Ensemble of Data-Driven EBF model with Knowledge Based AHP Model for Slope Failure Assessment in GIS Using Cluster Pattern Inventory

Omar F. ALTHUWAYNEE and Biswajeet PRADHAN, Malaysia

Key words: Landslide, GIS, Spatial Pattern Analysis, Nearest Neighbor Index, Ensemble method, Malaysia.

SUMMARY

A novel methodology is proposed to predict rainfall-induced susceptibility map in Kuala Lumpur city and surrounding areas using geographic information system (GIS). A landslide inventory map consisting of 220 landslide locations represent the only dependent factor, were collected using historical landslide data, and Fourteen landslide conditioning factors prepared to represent the independents factors in the analysis. The methodology started by preparing the inventory to be tested for randomness using nearest neighbour index (NNI) method, then the cluster pattern landslides were extracted, in addition to the random cluster landslides that both used separately as training data for comparison purpose. An ensemble model was developed using evidential belief function (EBF) and Analytic Hierarchy Process (AHP) model. Later, the prediction accuracy of area under the curve (AUC) used to test the ensemble prediction rate, the AUC showed 83.5% with cluster locations, and 82% with random pattern locations. As a conclusion, The proposed ensemble model managed to optimize the input layers, which can be served as major research advancement in areas where landslide inventory map is scarce.

Ensemble of Data-Driven EBF model with Knowledge Based AHP Model for Slope Failure Assessment in GIS Using Cluster Pattern Inventory

Omar F. ALTHUWAYNEE and Biswajeet PRADHAN, Malaysia

ABSTRACT

In this study, an ensemble model was developed using a data-driven evidential belief function (EBF) and knowledge-based Analytic Hierarchy Process (AHP) model. The ensemble model was developed to overcome the subjectivity of the expert opinion in AHP model, as a semi-quantitative model and also to find the inter-relationship importance between landslide causative factors. Firstly, two different inventory patterns were used to compare its prediction accuracy, i.e. 1st random pattern, 2nd cluster pattern. For inventory mapping, a total of 220 landslide locations were collected using historical landslide location data, and classified into training and testing data. The training data were tested for randomness in previous study using nearest neighbour index (NNI) technique. The test results show a large percentage of cluster patterns in training data (88%). The cluster locations were used to train 14 landslide conditioning factors derived from various sources: topographic derived parameters, lithology, normalized difference vegetation index (NDVI) and landuse/landcover map. For model validation, an area under the curve (AUC) of ensemble prediction map, showed 83.5% with cluster locations, and 82% with random pattern locations. The proposed methodology enhanced the previous research's results to predict rainfall-induced susceptibility map in Kuala Lumpur city and surrounding areas using geographic information system (GIS). Based on the findings, one can infer that the clustered data can be effectively used as training data with ensemble model instead of random selection technique. As a conclusion, the final result can provide a valuable scientific basis for spatial decision making in planning and urban management studies.

1. INTRODUCTION

Landslides in Malaysia mostly occur during the heavy monsoon season. Also, anthropogenic factors such as deforestation and unplanned developmental play important roles in initiation of the landslides. Spatial pattern in landslide inventory plays a vital role in predictive analysis. Generally, landslides are frequently distributed in cluster pattern groups both in space and time than a disperse pattern (Jarman, 2006). Cluster pattern can be described as, high density of events occurring in specific location than other locations. Moreover, the random simulation test of data distribution should reject the hypothesis of independency among the events. In an earlier paper, Keeper, (1984) concluded that landslides triggered by earthquakes, have more tendencies to occur at well-defined location around the epicenter. In this paper, a second-order statistics i.e. Nearest Neighbor Index (NNI) (Clark and Evans, 1954), was used to extract the average mean distance for pattern of landslides events in Kuala Lumpur and surrounding areas.

In modeling process, such spatial associations need to be performed in parallel; first, landslide events with each single landslide causative factors, second, the inter-relationship between all the causative factors which are non-linear in nature. Therefore, there exist some drawbacks of using the bivariate analysis. This is because bivariate analysis describes the spatial relationship between known landslides with each single causative factor, but is not considered as the inter-relationship between the causative factors together simultaneously. Moreover, it owes much to the inappropriate ranking or optimizing the important landslide causative factors. On the contrary, in multivariate analysis, the aforementioned limitations are not evident, and the inter-relationship is clearly found and interpreted as factors weights. Consequently, an ensemble methodology was developed using data-driven Evidential Belief Function (EBF) model to find the relationship between each single conditioning factor and landslide locations, then a knowledge-based Analytic Hierarchy Process (AHP) model was used to find the inter-relationships importance between each conditioning factors. Finally, both the models were ensemble together to eliminate the drawback of individual models when applied separately. The main objective of this paper is to compare the prediction accuracy of the application of ensemble methodology by using cluster pattern and random pattern landslides locations. In order to test the proposed ensemble mode, a landslide-prone area of Kuala Lumpur and vicinity areas from tropical Malaysia was chosen.

2. STUDY AREA

Kuala Lumpur and vicinity areas, plays a major role in economic and social development in Malaysia. During the monsoon, the area receive high amount of precipitation that weakness the slope stability (Pradhan, 2011; Pradhan and Lee, 2007). The study area is enclosed geographically between 2°56'N to 3°20'N latitude and 101°29'E to 101°50'E longitude, with approximate area of 1975 km^2, (Fig.1).
The major types of landover in the study area comprises of settlement, peat swamp forest, and abandoned mining, grassland and few shrub areas. The overall temperature of the area ranges between 29 to 32° C. The average precipitation varies from 58 to 240 (mm/month), which trap large amount of water leading to a high pore-water pressure that decreases the shear stress stability (Malaysian Meteorological Services Department). Also, the deforestation activity play some role in destabilizing the slopes (Evett et al., 2006). More about geological and geomorphological characteristics of the study area can be seen in Althuwaynee et al. (2012a, b).

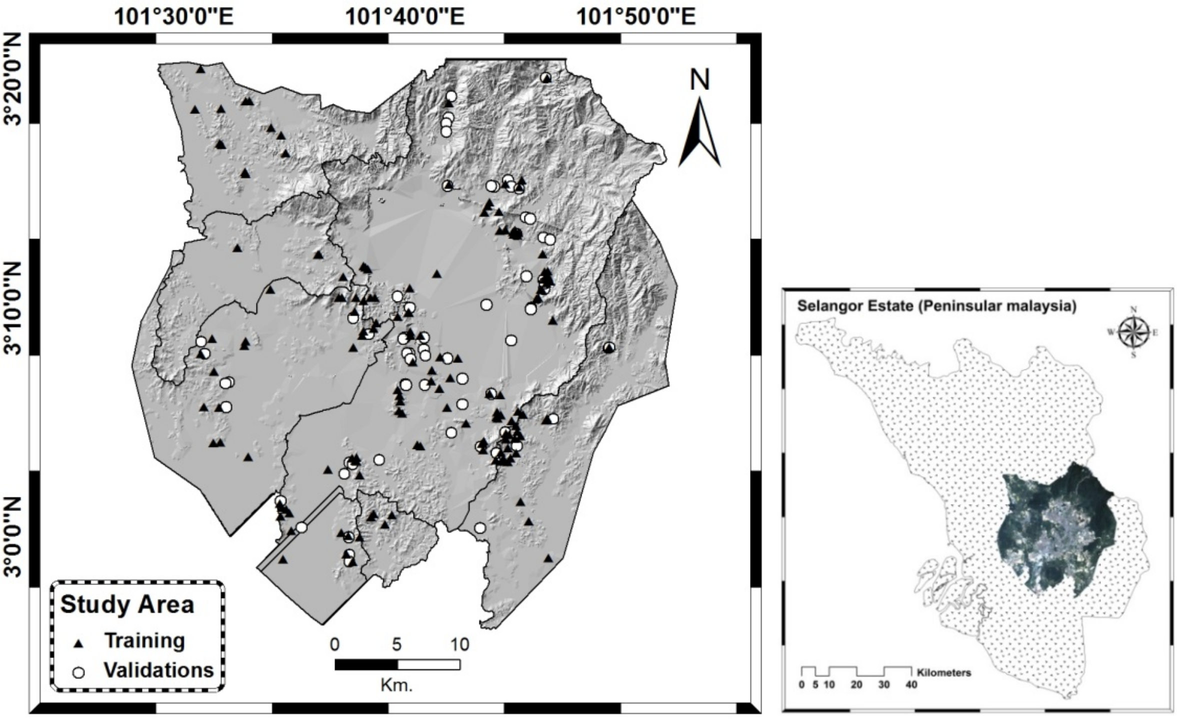


Figure 1. Study area

2.1 Spatial data

Since this is a follow-up article toAlthuwaynee et al., 2012a; 2012b; therefore the basic data and other characteristics of the study area can be referred to those aforementioned articles. The landslide inventory consists of 219 landslides, was collected over the past 25 years mainly by using various sources including remotely sensed sensors:

- The landslide inventory map was prepared based on archived 1:5000–1:50,000 aerial photos, SPOT 5 panchromatic satellite image and archived landslide location map and previous reports.
- A 1:25,000 scale topographic map was used to model a digital elevation model,
- A geological map at scale 1:63,000 was used to produce the lithological map and distance from the faults.
- The soil map at the scale 1:100,000 were used to extract the soil properties.
- The precipitation map was prepared using the past 29 years (1981–2010) of rainfall data.
- An Enhanced Thematic Mapper (ETM+) satellite data was used to extract the land cover and NDVI (Normalize difference vegetation index) maps.

3. METHODOLOGY

The methodology employed here is built up from three main stages, (Fig.2):

1. The NNI method tested the spatial pattern of landslides, and the result showed a clustered pattern tendency. Now, the output will be used as a training dataset in next stage.
2. EBF as bivariate analysis model was used to quantifying the relationship between each individual causative factors and landslide locations using cluster pattern inventory.
3. An analytical hierarchy process (AHP), was used (Saaty, 1977) to quantify the EBF results and determine each predictors importance.

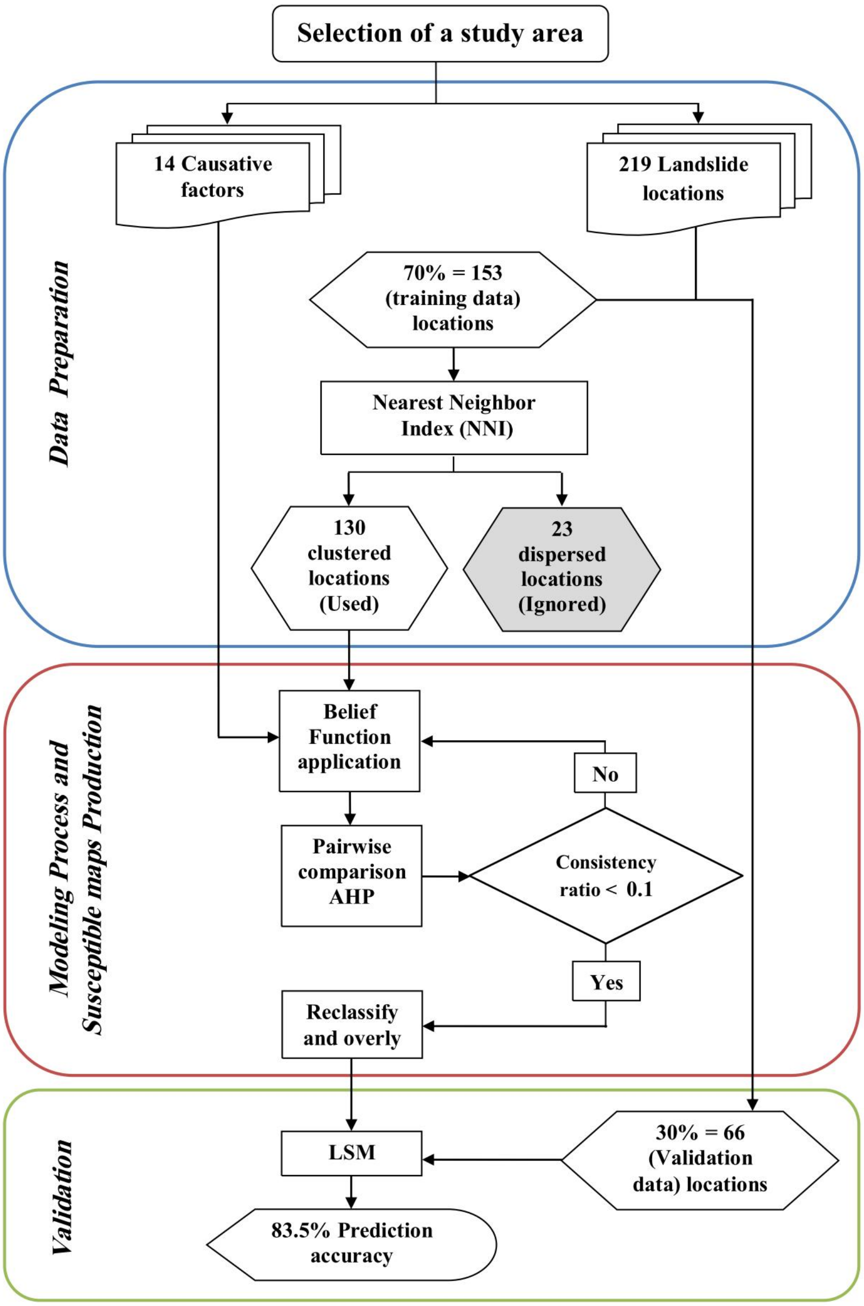


Figure 2. Flow chart of the proposed methodology used in this study

3.1 Spatial association analysis for conditioning factors using EBF

The EBF as bivariate model was applied to measure the spatial association between the landslides and conditioning factors. A detail description and algorithm about the EBF can be seen in (Althuwaynee et al., 2012a). Eq. (1) shows the algorithm behind belief map (Bel) result, which represent the probability result, which was constructed based on, (L) numbers of multiple spatial data layers in which each layer is considered as an evidence; and Eij represents the probability values of Bel found for each factor using Eq. (2).

$$\lambda(\mathrm{Tp})\mathrm{Eij} = \frac{\mathrm{N(L \cap Eij)/N(L)}}{\mathrm{N(Eij) - N(L \cap Eij)/N(A) - N(L)}} = \frac{\mathrm{N}}{\mathrm{D}} \qquad (1)$$

where:

i:	Amount of layers
j:	Class attribute
N (L∩Eij)	Number of landslide pixels in domain
N (L):	Total number of landslide, or ∑ N (L∩Eij)
N (Eij):	Number of pixel in domain
N (A):	Total number of pixels in domain, or ∑ N(Eij).
N:	Proportion of landslide occurrence
D:	Proportion of non-landslide area

$$\mathrm{Bel} = \frac{\lambda(\mathrm{Tp})\mathrm{Eij}}{\sum \lambda(\mathrm{Tp})\mathrm{Eij}} \qquad (2)$$

Where:

Bel: belief result map, which represent the probability map

3.2 Weighting of causative factors by AHP integration

Generally, each spatial causative factor carries a different degree of negative or positive effect on prone slopes stability. Therefore, the quantified conditioning factors of belief (Bel), acts as an input data for pair-wise analysis instead of classic common 9-point pair-wise rating scale of experts' opinion. This analysis process was carried through three steps following to the work of (Ghosh et al., 2011). In the first step, predictor rating (PR) for every spatial factor was derived based on degree (importance) of spatial relationship with landslide training dataset. For this purpose, Eq. (3) was used to find the absolute difference between the maximum and minimum value of Index of spatial association (SA), and then was divided that by lowest absolute difference of all factors.

$$\mathrm{PR} = (\mathrm{SA_{max}} - \mathrm{SA_{min}})/(\mathrm{SA_{max}} - \mathrm{SA_{min}})_{\mathrm{min}} \qquad (3)$$

where:

SA: Index of spatial association (Bel)

In the second step, the fractional predictor was converted into integer weight by dividing the fractional weights by the smallest weight among all the predictors' fractions. Subsequently, in the third step, the consistency ratio (CR) was used to evaluate if the matrix judgments of expert's opinion was randomly generated or not (Ishizaka and Labib, 2009; Saaty, 1977). Basically, if CR value is greater than 0.1 then it represents the limit of inconsistency according to Eq. (4).

$$\mathrm{CR} = \mathrm{CI}/\mathrm{RI} \tag{4}$$

where:

RI: Average of the resultant consistency index, depends on the order of the given matrix

CI: Consistency index

Integer weights were extracted and the final map was produced using Eq. (5) by applying weighted multi-class index overlay method in ArcGIS software (Bonham-Carter, 1994).

$$\mathrm{S} = \sum_{l}^{n}(\mathrm{F_{ji}} \times \mathrm{W_i})/\sum_{l}^{n}\mathrm{W_i} \tag{5}$$

where:

S: Predictive model of susceptibility map

Wi: Integer weights

Finally, the results were evaluated by using the landslides locations which was not used during the model building process.

4. RESULTS AND DISCUSSION

The study hypothesize state that using the cluster pattern training landslide data is much worth than using the random pattern one. The ensemble methodology was successfully applied using the 130 cluster pattern landslides tested by NNI, with 14 causative factors (such as slope, aspect, land cover, soil type, lithology, altitude, NDVI, curvature, surface roughness, SPI (Stream Power Index), distance from road, distance from faults, distance from drains, and precipitation).

The efficiency of AHP model as sole knowledge based model has some weakness owing to the subjectivity of the ranking of the causative parameters. Therefore to overcome this drawback, a methodology is proposed by integrating data-driven EBF as quantitative model through a novel ensemble methodology.

Subsequently, the belief weights were obtained and listed in table 1, and the parameters weights were compared by pair-wise comparison method. The pair-wise comparison shows the priority of (roughness, elevation, SPI, lithology, and distance from faults) as the most effective prediction factors as shown in table 2.

Table 1. Belief weights of every class of categorical spatial factor

factor	Factor class	Bel.	Min.	Max.	[Max - Min]	Min tot.	PR
Slope	0 - 10	0.22	0.22	0.46	0.24	0.13	1.75
	10 -25.	0.46					
	26 - 35	0.32					
	36 - 85	0.00					
Soil	RGM	0.16	0.00	0.19	0.19		1.44
	STP	0.08					
	DLD	0.08					
	LAA_COL	0.12					
	MUN_SBN	0.00					
	SDG_MUN	0.19					
	SDG_KDH_DRN	0.06					
	ISA	0.06					
	SDG_MUN_KDH	0.05					
	MCA	0.00					
	SDG_MUN_SBN	0.17					
	SLR_BRH	0.03					
	KDH_SDG_MUN	0.00					

Table 2. Estimated eigenvectors of the pair-wise rating matrix and weights of predictors

Predictor	Slope	Aspect	Curve	Rough	Elev.	NDVI	SPI	Dis_road	Dis_drain	Litho.	Soil	Landcov.	Prec.	Dis_fault	∑sum	Fractional weight	Integer weight
Slope	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.76	0.055	**18**
Aspect	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.44	0.031	10
Curve	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.71	0.051	16
Rough	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	1.74	0.124	**40**
Elev.	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	1.54	0.110	**35**
NDVI	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.53	0.038	12
SPI	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	1.49	0.107	**34**
Dis_road	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	1.05	0.075	**24**
Dis_drain	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.55	0.039	13
Litho.	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	1.36	0.097	**31**
Soil	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.63	0.045	14
Landcov.	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	1.04	0.074	**24**
Prec.	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.96	0.068	**22**
Dis_fault	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	1.20	0.085	**27**
∑sum	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	14.00	1.00	320.94

The consistency ratio (Ergu et al., 2011; Finan and Hurley, 1997) shows 0.01 value, which is reasonably good accuracy value, which reflect the high accuracy of ranking consistency between the factors (Fig. 3).

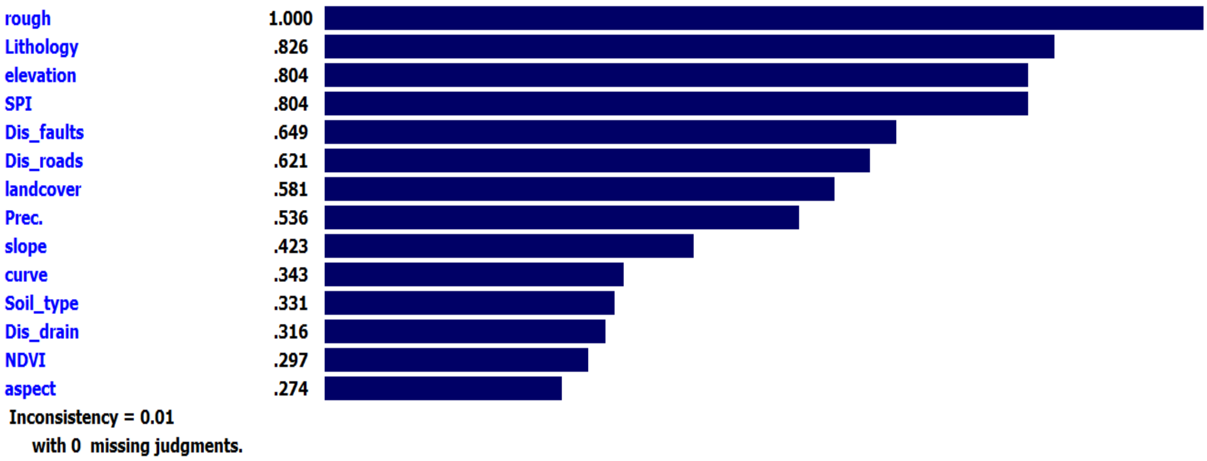
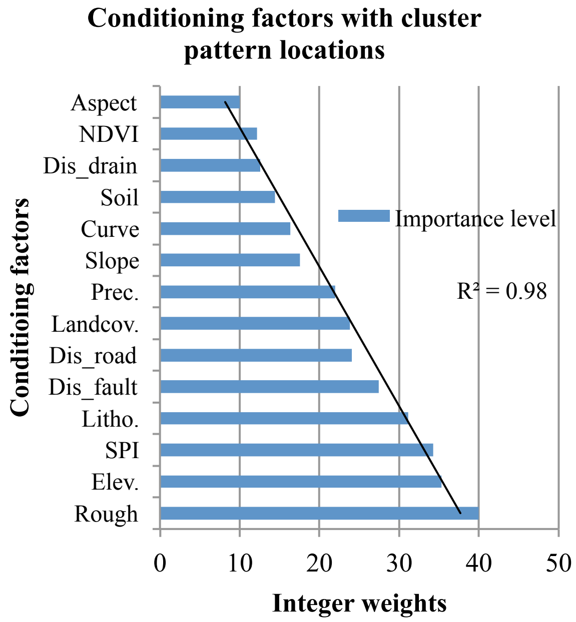


Figure 3. Consistency ratio result.

The current results shows a good level of agreement with previous random analysis results in importance sequence of conditioning factors, but also some differences are noticed (Fig. 4):

1. Using cluster data produced a linear ranking level (R^2 = %98), which effectively help to understand the comparison relationships between the conditioning factors.
2. The distance from faults shows a big leap in level of importance, from 12 in previous results to 27 currently, which reflect a direct relationship with cluster data, as the majority of landslide events accumulate near the fracture faults.

In order to validate the abovementioned findings, it needs to be explored and test on different study areas.



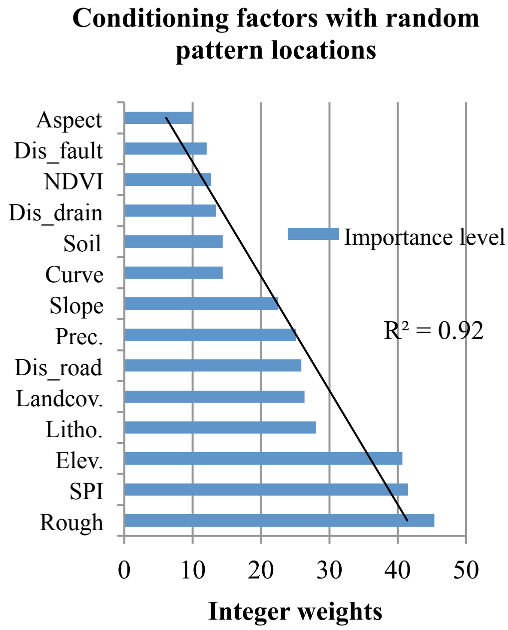


Figure 4. Conditioning factors and its importance level with different data patterns.

Finally, the landslide susceptibility map was produced, (Fig.5a). Furthermore, the current methodology was compared with previous researches conducted on the same study area, and proved its validity (Fig. 5b). The prediction result showed 83.5% accuracy using ensemble methodology with cluster locations while in case of random locations showed 82%. Also the authors concluded that that the ensemble methodology performed a satisfactory prediction results, with comprehensive and simple ranking procedure. It's possible to produce a reliable landslide susceptibility model using only a limited number of predisposing factors, which contribute towards higher conditional independence of the predisposing factors.

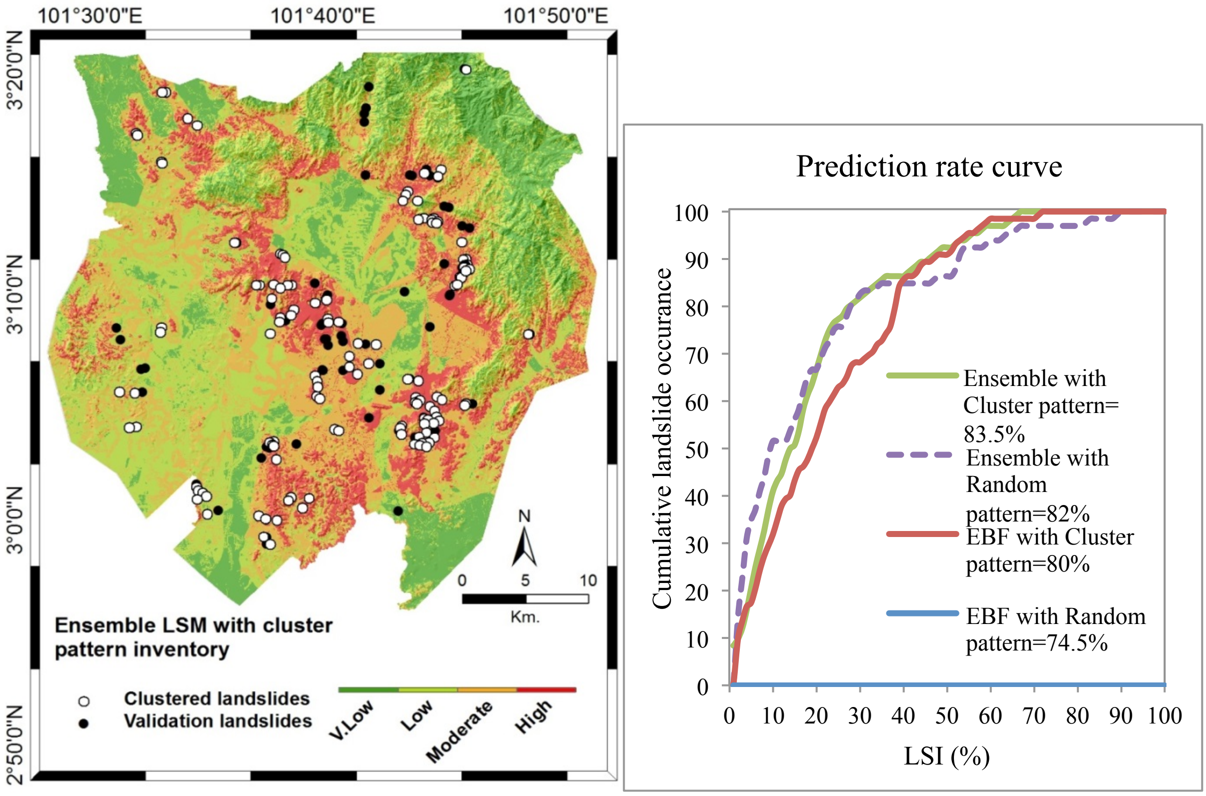


Figure 5. (a) Ensemble prediction map and (b) Models performance result.

5. CONCLUSION

The highly hazardous nature of landslides and its' disastrous consequences especially on the socio-economic sectors, become an indisputable fact worldwide, and topped the priorities of planners to develop solutions in short-and long-term basis. In recent years, landslide prediction mapping techniques has moved by bounds and leaps by utilizing the computation power of GIS.

Due to some shortening of the AHP and EBF model when applied individually in landslide susceptibility mapping, it can be overcome by using ensemble techniques. In this paper an ensemble model application was built based on two different inventories pattern, 1st random pattern, and 2nd cluster pattern. During the assessment, the landslide conditioning factors were categorized and ranked based on belief (Bel) association result. The results indicated that the spatial association between the bivariate EBF and the pair-wise comparison of AHP showed higher prediction accuracy in case of cluster pattern than random one. Moreover, the ensemble model managed to optimize the input layers, which can be served as major research advancement in areas where landslide inventory map is scarce.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from the UPM-RUGS project grant, vote number: 9344100 with additional support from Malaysia Lincense Surveyor Board under the special FIG incentive scheme.

REFERENCES

Althuwaynee, O.F., Pradhan, B., Lee, S., 2012a. Application of an evidential belief function model in landslide susceptibility mapping. Computers & Geosciences, 44, 120-135.

Althuwaynee, O.F., Pradhan, B., Mahmud, A.R., Yusoff, Z.M., 2012b. Prediction of slope failures using bivariate statistical based index of entropy model, Humanities, Science and Engineering (CHUSER), 2012 IEEE Colloquium on. IEEE, pp. 362-367.

Bonham-Carter, G., 1994. Geographic information systems for geoscientists: modelling with GIS. Pergamon press.

Clark, P.J., Evans, F.C., 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology, 35, 445-453.

Ergu, D., Kou, G., Peng, Y., Shi, Y., 2011. A simple method to improve the consistency ratio of the pair-wise comparison matrix in ANP. European Journal of Operational Research, 213, 246-259.

Evett, S.R., Tolk, J.A., Howell, T.A., 2006. Soil profile water content determination: Sensor accuracy, axial response, calibration, temperature dependence, and precision. Vadose Zone Journal, 5, 894.

Finan, J.S., Hurley, W.J., 1997. The analytic hierarchy process: does adjusting a pairwise comparison matrix to improve the consistency ratio help? Computers & operations research, 24, 749-755.

Ghosh, S., Carranza, E.J.M., van Westen, C.J., Jetten, V.G., Bhattacharya, D.N., 2011. Selecting and weighting spatial predictors for empirical modeling of landslide susceptibility in the Darjeeling Himalayas (India). Geomorphology, 131, 35-56.

Ishizaka, A., Labib, A., 2009. Analytic hierarchy process and expert choice: Benefits and limitations. OR Insight, 22, 201-220.

Jarman, D., 2006. Large rock slope failures in the Highlands of Scotland: characterisation, causes and spatial distribution. Engineering Geology, 83, 161-182.

Keeper, D.K., 1984. Landslides caused by earthquakes. Geological Society of America Bulletin, 95, 406-421.

Pradhan, B., 2011. Manifestation of an advanced fuzzy logic model coupled with Geo-information techniques to landslide susceptibility mapping and their comparison with logistic regression modelling. Environmental and Ecological Statistics, 18, 471-493.

Pradhan, B., Lee, S., 2007. Utilization of Optical Remote Sensing Data and GIS Tools for Regional Landslide Hazard Analysis Using an Artificial Neural Network Model. Earth Science Frontiers, 14, 143-151.

Saaty, T.L., 1977. A scaling method for priorities in hierarchical structures. Journal of mathematical psychology, 15, 234-281.

BIOGRAPHICAL NOTES

Dr. Biswajeet Pradhan is an Associate Professor at Department of Civil Engineering, University Putra Malaysia. Dr. Pradhan has more than 15 years of research, teaching, and industrial experience and has published over 170 research articles in referred technical international journals, 11 book chapters and 3 books. He specializes in remote sensing, GIS application and soft computing techniques in natural hazard and environmental problems. He is in the editorial board of many ISI journals. Dr. Pradhan is the Editor of Arabian Journal of Geosciences (ISI); Disaster Advances (ISI); Landslides (ISI); Central European Journal of Geosciences (ISI). He is the recipient of prestigious Alexander von Humboldt Research Fellowship and German Deutscher Akademischer Austausch Dienst (DAAD) and Saxony Scholarship holder, all from German Government.

CONTACT

Assoc. Prof. Dr. Biswajeet Pradhan*
Department of Civil Engineering, Faculty of Engineering,
University Putra Malaysia, 43400, Serdang, Selangor Darul Ehsan, Malaysia
Email. biswajeet24@gmail.com or biswajeet@lycos.com (corresponding author)