A novel hybrid integration model using support vector machines and random subspace for weather-triggered landslide susceptibility assessment in the Wuning area (China)

Haoyuan Hong[1,2,3], Junzhi Liu[1,2,3], A-Xing Zhu[1,2,3], Himan Shahabi[4], Binh Thai Pham[5*], Wei Chen[6*], Biswajeet Pradhan[e7,8], , Dieu Tien Bui[9]

[1]Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing, China

[2]State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing, China

[3]Jiangsu Center for Collaborative Innovation in Geographic Information Resource Development and Application, Nanjing, China

[4]Department of Geomorphology, Faculty of Natural Resources University of Kurdistan Sanandaj, Iran

[5]Department of Geotechnical Engineering, University of Transport Technology, Hanoi, Vietnam

[6]College of Geology and Environment, Xi'an University of Science and Technology, Xi'an, China

[7]Department of Civil Engineering, Geospatial Information Science Research Center (GISRC), Faculty of Engineering, University Putra Malaysia, Serdang, Malaysia

[8]School of Systems, Management and Leadership, Faculty of Engineering and Information Technology, University of Technology, Sydney, Ultimo, Australia

[9]Geographic Information System group, Department of Business and IT, University College of Southeast Norway, Bøi, Telemark, Norway

*Corresponding author: azhu@wisc.edu (A-Xing Zhu)

Wei Chen

chenwei.0930@163.com; chenwei0930@xust.edu.cn

Abstract This study proposed a hybrid modelling approach using two methods: support vector machines and random subspace to create a novel model namely Random SubSpace based Support Vector Machines (RSSVM) for landslide susceptibility assessment. The newly developed model was then tested at Wuning area, China to produce landslide susceptibility map. To achieve the objective of the study, first a spatial dataset was constructed that includes a landslide inventory map consisting of 445 landslide locations; and various landslide conditioning factors, such as slope, aspect, altitude, topographic wetness index (TWI), stream power index (SPI), sediment transport index (STI), soil, lithology, normalized difference vegetation index (NDVI), land use, rainfall, distance to roads, distance to rivers, distance to faults. Next, the result of RSSVM model was

validated using statistical index-based evaluations and receiver operating characteristic curve method. Next, to assess the performance of the proposed RSSVM model, a comparison analysis was performed with other existing methods such as ANN (Artificial Neural Network), NB (Naïve Bayes) and SVM (Support Vector Machine). In general, the performance of the RSSVM model showed higher than the rest of the models for spatial prediction of landslides. The AUC results of the applied models are: RSSVM (AUC=0.857), followed by MLP (AUC=0.823), SVM model (AUC=0.814) and NB model (AUC=0.783). The present study indicates that RSSVM can be used for spatial prediction of landslides and the result is very useful for the lcoal government and people lived in the Wuning area.

Keywords: Landslides; GIS; Support Vector Machines; Random SubSpace

1. Introduction

Landslide is considered as one of the most widely distributed mass movements in mountainous areas all over the world (Pham et al., 2016a). Due to its widespread submerge, unexpected, and seasonal characteristics (Vilimek and Smolikova, 2015), landslides always lead to huge risks towards human life and economic (Kritikos and Davies, 2015), especially in industrial and other social activity such as mining factory, water resources facilities, and hydropower stations (Zhang et al., 2015a). Therefore, over the last two decades many scientists have engaged themselves to predict the landslide locations spatially which is critical for development and planning purpose (Posner and Georgakakos, 2015). In the formation of landslide, there are many factors play key role, some are natural factors, and some are human factors. Accroding to Bandini et al. (2015), the human factors are very important and cannot be ignored (Bandini et al., 2015). It is also observed that climate change and landslide activities have some significant relationship (Sewell et al., 2015). Some recent studies have revealed that submarine landslides and turbidity currents may become more likely due to future rapid global warming (Chang and Huang, 2015). However, some research have indicated that rapid climate change does not necessarily increase the current activity, and thus do not provide a direct evidence as a primary triggering factor (Clare et al., 2015). In fact, heavy precipitation and snow melting have been witnessed as important indirect triggering factors, due to complex hydrogeological relations and an isolated groundwater system. As such it is difficult to determine the triggering parameter especially when it is principally associated with lateral river erosion in the landslide toe (Kirschbaum et al., 2015).

Preparation of a good quality landslide inventory map is very critical in landslide studies. There is still an argument and debate among the landslide researchers whether traditional detection of landslides, for example through direct field surveys or remote sensing based analysis is better. Both of the aforementioned approach has some merits and demerits such as in the case of field investigation which is highly time consuming and very laborious act, on the other hand satellite image processing techniques also need rigorous validation steps to produce an accurate landslide inventory map (Lin et al., 2015). High-resolution imagery is an indispensable data source to model

and validate landslide features appropriately (Akcay, 2015). Similarly, selecting landslide conditioning factors is also important and a complex process. In this regard, statistical index method is generally used for the validation of different landslide conditioning factor sets. It describe how strong every single landslide conditioning factor improves or reduces the objective function, then is used to find a better input conditioning factor set by excluding some factors (Meinhardt et al., 2015).

In recent years, an extensive application of GIS and RS technologies have been used to assess landslide susceptibility, hazard and risk areas (Pham et al., 2016d; Tan et al., 2015). Meanwhile, there are a lot of new methods and techniques have been used in landslide susceptibility mapping, for example, step-wise weight assessment ratio (Dehnavi et al., 2015), multivariate adaptive regression spline (Wang et al., 2015b), artificial neural network (Pradhan and Lee, 2010; Gelisli et al., 2015; Pham et al., 2015b; Arnone et al., 2016), random forest (Trigila et al., 2015), multi-criteria evaluation (Ahmed, 2015), spatial multicriteria evaluation (SMCE) (Gaprindashvili and Van Westen, 2016). However, there still remains a question on the selection of the best method in landslide susceptibility assessment (Goetz et al., 2015; Pham et al., 2015a). So, in order to get the most suitable and optimum model, many scientist have used various models in geographically located different study area worldwide (Hong et al., 2015; Pham et al., 2016c; Tien Bui et al., 2016).

The main objective of the present study is to explore potential application of the novel hybrid integration approach of support vector machines and random subspace named RSSVM for landslide susceptibility assessment. Random subspace is an integrated algorithms, whereas RSSVM is a SVM based Classifier. Performance of the RSSVM model has been evaluated using statistical index-based evaluation and receiver operating characteristic curve technique. Besides, other method such as ANN (Artificial Neural Network), NB (Naïve Bayes) and SVM (Support Vector Machine), have been applied and the results were compared with the result of the RSSVM model. All the aforementioned models were applied in a study area located in the Jiangxi province (China). The analysis process has been carried out using Weka 3.7.12 and ArcMap10.0.

2. Models Background

2. 1. Support Vector Machines

Support vector machine (SVM) is a supervised learning method and is widely used in the statistical classification and regression analysis (Lee et al., 2015; Pham et al., 2016b). It lead to map vector to a higher dimension space, set up in this space there is a maximum separation hyper plane (Feng et al., 2016). On the opposite side of the hyper plane, it separates data with two parallel hyper planes with each other (Xu et al., 2015). Besides, it separates super flat surface to maximize the distance of two parallel hyper-planes (Bezak et al). Assuming that the distance between parallel hyper planes or gap, the greater the classifier of the total error is smaller (Wu et al., 2016). The main ideas of the SVM can be summarized as two points (Li et al., 2016; Ma et al., 2016):

1) It is to solve the problem of linear separable analysis, and in the case of linear inseparable, nonlinear mapping algorithms are generally used which makes a low-dimensional input space linear sample into a high dimensional feature space linear separable. So, to make the high dimensional feature space by using linear algorithm on the nonlinear characteristics of samples is possible (Alvioli and Baum, 2016).

2) It is based on structural risk minimization theory to construct the optimal hyper plane segmentation in the feature space (LaHusen et al., 2016).

3) More details about the SVM algorithm can be seen in XXXXX

2. 2. Random Subspace Ensemble

Random subspace is a classic integrated algorithm which was proposed by Ho in 1998 (Tin Kam, 1998). The algorithm is similar to the bagging algorithm and is randomly selected by the original training set to construct the training subset (Kuncheva et al., 2010). But the difference is that, Random subspace is randomly selected from the original training set of features (Skurichina and Duin, 2001, 2002). Then, the series features a subset of each sub-classifier training at the final forecast results obtained by a combination of voting methods (Sun and Zhang, 2007). The Performance of the result depends on integrated learning differences, bagging method sub-categories to get the difference between the sub-classification among different samples of each sub-classifier training, and random subspace ensemble learning method to take samples at different spatial characteristics to obtain differences between sub-classification (Zhang and Jia, 2007).

3. Novel Hybrid Integration of Support Vector Machines and Random SubSpace (RSSVM)

The novel hybrid integration of support vector machines and random subspace ensemble (RSSVM) for spatial prediction of landslides occurred in the study area is shown in Fig 1.

    (i)   ***Data collection and processing in GIS:*** The GIS database is generated for establishing and verifying the proposed integrated landslide model. The landslide inventory map with 445 landslides was randomly split in 70% (311 landslides shown in the yellow color) for training the models and the remaining 30% (134 landslides shown in the red color) was used for validation purpose (Figure 2). The training and validating sets were utilized to construct the landslide model whereas the testing set is employed to verify the RSSVM model. For this purpose, fourteen landslide conditioing factors were converted from categorical classes into continuous values by employing the frequency ratio method (Bui et al., 2015). Subseqiently, all the 14 landslide conditioing factor maps and landslide and non-landslide data were converted to a raster format with a resolution of 25 m.

    (ii)   ***Random SubSpace ensemble (RSS) optimization:*** The algorithm is similar to the bagging algorithm are randomly selected by the original training set to construct the training subset (Kuncheva et al., 2010). But the difference is that, Random subspace is randomly selected from the original training set of features (Skurichina and Duin, 2001, 2002). Random sub space (RSS) can improve the performance of the SVM classifier

4

through dividing the dataset with a high dimensional feature space to smaller datasets, for example in training data. Generally, these ensemble data mining classifiers have a better accuracy than a single predictor. These methods, firstly, produce via training the dataset base classifications. Secondly, the real classification was done by combining the results of base classifiers with the previous base classifier (Piao et al., 2015). In other words, some classifiers are trained on random feature subsets of the data and the decisions of these classifiers were combined. After dividing the original feature space (FR) into L feature subsets (FS) of p-dimensionality, each subset is submitted to a base classifier (BC) in the ensemble. The final decision on the class of the document is obtained by combining the decisions of these BCs using a combining rule such as majority vote. Since the feature subsets submitted to the BCs are different, parallel ensemble is a natural choice (Kuncheva and Plumpton, 2010). By evaluating the fitness of each solution, the algorithm gradually explores the solution space and seeks for the most desirable set of model's parameters.

(iii) ***Training support vector machines model (SVM):*** Since, SVM is a supervised learning method, a training dataset with input factors and corresponding desired class labels must be provided. Based on the training dataset, SVM maps the input data from the original input space into a high dimensional feature space using RBF kernel function. Accordingly, the algorithm constructs a hyper plane that can separate the input data of influencing factors into two distinctive decision regions: "landslide" ($Y=1$) and "non-landslide" ($Y=-1$). When the training process is accomplished, the resulting SVM for classifying input patterns is demonstrated.

(iv) ***The optimized RSSVM model:*** the SVM-based optimization process continues until the maximum number of generation is reached. When the searching process terminates, a desirable set of the RSSVM tuning parameters is identified and it is used to construct a classification model for spatial prediction of landslides of the study area. The proposed model is ready to predict unseen input patterns in the validation dataset.

(v) ***Comparison with data mining techniques:*** The obtained results from novel hybrid integration of support vector machines and random subspace ensemble (RSSVM) were compared with other well known data mining techniques such as MLPN (Multiple Perceptron Neural Network), NB (Naïve Bayes) and SVM (Support Vector Machine) for spatial prediction of landslides in the study area. Finally, the proposed model is ready to predict unseen input patterns in the validation dataset.
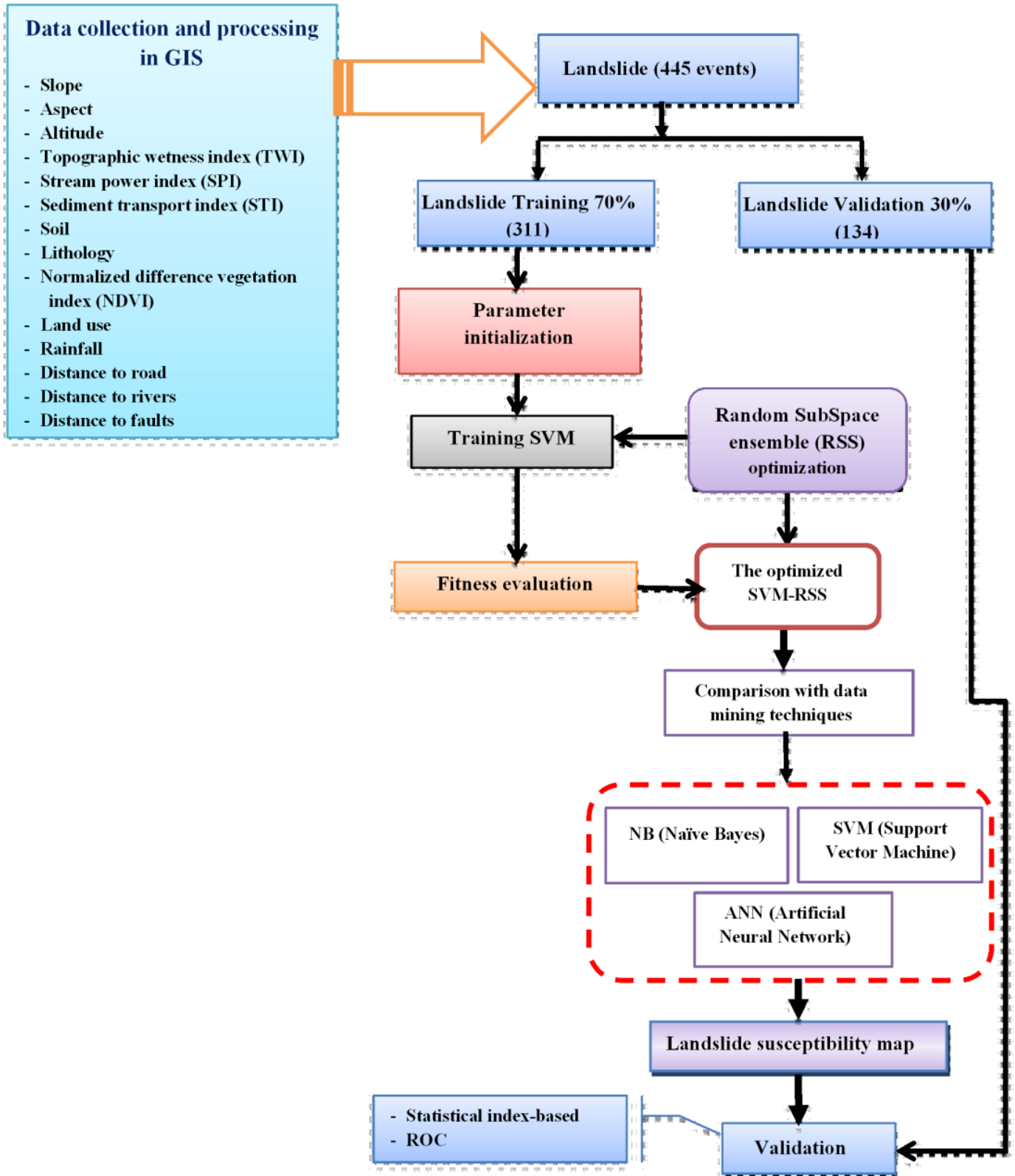
Fig 1 The flow chart of the proposed RSSVM model

4. Experiment and analysis

4. 1. Description of study region

The study area is about 3,507 km^2 and covers the whole area of the Wuning district. It is located in the West part of the Jiangxi province that belongs to the South mountainous area of Mubu, between longitudes 114°29′E and 115°27′E, and latitudes 28°53′N and 29°35′N (Fig 2). According to the report of geological survey results, Wuning area has found a total of 30 kinds of minerals such as Iron, manganese, lead, tungsten, mercury, limestone, and stone and so on. Meantime, in recent years, there are lot of mining factory had been built in the Wuning area.

The study area is situated in the subtropics monsoon climate region. The rainy season is normally from April to August. The total average rainfall in three month (April to June) was about 638 mm. The largest daily rainfall is more than 100 mm during the rainy season. The dry season is generally from September to January with average rainfall is around 65.6 mm/month. The annual average rainfall ranges between 578 and 1183 mm with average of 162.4 rainy days. The annual average temperature is 16.6 °C, i.e., the highest and lowest average temperatures are 39.1 and -5.2 °C, respectively. The annual average sunshine is 1700.5 hours.
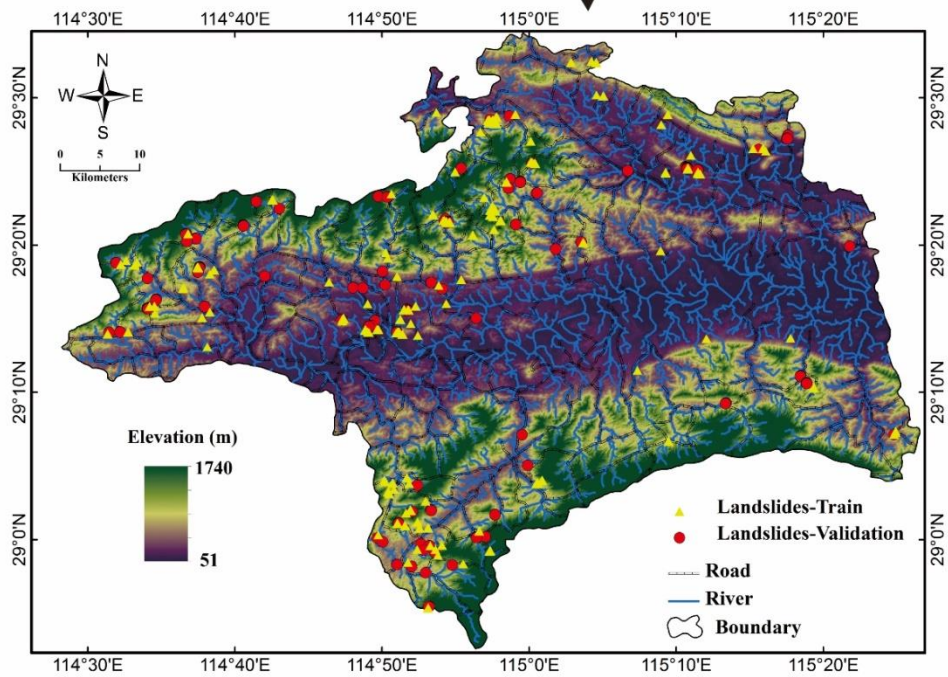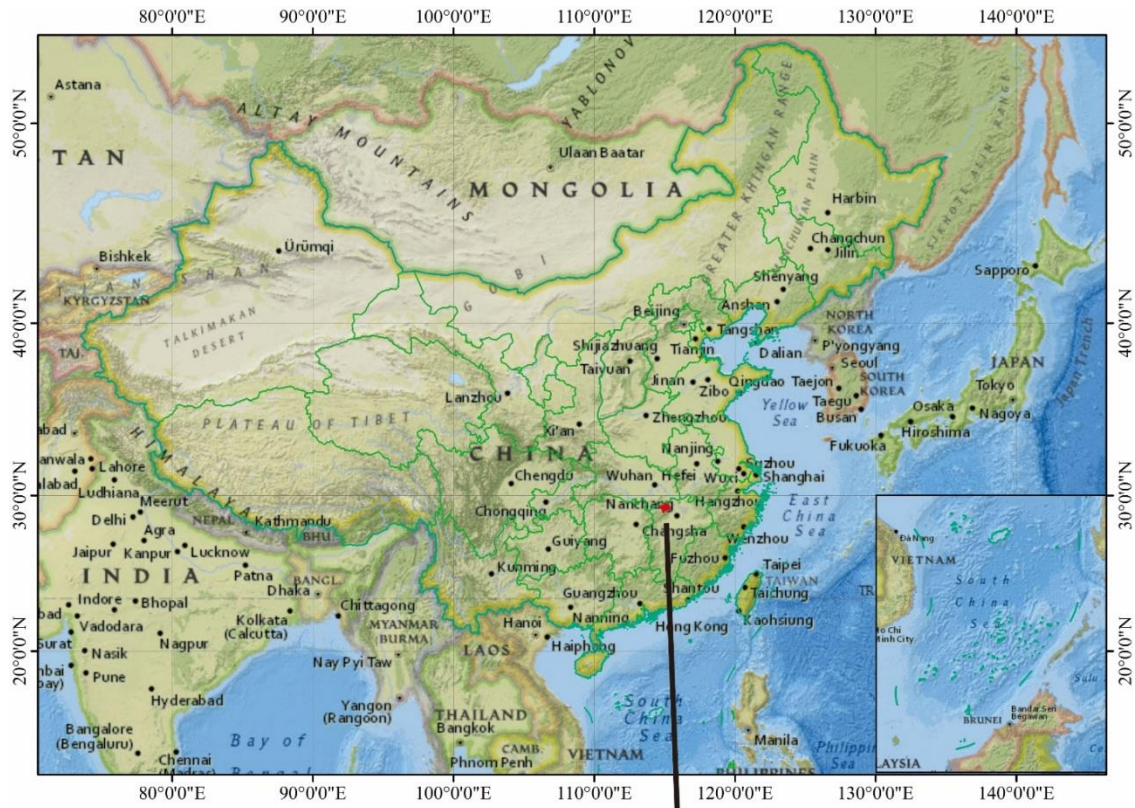
4. 2. Data Preparation

4. 2. 1. Historical landslide events

There are a lot of different methods and techniques being used while constructing landslide inventory map. Those methods are field survey, satellite image interpretation, aerial photograph, and historical landslide records (Wang et al., 2015a). However, till date, there are no agreement in the among the scientist on the best suitable solo method for producing an accurate landslide inventory map (Uhlemann et al., 2016).

In this research, by combining the field survey with satellite image the landslide inventory map was produced. The landslide inventory map for the Wuning area had a total of 445 landslide locations which was provided by the Department of Land and Resources (http://www.cgs.gov.cn), the Meteorological Bureau (http://www.weather.org.cn), the Jiangxi Province, China (Fig 3). The landslide inventory map shows that the volume of the smallest landslide is 20 m^3, the largest is 96,000 m^3, and the average is 1761.3 m^3. Four large-sized landslides (>1000 m^3) have occurred in the study area that has affected 254 people. These landslides are accounted for 8.1% of the total number of landslides. Around 16.0% of the landslides are medium-sized (200–1000 m^3) which has affected 121 people. Small-sized landslides (<200 m^3) has affected 551 people and that accounts for 75.9% of the total landslides. All the landslides occurred during and after the heavy rainfalls. Around 38.5% of the landslides were reported when the measured rainfalls were around 100 mm per day, whereas the other landslides occurred when the daily rainfall was larger than 105 mm.
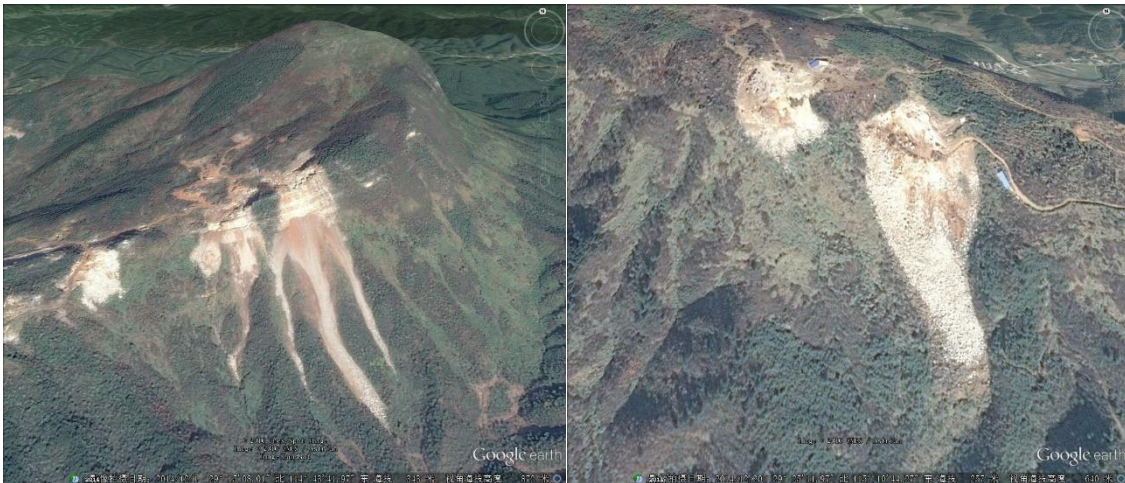

Fig 3 Google Earth images showing landslides of the study area

4. 2. 2. Landslide conditioning parameters

In this study, a total of fourteen landslide conditioning factors were selected based on the analysis of the landslide inventory map. They are slope, aspect, altitude, topographic wetness index (TWI),

stream power index (SPI), sediment transport index (STI), soil, lithology, normalized difference vegetation index (NDVI), landuse, rainfall, distance to roads, distance to rivers, distance to faults.

A digital elevation model (DEM) for the study area was generated from the ASTER GDEM Version 2 (http://gdem.ersdac.jspacesystems.or.jp). A resolution of the DEM is transferred to 25 × 25 m and based on the DEM, slope, altitude, aspect, TWI, SPI, STI were extracted. The slope map was constructed with 6 categories namely: 0–8, 8–15, 15–25, 25–35, 35–45, >45 (Fig 4a). The aspect map (Fig 4b) was constructed with nine groups. The altitude map was constructed with 5 categories as <150, 150–250, 250–400, 400–700, 700–1000,>1000 (Fig 4c). The TWI map in this study was constructed with five classes as <5, 5–7, 7–9, 9–11, >11, (Fig 4d). The SPI and STI maps were constructed into five categories (Fig 4e and Fig 4f) namely<20, 20–40, 40–70, 70–100, >100 and <5, 5–20, 20–40, 40–60, >60 respectively.

The soil map (Fig 4g) was compiled in 1995 by the Institute of Soil Science, Chinese Academy of Sciences (ISSCAS) from data of the Office for the Second National Soil Survey of China (http://www.issas.ac.cn).

The lithology map for the study area was obtained from China Geology Survey (http://www.cgs.gov.cn). The lithology map (Fig 4h) was categorized into eight groups (A, B, C, D, E, F, G, H, I and J).
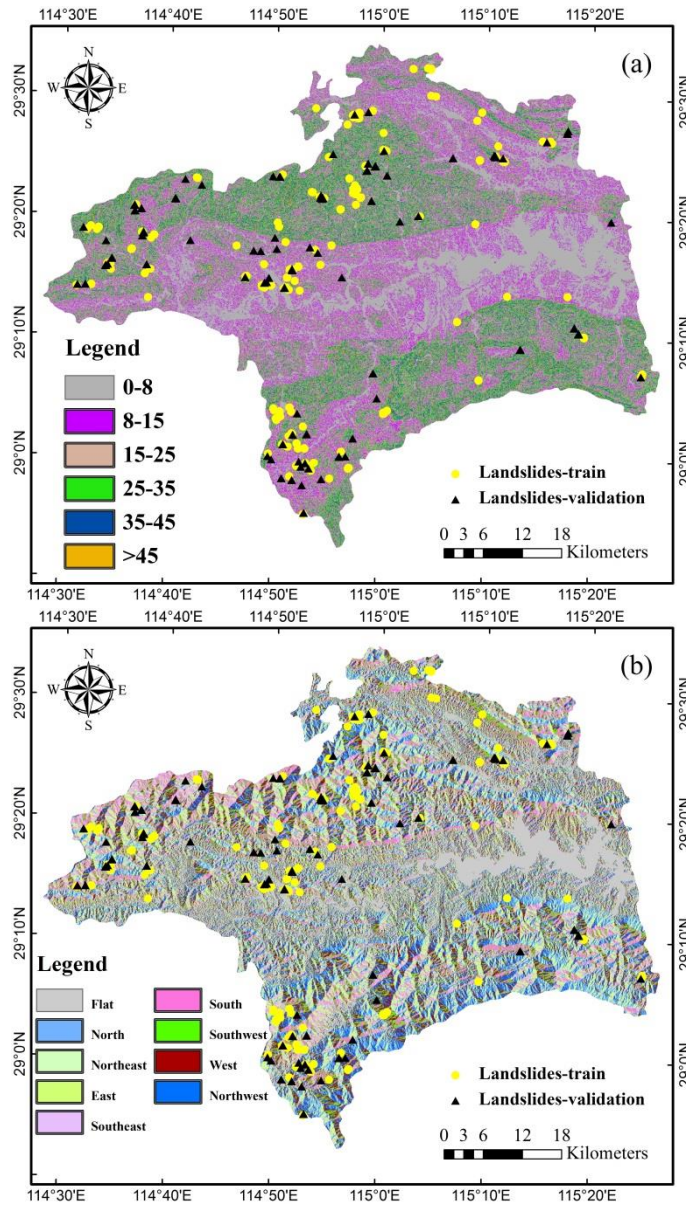
The NDVI map with five classes such as <0.05, 0.05–0.1, 0.1–0.15, 0.15–0.2, >0.2 for the study area (Fig 4i) was obtained from the Landsat 7 ETM+ satellite images which were acquired on 10 December 1999. These images were obtained from Computer Network Information Center of Chinese Academy of Sciences (http://www.gscloud.cn). The value of NDVI was calculated using the common formula NDVI = (IR − R) / (IR + R), where IR and Red are the near infrared (NIR) and red bands, which are from 0.7 to 1 lm and 0.6 to 0.7 lm of the electromagnetic spectrum.
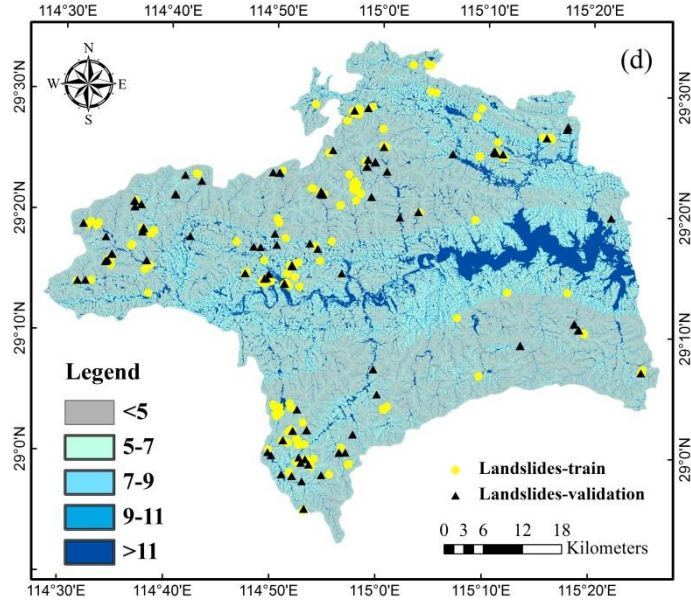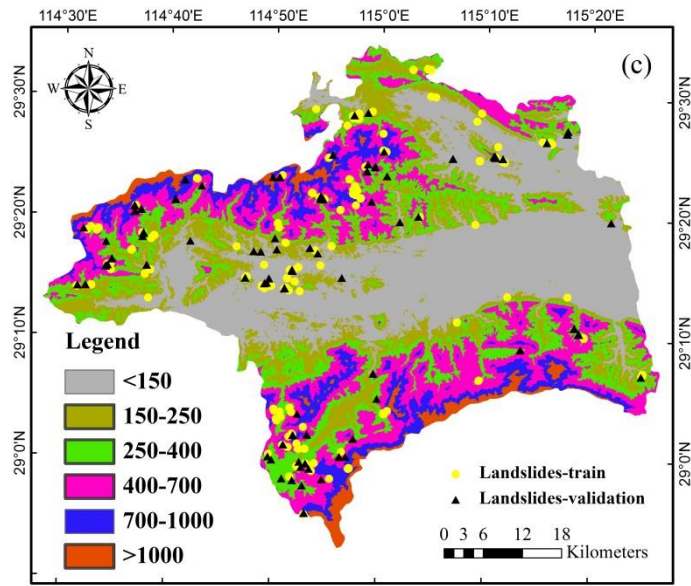
Land use map was extracted from Landsat 7 ETM+ satellite images, six classes were identified such as water, residential area, forest land, bare land, farm land, and grass land. Maximum likelihood supervised method was used for the classification with the classification accuracy of 90.7% (Fig 4j).
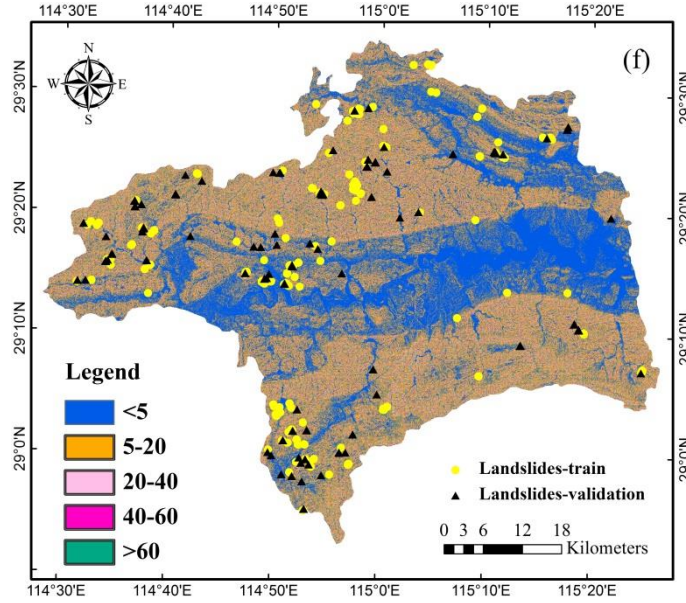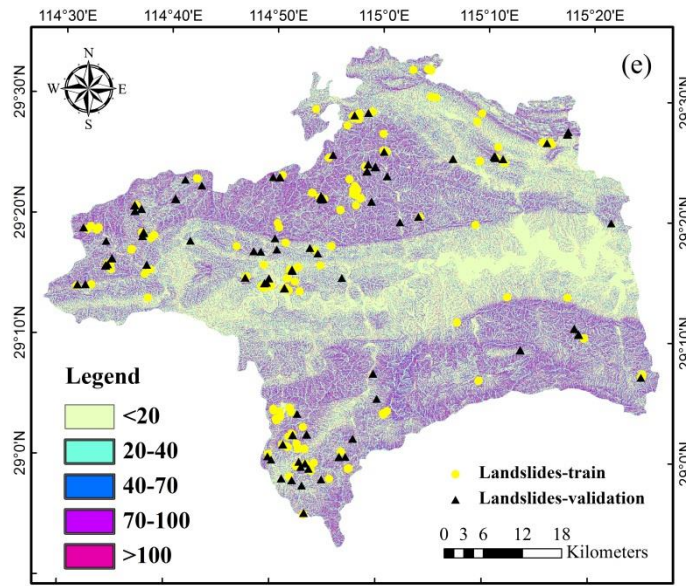
Regarding rainfall, the mean annual precipitation for the period 1960–2012 at 37 rainfall stations was used to create the rainfall map which is divided into five classes: <850, 850–950, 950–1050, 1050–1150, >1150 (Fig 4k) using inverse distance weighted method. The precipitation data were extracted from a database from the government of Jiangxi Province Meteorological Bureau (http://www.weather.org.cn). A brief description the preparation of each of the landslide conditioning factors is given below and is shown in Table.
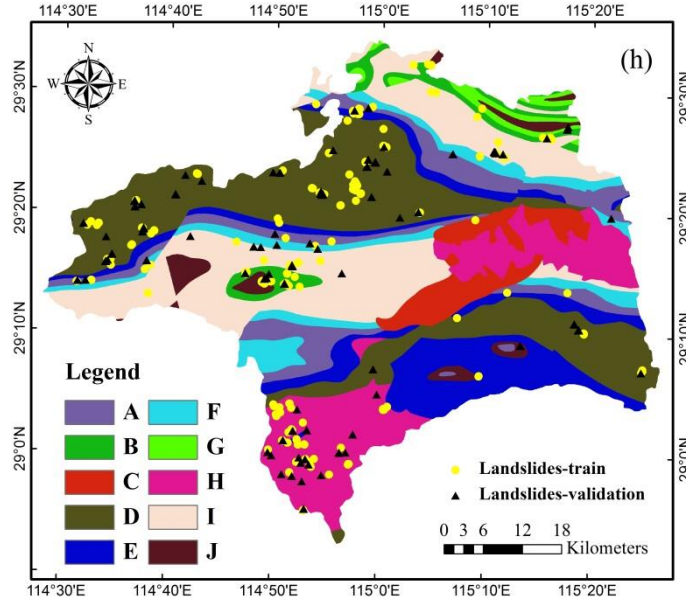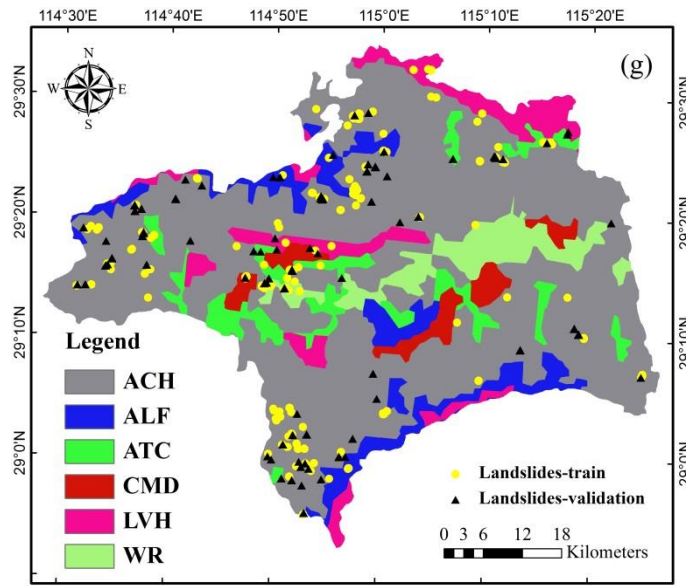
Road and river networks were constructed into five groups categories that undercut slopes lager than 15° were extracted from the topographic map at a scale of 1:50,000 and used to construct the distance to road maps (Fig 4l) and the distance to river maps (Fig 4m), the group was <50, 50–100, 100–150, >150, respectively. The fault lines were extracted from the geological map at a

scale of 1:200,000 and were used to construct the distance to faults map with 7 categories: <200, 200–400, 400–700, 700–1000, >1000 (Fig 4n).

14

(k)

Legend
- ■ <850
- ■ 850-950
- ■ 950-1050
- ■ 1050-1150
- ■ >1150

- ● Landslides-train
- ▲ Landslides-validation

0 3 6 12 18
Kilometers

(l)

Legend
- ■ <50
- ■ 50-100
- ■ 100-150
- ■ >150

- ● Landslides-train
- ▲ Landslides-validation
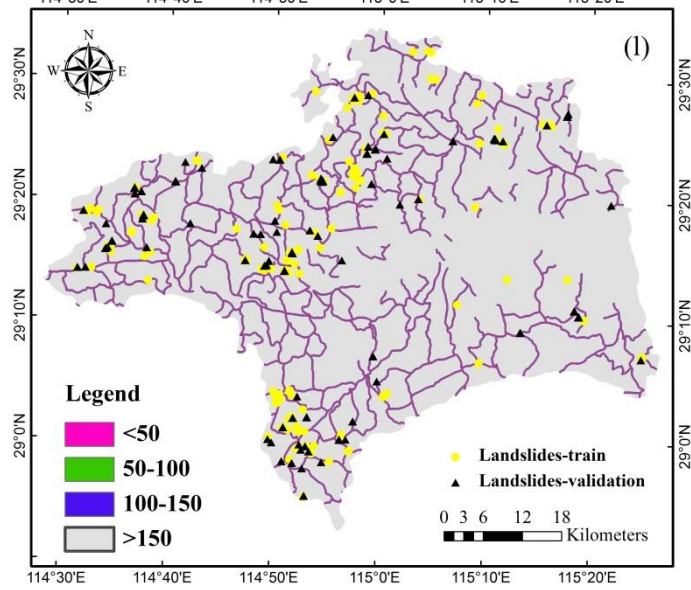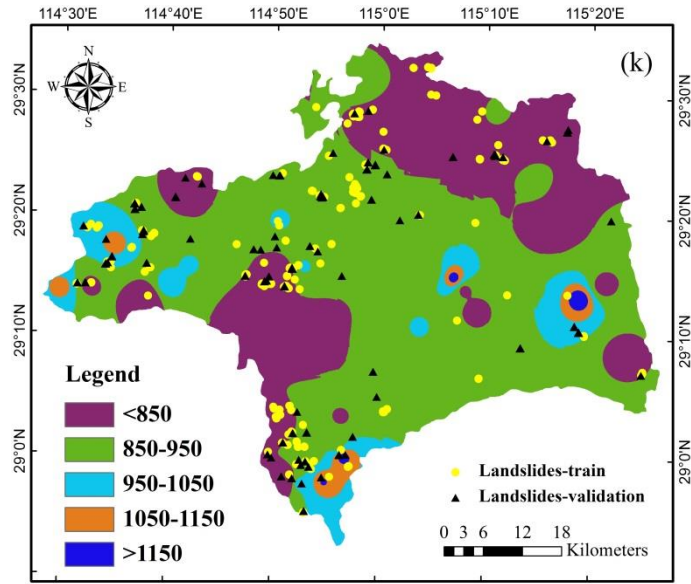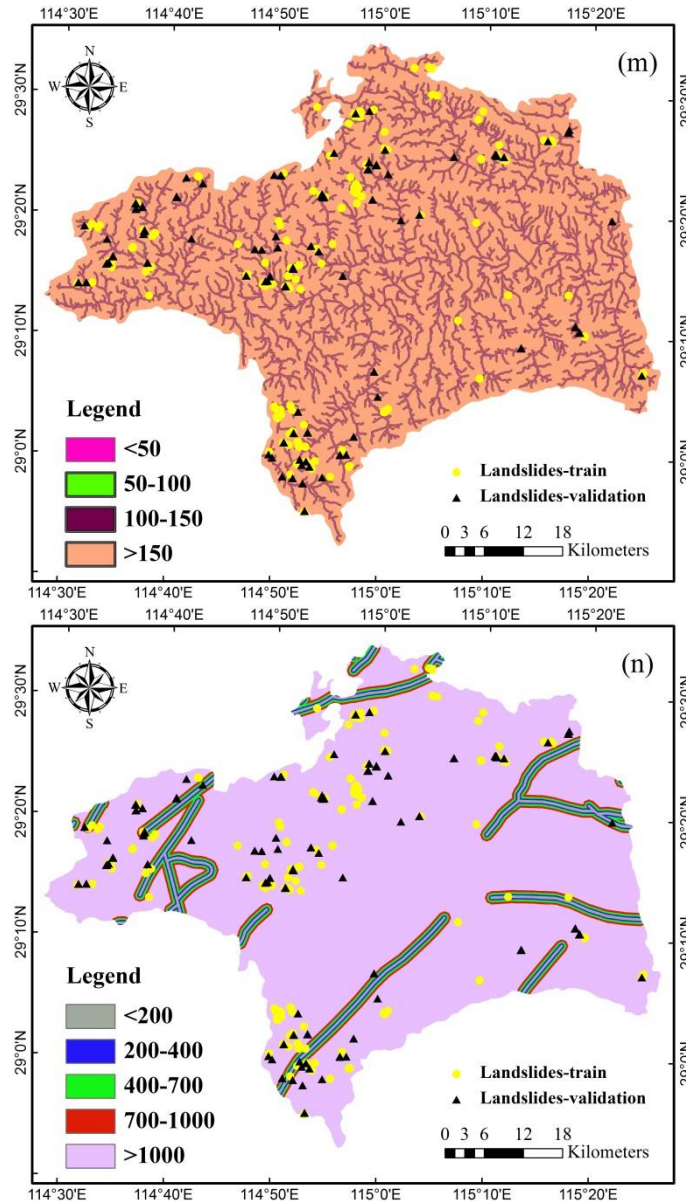
0 3 6 12 18
Kilometers

16

Fig 4 Landslide conditioning factor maps of the study area: (a) slope, (b) aspect, (c) altitude, (d) topographic wetness index (TWI), (e) stream power index (SPI), (f) sediment transport index (STI), (g) soil, (h) lithology, (i) normalized difference vegetation index (NDVI), (j) landuse, (k) rainfall, (l) distance to roads, (m) distance to rivers, (n) distance to faults.

4. 3. Results and discussions

4. 3. 1. Feature selection of Linear Support Vector Machine (LSVM)

Linear Support Vector Machine is a classifier of Support Vector Machine; and it is widely used in data mining (Chou et al., 2016). Similar to SVC with parameter kernel='linear', but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples (Franz et al., 2016). This class

17

supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme (Fan et al., 2016; Lora et al., 2016)

It is very important to evaluate the predictive capability of preparing training dataset using fourteen landslide conditioning factors. In this study, the Linear Support Vector Machine method with tenfold cross-validation has been used for this purpose. Figure 5 presents the predictive capability of landslide conditioning factors in this study. It is shown that slope angle has the best predictive capability in landslide models (AM = 13.8). Rainfall also has very high contribution in landslide models (AM = 12.8). TWI (AM = 12) and STI (AM = 11.4) has relatively high contribution in landslide models as well. Slope aspect (AM = 10) and distance to road (AM = 9) has the moderate contribution in the modelling. Other factors such as distance to faults (AM = 7.3) and distance to rivers (AM = 7.2) had almost similar predictive capabilities. Besides, altitude (AM = 5.4), land use (AM = 4.9), lithology (AM = 4.9) had shown similar contribution. On the contrary, NDVI (AM = 2.3), SPI (AM = 2.1) have low predictive capability, especially soil type (AM = 1.9) has the lowest contribution.

All in all, all fourteen landslide conditioning factors had contribution to landslide models (AM>0). All of them wre important in this study, so, all these factors were used in landslide susceptibility modeling.

Fig 5 Predictive capability of landslide conditioning factors using LSVM method

4. 3. 2. Preparation of dataset and training the RSSVM model

Performance of the RSSVM model significantly depends on the selection of the calculating parameter which is the number of iterations. Thus, a test of the performance of the RSSVM model has been accomplished with different number of iterations to filter the optimal parameter. For this purpose, ROC curve method has been used to evaluate the performance of the RSSVM.

Figure 6 shows the analysis result of ROC curve with various numbers of iterations for training and testing the RSSVM model. It can be found that the number of iterations equals to 14 resulting the best performance of the RSSVM model. Thus, in the present study, the number of iterations has been set to 14 for training the RSSVM classifier in novel classifier framework.
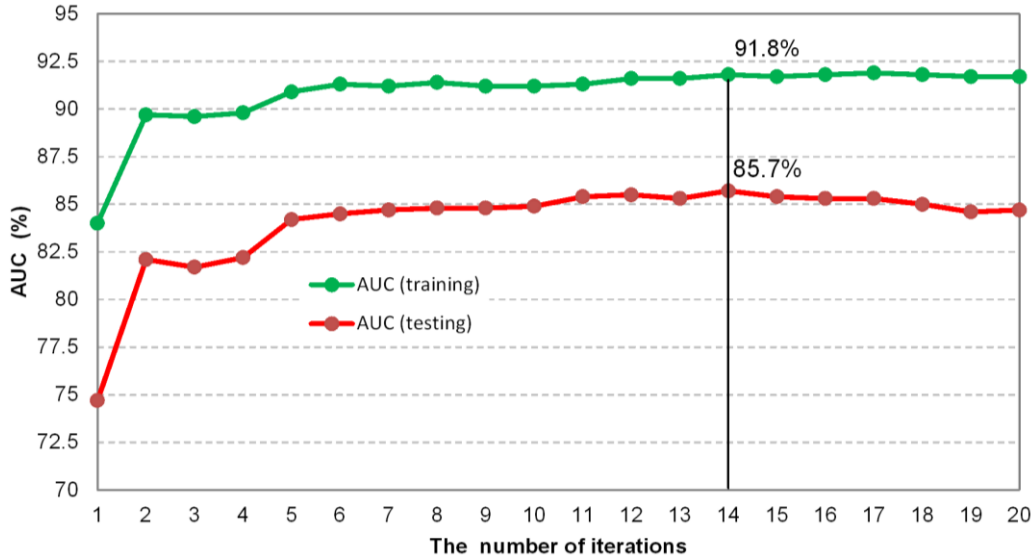

Figure 6 Analysis of ROC curve with various number of iterations for training the RSSVM model

4. 3. 2. Validation of predictive capability of the RSSVM model

Performance of the RSSVM model for landslide susceptibility assessment using statistical index-based evaluations is shown in Table 1. It can be seen that the RSSVM model achieved good classification of both landslide and non-landslide pixels. Out of these, the performance of the RSSVM model for classification of landslide pixels (specificity=79.93%) is better than those of non-landslide pixels (specificity=78.18%).

Besides, in this study, receiver operating characteristic (ROC) curve have also been applied to evaluate the general performance of the RSSVM model. The ROC curve is widely used in landslide susceptibility mapping. In general, the AUC value varies from 0.5 To 1, if the AUC value equals to 1, the result of landslide model is excellent, otherwise, if the AUC value equals to 0.5, the result of landslide model is imprecise. Figure 7 and Figure 8 show the results of the RSSVM model for landslide susceptibility assessment using ROC curve technique. In this study, the RSSVM model executed very well based on the analysis of the ROC curve (AUC = 0.918), additionally, the ROC curve for testing the RSSVM model is 0.857 which is reasonably satisfactory.

Table 1 Performance of the RSSVM model using statistical index based evaluations

No	Parameter	Training	Testing
1	TP	397.00	172.00

2	TN	462.00	172.00
3	FP	116.00	48.00
4	FN	51.00	48.00
5	Positive predictive value (%)	77.39	78.18
6	Negative predictive value (%)	90.06	78.18
7	Sensitivity (%)	88.62	78.18
8	Specificity (%)	79.93	78.18
9	Accuracy (%)	83.72	78.18

AUC = 0.918
Standard Error = 0.00845
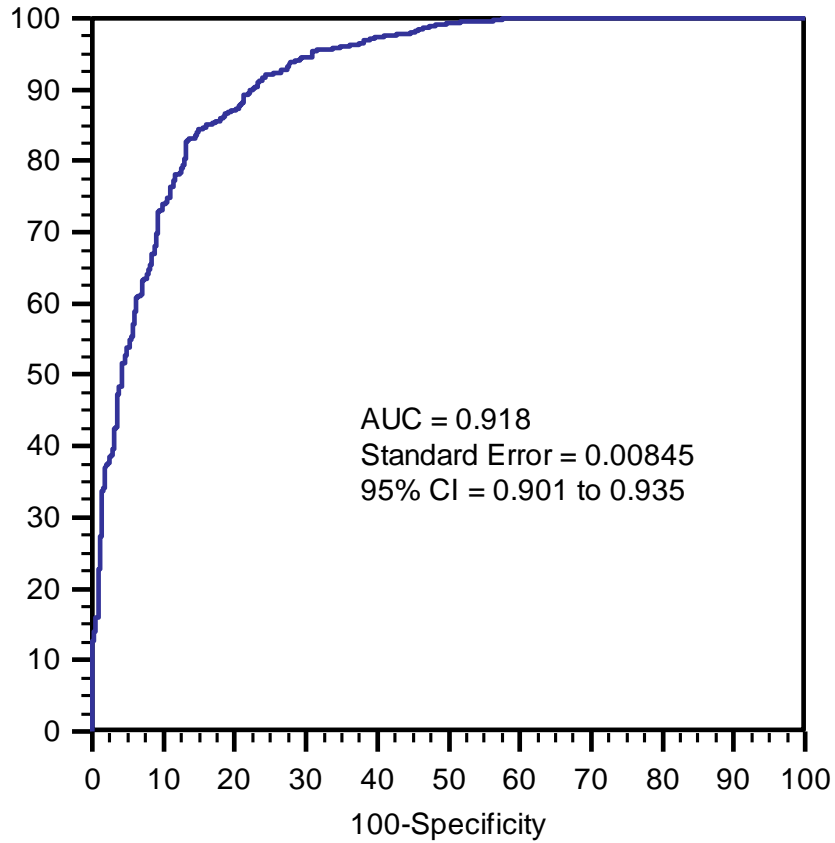95% CI = 0.901 to 0.935

100-Specificity

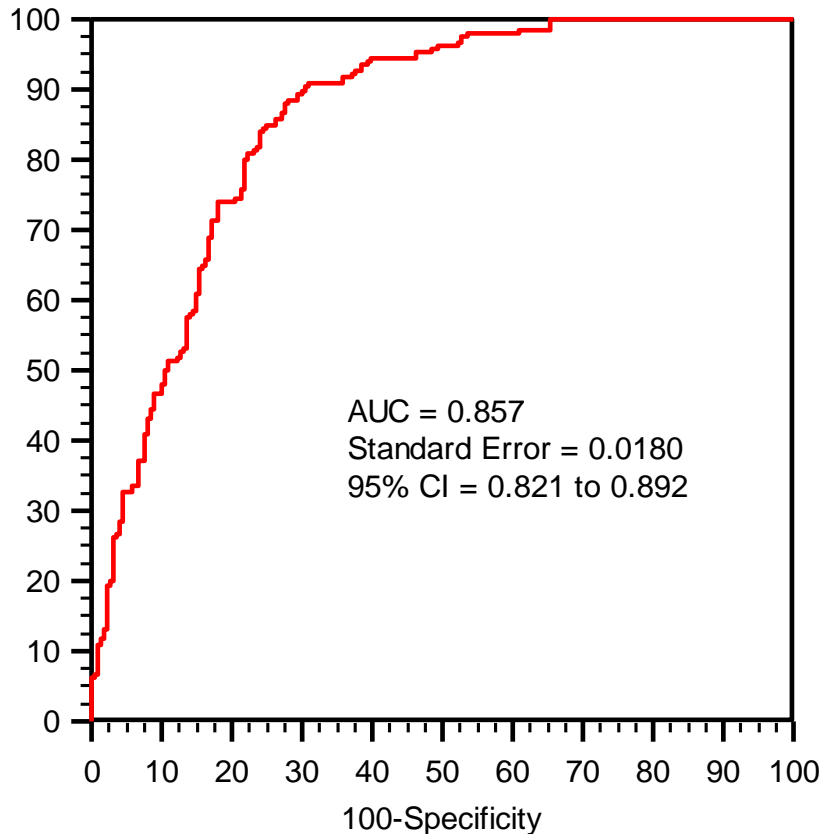**Fig 7** Analysis of ROC curve for training the RSSVM model

Fig 8 Analysis of ROC curve for testing the RSSVM model

4. 3. 3. Comparison of the RSSVM model with popular landslide models

In this study, other popular well known landslide models such as Multiple Perceptron Neural Network, Naïve Bayes, and Support Vector Machine, have been applied and compared with the result of the proposed hybrid model.

Multiple Perceptron Neural Network (MLPN): artificial neural networks are an information processing technique based on the way biological nervous systems, such as the brain, process information (Satorra and Bentler, 2001). The true power and advantage of neural networks lies in their ability to represent both linear and non-linear relationships and to learn these relationships directly from the data being modeled (Gutiérrez and Lizaga, 2016). Traditional linear models are simply inadequate when it comes to modeling data that contains non-linear characteristics (Ye and Chen, 2001). Neural networks are being applied to an increasing large number of real world problems (Wang et al., 2016). Their primary advantage is that they can solve problems that are too complex for conventional technologies; problems that do not have an algorithmic solution or for which an algorithmic solution is too complex to be defined (Leung et al., 2000). In general, neural networks are well suited to problems that people are good at solving, but for which computers generally are not. These problems include pattern recognition and forecasting, which requires the recognition of trends in data (Ogneva-Himmelberger et al., 2009). The most common neural

network model is the multiple perceptron. This type of neural network is known as a supervised network because it requires a desired output in order to learn (Moore and Sawyer, 2016). The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown (Webster et al., 2016). The multiple perceptron neural network and many other neural networks learn using an algorithm called back propagation (Shi et al., 2016). With back propagation, the input data is repeatedly presented to the neural network. With each presentation the output of the neural network is compared to the desired output and an error is computed (Ciurleo et al., 2016). This error is then fed back (back propagated) to the neural network and used to adjust the weights such that the error decreases with each iteration and the neural model gets closer and closer to producing the desired output (Gutiérrez and Lizaga, 2016). This process is known as "training".

Naïve Bayes (NB): Bayesian network is a directed acyclic graph with probability annotation, each node in the graph represent a random variables, two nodes in the graph if there is an arc, said the two nodes corresponding to the probability of a random variable is dependent, conversely indicates that two independent random variables is conditions (Pirdavani et al., 2014a; Pirdavani et al., 2014b). An arbitrary nodes in the network of X has a corresponding Conditional Probability Table (Conditional aim-listed Probability Table, CPT), nodes X in his father to get all the possible values of the Conditional Probability (Chen et al., 2012). If nodes are without father X, then X CPT for the prior probability is distribution (Wei and Qi, 2012). Bayesian network structure of the nodes and the CPT define the probability distribution of each variable in the network. Naive bayes model has been originated through classical mathematics theory and has a solid mathematical foundation, and the stability of the classification efficiency (Zhang et al., 2011). At the same time, the NBC model requires only a few estimated parameters, less sensitive to missing data, algorithm is simpler (Zhang and Mei, 2011). In theory, NBC model of minimum error rate is compared with other classification methods (Hadayeghi et al., 2010). But in fact is not always the case, this is because the NBC model assumptions are independent of each other between attribute, this assumption is often not established in practical application, this brought certain influence to the correct classification of NBC model (Koutsias et al., 2010). More in the number of attributes or attribute correlation between larger, NBC model classification efficiency than the decision tree model. If the attribute correlation is small, NBC model the performance of the most good (Sharma et al., 2011). Naive Bayes theorem is a kind of unsupervised learning with no iteration, learning efficiency is high, and easy to implement under the large sample size will have better performance (Lukawska-Matuszewska and Urbanski, 2014). But because of the conditional independence assumption is too strong, assumption features, associated with the condition of the characteristics of the input vector of scenarios do not apply (Martinez-Fernandez et al., 2013).

Using the results from analyzing the performance of ROC curve for different landslide models (Figure 9), it can be found that the RSSVM model (AUC = 0.857) has the highest performance,

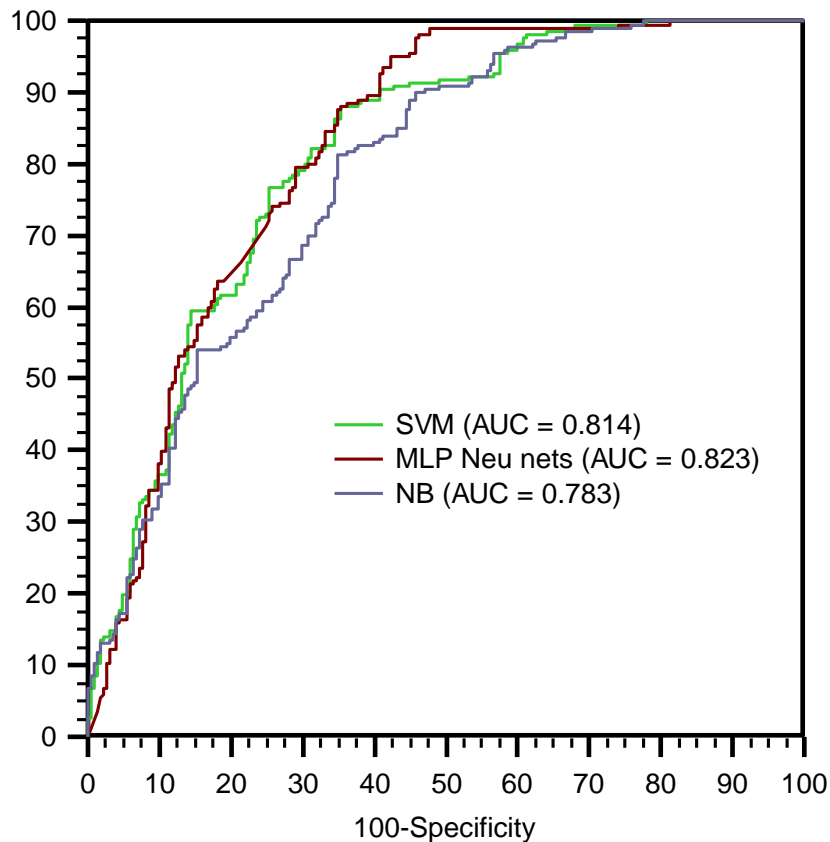followed by the SVM model (AUC = 0.814), MLPN model (AUC = 0.823) and NB model (AUC = 0.783) respectively.

Fig 9 Comparison of predictive capability of landslide models

4. 3. 4. Delineating landslide susceptibility maps

In this study, 311 landslides were used for training data (70%) and 134 landslides were used for validation data (30%). First, the training data set of the SVM and NB models were run in the Weka 3.7.12 software. The polygon of the study area in ArcGIS 10.2 was then converted to raster with a pixel size of 25×25 meter similar for all conditioning factors. The raster polygon was sampled in the GIS with all conditioning factors, and this layer is the test dataset to use in the Weka for landslide susceptibility mapping. Consequently, according to the samples for each point raster of the study area, a probability of landslide incidence was obtained and transferred to the GIS and landslide susceptibility map for the RSSVM model produced. Ultimately, the range of values of the susceptibility map was classified into five categories based on natural break classification method including very low susceptibility (VLS), low susceptibility (LS), moderate susceptibility (MS), high susceptibility (HS) and very high susceptibility (VLS) (Dai and Lee, 2002). Landslide susceptibility map of the RSSVM models are given in Figure 10.

Fig 10 Landslide susceptibility map using the RSSVM model

Table 5 Landslide density on landslide susceptibility map

Class	LSIs intervals	Pixels of class	Pixels of landslides	% Class	% Landslides	Landside Density (LD)
Very low	(0.019 - 0.153)	1431546	0	25.47	0.00	0
Low	(0.153 - 0.301)	1330357	7	23.67	3.57	0.15
Moderate	(0.301 - 0.459)	1145280	23	20.37	11.73	0.58
High	(0.459 - 0.622)	984530	56	17.51	28.57	1.63
Very high	(0.622 - 0.939)	729472	110	12.98	56.12	4.32

5. Conclusions

Landslide is a most dangerous and huge destructive disaster all over the world. Therefore, landslide susceptibility mapping is very important for local government and town planners. Many scientists have utilized different methods to carry out landslide susceptibility map in various regions worldwide. However, until now there is still no agreement with the best method or model in landslide susceptibility modeling. So, the current study aimed to find some new ensemble method to achieve this target.

In this study, a novel hybrid integration approach was used by integrating support vector machines and random subspace for landslide susceptibility assessment. This is an integration of support vector machines and random subspace ensemble. The result shows that the RSSVM model performs very well in landslide susceptibility mapping in Wuning area (China). The predictive capability of a base classifier of RSSVM is significant improved through the newly proposed RSSVM model. In comparison with ANN (Multiple Perceptron Neural Network), NB (Naïve Bayes) and SVM (Support Vector Machine), the RSSVM model has the best performance. Overall, the landslide susceptibility map of these models is very useful for land use planning and decision making in Wuning area.

Acknowledgement

References

Ahmed, B., 2015. Landslide susceptibility mapping using multi-criteria evaluation techniques in Chittagong Metropolitan Area, Bangladesh. Landslides 12(6) 1077-1095.

Akcay, O., 2015. Landslide Fissure Inference Assessment by ANFIS and Logistic Regression Using UAS-Based Photogrammetry. Isprs International Journal Of Geo-Information 4(4) 2131-2158.

Arnone, E., Francipane, A., Scarbaci, A., Puglisi, C., Noto, L.V., 2016. Effect of raster resolution and polygon-conversion algorithm on landslide susceptibility mapping. Environmental Modelling & Software 84(C) 467-481.

Alvioli, M., Baum, R.L., 2016. Parallelization of the TRIGRS model for rainfall-induced landslides using the message passing interface. Environmental Modelling & Software 81 122-135.

Bandini, A., Berry, P., Boldini, D., 2015. Tunnelling-induced landslides: The Val di Sambro tunnel case study. Engineering Geology 196 71-87.

Bezak, N., Šraj, M., Mikoš, M., Copula-based IDF curves and empirical rainfall thresholds for flash floods and rainfall-induced landslides. Journal of Hydrology.

Bui, D.T., Pradhan, B., Revhaug, I., Nguyen, D.B., Pham, H.V., Bui, Q.N., 2015. A novel hybrid evidential belief function-based fuzzy logic model in spatial prediction of rainfall-induced shallow landslides in the Lang Son city area (Vietnam). Geomatics, Natural Hazards and Risk 6(3) 243-271.

Chang, K.T., Huang, H.C., 2015. Three-dimensional analysis of a deep-seated landslide in central Taiwan. Environmental Earth Sciences 74(2) 1379-1390.

Chen, G., Zhao, K.G., McDermid, G.J., Hay, G.J., 2012. The influence of sampling density on geographically weighted regression: a case study using forest canopy height and optical data. International Journal Of Remote Sensing 33(9) 2909-2924.

Chou, H.-T., Lee, C.-F., Lo, C.-M., 2016. The formation and evolution of a coastal alluvial fan in eastern Taiwan caused by rainfall-induced landslides. Landslides 1-14.

Ciurleo, M., Calvello, M., Cascini, L., 2016. Susceptibility zoning of shallow landslides in fine grained soils by statistical methods. Catena 139 250-264.

Clare, M.A., Talling, P.J., Hunt, J.E., 2015. Implications of reduced turbidity current and landslide activity for the Initial Eocene Thermal Maximum - evidence from two distal, deep-water sites. Earth And Planetary Science Letters 420 102-115.

Dai, F.C., Lee, C.F., 2002. Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. Geomorphology 42(3–4) 213-228.

Dehnavi, A., Aghdam, I.N., Pradhan, B., Varzandeh, M.H.M., 2015. A new hybrid model using step-wise weight assessment ratio analysis (SWAM) technique and adaptive neuro-fuzzy inference system (ANFIS) for regional landslide hazard assessment in Iran. Catena 135 122-148.

Fan, L., Lehmann, P., Or, D., 2016. Effects of soil spatial variability at the hillslope and catchment scales on characteristics of rainfall-induced landslides. Water Resources Research 52(3) 1781-1799.

Feng, Z.-y., Lo, C.-M., Lin, Q.-F., 2016. The characteristics of the seismic signals induced by landslides using a coupling of discrete element and finite difference methods. Landslides 1-14.

Franz, M., Carrea, D., Abellán, A., Derron, M.-H., Jaboyedoff, M., 2016. Use of targets to track 3D displacements in highly vegetated areas affected by landslides. Landslides 1-11.

Gaprindashvili, G., Van Westen, C.J., 2016. Generation of a national landslide hazard and risk map for the country of Georgia. Natural Hazards 80(1) 69-101.

Gelisli, K., Kaya, T., Babacan, A.E., 2015. Assessing the factor of safety using an artificial neural network: case studies on landslides in Giresun, Turkey. Environmental Earth Sciences 73(12) 8639-8646.

Goetz, J.N., Brenning, A., Petschko, H., Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Computers & Geosciences 81 1-11.

Gutiérrez, F., Lizaga, I., 2016. Sinkholes, collapse structures and large landslides in an active salt dome submerged by a reservoir: The unique case of the Ambal ridge in the Karun River, Zagros Mountains, Iran. Geomorphology 254 88-103.

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010. Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. Accident Analysis And Prevention 42(2) 676-688.

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

Kirschbaum, D., Stanley, T., Zhou, Y.P., 2015. Spatial and temporal analysis of a global landslide catalog. Geomorphology 249 4-15.

Koutsias, N., Martinez-Fernandez, J., Allgower, B., 2010. Do Factors Causing Wildfires Vary in Space? Evidence from Geographically Weighted Regression. Giscience & Remote Sensing 47(2) 221-240.

Kritikos, T., Davies, T., 2015. Assessment of rainfall-generated shallow landslide/debris-flow susceptibility and runout using a GIS-based approach: application to western Southern Alps of New Zealand. Landslides 12(6) 1051-1075.

Kritikos, T., Robinson, T.R., Davies, T.R.H., 2015. Regional coseismic landslide hazard assessment without historical landslide inventories: A new approach. Journal Of Geophysical Research-Earth Surface 120(4) 711-729.

Kuncheva, L.I., Plumpton, C.O., 2010. Choosing Parameters for Random Subspace Ensembles for fMRI Classification, In: El Gayar, N., Kittler, J., Roli, F. (Eds.), Multiple Classifier Systems: 9th International Workshop, MCS 2010, Cairo, Egypt, April 7-9, 2010. Proceedings. Springer Berlin Heidelberg: Berlin, Heidelberg, pp. 54-63.

Kuncheva, L.I., Rodríguez, J.J., Plumpton, C.O., Linden, D.E., Johnston, S.J., 2010. Random subspace ensembles for fMRI classification. Medical Imaging, IEEE Transactions on 29(2) 531-542.

LaHusen, S.R., Duvall, A.R., Booth, A.M., Montgomery, D.R., 2016. Surface roughness dating of long-runout landslides near Oso, Washington (USA), reveals persistent postglacial hillslope instability. Geology 44(2) 111-114.

Lee, C.-F., Lo, C.-M., Chou, H.-T., Chi, S.-Y., 2015. Landscape evolution analysis of large scale landslides at Don-Ao Peak, Taiwan. Environmental Earth Sciences 75(1) 1-19.

Leung, Y., Mei, C.-L., Zhang, W.-X., 2000. Testing for spatial autocorrelation among the residuals of the geographically weighted regression. Environment and planning A 32(5) 871-890.

Li, G., West, A.J., Densmore, A.L., Hammond, D.E., Jin, Z., Zhang, F., Wang, J., Hilton, R.G., 2016. Connectivity of earthquake‐triggered landslides with the fluvial network: Implications for

landslide sediment transport after the 2008 Wenchuan earthquake. Journal of Geophysical Research: Earth Surface.

Lin, C.H., Jan, J.C., Pu, H.C., Tu, Y., Chen, C.C., Wu, Y.M., 2015. Landslide seismic magnitude. Earth And Planetary Science Letters 429 122-127.

Lora, M., Camporese, M., Troch, P.A., Salandin, P., 2016. Rainfall-triggered shallow landslides: infiltration dynamics in a physical hillslope model. Hydrological Processes n/a-n/a.

Lukawska-Matuszewska, K., Urbanski, J.A., 2014. Prediction of near-bottom water salinity in the Baltic Sea using Ordinary Least Squares and Geographically Weighted Regression models. Estuarine Coastal And Shelf Science 149 255-263.

Ma, H.-R., Cheng, X., Chen, L., Zhang, H., Xiong, H., 2016. Automatic identification of shallow landslides based on Worldview2 remote sensing images. Journal of Applied Remote Sensing 10(1) 016008-016008.

Martinez-Fernandez, J., Chuvieco, E., Koutsias, N., 2013. Modelling long-term fire occurrence factors in Spain by accounting for local variations with geographically weighted regression. Natural Hazards And Earth System Sciences 13(2) 311-327.

Meinhardt, M., Fink, M., Tunschel, H., 2015. Landslide susceptibility analysis in central Vietnam based on an incomplete landslide inventory: Comparison of a new method to calculate weighting factors by means of bivariate statistics. Geomorphology 234 80-97.

Moore, Z.T., Sawyer, D.E., 2016. Assessing post-failure mobility of submarine landslides from seismic geomorphology and physical properties of mass transport deposits: An example from seaward of the Kumano Basin, Nankai Trough, offshore Japan. Marine Geology 374 73-84.

Ogneva-Himmelberger, Y., Pearsall, H., Rakshit, R., 2009. Concrete evidence & geographically weighted regression: a regional analysis of wealth and the land cover in Massachusetts. Applied Geography 29(4) 478-487.

Pham, B.T., Bui, D.T., Prakash, I., Dholakia, M., 2016a. Evaluation of predictive ability of support vector machines and naive Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS. Journal of Geomatics 10(1) 71 - 79.

Pham, B.T., Pradhan, B., Tien Bui, D., Prakash, I., Dholakia, M.B., 2016b. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). Environmental Modelling & Software 84 240-250.

Pham, B.T., Tien Bui, D., Dholakia, M.B., Prakash, I., Pham, H.V., 2016c. 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. Geotechnical and Geological Engineering 34(1) 1-18.

Pham, B.T., Tien Bui, D., Indra, P., Dholakia, M.B., 2015a. Landslide susceptibility assessment at a part of Uttarakhand Himalaya, India using GIS – based statistical approach of frequency ratio method. International Journal of Engineering Research & Technology 4(11) 338-344.

Pham, B.T., Tien Bui, D., Pourghasemi, H.R., Indra, P., Dholakia, M.B., 2015b. 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. Theoretical and Applied Climatology 122(3-4) 1-19.

Pham, B.T., Tien Bui, D., Prakash, I., Dholakia, M.B., 2016d. Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS. Natural Hazards 83(1) 1-31.

Piao, Y., Piao, M., Jin, C.H., Shon, H.S., Chung, J.-M., Hwang, B., Ryu, K.H., 2015. A New Ensemble Method with Feature Space Partitioning for High-Dimensional Data Classification. Mathematical Problems in Engineering 2015.

Pirdavani, A., Bellemans, T., Brijs, T., Kochan, B., Wets, G., 2014a. Assessing the road safety impacts of a teleworking policy by means. of geographically weighted regression method. Journal Of Transport Geography 39 96-110.

Pirdavani, A., Bellemans, T., Brijs, T., Wets, G., 2014b. Application of Geographically Weighted Regression Technique in Spatial Analysis of Fatal and Injury Crashes. Journal Of Transportation Engineering 140(8).

Posner, A.J., Georgakakos, K.P., 2015. Normalized Landslide Index Method for susceptibility map development in El Salvador. Natural Hazards 79(3) 1825-1845.

Pradhan, B., Lee, S., 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environmental Modelling & Software 25(6) 747-759.

Satorra, A., Bentler, P.M., 2001. A scaled difference chi-square test statistic for moment structure analysis. Psychometrika 66(4) 507-514.

Sewell, R.J., Parry, S., Millis, S.W., Wang, N., Rieser, U., DeWitt, R., 2015. Dating of debris flow fan complexes from Lantau Island, Hong Kong, China: The potential relationship between landslide activity and climate change. Geomorphology 248 205-227.

Sharma, V., Kilic, A., Irmak, S., Kabenge, I., 2011. Application of GIS and Geographically weighted regression to evaluate the spatial non-stationarity relationship between precipitation vs. irrigated and rainfed maize and soybean yield. Transactions of the Asabe 54(3) 953-972.

Shi, J.S., Wu, L.Z., Wu, S.R., Li, B., Wang, T., Xin, P., 2016. Analysis of the causes of large-scale loess landslides in Baoji, China. Geomorphology 264 109-117.

Skurichina, M., Duin, R.P., 2001. Bagging and the random subspace method for redundant feature spaces, Multiple Classifier Systems. Springer, pp. 1-10.

Skurichina, M., Duin, R.P., 2002. Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis & Applications 5(2) 121-135.

Sun, S., Zhang, C., 2007. The selective random subspace predictor for traffic flow forecasting. Intelligent Transportation Systems, IEEE Transactions on 8(2) 367-373.

Tan, Y.M., Guo, D., Xu, B., 2015. A geospatial information quantity model for regional landslide risk assessment. Natural Hazards 79(2) 1385-1398.

Tien Bui, D., Pham, B.T., Nguyen, Q.P., Hoang, N.-D., 2016. 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. International Journal of Digital Earth 1-21.

Tin Kam, H., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8) 832-844.

Trigila, A., Iadanza, C., Esposito, C., Scarascia-Mugnozza, G., 2015. Comparison of Logistic Regression and Random Forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). Geomorphology 249 119-136.

Uhlemann, S., Smith, A., Chambers, J., Dixon, N., Dijkstra, T., Haslam, E., Meldrum, P., Merritt, A., Gunn, D., Mackay, J., 2016. Assessment of ground-based monitoring techniques applied to landslide investigations. Geomorphology 253 438-451.

Vilimek, V., Smolikova, J., 2015. Scientific research for landslide risk analysis and international education for mitigation and preparedness. Landslides 12(6) 1227-1231.

Wang, G.Q., Bao, Y., Cuddus, Y., Jia, X.Y., Serna, J., Jing, Q., 2015a. A methodology to derive precise landslide displacement time series from continuous GPS observations in tectonically active and cold regions: a case study in Alaska. Natural Hazards 77(3) 1939-1961.

Wang, L.J., Guo, M., Sawada, K., Lin, J., Zhang, J.C., 2015b. Landslide susceptibility mapping in Mizunami City, Japan: A comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models. Catena 135 271-282.

Wang, Y., Song, C., Lin, Q., Li, J., 2016. Occurrence probability assessment of earthquake-triggered landslides with Newmark displacement values and logistic regression: The Wenchuan earthquake, China. Geomorphology 258 108-119.

Webster, J.M., George, N.P.J., Beaman, R.J., Hill, J., Puga-Bernabéu, Á., Hinestrosa, G., Abbey, E.A., Daniell, J.J., 2016. Submarine landslides on the Great Barrier Reef shelf edge and upper slope: A mechanism for generating tsunamis on the north-east Australian coast? Marine Geology 371 120-129.

Wei, C.H., Qi, F., 2012. On the estimation and testing of mixed geographically weighted regression models. Economic Modelling 29(6) 2615-2620.

Wu, X., Benjamin Zhan, F., Zhang, K., Deng, Q., 2016. Application of a two-step cluster analysis and the Apriori algorithm to classify the deformation states of two typical colluvial landslides in the Three Gorges, China. Environmental Earth Sciences 75(2) 1-16.

Xu, Q., Li, Y., Zhang, S., Dong, X., 2015. Classification of large-scale landslides induced by the 2008 Wenchuan earthquake, China. Environmental Earth Sciences 75(1) 1-12.

Ye, N., Chen, Q., 2001. An anomaly detection technique based on a chi‑square statistic for detecting intrusions into information systems. Quality and Reliability Engineering International 17(2) 105-112.

Zhang, C.S., Tang, Y., Xu, X.L., Kiely, G., 2011. Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Applied Geochemistry 26(7) 1239-1248.

Zhang, H.G., Mei, C.L., 2011. Local least absolute deviation estimation of spatially varying coefficient models: robust geographically weighted regression approaches. International Journal Of Geographical Information Science 25(9) 1467-1489.

Zhang, J.Q., Gurung, D.R., Liu, R.K., Murthy, M.S.R., Su, F.H., 2015a. Abe Barek landslide and landslide susceptibility assessment in Badakhshan Province, Afghanistan. Landslides 12(3) 597-609.

Zhang, S.J., Xie, X.P., Wei, F.Q., Chernomorets, S., Petrakov, D., Pavlova, I., Tellez, R.D., 2015b. A seismically triggered landslide dam in Honshiyan, Yunnan, China: from emergency management to hydropower potential. Landslides 12(6) 1147-1157.

Zhang, X., Jia, Y., 2007. A linear discriminant analysis framework based on random subspace for face recognition. Pattern Recognition 40(9) 2585-2591.