Landslide Susceptibility Assessment Using C5.0 Model in Wanzhou District, China

Xia Hui[1] Yin Kunlong[2] Gui Lei[3] Chen Lixia[4] Zhang Wen[5] Fu Xiaolin[6]

[1]Graduate, [2]Professor, [3]Ph.D, [4]Associate Professor, China University of Geosciences, Wuhan, China
[5]Engineer, Geo-hazard Prevention and Control Centre, Chongqing, China
[6]Senior engineer, China Institute of Geological Environment Monitoring, Beijing, China
Corresponding author's E-mail: yinkl@126.com

Abstract

The main purpose of this study is to produce landslide susceptibility map in Wanzhou District, China using C5.0 model and analyse the performance of the model. A landslide database of 664 landslides was built by field geological surveys. To assess the landslide susceptibility assessment, 13 conditioning factors were extracted from the landslide database (i.e. elevation, slope angle, aspect, terrain ruggedness index, plan curvature, profile curvature, lithology, distance from structure, bedding structure, distance from rivers, topographic wetness index, stream power index and land use). The landslide sample data was randomly divided into two groups: 80% (training data) and 20% (validating data). C5.0 model was used to analysis the relationships between landslides and conditioning factors. The performances of the method were evaluated by the receiver operating characteristic (ROC) curve. The area under the relative operating characteristic curve (AUC) values for success rates were 0.974, prediction rates were 0.964 of C5.0 model. From the results of the study, C5.0 model has efficient and reliable performances in landslide susceptibility assessment.

Keywords: Landslide; Susceptibility assessment; C5.0; ROC curve.

1. INTRODUCTION

Landslide is one of the most destructive geological disasters all over the world because of wide distribution, high frequency, and fast movement in some cases. To alleviate damage from landslide hazards, it is crucial to evaluate these landslides, including landslide susceptibility, hazard and risk, and to assess their potential exposure to landslides and improved disaster preparedness (Felicísimo et al. 2012). The hypothesis of landslide susceptibility assessment is that landslides may occur in the future with a similar environment where landslides occurred in the past (Trigila et al. 2015).

Landslide susceptibility analysis methodologies proposed in the previous literature could be classified into three categories: statistical, deterministic, and heuristic (Ciurleo et al. 2016). With the rapid development of Geographical Information Systems and computing technology, the statistical methods have become more popular for landslide susceptibility assessment, including weights of evidence (Teerarungsigul et al. 2015), decision tree method (Tsangaratos and Ilia 2015), logistic regression (Erener et al. 2016), support vector machine (Kavzoglu et al. 2013) and artificial neural network (Dou et al. 2015). Kavzoglu (2013) mapped landslide susceptibility using multi-criteria decision analysis, logistic regression, and support vector regression and found that the performance of MCDA and SVR methods were more effectively than the LR method. Felicísimo (2012) compared multiple adaptive regression splines, logistic regression, maximum entropy, and classification and regression trees for landslide susceptibility assessment. The results of their study show that RLM and MAXENT were more stable than CART.

However, there are relative few studies that the performance of the C5.0 model. C4.5 algorithm, which is a kind of the decision tree method, was used for landslide susceptibility assessment (Yeon et al.

2010). C5.0 has never used to assess landslide susceptibility as an extension version of C4.5. Therefore, this study evaluated the performance C5.0 for landslide susceptibility assessment in Wanzhou District, China.

2. INFORMATION OF LANDSLIDES

2.1. Study area

Wanzhou is located in the middle of Three Gorges Reservoir, at northeast edge of Chongqing Municipality (Fig. 1). The study site lies between latitude 30°24′25″N and 31°14′58″N and longitude 107°55′22″E and 108°53′25″E. Wanzhou belongs to subtropical monsoon climate zone, with an annual mean precipitation of 1,194.4 mm (981.9mm lowest, 1,641.4 mm highest). Geologically, it borders TieFeng mountain anticline in the north, FangDou mountain anticline in the south, located in the shaft portion of Chuandong fold bundle and Wanxian synclinorium NE. The platform of Wanzhou city is comprised of floodplain, terraces and hilly tectonic denudation. The main exposed strata in the study area are Jurassic stratum and Triassic stratum. The main lithology in the study zone is consisted of purple sandstone, gray feldspar quartz sandstone, argillaceous siltstone, silty mudstone and mudstone. Furthermore, mudstone contains large amounts of montmorillonite, which has very well performance of expansion. Some of the strata in this area are horizontal practically, because the strata near the axial surface of the syncline are very soft. According to those factors, shallow translational landslides often happen in certain season and 664 historical landslide points were detected by field surveys.

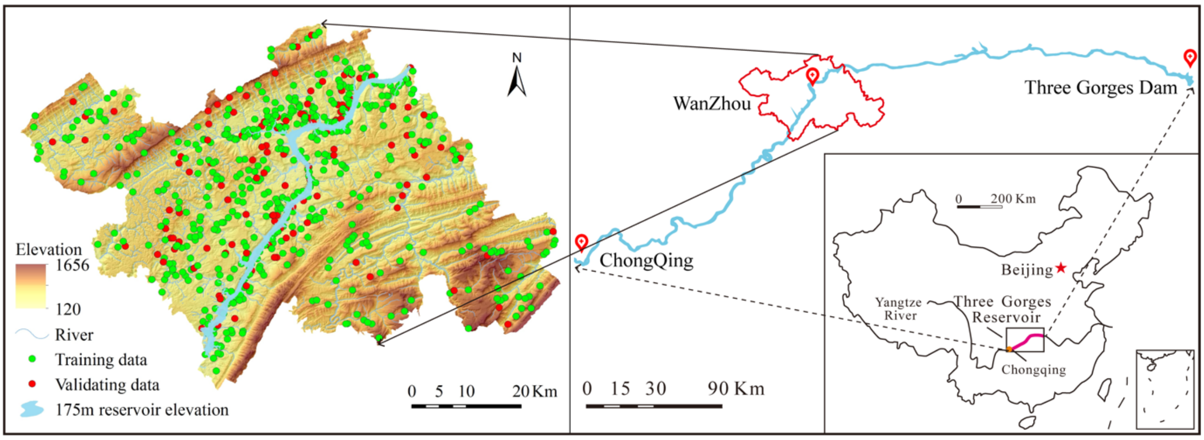

Figure 1. Location of the studied area and distribution of landslides in Wanzhou District, China.

2.2. Controlling factors

In the landslide susceptibility assessment, it is a crucial step that the selection of controlling parameters as input variables in the statistical models. It is also significantly influence the quality, and accuracy of the statistical model results (Trigila et al. 2015). In the study, 1: 20,000-scale tectonic map and 1: 50,000-scale digital topographic map were used to extract the controlling factors. Based on the software of ArcGIS, 13 factors were extracted from the database: elevation (Fig. 1), slope angle, aspect, terrain ruggedness index, lithology, plan curvature, profile curvature, distance from structure, bedding structure, distance from rivers, topographic wetness index, stream power index and land use (Fig. 2). Considering the scale of historical landslides and the precision of susceptibility map, all landslide explanatory factors were converted into a 25×25-m raster files. Data matrix used to calculate the mathematics models was extracted by the raster layers of the variables. The landslide sample data was randomly divided into 80% for training data and 20% for validating data.

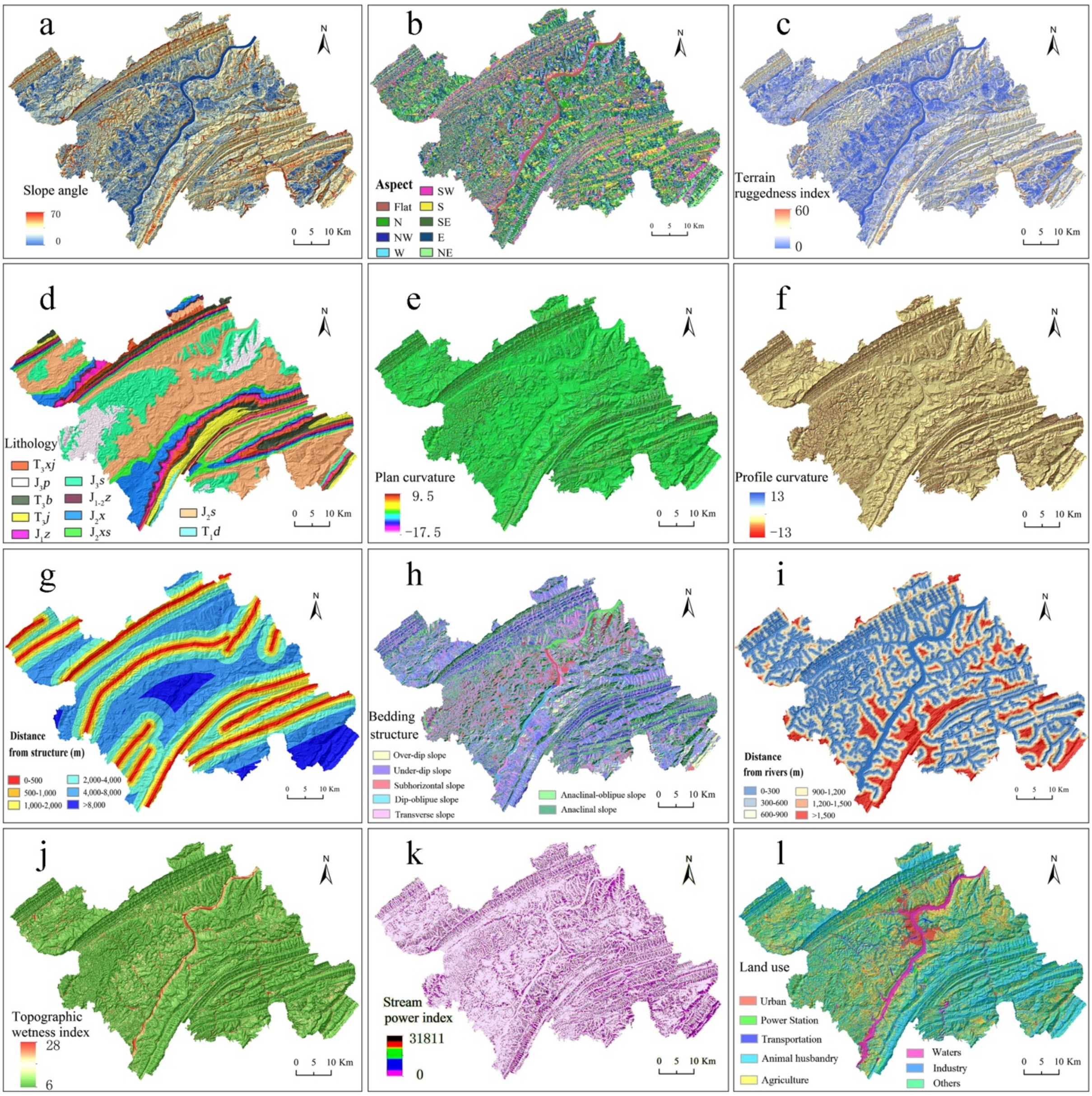

Figure 2. The map of landslide conditioning factors used in this study: (a)slope angle, (b)aspect, (c) terrain ruggedness index, (d)lithology, (e)plan curvature, (f)profile curvature, (g)distance from structure, (h)bedding structure, (i)distance from rivers, (j)topographic wetness index, (k)stream power index, (l)land use.

3. METHODOLOGY

The C5.0 algorithm was the development of C4.5 as an extension commercial version to improve the predictive ability of the model (Quinlan 1993). The information theory is mainly used to solve the problem of information transmission process. The information entropy is the mathematical expectation of amount of information, which is the average uncertainty before the information source sends out information. The information entropy equation is given as:

$$Ent(N) = \sum_{i=1}^{m} \frac{freq(C_i, N)}{|N|} \times \log_2 \left(\frac{freq(C_i, N)}{|N|} \right)$$

(1)

where N is the set of the sample data, C_i is a set of target variables (i = 1, 2, …, m, m is the number of classes), and $freq(C_i, N)$ is the relative probability of cases in C_i. *Ent(N)* is used to measure the average amount of information needed to identify the class of a case in N.

If the attribute variable N_i has n classification, then the conditional entropy of attribute variable N_i is defined as:

$$Ent_X(N) = \sum_{i=1}^{n} \frac{|N_i|}{|N|} \times Ent(N_i)$$

(2)

The information gain is used to measure the information that is obtained by divided N in according to the test X. The equation is as follows:

$$Gain(X) = Ent(N) - Ent_X(N)$$

(3)

The gain criterion could reflect the degree of information to eliminate random uncertainty.

4. RESULTS

4.1. C5.0 result

In this study, C5.0 model was constructed using the software of SPSS Modeler. Boosting and cross-validation were applied to improve the generalization ability of the model and prevent the model from over-fitting. At the tenfold boosting and cross-validation were used to construct and improve the accuracy of model. Eventually, a decision tree of depth 21 was obtained.

The landslide susceptibility index was calculated by brought all the raster data into the model created with the training data. The landslide susceptibility map produced by ArcGIS was shown in Figure 3. The very high and high susceptibility are mainly distributed in the interchange of Wanxian diagonal and Yangtze River, and the slope of between the side of the Fang Doushan anticline and the right bank of the Yangtze River. The elevation of these areas is mainly located in 200-400m. These areas are mainly exposed to Jurassic Shaximiao formation, Xintiangou formation and Suining formation.

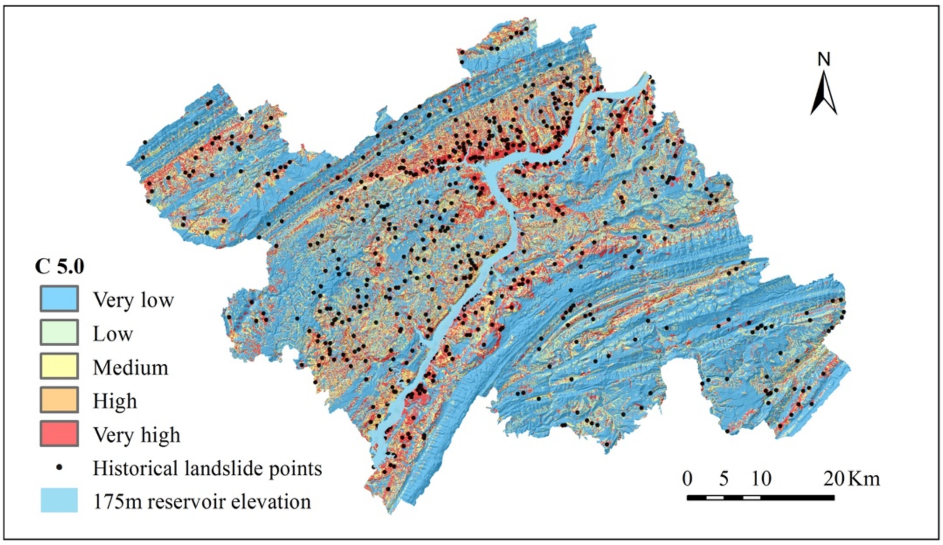

Figure 3. Landslide susceptibility map produced by C5.0 and MARS model.

4.2. ROC curve

The predictive performance of model was assessed by painting receiver operating characteristics (ROC) curve (David J. Goodenough 1973). In this study, the ROC curve was created using SPSS software package. The success-rate and prediction-rate curves from Wanzhou District are shown in Figure 4. The success-rate curves show that the area under the ROC curve (AUC) value was 0.974 (97.4%). In prediction rate curves, the AUC value was 0.964 (96.4%). The AUC values show that the C5.0 model gave very high success and prediction rates. It can be concluded the model for landside susceptibility mapping are reasonable because all success and prediction rate values above 0.7.

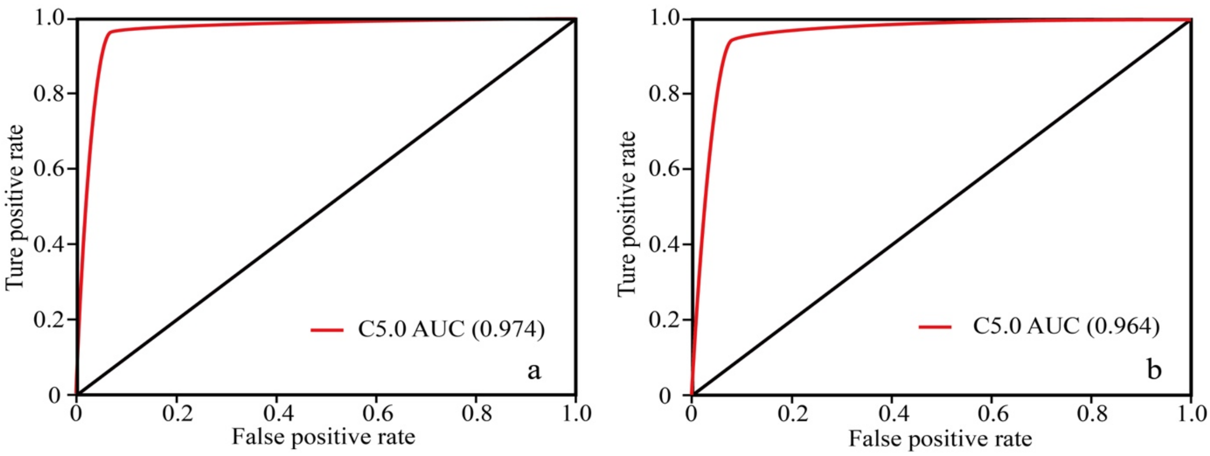

Figure 4. Success rate curves (a) and prediction rate curves (b) for the landslide susceptibility assessment produced in the study area.

5. CONCLUSIONS AND DISCUSSION

In this study, we have implemented C5.0 model for landslide susceptibility mapping at Wanzhou District, China. Thirteen landslide conditioning factors were extracted and used with the landslide database to produce the landslide susceptibility map. The performance of model was assessed by the area under receiver operating characteristics curve (AUC).

According to the landslide susceptibility maps produce by C5.0 model, the very high and high susceptibility zone are mainly distributed along the Yangtze River and some structural zones. The exposed strata are predominantly Jurassic Shaximiao formation, Xintiangou formation and Suining formation, which is consistent with the results of the data statistics.

The result of model comparison showed that the landslide susceptibility assessment produced by C5.0 model has very high success rate (97.4%) and prediction accuracy (96.4%). Therefore, it can be concluded that the C5.0 algorithm could be used efficiently to assess landslide susceptibility and the performance of C5.0 was accurately for landslide susceptibility mapping. According to the values of the AUC (>0.7), the landslide susceptibility map generated in this study area can be used for landslide hazard risk management and to guide for action on landslide mitigation by local policy makers, planners and engineers.

6. ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundations of China (No. 41572292, No. 41641012). We greatly thank Geological Hazard Monitoring Station of Wanzhou for providing the valuable data.

REFERENCES

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

David J. Goodenough PD, Kurt Rossmann, Ph.D.,and Lee B. Lusted (1973). Radiographic Applications of Receiver Operating Characteristic (ROC) Curvesl, Radiology, 110, 89-95.

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

Erener A, Mutlu A, Sebnem Düzgün H (2016). A comparative study for landslide susceptibility mapping using GIS-based multi-criteria decision analysis (MCDA), logistic regression (LR) and association rule mining (ARM), Engineering Geology, 203, 45-55.

Felicísimo ÁM, Cuartero A, Remondo J, Quirós E (2012). Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study, Landslides, 10, 175-189.

Kavzoglu T, Sahin EK, Colkesen I (2013). Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression, Landslides, 11, 425-439.

Quinlan JR (1993). C4.5: Programming for machine learning, Morgan Kaufmann Publishers, 21-22.

Teerarungsigul S, Torizin J, Fuchs M, Kühn F, Chonglakmani C (2015). An integrative approach for regional landslide susceptibility assessment using weight of evidence method: a case study of Yom River Basin, Phrae Province, Northern Thailand, Landslides, 13, 1151-1165.

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.

Tsangaratos P, Ilia I (2015). Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece, Landslides, 13, 305-320.

Yeon Y-K, Han J-G, Ryu KH (2010). Landslide susceptibility mapping in Injae, Korea, using a decision tree, Engineering Geology, 116, 274-283.