geology*, 79(3-4):251–266.
- Yin Y, Wang F, Sun P (2009) Landslide hazards triggered by the 2008 Wenchuan earthquake, Sichuan, China. *Landslides* 6(2):139–152
- Yin Y, Zhang Y, Ma Y, Hu D, Zhang Z (2010) Research on major characteristics of geohazards induced by the Yushu Ms7. 1 earthquake. *J Eng Geol* 18(3):289–296

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Immediate publication on acceptance
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ springeropen.com
