Comparing the Performance of a Logistic Regression and a Random Forest Model in Landslide Susceptibility Assessments. the Case of Wuyaun Area, China

Haoyuan Hong, Paraskevas Tsangaratos, Ioanna Ilia, Wei Chen, and Chong Xu

Abstract

The main objectives of the study was to apply a Logistic Regression and a Random Forest model for the construction of a landslide susceptibility map in the Wuyuan area, China, and to compare their results by performing non-parametric and linear regression analysis. Thirteen landslide variables were analyzed, namely: lithology, soil, slope, aspect, altitude, topographic wetness index, stream power index, stream transport index, plan curvature, profile curvature, distance to roads, distance to rivers and distance to faults, while 255 sites classified as landslide and 255 sites classified as non-landslide were separated into a training dataset (70%) and a validation dataset (30%). The comparison and validation of the outcomes of each model were achieved using statistical evaluation measures, the receiving operating characteristic and the area under the success and prediction rate curves. The presence of linear correlation between the two models was estimated by performing a simple linear regression analysis. The most accurate model was Random Forest, which identified correctly 98.32% of the instances during the training phase, followed by Logistic Regression (87.43%). During the validation phase, the Random Forest achieved a classification accuracy of 85.52%, while Logistic Regression model achieved an accuracy of 80.92%. The area under the success and prediction rate curves for the Random Forest were calculated to be 0.9805 and 0.9324, respectively, while the Logistic Regression model showed as slightly lower predictive performance, 0.9372 and 0.8903 respectively. Finally, by performing a non-parametric analysis, the two models were found to be significantly different. Strong evidence of linear relationship between the two models exist, having a p-value less than 0.0001 at a 95% confidence level and an R^2 value estimated to be 0.6993 indicating that 69.93% of the variability in the Logistic Regression model can be explained by variation in the Random Forest model.

Keywords

Landslide susceptibility assessment • Logistic regression • Random forest • Wuyaun area • China

H. Hong · C. Xu
Key Laboratory of Active Tectonics and Volcano,
Institute of Geology, China Earthquake Administration,
#1 Huayanli, Chaoyang District, 9803, Beijing 100029
People's Republic of China
e-mail: hong_haoyuan@outlook.com

C. Xu
e-mail: xc11111111@126.com; xuchong@ies.ac.cn

P. Tsangaratos (✉) · I. Ilia
Department of Geological Studies, National Technical University
of Athens, School of Mining and Metallurgical Engineering,
Zografou Campus, Heroon Polytechniou 9,
15780 Zografou, Greece
e-mail: ptsag@metal.ntua.gr

I. Ilia
e-mail: gilia@metal.ntua.gr

W. Chen
College of Geology and Environment, Xi'an University of Science
and Technology, Xi'an 710054, People's Republic of China
e-mail: chenwei.0930@163.com

M. Mikoš et al. (eds.), *Advancing Culture of Living with Landslides*,
DOI 10.1007/978-3-319-53498-5_118

Introduction

For the past three decades, the analysis of landslide phenomenon has been a subject of research mainly because of the efforts of the scientific community to mitigate the negative effects of their manifestation. According to Korup and Stolle (2014), the prediction of the spatial distribution of landslides is by far the most investigated topic with the development and application of numerous methods and techniques.

In general, the methods and techniques that are used in landslide susceptibility assessments could be classified into two main approaches; the data-driven approach that is based on the exploration of data and the knowledge-driven approach that is based on the assessment of knowledge. Knowledge–driven approach incorporates methods that are based on the site specific experience of experts with the landslide susceptibility determined directly in the field or by combining different layered index maps, while data–driven approach incorporates methods that perform statistical and probabilistic analysis or follow deterministic approaches (Pourghasemi et al. 2012; Chen et al. 2016; Ilia and Tsangaratos 2016). In recent years, the implementation of these methods has been aided by the technology of Geographical Information System (Akgun et al. 2012; Hong et al. 2016). For both approaches the validation and comparison of the performance of the produced models are based on statistical evaluation measures, the receiving operating characteristic and the area under the success and prediction rate curves (Youssef et al. 2015; Pham et al. 2016). However, there are only few studies found that performed additional test to confirm if any significant statistical differences among the models exist (Tien Bui et al. 2016).

In this context, the present study applied two data-driven methods, a Logistic Regression (LR) and a Random Forest (RF) algorithm for the construction of a landslide susceptibility map in the Wuyuan area, China. The comparison of the outcomes of the LR and RF model was based on non-parametric and linear regression analysis. Specifically, the Wilcoxon signed-rank test was utilized to confirm significant statistical difference among the models and linear regression analysis in order to analyze the potential relationship between the two models. The computation process was carried out using Rstudio, an integrated development environment for R language and ArcGIS 10.1 for compiling the data and producing the landslide susceptibility maps.

Study Area

The Wuyuan area is located in the Northeast of the Jiangxi Province, China, covering an area of approximately 2947.5 km^2, with altitude ranging between 13 and 1631 m above sea level (Fig. 1).

According to the Jiangxi Province Meteorological Bureau, the average annual rainfall is 1961.8 mm (estimated for the period 1960–2013) while the average annual temperature is 16.7 °C. The rainy season is from March to August, while April to June accounts the 62.6% of the rainy season rainfall and 46.7% of the yearly rainfall. The study area is comprised of approximately 47.0% forest land, 3.9% farmland, 7.2% residential, 0.1% bare land, 39.5% grass land, and 2.0% water bodies.

Concerning the geological settings, more than 46 geologic groups and units are recognized. The main lithology units involved limestone formations, inkstone, granite, marble, kaolin, potash feldspar, building sand, sandstone, quartz, slate, clay, quartz and silica grouped into eight lithological units.

Materials and Methods

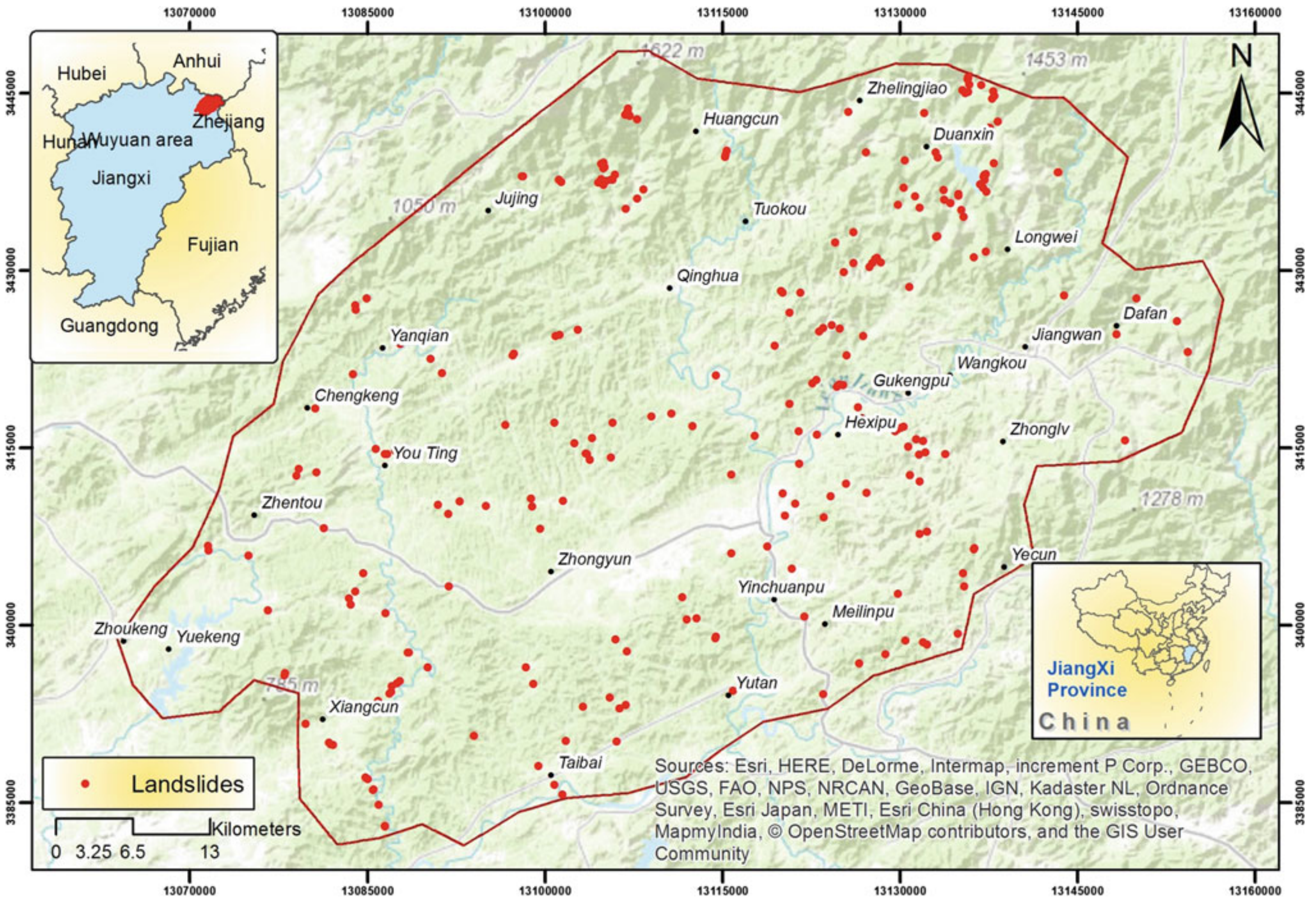
Spatial Data

The inventory database included information about the location, features and abundance of landslide and non-landslide areas. The identification and acceptance of those areas was based on historical information concerning landslide incidence, the interpretation of aerial photos, the use of satellite imagery and extensive field observations. The landslide inventory map for the Wuyuan area with 255 landslide locations was provided by the Jiangxi Department of Land and Resources and the Jiangxi Meteorological Bureau (Fig. 1). The landslide inventory map consists of 115 rotational slides and 140 translational slides. An analysis of the landslide inventory map shows that the size of the smallest landslide is 13.5 m^2, the largest is 9000 m^2, and the average is 580.9 m^2.

Concerning the non-landslide areas they were identified by using airborne imagery and extensive field investigation. A total of 510 sites, 255 landslide and 255 non-landslide areas was recorded. As proposed by the methodology, training and validating data sets were randomly produced from the total number of landslide and non-landslide areas. The first data set contained an initial number of data that equaled to approximately 70% of the total number of landslide and non-landslide.

Thirteen landslide related variables were analyzed, namely: lithology, soil, slope, aspect, altitude, topographic wetness index (TWI), stream power index (SPI), stream transport index (STI), plan curvature, profile curvature, distance to roads, distance to rivers and distance to faults.

A digital elevation model (DEM) of grid size 25 × 25 m generated from 1:50,000 scale topographic maps was used, to construct the layers of slope, aspect, altitude, TWI, SPI, STI, plan curvature and profile curvature. Lithology was



Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Fig. 1 Study area

obtained from the China Geology Survey, while the tectonic features were extracted using the geological map. The lithology map was reconstructed by classifying the geological formations into eight groups. The soil map was compiled by the Institute of Soil Science, Chinese Academy of Sciences (ISSCAS) Nanjing, and the classification scheme that was used according to the FAO-UNESCO classification. Road and river maps were digitized from 1:50,000 scale topographic maps.

Methods of Analysis

Logistic Regression

LR is among those statistical methods that have been proven to be highly reliable when performing a landslide susceptibility assessment (Dai et al. 2002; Ayalew and Yamagishi 2005; Yesilnacar and Topal 2005; Gorsevski et al. 2006; Lee and Pradhan 2007; Yilmaz 2010; Xu et al. 2013; Wang et al. 2013; Tsangaratos and Ilia 2016). The independent variables are considered as predictors of the dependent variable and can be measured on a nominal, ordinal, interval or ratio scale, while the dependent variable is in a binary format. The relationship between the dependent variable and independent variables is nonlinear (Yesilnacar and Topal 2005).

LR is a special case of a generalized linear model; however, it is based on quite different assumptions concerning the relationship between the dependent and independent variables from those followed by linear regression models. The conditional distribution is a Bernoulli distribution rather than a Gaussian distribution, since the dependent variable has the form of a binary variable (presence or absence of landslides).

In logistic regression analysis the relationship between the occurrence and its dependency on several variables can be expressed by the following equation:

$$p = \frac{1}{1 + e^{-z}} \tag{1}$$

where p is the probability of a landslide occurrence.

The probability can take values from 0 to 1 on an S-shaped curve and z is the linear combination of a set of landslide related variables. Logistic regression involves fitting an equation of the following form to the data:

$$z = b_0 + b_1 x_1 + b_2 x_2 + \cdots + + b_n x_n \tag{2}$$

where b0 is the intercept of the model, the b_i (i = 0, 1, 2, …, n) is the slope coefficients of the logistic regression model, and x_i (i = 0, 1, 2, …, n) are the independent variables.

The linear model formed is then a logistic regression of presence or absence of landslides (present conditions) on the independent variables (pre-failure conditions).

Random Forest

Random Forest (RF) is an ensemble learning method, which is based on the generation of several classification trees, which are aggregated to estimate a classification (Breiman et al. 1984; Breiman 2001). The algorithm exploits random binary trees which use a subset of observations through bootstrapping techniques: from the original data set a random selection of training data is sampled and used to build the model, the data not included are referred to as out-of-bag (OOB) (Breiman 2001). According to Hansen and Salamon (1990) an ensemble method, such as RF, is more accurate than individual members if only data appear random and are diverse. In the case of RF, diversity is achieved by resampling the data with replacement and randomly changing the predictive factor over the different tree induction processes (Youssef et al. 2015).

One of the main advantages of RF is the ability to avoid over-fitting and growing a large number of random forest trees where it does not create a risk of over-fitting (e.g., each tree is a completely independent random experiment). The RF algorithm data does not need to be rescaled, transformed, or modified. It has resistance to outliers in predictors and automatically handles the missing values (Catani et al. 2013).

Linear Regression Analysis and Inferential Statistics

The Wilcoxon signed-rank test is a non-parametric test that is used to compare two sets of scores that come from the same population (Wilcoxon 1945). The Wilcoxon signed-rank test has a null hypothesis that there is no significant statistical difference between the performances of two or more models (Tien Bui et al. 2016). Furthermore, a linear regression analysis was used in order to analyze the potential relationship between the two models. The R^2 value also known as the coefficient of multiple determination is a numeric measure of how much of the variation in the response variable (in the Y-axis) can be explained by variation in the predictor variable (in the X-axis).

Results

Logistic Regression Model

The training dataset was evaluated using Cox and Snell R^2 and Nagelkerke R^2 tests (Table 1) indicating a good performance, while the accuracy of classification during the training and validation phase was also calculated.

The outcomes of the experiment showed that 87.43% of the instances during the training phase were correctly classified. During the validation phase, the LR model achieved an accuracy of 80.92%. The area under the success and prediction rate curve for the model was calculated 0.9372 and 0.8903, respectively.

The logit of f(x) function was calculated for all of the grids of the Wuyaun County, in which zero (0) corresponds to no susceptibility and one (1) to total susceptibility. Based on constant values that were calculated, the logistic regression was compiled according to Eq. (2), while the possibility of landslide occurrence in each grid was calculated from Eq. (1) the outcome of which produced the landslide susceptibility map (Fig. 2).

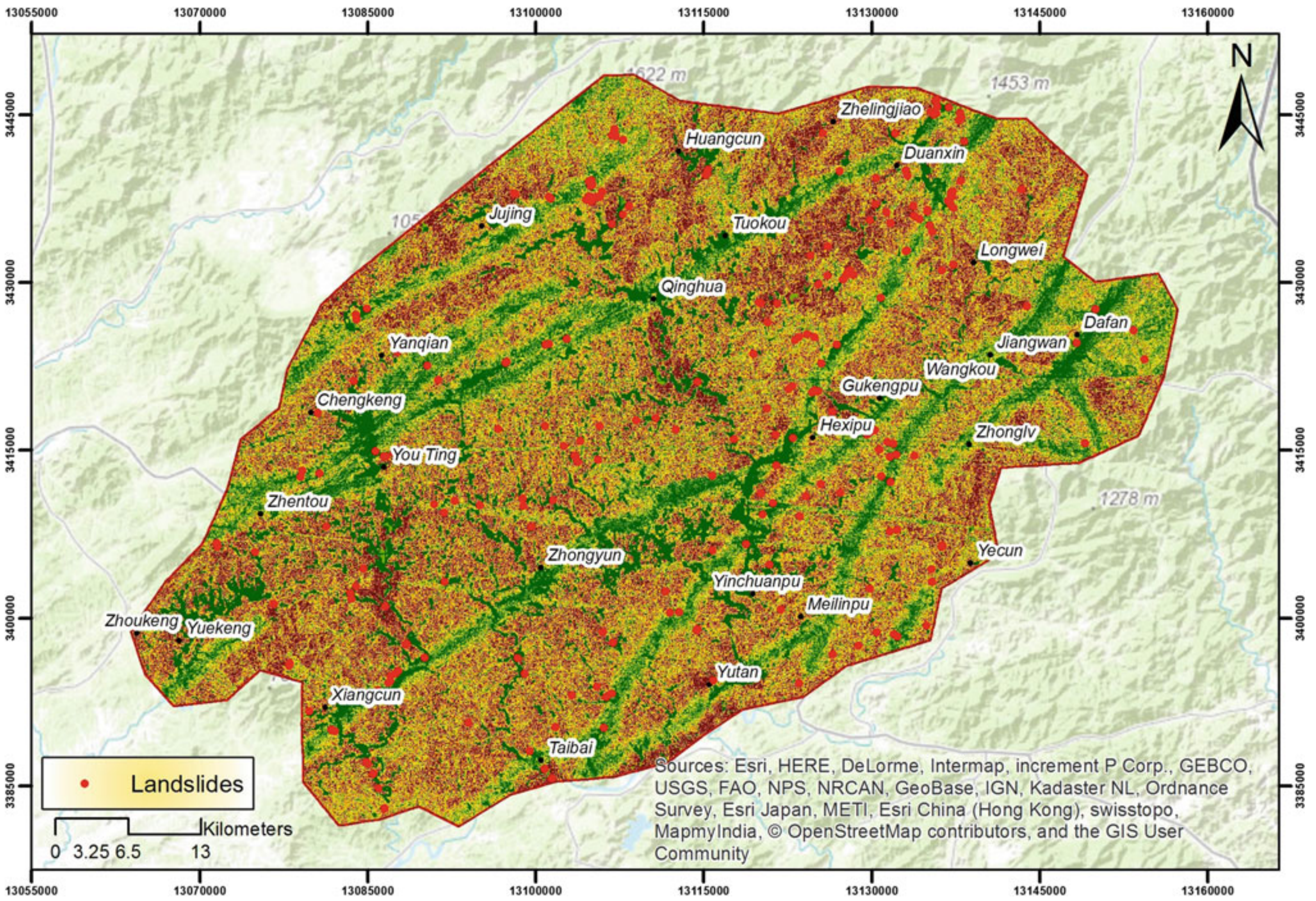
From the outcomes of the LR analysis it was induced that the variables SPI, STI, distance to rivers and distance to faults affect the LR function positively, while the highest b coefficient is allocated to distance to faults and SPI, which was 0.7167 and 0.6980, respectively. The rest of the variables have a negative effect on the landslide occurrence as they have negative b coefficients.

Random Forest Model

To implement successively the RF method, there is a need to estimate the minimum number of trees required to minimize the Out-Of-Bag error and also the need to estimate the number of variables randomly sampled as candidates at each split. From the conducted analysis it was estimated that the

Table 1 The overall statistics

−2 log likelihood	Cox and snell R^2	Nagelkerke R^2
295.69	0.48	0.64



Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Fig. 2 Landslide susceptibility logistic regression model

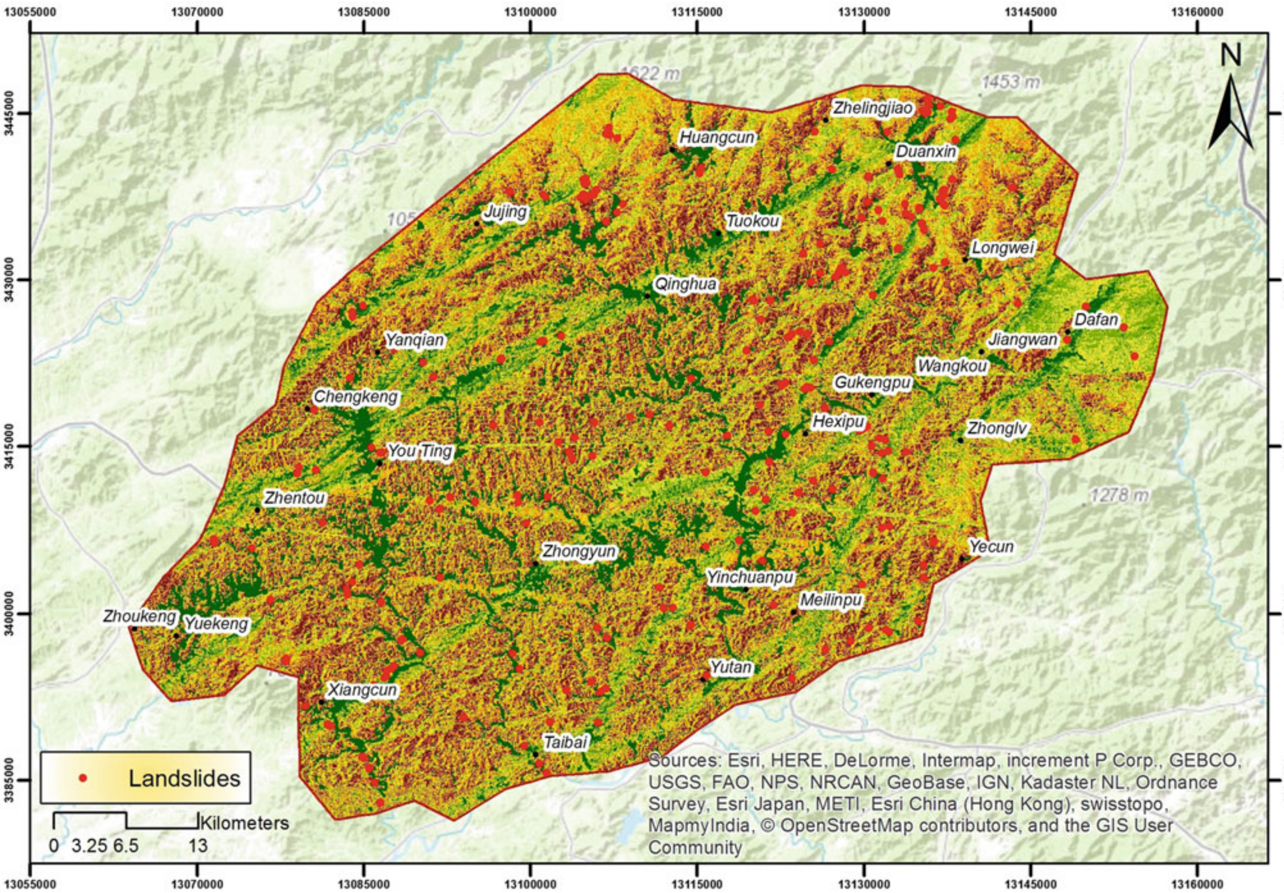
optimal performance was achieved for the RF model by using two (2) random variables at each split and 500 trees.

After the training phase ended, some extra information about the influence of each variable has on the overall landslide susceptibility analysis followed by the RF method, was gained. Specifically, the analysis calculated ordered the variables by the mean decrease accuracy and the mean decrease Gini. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting RF model, while the mean decrease in accuracy a variable causes is determined during the Out-Of-Bag error calculation phase. The more the accuracy of the RF due to the exclusion of a variable, the more important that variable is assumed, thus variables with a large mean decrease in accuracy are more important. According, to those two metrics, the most important variable is altitude followed by TWI and lithology.

Figure 3 illustrates the landslide susceptibility map constructed according to the RF method.

Linear Regression Analysis and Inferential Statistics

Performing the Wilcoxon signed-rank test at a 95% significant level the p-value was estimated to be 0.000 (less than 0.05), while the z value (−7.302) exceeded the critical values of z (−1.96 and +1.96) indicating that the performance of the susceptibility models was significantly different. In order to assess further the landslide susceptibility values that the two models produced, 1000 random points which covered the entire research area where generated and their susceptibility values were obtained. Descriptive statistics revealed that LR model had a mean value of 0.7639 and a standard deviation value of 0.2867, while RF model had a mean value of 0.7385 and a standard deviation value of 0.2733. In addition, the produced susceptibility values of the LR model were higher than the values of the RF model in 621 out of the total 1000 points. Regarding the linear regression analysis and the performed Analysis of Variance, it revealed that a strong evidence of linear relationship between the two landslide



Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Fig. 3 Landslide susceptibility random forest model

susceptibility maps exist, having a p-value less than 0.0001 at a 95% confidence level and an R^2 value estimated to be 0.6993. The R^2 value indicates that 69.93% of the variability in the LR model can be explained by variation in the RF model (Fig. 4).

Discussion

Concerning the produced landslide susceptibility map from the RF model, the very high susceptibility class was estimated to cover 18.70% of the total research area, while the relative landslide density for the high and very high landslide susceptibility class was estimated to be 77.82%. Respectively, for the LR model, the very high susceptibility class was estimated to cover the 20.82% of the total research area, while the relative landslide density for the high and very high landslide susceptibility class was estimated to be 73.06%.

From the visual analysis of the landslide susceptibility map produced by the LR model, high and very high susceptible zones are located at the west and east mountainous areas, while the central area is characterized by very low to low susceptibility values. It is clear that the spatial pattern of the landslide susceptibility follows the distribution of the elevation and slope observed in the study area, since lowlands are characterized by very low to low susceptibility values. One can also observe a strong association between the lithological coverage and the landslide susceptibility values. Similar outcomes are observed from the visual analysis of the landslide susceptibility map produced by the RF model. It seems that the spatial distribution of landslide susceptibility values follow the pattern of altitude, lithology and the distance to river network. High and very high susceptible zones are located along the road network mainly at the west and east mountainous areas, while the central area is characterized by very low to low susceptibility values.

Concerning the performance of the two models, it was induced that the two models gave similar results in classifying incidence, however they were significantly different. Also, the linear regression analysis revealed a strong evidence of linear relationship between the landslide susceptibility maps produced by the two models. Both models have estimated different variables as the most important during the

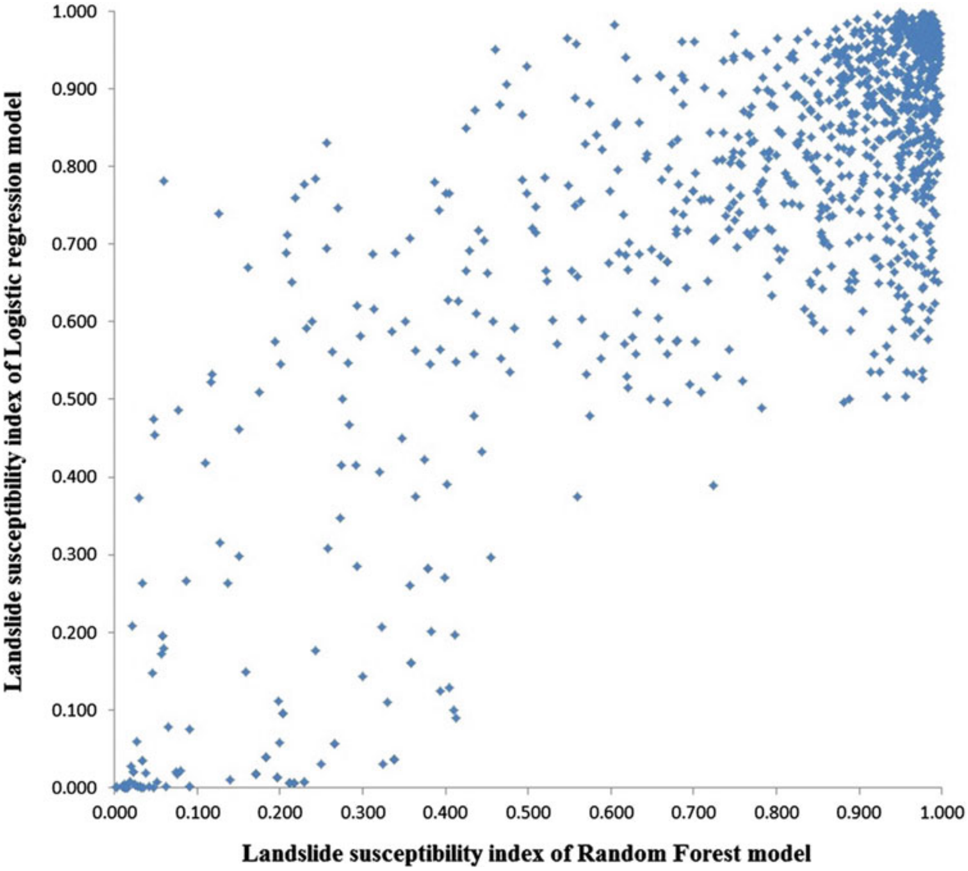


Fig. 4 Linear regression analysis

training phase except of the variable distance to faults that can be clearly seen in the produced susceptibility maps (Figs. 2 and 3).

Conclusions

In the present study, a LR and a RF model was applied for the construction of landslide susceptibility maps in the Wuyuan area, China. A total of 13 conditional factors were analyzed, namely, lithology, soil, slope, aspect, altitude, TWI, SPI, STI, plan curvature, profile curvature, distance to roads, distance to rivers and distance to faults. The landslide inventory database contained 255 locations that were divided into two subsets, one for training (70% of the total number of areas) and one for validating the model. The database was enriched with 255 locations of non-landslide areas that also were partitioned into training and validating datasets.

According to the analysis performed by the RF model, the most important variable was aspect followed by distance to faults and TWI. The analysis performed by the LR classifier showed that SPI, STI, distance to rivers and distance to faults affected the LR function positively, while the highest b coefficient was allocated to the variables distance to faults and STI.

The comparison of the two models revealed that the RF slightly outperformed the LR model with the area under the success and prediction rate curve calculated to be 0.9805 and 0.9324 respectively (for the RF model), and 0.9372 and 0.8903 respectively (for the LR model).

Concerning the potential linear relationship between the two models, the analysis revealed a strong evidence of linear relationship with 69.93% of the variability in the LR model explained by variation in the RF model.

Acknowledgements This research was supported by the National Natural Science Foundation of China (41472202, 41202235 and 41401032) and the General Program of Jiangxi Meteorological Bureau.

References

Akgun A, Kincal C, Pradhan B (2012) Application of remote sensing data and GIS for landslide risk assessment as an environmental threat to Izmir city (west Turkey). Environ Monit Assess 184:5453–5470

Ayalew L, Yamagishi H (2005) The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology 65:15–31

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breiman L, Freidman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. CRC Press, Wadsworth

Catani F, Lagomarsino D, Segoni S, Tofani V (2013) Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat Hazards Earth Syst Sci 13:2815–2831

Chen W, Chai H, Zhao Z, Wang Q, Hong H (2016) Landslide susceptibility mapping based on GIS and support vector machine models for the Qianyang County, China. Environ Earth Sci 75:474. doi:10.1007/s12665-015-5093-0

Dai FC, Lee CF, Ngai YY (2002) Landslide risk assessment and management: an overview. Eng Geol 64(1):65–87

Gorsevski PV, Gessler PE, Foltz RB, Elliot WJ (2006) Spatial prediction of landslide hazard using logistic regression and ROC analysis. Trans GIS 10(3):395–415

Hansen L, Salamon P (1990) Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 12:993–1001

Hong H, Naghibi SA, Pourghasemi H, Pradhan B (2016) GIS-based landslide spatial modeling in Ganzhou City, China. Arab J Geosci 9:112. doi:10.1007/s12517-015-2094-y

Ilia I, Tsangaratos P (2016) Applying weight of evidence method and sensitivity analysis to produce a landslide susceptibility map. Landslides 13(2):379–397

Korup O, Stolle A (2014) Landslide prediction from machine learning. Geol Today 30(1):26–33

Lee S, Pradhan B (2007) Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. Landslides 4(1):33–41

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

Pourghasemi HR, Mohammady M, Pradhan B (2012) Landslide susceptibility mapping using index of entropy and conditional probability models in GIS: Safarood Basin, Iran. CATENA 97:71–84

Tien Bui D, Tuan TA, Klempe H, Pradhan B, Revhaug I (2016) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides 13:361–378

Tsangaratos P, Ilia I (2016) Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: the influence of models complexity and training dataset size. CATENA 145:164–179

Wang L, Sawada K, Moriguchi S (2013) Landslide susceptibility analysis with logistic regression model based on FCM sampling strategy. Comput Geosci 57:81–92

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bull 1(6):80–83

Xu C, Xu X, Dai F, Wu Z, He H, Shi F, Wu X, Xu S (2013) Application of an incomplete landslide inventory, logistic regression model and its validation for landslide susceptibility mapping related to the May 12, 2008 Wenchuan earthquake of China. Nat Hazards 68:883–900

Yesilnacar E, Topal T (2005) Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). Eng Geol 79(3–4):251–266

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

Youssef AM, Pourghasemi HR, Pourtaghi Z, Al-Katheeri MM (2015) Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir region, Saudi Arabia. Landslides doi:10.1007/s10346-015-0614-1