Landslide susceptibility mapping using decision-tree based CHi-squared automatic interaction detection (CHAID) and Logistic regression (LR) integration

Omar F. Althuwaynee, Biswajeet Pradhan*, Noordin Ahmad

Department of Civil Engineering, Faculty of Engineering, University Putra Malaysia, 43400, Serdang, Selangor DarulEhsan, Malaysia

*Email biswajeet24@gmail.com or biswajeet@lycos.com (Corresponding author)

Abstract. This article uses methodology based on chi-squared automatic interaction detection (CHAID), as a multivariate method that has an automatic classification capacity to analyse large numbers of landslide conditioning factors. This new algorithm was developed to overcome the subjectivity of the manual categorization of scale data of landslide conditioning factors, and to predict rainfall-induced susceptibility map in Kuala Lumpur city and surrounding areas using geographic information system (GIS). The main objective of this article is to use CHI-squared automatic interaction detection (CHAI) method to perform the best classification fit for each conditioning factor, then, combining it with logistic regression (LR). LR model was used to find the corresponding coefficients of best fitting function that assess the optimal terminal nodes. A cluster pattern of landslide locations was extracted in previous study using nearest neighbor index (NNI), which were then used to identify the clustered landslide locations range. Clustered locations were used as model training data with 14 landslide conditioning factors such as; topographic derived parameters, lithology, NDVI, land use and land cover maps. Pearson chi–squared value was used to find the best classification fit between the dependent variable and conditioning factors. Finally the relationship between conditioning factors were assessed and the landslide susceptibility map (LSM) was produced. An area under the curve (AUC) was used to test the model reliability and prediction capability with the training and validation landslide locations respectively. This study proved the efficiency and reliability of decision tree (DT) model in landslide susceptibility mapping. Also it provided a valuable scientific basis for spatial decision making in planning and urban management studies.

1. Introduction
Landslide hazards have a large risk impact, particularly on main cities, especially in the tropical developing countries, mainly because of rapid urbanization trends. Man-made activities have a vital role in disturbing the soil of hill slopes and thus increase infrastructure vulnerability index and fatality ratio[1].

The population of Malaysia increased from 6 million in 1960 to more than 28 million in 2012 (Figure 1). The above statistics shows that the population had changed 256% during the last 50 years as reported by the Department of Statistics, Malaysia. This consistently allied with excessive exploitation of lands, especially in Kuala Lumpur and commercial regions in Selangor estate, by
establishing new settlements, roads, and other infrastructure projects. Despite of the fact that abundance of relevant studies have been carried out in this study area, landslide is still a recurring hazard.

Figure 1. Variation of the population in Malaysia between 1960 to 2012

DT techniques, is popularly known as a multivariate automatic classification approach in landslide literature [2]. Some of the recently developed algorithms of decision tree pattern recognition are; Classification and Regression Tree (CART) [3], Improved algorithm ID3 “Iterative Dichotomizer” C4.5[4], CHAID[5] are used for prediction and classification purposes, [6]. A final report published in B.C. Ministry of forest, [7] have considered the road design parameters such as; width, angle, slope, cut, and fill of the road, that analysed using CHAID, to study their effects on road landslide vulnerability.

Multivariate statistical CHAID method produces an estimated significant chi-square level of interaction between conditioning factors and landslides. Pre-processed arbitrary classification of conditioning factors and tree pruning task are not required. It is worth to mention that, CHAID application considered a novel technique in landslide hazard mapping.

Multivariate statistical approaches, such as discriminant analyses and LR models have been successfully employed in landslide susceptibility mapping [8, 9]. The LR method is able to perform variable regression even if the variable has an abnormal distribution [10]. On the contrary, it requires few data pre-processing on theoretical assumptions than discriminant analysis like a bivariate classification [11]. This will raise the uncertainty due to classification mechanism. Therefore, an auto classification technique is preferable.

The current research aims to perform landslide susceptibility mapping for the part of the in Kuala Lumpur city and surrounding (Malaysia) using ensemble CHAID and LR model. The possibility of adopting this theory must be appreciated in a medium-scale landslide prediction assessment.

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 cause instability in slopes [12]. The study area is enclosed between 2°56’N to 3°20’N latitude and 101°29’E to 101°54’0E longitude, with approximate area of 1975 km² (Figure 2).

The major types of land cover 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 vary from 58 to 240 (mm/month), which trap large amount of water periodically, leading to a high pore-water pressure that decreases the shear stress of slopes (Malaysian Meteorological Services Department). More about geological and geomorphological characteristics of the study area can be seen in Althuwaynee et al. [13].
3. Methodology
Two multivariate approaches (CHi-squared Automatic Interaction Detection CHAID and LR) were used in ensemble model. The methodology employed here was applied in three main stages:
1. The NNI method tested the spatial pattern of landslides, and the result showed a clustered pattern tendency. Hence, the output will be used as a training dataset.
2. Running the CHAID model, then observing the terminal nodes which carry the lowest chi-square values among the classified layers.
3. Terminal nodes were used as input dataset in LR model, the resultant probability coefficients of LR analysis, helped to find the best fitting function.
4. Finally, all the LR probability coefficients were combined with their corresponding factors, then the landslide probability map LSM was produced.

3.1. Data
Since this article follow-up article to [13, 14], the basic data and other characteristics of the study area can be referred to these articles. The landslide inventory consists of 219 landslides, were collected over the past 25 years, mainly by using remote sensing sources like aerial photos and SPOT 5 images. Landslide conditioning factors used in the modelling were derived from topographic, soil and geology maps.

3.2. CHi-squared Automatic Interaction Detection (CHAID)
CHAID model is considered as one of the main DT techniques, has been popularly used in many regression. Some tangible characteristics and applications of CHAID method can be found in [15]. The major enhanced characteristics of CHAID are; categorical or ordinal data which can be modelled. Both dependent variable and conditioning factors were presented as nominal, ordinal or continuous data. Althuwaynee et al.[9] discussed the CHAID approach in more detail.

Figure 2. Study area.
The performance of classification iteration stops whenever there is no significant Chi-Square value between the dependent variable and conditioning factors. For that reason, highest Chi-Square value nodes comes first in the resultant tree, whereas, the terminal nodes carry the lowest Chi-Square values. CHAID method considers a statistical test that copes with data type and target nature. Like categorical data, Equation 1 of Pearson chi–squared will be used:

$$X^2 = \sum_{j=1}^{L} \sum_{i=1}^{I} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

(1)

where,

$$n_{ij} = \sum_{n \in D} f(n(x_n = i \cap y_n = j))$$

$$m_{ij}$$ is the observed cell frequency and

$$n_{ij}$$ is the expected cell frequency for ($$x_n = i, y_n = j$$) following the independence model.

The corresponding p-value given by $$p = \Pr(x^2 > x^2)[16]$$. The output of the CHAID method has the tree structure which consists of branches that represent automated classification of each significant conditioning factor, with a specific percentage of landslides (1) and non-landslides (0), in each node. Top of the tree is composed by a unique cell carry node (# 0) which represents the dependent variable information, such as the total number of training sites (landslides and non-landslides). The rest of the tree is formed by nodes representing the conditioning factors classification, which carries certain information such as node number, class interval limit, counts of landslides and non-landslides.

Pre-setting of the model’s criteria is essential, because that will positively or negatively affect the tree size and processing time. Here, the main criteria components are; growth limit, merging value, and parent and child nodes[17]. The minimum value of each criterion represents the limit of stopping the tree growth process.

### 3.3 Logistic regression (LR)

Logistic regression (LR) relies on measuring the results with dichotomous variables such as 1 and 0 or true and false. Moreover, LR builds a statistical model to predict the logit transformation of dependent variable (landslide) occurrence probability [8, 18].

LR compute the changes in the likelihood that falls in each category to find the best fitting model to describe the relationship between landslide conditioning factors, then ranking the factors according to the highest numerical code among dependents [19]. Equation 2 was used to fit the dependent variables:

$$Z = b_0 + b_1 x_1 + \ldots + b_p x_p,$$

(2)

where, $$Z$$ represents the linear combination of independent variables, absence or presence of a landslide, and variables value from $$-\infty$$ to $$+\infty$$. By contrast, fewer parameters with fewer cells are advisable for short and reliable operation process and reasonable equation limits.

The probability (p), estimate the probability of occurrence, denotes any pixel that is susceptible to slope failure, and it can be represented as the conditional probability in the LR model by the following expression:

$$\text{Logit} (p) = \ln \left( \frac{p}{1-p} \right) = \frac{1}{1+e^{-z}},$$

(3)
where, \( p \) is the number of independent variables, \( b_0 \) is the constant or intercept of the equation, and \( b_1, b_2, \ldots, b_p \) are the coefficients of the independent variables \( x_1, x_2, \ldots, x_p \).

The independent binary variable with values of one or zero indicates the presence or absence of a category [20]. A forward stepwise logistic regression approach was used. Finally, the regression indices of the predictors were imported to ArcGIS software, to perform the final probability map using Equations 2 and 3.

**4. Results and discussion**

CHAID was successfully applied using the 130 cluster pattern landslides representing 2500 cells tested by NNI. A total of 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) represented the independent variables. Model criteria setting was carried out for optimizing the processing time and achieving the desirable results. Then Pearson Chi-square statistic test was used to control the classification significant level of CHAID result [9]. Most of the conditioning factors that showed a strong logical relationship with slope failure were classified by CHAID. The terminal nodes that represent the optimum decision consist of SPI, soil type, slope, aspect, roughness, distance from faults, distance from road, distance from drainage, land cover, elevation.

In the next step, terminal nodes were nominated according to the susceptibility level against landslide using LR method.

Multi-co linearity test had to be performed (i.e., ANOVA of inflation factor, tolerance, and co-linearity index). Test shows that all terminal spatial factors are not interdependent (Tol > 0.3) [21]. The maximum likelihood estimation method was used for coefficient extraction by an iterative process. The analysis was carried and the result showed a highest significant relationship was found between landslide occurrence and the following conditioning factors; SPI, soil type, land cover, distance from road, distance from fault, elevation, and distance from drainage.

The difference in the \(-2 \log \text{ likelihood} (-2 \text{LL})\) is considered as an indicator of model improvement over the null model. The lowest value of the \(-2 \text{LL}\) represents the best step fit of the model to the data and explains the value decrement until the final iteration step. Cox/Snell’s and Nagelkerke’s R-square was used to measure the usefulness of the model. A higher R-square value of Cox/Snell and Nagelkerke corresponds to a better model (Table 1).

**Table 1.** Statistics summary of LR model.

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square | Chi-square | df | Sig. |
|------|-------------------|----------------------|---------------------|------------|----|------|
| 12   | 3573.843<sup>b</sup> | .048                 | .064                | 21.973     | 8  | .005 |

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Table 2 shows the Wald statistics for each model, the highest Wald value, implies the most significant value for the occurrence of landslides. The P-value represents the anticipated probability of
rejecting the null hypothesis, and defined by statistical relationship between variables at the 95% confidence level. Hence if the P-value is less than 0.05 then the null hypothesis is likely not true.

Table 2. LR model statistics of the regression.

| Variables in the Equation | B     | S.E. | Wald  | df | Sig. | Exp(B) | 95% C.I. for EXP(B) |
|---------------------------|-------|------|-------|----|------|--------|---------------------|
| Step 12                   |       |      |       |    |      |        |                     |
| spi                       | .009  | .002 | 27.682| 1  | .000 | 1.009  | 1.006 1.012          |
| soil                      | .005  | .002 | 9.020 | 1  | .003 | 1.005  | 1.002 1.008          |
| road4                     | .014  | .003 | 19.267| 1  | .000 | 1.014  | 1.008 1.021          |
| lcover                    | .011  | .003 | 11.000| 1  | .001 | 1.011  | 1.004 1.018          |
| fault6                    | -.009 | .002 | 19.437| 1  | .000 | .991   | .987  .995            |
| drain1u                   | .004  | .001 | 8.261 | 1  | .004 | 1.004  | 1.001 1.006          |
| elev3                     | .007  | .002 | 8.819 | 1  | .003 | 1.007  | 1.002 1.011          |
| fault2                    | -.010 | .004 | 6.908 | 1  | .009 | .990   | .983  .997            |
| road2                     | .006  | .002 | 9.341 | 1  | .002 | 1.006  | 1.002 1.010          |
| lcover                    | -.005 | .002 | 5.685 | 1  | .017 | .995   | .990  .999            |
| soil3                     | .004  | .002 | 5.342 | 1  | .021 | 1.004  | 1.001 1.007          |
| spi1u                     | -.004 | .002 | 4.714 | 1  | .030 | .996   | .993  1.000          |
| Constant                  | -1.505| .508 | 8.786 | 1  | .003 | .222   |                     |

Finally, the susceptibility map prepared by multiplying the $\hat{B}$ (probability weight) values, which represent the positive/ negative weight of factors in LR. The prediction rate curve test was carried out on the integrated landslide susceptible map LSM (Figure 3a) using the validation data. LSM accuracy showed AUC=70% (Figure 3b). The prediction value of CHAID classification showed 73.4% of landslides locations and 73.3% for non-landslide locations (Table 3). The reason behind that low prediction accuracy of LSM, is the large study area compared by number/size of evidence (training data). Large difference of study area to landslide inventory (landslide locations area to study area equal to 1:6,512), the classification tree show 151 of terminal nodes. Thus we recommend of using a reasonable ratio between size and number of landslide and scale of study area, good example of the reasonable ratio was found in previous research done by the same authors [9].

Table 3. CHAID prediction value.

| Risk Estimate | Std. Error | Classification |
|---------------|------------|----------------|
| Observed      | Predicted  |                |
| 0             | 1          | Percent Correct|
| 0             | 1843       | 673            | 73.3%       |
| 1             | 666        | 1836           | 73.4%       |
| Overall       | 50.0%      | 50.0%          | 73.3%       |
| Percentage    |            |                |

7th IGRSM International Remote Sensing & GIS Conference and Exhibition
IOP Conf. Series: Earth and Environmental Science 20 (2014) 012032
doi:10.1088/1755-1315/20/1/012032
5. Conclusion
Highly hazardous nature of landslides and disastrous consequences especially on the socio-economic sectors, has become an indisputable fact worldwide. This has led for the planners to develop solutions on a short-and long-term basis. In recent years, landslide prediction mapping s has moved by bounds and leaps by utilizing the computation power of GIS. The main objective of this research was to use the DT-based CHAID method to perform the best classification fit for each landslide conditioning factor, then terminal nodes of the CHAID tree were integrated with LR model to find the corresponding coefficients of best fitting function, and also to assess the optimal terminal node. The classification prediction of CHAID showed 73% of the total landslide location successfully predicted. Then LSM was produced by integrating the terminal nodes in the LR model. Prediction rate curves test was used to evaluate the prediction performance of LSM, and it achieved AUC=0.70. As a conclusion, the model showed a good fit for logical automatic classification for landslide conditioning factors, without any interference from experts. Moreover, to facilitates the design productivity for roads embankment, slope rehabilitations and design, which may be useful in decision making.

Acknowledgement
The authors gratefully acknowledge the financial support from the UPM-RUGS project grant number: 6300114.

References
[1] P. Aleotti, "A warning system for rainfall-induced shallow failures," Engineering Geology, vol. 73, pp. 247-265, 2004.
[2] D. T. Bui, et al., "Regional prediction of landslide hazard using probability analysis of intense rainfall in the Hoa Binh province, Vietnam," Natural Hazards, vol. 66, pp. 707-730, 2013.

[3] Y.-K. Yeon, et al., "Landslide susceptibility mapping in Injae, Korea, using a decision tree," Engineering Geology, vol. 116, pp. 274-283, 2010.

[4] J. R. Quinlan, C4. 5: programs for machine learning vol. 1: Morgan kaufmann, 1993.

[5] J. Magidson, SPSS for Windows CHAID release 6.0: SPSS Incorporated, 1993.

[6] A. S. Bozkir and E. A. Sezer, "Predicting food demand in food courts by decision tree approaches," Procedia Computer Science, vol. 3, pp. 759-763, 2011.

[7] Terratech, "Inventory of forest landslide occurrence in the kamloops forest region," Terratech Western profile consultants LTD1994.

[8] B. Pradhan, "Landslide susceptibility mapping of a catchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches," Journal of the Indian Society of Remote Sensing, vol. 38, pp. 301-320, 2010.

[9] O. F. Althuwaynee, et al., "A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping," Catena, vol. 114, pp. 21-36, 2014.

[10] S. Menard, "Coefficients of determination for multiple logistic regression analysis," American Statistician, vol. 54, pp. 17-24, 2000.

[11] L. Ayalew and H. Yamagishi, "The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan," Geomorphology, vol. 65, pp. 15-31, February 1, 2005.

[12] B. Pradhan, "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, vol. 18, pp. 471-493, 2011.

[13] O. F. Althuwaynee, et al., "Application of an evidential belief function model in landslide susceptibility mapping," Computers & Geosciences, vol. 44, pp. 120-135, 2012.

[14] O. F. Althuwaynee, et al., "Prediction of slope failures using bivariate statistical based index of entropy model," in Humanities, Science and Engineering (CHUSER), 2012 IEEE Colloquium on, 2012, pp. 362-367.

[15] StatisticsSolutions. (2012). CHAID. Available: http://www.statisticssolutions.com/academic-solutions/resources/directory-of-statistical-analyses/chaid/

[16] S. Baker and R. D. Cousins, "Clarification of the use of CHI-square and likelihood functions in fits to histograms," Nuclear Instruments and Methods in Physics Research, vol. 221, pp. 437-442, 4/1/ 1984.

[17] I. S. Naftulin and O. Y. Rebrova, "Application of C&RT, CHAID, C4. 5 and WizWhy algorithms for stroke type diagnosis," in Artificial Intelligence and Soft Computing, 2010, pp. 651-656.

[18] K. Devkota, et al., "Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugu-­-Narayanghat road section in Nepal Himalaya," Natural Hazards, vol. 65, pp. 135 -165, 2013.

[19] S. Bai, et al., "Combined landslide susceptibility mapping after Wenchuan earthquake at the Zhouqu segment in the Bailongjiang Basin, China," Catena, vol. 99, pp. 18-25, 2012.

[20] G. C. Ohlmacher and J. C. Davis, "Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA," Engineering Geology, vol. 69, pp. 331-343, 2003.

[21] M. Van Den Eeckhaut, et al., "Statistical modelling of Europe-wide landslide susceptibility using limited landslide inventory data," Landslides, vol. 9, pp. 357-369, 2012.