ISSN No:-2456-2165

Spatial Prediction of Landslides using Time Series Analysis and Support Vector Machine

Divya P, Geetha D S, Sivagami H, Sivagami S
Department of Information Technology
Velammal Institute of Technology
Chennai, India

Abstract:- Landslides are more prone in high altitude regions and determining them is challenging due to their unforeseen and sudden occurrence. As the technology is, improving day by day, the landslides are determined based on their spatial extent and the socio-economic losses can thereby be reduced. Many factors such as altitude, rainfall level, ground water level, reservoir water level, latitude, longitude, etc. ascertain their occurrence. One of the supervised Machine learning approaches called the Support Vector Machine (SVM) is used to predict whether there is a high probability of landslide occurrence in the given region. Time Series Analysis is used to find the direction and periodic propagation of landslides and the total amount of their deformation. Furthermore, Genetic Algorithm (GA) is for the further optimization and to reduce the mean square error of landslide susceptibility mapping. The prediction and validation results reveal that the proposed model can help in land use planning for shrinking the losses.

Keywords:- *Landslides; Support Vector Machine (SVM); Genetic algorithm (GA); prediction; trend component; periodic component.*

I. INTRODUCTION

Technology is advancing day by day and the development of technology has started long ago and has been there for a long as man can remember. The role of technology and its emergence in all the fields is helping in recovery from natural disasters. Natural disasters are sudden happenings, which cannot be stopped but can be predicted and prevention from losses is possible with enough knowledge and preparedness. Among all the natural disasters landslide is one from which the high altitude regions get affected adversely. There are many factors, which contribute to their occurrence. The major factors are altitude, slope, rainfall level, ground water level, reservoir water level, distance from road, human activities, topography, geology, lithology etc. Altitude is the height of the region from the sea level. If the region is at high altitude, then there is a higher probability of landslide occurrence. Many of the landslides are induced by rainfall because rainfall increases the porosity of the soil and lead them to erode. The mountainous areas with more rainfall are highly susceptible to landslide. The ground water level and the reservoir water level affect the lithological features of the region and cause the soil to loosen which thereby results in landslide. The topography of a region also plays a vital role in determining which part of the high altitude area will slide first and which will be affected more. The lithological factors are the soil features, its water content, chemical nature etc. After landslide occurrence, the deformation of the hilly area depends on the intensity of the landslides. The stochastic nature of these environmental factors, the nature of landslides and their nonlinear interrelationships make their prediction a challenging task. Several datamining approaches to landslide determination are suggested. Datamining in landslide prediction takes into account the spatial extent of the high altitude regions. The mining of landslide datasets involves the preparation of spatial database, pre-processing the data, finding out the appropriate influencing factors, analysis, prediction and validation of landslide models.

II. SURVEY ON PREVIOUS WORKS

A wide range of different methods and techniques have been used for landslide susceptibility modelling, such as the k-nearest neighbour, artificial neural networks, decision tree, logistic regression, boosted tree, Naive Bayesian classification, support vector machine, etc. Chung et al. (2003) used multi – layered spatial database, which contains all the influencing factors such as slope, lithological features, geomorphologic factors, etc. for determining the occurrence of landslides based on the obtained prediction images of the target area. Chang et al. (2007) applied the multisource data fusion approach for land slide classification using generalized positive Boolean functions. Jeremy et al. (2009) have used spatial data mining and geographic knowledge discovery to effectively analyse and predict the landslide consequence based on the available voluminous data obtained by remote sensing and other Global Positioning System (GPS) techniques. Choi et al. (2009) have applied the neural network model in landslide susceptibility mapping and have validated the model using the existing landslide data They have applied the neural network model at three study areas in Korea and have cross-applied their weight for landslide susceptibility mapping to achieve a reasonable prediction accuracy (81.36%).Cadan (2010) have used a hybrid learning method namely adaptive neuro - fuzzy inference system to analyse the external factors and to produce landslide susceptibility maps based on the obtained aerial photographs and satellite images of the target area. Five prediction models were developed using Sugeno approach for generating if - then rules which in turn predicts the landslide. The verification results showed that the model 5 has the highest accuracy. Pradhan and Lee (2010) have compared three landslide susceptibility maps generated by frequency ratio, multivariate logistic regression, and neural network model for the Penang Island

ISSN No:-2456-2165

and Selangor area in Malaysia. Dino et al. (2012) have devised a new hybrid approach of integrating K-nearest neighbour and support vector machine, which is the SVM-NN classification. This approach reduces the impact of prediction on the parameters. Omar et al. (2014) had used an ensemble algorithm of data mining Decision Tree based Chi-Squared Automatic Iteration Detection (CHAID).This algorithm predicts landslide by producing a multi - branched decision tree based on 13 conditioning factors and thereby achieving prediction rate of 79%. Hamid et al. (2015) conducted landslide assessment studies by comparing three data mining models namely Functional Trees (FT), Multilayer Perceptron Neural Networks (MLP Neural Nets) and Naïve Bayes (NB). The Area under Curve (AUC) was drawn and found that the MLP Neural Nets showed better accuracy than the other two models. Zhigang et al. (2016) proposed a new approach to establish landslide - forecasting model based on artificial neural networks (ANN) with random hidden weights. A lower - upper bound estimation (LUBE) method is used to construct ANN based prediction intervals (PI).A hybrid evolutionary algorithm combining particle swarm optimization(PSO) and gradient search algorithm (GSA) is utilized to optimize the output weights. Hyun and Saro (2017) used sophisticated data mining techniques namely artificial neural networks (ANN) and boosted tree (BT) for acquiring landslide susceptibility model. They found that these two models showed the validation result of 82.25% and 90.79% respectively. The main difference between above shown models and our proposed model is that the success rate of the present study is better than other devised models.

III. METHODOLOGY

A. *Proposed Approach*

The present study mainly focuses on improving the efficiency of landslide prediction. The architecture of the corresponding model is shown in Fig. 1.This prediction model uses three of the data mining algorithms such as Least Square Support Vector Machine (LSSVM), Genetic Algorithm (GA) and Time Series Analysis (TSA). The external factors (lithology, rainfall, ground water level, reservoir water level, topography, human activities, etc.) determine the vulnerability and consequences of landslide occurrence. The GA-LSSVM with TSA model has strong generalization ability and can effectively overcome the limitations of other methods including small sample sizes, high dimensionality and nonlinearity. The Genetic Algorithm is used to choose the appropriate factors affecting landslides. The selection of these parameters accurately helps in finding the probability of landslides effectively. The LSSVM stand a major prediction strategy in suspecting landslide occurrence. Based on the training and testing data, the areas with high and low probability for landslide occurrence are found. Time Series Analysis is used to determine the direction of landslide propagation thereby finding the amount of deformation by comparing the study area before and after landslide. The training data and the external driving factors are given as input to the prediction

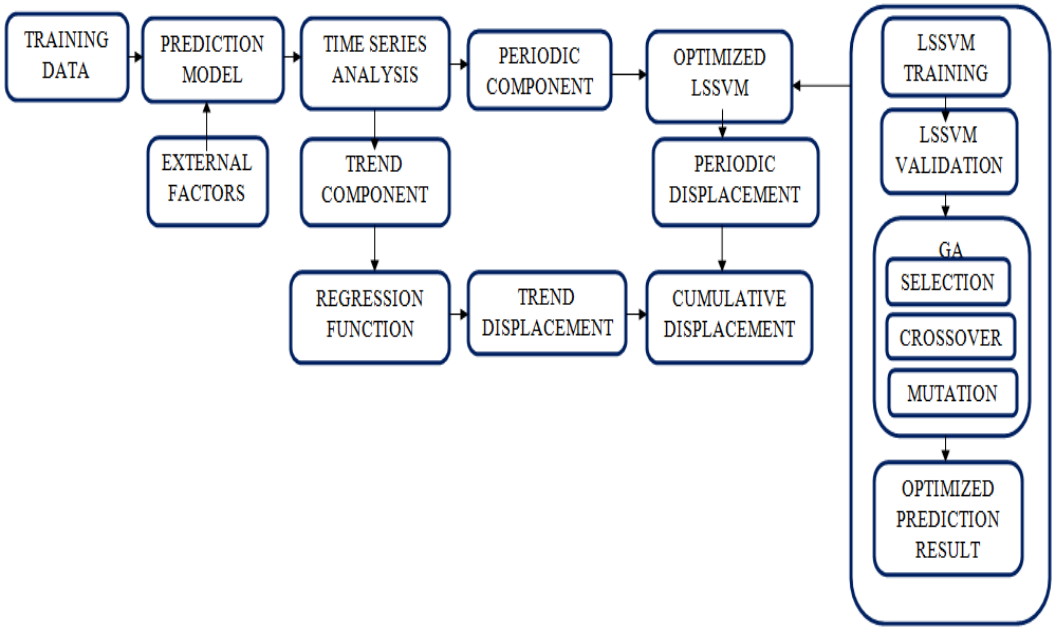

Fig 1:- Overall Design Architecture of GA – LSSVM Model with TSA

model. From the given data, TSA is done which results in finding out the trend and periodic components. The GA model optimizes the prediction model by choosing the appropriate parameters. The results of these models are given to LSSVM, which finally suspects the occurrence of landslides. The amount of deformation of the areas affected by landslides is found from the results of Time Series Analysis.

B. *Time Series Analysis*

Time Series analysis is done to forecast data based on time. The landslide data obtained through spatial data mining is incomplete and highly variable due to the conditioning factors, which cannot be monitored accurately.

ISSN No:-2456-2165

Time series analysis overcomes this shortcoming by periodically monitoring the data series. For any data (L_{tsa}), there are three components, which determine the nature and influence of them. To figure out the deformation of landslides the displacement of the land is found. As the landslides are non – linear in general, the direction of their future movements is determined by this prediction. The three components, which influence them, are; trend component (T_{tsa}), seasonal component (S_{tsa}) and a random error component (RE_{tsa}). This is given as in:

$$L_{tsa} = T_{tsa} + S_{tsa} + RE_{tsa}$$

Based on this the below algorithm for Time series analysis is formulated as follow

- *Algorithm Time Series Analysis*

- *Input*

Training samples of landslide data with their respective time of occurrence.

- *Output*

Total displacement (L_{tsa}) for the given landslide data.

T_{tsa} – Trend component of displacement
S_{tsa} – Seasonal component of displacement
RE_{tsa} – Random error

```
for sample in sorted(all_samples, key=lambda o: o.date):
   input_data, expected_prediction = sample
# Test on current test slice.
   actual_prediction = predictor. predict (input_data)
   errors.append(expected_prediction == actual_prediction)
   # Re-train on all "past" samples relative to the current time slice.
   training_samples.append (sample)
predictor = Predictor. train(training_samples)
for samples with actual_prediction != NULL:
```

$$L_{tsa} = T_{tsa} + S_{tsa} + RE_{tsa}$$

C. LSSVM Model

Least Square Support Vector Machine (LSSVM)is a supervised machine learning approach used for classification and regression analysis.It is a version of Support Vector Machine (SVM), which enhances the performance by reducing the mean square error and to resolve the limitations of SVM which includes small sample sizes, high dimensionality and nonlinearity. The training data samples are plotted through non – linear mapping. Then for that mapped data the regression function (f (x)) is determined as in:

$$f(x) = W^T \alpha(x) + \beta$$

where W^T is the weight vector, α (x) is the mapping function and β is an additional random constant.

D. Genetic Algorithm

GA is a family of population-based search algorithms for optimization problems. They maintain a set of solutions known as population. In each generation, it generates a new population from the current population using a given set of genetic operators known as crossover and mutation. It then replaces the inferior solutions by superior newly generated solutions to get a better current population. They use a parallel, random and adaptive searching based on natural biological selection and optimization. In landslide prediction, the GA is used to optimally choose the conditioning parameters and thereby improving the efficiency. All the parameters with both minor and major effects are taken into account in this model and so this algorithm achieves a fitness ratio of about 90%. The process of GA is diagrammatically shown in Fig. 2.The algorithm proceeds by improving the fitness ratio and the iteration stops when the ratio reaches the predetermined threshold value.

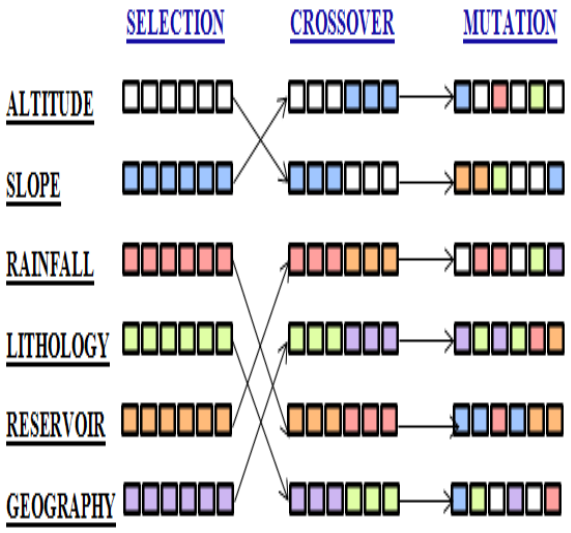

Fig 2:- Process of Genetic Algorithm

E. *Data Flow*

The prediction model suspects landslide as shown in the Fig. 3. The obtained spatial data of the area under study is preprocessed and divided into training and testing datasets. The training data with external conditioning factors and test data is given as input to the timeseies analysis (TSA) model from which the trend component and the periodic component is obtained. These two components determine the direction of landslide propagation and the amount of displacement based on the mathematical regression function.Simultaneously, the external factors are processed in genetic algorithm (GA) model and the appropriate parameters are found.With the time analysed data and the chosen parameters the LSSVM model is performed. The LSSVM model uses hyperplane to find the high and low probable areas which are to be affected by landlides from the given data. Based on the necessary steps are taken to avoid unnecessary losses.

ISSN No:-2456-2165

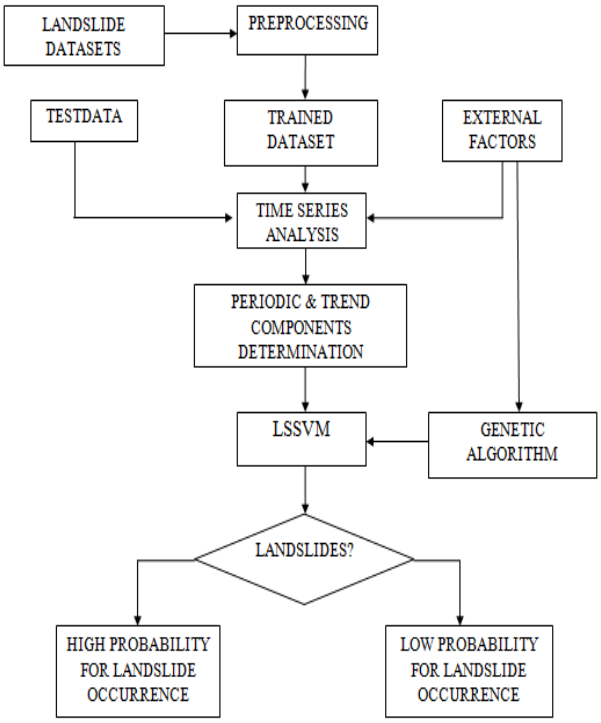

Fig 3:- Steps Involved in Landslide Prediction

IV. EXPERIMENTAL ANALYSIS &RESULTS

The experiment is performed with the sample datasets of North America for the year 2007 to

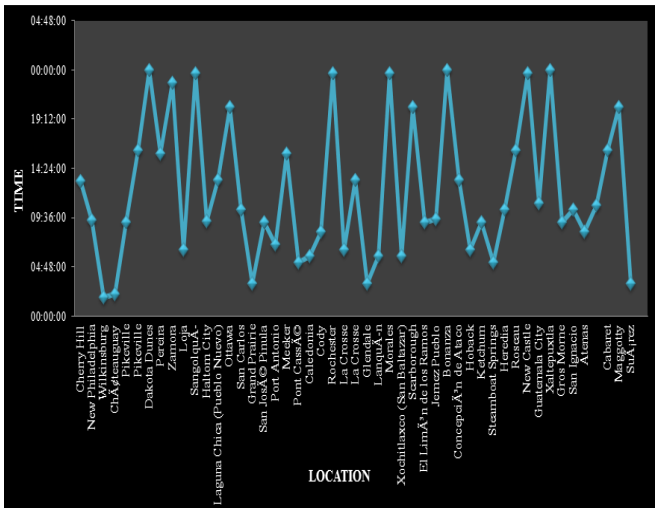

Fig 4:- Results of Time Series Analysis

2016 from Kaggle.com. The landslide vulnerable areas are found by implementing the three algorithms as mentioned in the architecture. The datasets are first analysed through time series from which the direction of landslides and their location at respective times is found. The result of TSA for the given datasets is shown in Fig 4. Then genetic algorithm is performed and found that the altitude, slope and rainfall of the region stand major influencing factors. Based on the above results, the LSSVM model is done and the regions are differentiated into high and low probable landslides. The obtained results of the above data from LSSVM model is shown in Fig.5

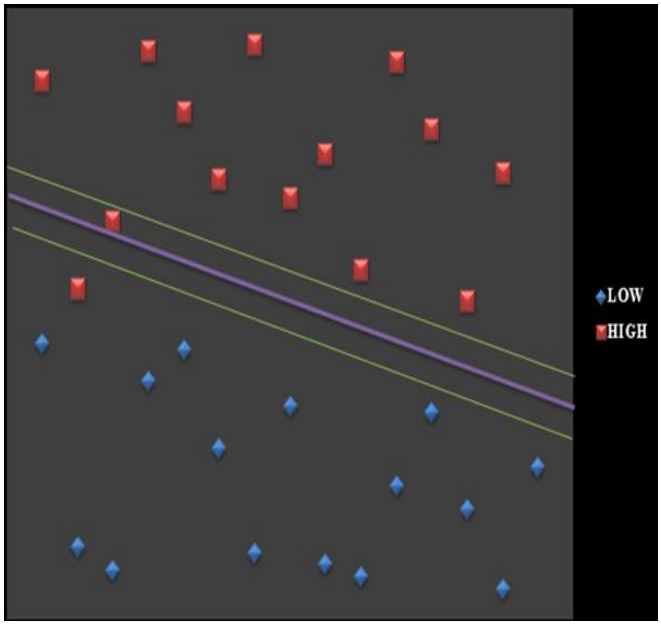

Fig 5:- Results of LSSVM Model Indicating Regions Having High& Low Probability for Landslides

V. CONCLUSION

There is huge volume of spatial data available worldwide. The proper handling of these data is essential in obtaining required results. Due to the widespread application of geographic information system (GIS) and global positioning system (GPS) technology, the access to high quality data is easily possible. Spatial data mining emerged as a result for utilizing the available landscape to satisfy our needs and to determine the natural land disasters prior to their occurrence. The unforeseen and vigorous nature of landslides and its consequences is an indisputable fact in all high altitude regions and so landslide prediction has become very important to avoid unnecessary losses. Many research have been taking place to predict landslides for which spatial data mining is highly helpful. To quantify the uncertainty and nonlinearity of landslides, many methods such as single layer artificial neural networks (S-ANN), multilayer artificial neural networks (M-ANN), Chi – squared Automatic Interaction Detection (CHAID), Naïve Bayes classification (NB) and boosted tree (BT) with prediction accuracies of 81.36%, 82.25%, 79%, 65.45% and 90.79% have been developed.

To improve the efficiency and to reduce the number of driving parameters the present approach of GA-LSSVM with TSA is proposed. The major objective of this study was to suspect landslides

ISSN No:-2456-2165

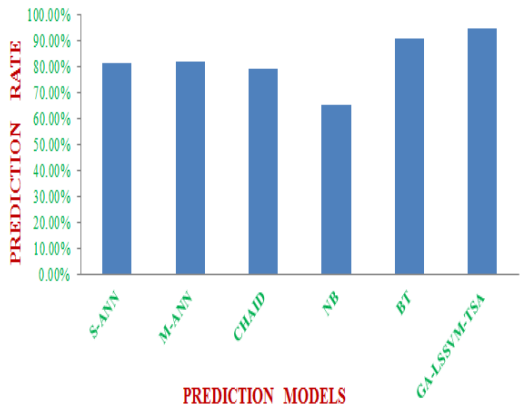

Fig 6:- Graph Comparing the Prediction Capabilities of Various Data mining Models

Nonlinearly in an optimized manner. Six landslide-conditioning factors such as lithology, altitude, slope, reservoir water level, rainfall and geography were exploited to accurately determine the landslide occurrence. The prediction capability was validated and compared with the prediction rates of other models mentioned above. This comparison is shown in the above graph Fig.6. Among all the models, GA – LSSVM with TSA showed better results. Thus, this method can be a better choice in land use planning and in getting rid of unnecessary losses.

REFERENCES

[1]. Zêzere, J.; Pereira, S.; Melo, R.; Oliveira, S.; Garcia, R. Mapping landslide susceptibility using data-driven methods. Science of the Total Environment 2017, 589, 250-267.

[2]. Catani F, Lagomarsino D, Segoni S, Tofani V (2013)Landslide susceptibility estimation by random forestsAmerican Journal of Geographic Information System 2017, 6(1A): 1-13 13 technique: sensitivity and scaling issues. Nat Hazards Earth Syst Sci 13:2815–2831.

[3]. Pourghasemi, HR., Kerle, N., 2016.Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran. Environ Earth Sci (2016) 75:185

[4]. Tsangaratos, P., Ilia, I., Hong, H., Chen, W., Xu, C., 2016. Applying Information Theory and GIS-based quantitative methods to produce landslide susceptibility maps in Nancheng County, China. Landslides, DOI: 10.1007/s10346-016-0769

[5]. Tien Bui, D.; Pradhan, B.; Nampak, H.; Quang Bui, T.; Tran, Q.-A.; Nguyen, Q.P. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibility modelling in a high-frequency tropical cyclone area using gis. Journal of Hydrology 2016, 540, 317-330.

[6]. Turner, D., Lucieer, A., and de Jong, S.: Time series analysis of landslide dynamics using an unmanned 1 aerial vehicle (UAV), Remote Sens-Basel, 7, 1736-1757, 10.3390/rs70201736, 2015.

[7]. Dou, J.; Tien Bui, D.; P. Yunus, A.; Jia, K.; Song, X.; Revhaug, I.; Xia, H.; Zhu, Z. Optimization of causative factors for landslide susceptibility evaluation using remote sensing and gis data in parts of japan, *PLoS One* 2015, *10*, e0133262.

[8]. Youssef, AM., Pourghasemi, HR., Pourtaghi, ZS.,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.

[9]. Shah, A.D.; Bartlett, J.W.; Carpenter, J.; Nicholas, O.; Hemingway, H. Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study. American Journal of Epidemiology 2014, 179, 764-774.

[10]. Sdao F., Lioi, DS., Pascale, S., Caniani, D., Mancini, IM., 2013. Landslide susceptibility assessment by using a neuro-fuzzy model: a case study in the Rupestrian heritage rich area of Matera. Nat. Hazards Earth Syst. Sci., 13, 395–407.

[11]. Pourghasemi, HR., Jirandeh, AG., Pradhan, B., Xu, C, Gokceoglu, C. 2013. Landslide susceptibility mapping using support vector machine and GIS at the Golestan province, Iran. J Earth Syst Sci. 122:349–369.

[12]. Yao, W., Zeng, Z. G., Lian, C., and Tang, H. M.: Ensembles of echo state networks for time series prediction. In: Sixth international conference on advanced computational intelligence, Hangzhou, China, 299-304, 2013.

[13]. Xu, H., and Chen, G.: An intelligent fault identification method of rolling bearings based on LSSVM optimized by improved PSO, Mech Syst Signal Pr, 35, 167-175, 10.1016/j.ymssp.2012.09.005, 2013.

[14]. Pradhan, B. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using gis. *Computers & Geosciences* 2013, *51*, 350-365.

[15]. Mohammady, M., Pourghasemi, HR., Pradhan, B., 2012.Landslide susceptibility mapping at Golestan Province, Iran: a comparison between frequency ratio, Dempster-Shafer, and weights-of-evidence models.Asian Earth Sci 61:221–236

[16]. Vahidnia, MH., Alesheikh, AA., Alimohammadi, A., Hosseinali, F., 2010. A GIS-based neuro-fuzzy procedure for integrating knowledgeand data in landslide susceptibility mapping. Computers & Geosciences, 36(29), 1101–1114.

[17]. Yeon, YK., Han, JG., Ryu, KH., 2010. Landslide susceptibility mapping in Injae, Korea, using a decision tree. Engineering Geology, 16(3–4), 274–283.

[18]. Saito, H., Nakayama, D., Matsuyama, H., 2009. Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: the Akaishi mountains, Japan. Geomorphology, 109(3–4), 108–121.

[19]. A. Akgün, S. Dag, and F. Bulut, “Landslide susceptibility mapping fora landslide-prone area (Findikli, NE of Turkey) by likelihood-frequencyratio

and weighted linear combination models," Environ. Geol., vol. 54,no. 6, pp. 1127–1143, May 2008.
[20]. Francke, T.; López-Tarazón, J.; Schröder, B. Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. Hydrological Processes 2008, 22, 4892-4904.
[21]. Chung, C.-J.; Fabbri, A.G. Predicting landslides for risk analysis — spatial models tested by a cross-validation technique. *Geomorphology* 2008, *94*, 438-452.
[22]. Fell, R.; Corominas, J.; Bonnard, C.; Cascini, L.; Leroi, E.; Savage, W.Z. Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning. *Engineering Geology* 2008, *102*, 99-111.
[23]. Yin, X., and Yu, W.: The virtual manufacturing model of the worsted yarn based on artificial neural networks and grey theory, Appl Math Comput, 185, 322-332, 10.1016/j.amc.2006.06.117, 2007.
[24]. S. Lee and B. Pradhan, "Landslide hazard mapping at Selangor, Malaysiausing frequency ratio and logistic regression models," Landslides, vol. 4,no. 1, pp. 33–41, Mar. 2007.
[25]. S. Lee and B. Pradhan, "Probabilistic landslide risk mapping at PenangIsland, Malaysia," J. Earth Syst. Sci., vol. 115, no. 6, pp. 661–672,Dec. 2006.
[26]. B. Pradhan, R. P. Singh, and M. F. Buchroithner, "Estimation of stressand its use in evaluation of landslide prone regions using remote sensingdata," Adv. Space Res., vol. 37, no. 4, pp. 698–709, 2006.
[27]. S. Lee and M. J. Lee, "Detecting landslide location using KOMPSAT 1and its application to landslide-susceptibility mapping at the Gangneungarea, Korea," Adv. Space Res., vol. 38, no. 10, pp. 2261–2271, 2006.
[28]. Vandenbergh, F., and Engelbrecht, A. P.: A study of particle swarm optimization particle trajectories, Inform Sciences, 176, 4 937-971, 10.1016/j.ins.2005.02.003, 2006.
[29]. Wang, Y., Yin, K. L., and An, G. F.: Grey correlation analysis of sensitive factors of landslide, Rock Soil Mech, 25, 91-93, 2004(in8 Chinese).
[30]. Wang, J. F.: Quantitative prediction of landslide using S-curve, Chin J Geol Hazard Control, 14, 3-10, 2003(in Chinese).
[31]. F. Guzzetti, A. Carrarra, M. Cardinali, and P. Reichenbach, "Landslidehazard evaluation: A review of current techniques and their applicationin a multi-scale study, Central Italy," Geomorphology, vol. 31, no. 1–4,pp. 181–216, Dec. 1999.
[32]. Gokceoglu, C.; Aksoy, H. Landslide susceptibility mapping of the slopes in the residual soils of the mengen region (turkey) by deterministic stability analyses and image processing techniques.*Engineering Geology* 1996, *44*, 147-161.
[33]. Vapnik, V.: The nature of statistical learning theory, Springer Verlag, New York, 1995.