International Journal of Earth, Energy and Environmental Sciences
ISSN: 2517-942X
Vol:3, No:11, 2009

Integrating Decision Tree and Spatial Cluster Analysis for Landslide Susceptibility Zonation

Chien-Min Chu, Bor-Wen Tsai, Kang-Tsung Chang

***Abstract*—**Landslide susceptibility map delineates the potential zones for landslide occurrence. Previous works have applied multivariate methods and neural networks for mapping landslide susceptibility. This study proposed a new approach to integrate decision tree model and spatial cluster statistic for assessing landslide susceptibility spatially. A total of 2057 landslide cells were digitized for developing the landslide decision tree model. The relationships of landslides and instability factors were explicitly represented by using tree graphs in the model. The local Getis-Ord statistics were used to cluster cells with high landslide probability. The analytic result from the local Getis-Ord statistics was classed to create a map of landslide susceptibility zones. The map was validated using new landslide data with 482 cells. Results of validation show an accuracy rate of 86.1% in predicting new landslide occurrence. This indicates that the proposed approach is useful for improving landslide susceptibility mapping.

***Keywords*—**Landslide susceptibility Zonation, Decision tree model, Spatial cluster, Local Getis-Ord statistics.

I. Introduction

Over the past 30 years, landslides have caused about 10,000 deaths and more than 10-billion USD in damage in Asia [1]. Predicting landslide distribution has thus become important in preventing disasters. Different models and techniques have been developed or tested to predict landslide occurrence. The aim of these models is to analyze the relationships between instability factors and the distribution of landslides in order to predict the slope failure probabilities in a specific area. Generally, statistical classification methods are used to estimate landslide susceptibility over large and complex areas [2]. These models based on multiple regression [3], discriminant function [3-5], and logistic regression [6-10] allow us to build algorithms, which generate a rule to predict landslides. Most multivariate statistical methods are used to quantify the linear relationships between predictor variables and response variables. However, the multivariate methods have some disadvantages. Firstly, the relationships between environmental data may be nonlinear and involve high-order interactions[11]. The commonly used multivariate methods often fail to fit model perfectly from such data. Next, the basic assumptions should be considered when setting statistical parametric models. For instance, it is usually assumed that the sample distribution of a given variable in a model must be normal, but this assumption does not always fit well in landslide studies [12]. Finally, the result of these models describes the correlations between the instability factors and landslide events by the estimated coefficients in an equation. Van Asch *et al.*[13] pointed out that, to represent slope movements adequately, a computer model must include the local characterization of the geometry and internal structure. Using one equation generating all complicated interactions and smoothing the spatial variety in the entire area cannot represent small or local landslides in a specific area [14].

Recently, researchers have used several applications of Neural Networks (NNs) in landslide modeling. NNs are designed for capturing highly complex nonlinear relationships and high-order interactions, which need not necessarily be pre-specified. These studies have demonstrated that NNs are effective tools for analyzing landslide susceptibility [9, 11-12, 15-16]. NNs are data-driven models that all relevant variables are allowed to interact and not constrained by basic statistical assumptions, and thus overcome the disadvantages of multivariate models. The 'black-box' or "hidden layers" method is suitable for estimating complicated interactions between variables, but lack of interpretation is its shortcoming[17]. For landslide hazard managements, the information of interactions between mass movement distribution and conditioning factors is important for decision-making.

In this paper, we propose a tree-based method, the Classification and Regression Trees (CART) [18], and apply in the method to landslide modeling. CART can analyze the complex data structure well in large dataset and illustrate the interactions between landslides and factors. Unlike regression methods, which try to identify a general relationship between instability factors and landslides, CART uses binary recursive partitioning procedures to divide all dataset into small homogeneous subsets by the significant factor. CART has been successfully applied in many different fields, such as medical diagnosis [19], ecology [20], remote sensing [21-22], and soil distribution [23-24].Furthermore, a spatial cluster method is used to compute the result of landslide probability derived from CART to identify high landslide susceptibility zones.

Chien-Min Chu, Ph.D. candidate, is with the Department of Geography, National Taiwan University, Taipei, Taiwan. (phone: 886-2-33665838; fax:886-2-23622911; e-mail: d91228003@ ntu.edu.tw).

Bor-Wen Tsai, associate professor, is with the Department of Geography, National Taiwan University, Taipei, Taiwan. (e-mail: tsaibw@ntu.edu.tw).

Kang-Tsung Chang, professor, is with the Department of Leisure and Recreation Management, Kainan University, Taoyuan, Taiwan.(e-mail: ktchang@mail.knu.edu.tw)

II. Methods

A. Classification and Regression tree

The main idea in CART is to partition the dataset into homogeneous subgroups with respect to the same class. The complex data structure can be represented conveniently by a tree structure in which an internal node denotes a best split predictor variable, the branches of a node denote the criteria value of the split variable, and a leaf denote the final response class. In the tree structure, the paths from the root node (top node) to leaf (terminal node) show the decision rules that maximize the distinction among the classes and minimize the diversity in each class. Each recorder is assigned to one of the terminal nodes on the basis of the paths. The paths and nodes can be portrayed as a tree graph or translated into convenient if-then rules. Therefore, the symbolic tree graph or if-then rules can be easily interpreted by users.

The first step of CART analysis is to build a tree using splitting rules from the root node. Tree building begins at the root node, which includes all predictor variables of the learning dataset. It evaluates all possible splits for all predictor variables, and chooses the best node, which maximizes the 'purity' of two child nodes. The best predictor is chosen using an impurity measure. The purpose is to produce two subsets of the data which are as homogenous as possible[18]. For the binary dependent variable (landslide or non-landslide), the CART method uses the Gini impurity measure to decide the purity.

For a node t, the Gini index of impurity, g(t), is defined as

$$g(t) = \sum_{j \neq i} p(j|t)\, p(i|t) \tag{1}$$

where *i* and *j* are categories of the target variable. A two-class problem assigns all class *j* objects the value 1 and all other objects the value 2. Then the equation for the index reduces to

$$g(t) = 2\, p(1|t)\, p(2|t) \tag{2}$$

When all recodes in the node belong to only one category, which means the node is purity, the index equals 0.

To select the best predictor variable of a node, we score every possible variable and select the one with the best score, which represents the greatest reduction in impurity. For any node *t*, suppose that there is a candidate split *s* of the node, which divides it into the left division t_L and the right division t_R. The score is defined as

$$\phi(s,t) = g(t) - p_L g(t_L) - p_R g(t_R) \tag{3}$$

where P_L is the proportion of cases of child node *t* sent to the left, and P_R to child node on the right. We can define a candidate set *S* of binary *s* at each node. When it starts at the root node t_1, it looks for the division s*, among all possible *S*, with a greater reduction value of impurity.

$$\phi(s^*, t_1) = \max_{s \in S} \phi(s, t_1) \tag{4}$$

A perfect split *s* separates the dataset into two subgroups and causes $g(t_L) = g(t_R) = 0$. The recursive partitioning algorithm loops until it is impossible to continue, i.e. when only one case remains or when all the cases belong to the same class. A maximal tree will be produced when it grows until all terminal nodes are perfect purity. The maximal tree is generally overlearning or overfitting because of the random or noisy cases in the learning dataset. The CART method uses a "overgrow and prune back' procedure to get an optimum tree that is fitted to signal rather than noise. In this study, a 10-fold cross-validation was adopted to select the best tree size, which has the minimum cross-validation cost.

B. Spatial Cluster Analysis - Local Getis-Ord's G_i*

The local Getis-Ord's G_i* is useful for identifying individual members of local clusters by determining the spatial dependence and relative magnitude between an observation and neighboring observations [25]. The local Getis-Ord's G_i can be written as follows [26]:

$$Gi^* = \frac{\sum_{j=1}^{m} w_{ij} x_j}{\sum_{j=1}^{m} x_i} \tag{5}$$

where x_j is a landslide occurrence probability for th *j*th cell, and w_{ij} is the spatial weight parameter for the pair of cells *i* and *j* to represent proximity relations. In this study, we define adjacency using a queen continuity weight file, which is constructed based on cells that share common boundaries and vertices. A simple 0/1 matrix is formed, where 1 indicates that the cells having a common border or vertex and 0 otherwise.

A cell with high *Gi** indicates cells in its neighborhood have relatively high probability of slope failure. Conversely, a cell with low *Gi** suggests cells in its neighborhood have relatively low probability of slope failure. The Z score of the *Gi** indicates the level that a high or low probability concentration is significantly different from a random distribution[27]. A group of cells with high Z scores reveals a cluster of cells with high landslide susceptibility, and vice versa. A Z score near 0 indicates no cluster of either high or low values.

III. Study Area and variable

The 37.5 km^2 study area is located in the Shihmen Reservoir watershed in northern Taiwan (Fig. 1). Elevation in the area ranges from 850 m in the northwest to 2375 m asl in the southeast, with generally rugged topography. Slope gradient ranges from 0° to 66°.

To develop our landslide model, we used seven independent

data layers that are recognized as effective factors in landslide occurrence, including slope, aspect, profile curvature, plan curvature, curvature, wetness index, and geology (categorical). The spatial locations of landslides in four different years (1976, 1986, 1992 and 2004) have been digitized from aerial photographs. All data layers were converted to raster layers with a 40 m × 40 m cell size. A total of 2057 landslide cells were derived in the four years. From the prepared data layers, we took a random sample of 2,057 non-landslide cells with 2,057 landslide cells for developing the landslide model.

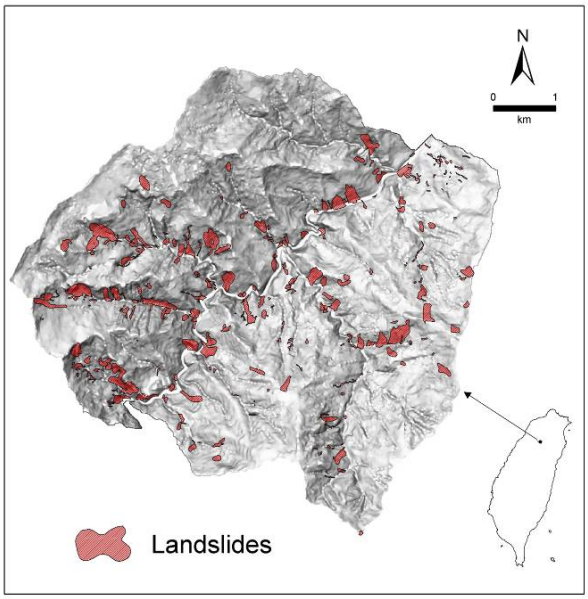

Fig. 1. The shaded relief map shows the spatial distribution of historical landslides in the study area in northern Taiwan.

IV. Result

A. Landslide tree model

Fig. 2 shows the best classification tree of the landslide cells in the training set. The tree shows a mean cross-validation error of 20.5%, calculated after 10 repetitions of the cross-validation procedure. That means that 79.5% of the data set is predicted correctly. The landslide tree model shows 16 terminal nodes and 16 classification rules. At each terminal node, the relevant decision is shown. Percentages are the proportion of cells in the training sample that are landslide at this point in the tree. For example, a terminal node with 60.4% signifies a branch of the tree in which 60.4% of the training cells are landslide.

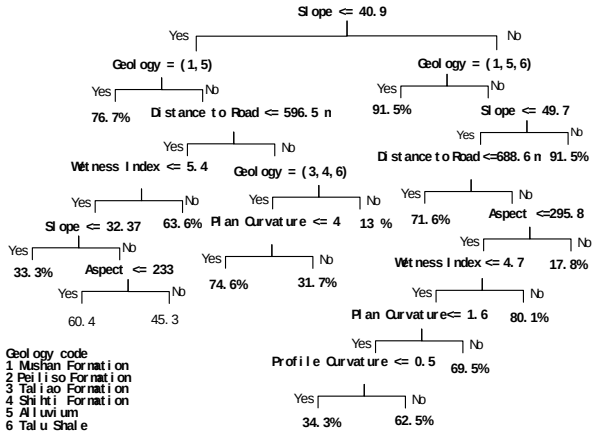

Fig. 2. Graphical representation of the landslide tree model.

B. Landslide probability map

Fig. 3 shows the computational result of the probability of landslide occurrence. Each cell is assigned a probability value by following the 16 classification rules. The darker cell indicates a higher probability of landslide occurrence.

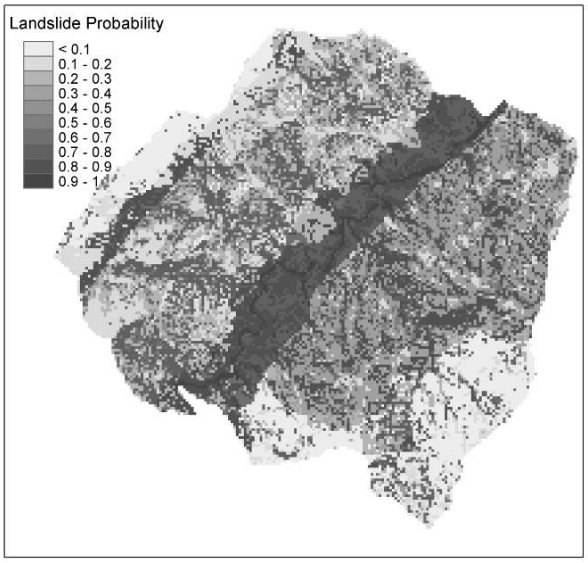

Fig. 3. The landslide probability map from the landslide tree model.

C. Landslide susceptibility zonation map

Fig. 4 represents the landslide susceptibility zonation map. It classifies the range of landslide susceptibility into four categories based on the Z scores of *Gi**: (a) Non-susceptible zone (< 0), (b) Low susceptible zone (0 - 1.65), (c) Moderate susceptible zone (1.65 - 1.96), and (d) High susceptible zone (> 1.96). Positive Z scores indicate spatial clustering of high values, whereas negative z scores indicate spatial clustering of low values.

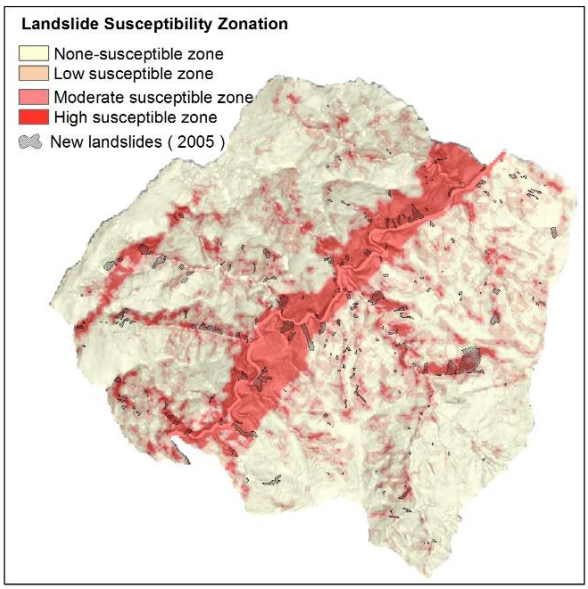

Fig. 4. Landslide susceptibility zonation map and the spatial distribution of new landslides (2005).

D. Model validation

New landslides in 2005 were used for validating the landslide susceptibility zonation derived from the landslide tree model and local Getis-Ord's G_i* statistics (Fig. 4). The total number of landslide cells in 2005 is 482. The landslide susceptibility zonation map correctly predicts 86.1% of the new landslide cells in the high, moderate, and low susceptible zones (Fig. 5).

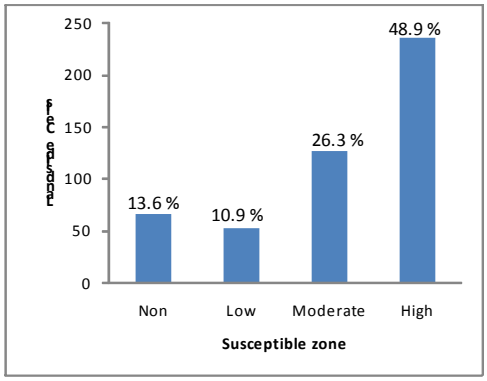

Fig. 5. The frequencies of new landslides calculated from the susceptible zones.

V. CONCLUSIONS

This study has effectively integrated the CART decision tree method and local Getis-Ord statistic to group high probability cells into high susceptible zones. The landslide decision tree model can clearly reveal the combination of instability factors related to slope failures and describe the relationships between variables. The landslide probability map derived from the landslide decision tree model delineates the probability of landslide occurrence at each cell. The spatial distribution of landslide probability is uneven, with abrupt changes within short distances. For the purpose of delineating landslide susceptibility zonation, the location Getis-Ord's algorithm is useful to map spatial clusters of high probability cells into high susceptibility zones. The results of model validation shows an accuracy rate of 86.1%.

This study has developed a new approach to integrate the decision tree model and spatial cluster method for landslide susceptibility mapping. The landslide decision rules provide useful information to diagnose slope stability. Moreover, the landslide susceptibility map based on the new integrated approach provides a management tool for regional planners working with landslide hazard.

ACKNOWLEDGMENT

This work was supported by the National Science Council, Department of Geography, National Taiwan University under project nos. NSC97-2621-M-002-026 and NSC97-2923-H-424-001-MY2.

REFERENCES

[1] EM-DAT, The OFDA/CRED International Disaster Database. 2008, Universit´e Catholique de Louvain - Brussels - Belgium.
[2] Guzzetti, F., P. Reichenbach, F. Ardizzone, M. Cardinali, and M. Galli, "Estimating the quality of landslide susceptibility models," Geomorphology. vol 1-2: pp. 166-184, 2006.
[3] Carrara, A., "Multivariate models for landslide hazard evaluation," Mathematical Geology. vol 3: pp. 403-426, 1983.
[4] Carrara, A., M. Cardinali, R. Detti, F. Guzzetti, V. Pasqui, and P. Reichenbach, "Gis Techniques and Statistical-Models in Evaluating Landslide Hazard," Earth Surface Processes and Landforms. vol 5: pp. 427-445, 1991.
[5] Reger, J.P., "Discriminant-Analysis as a Possible Tool in Landslide Investigations," Earth Surface Processes and Landforms. vol 3: pp. 267-273, 1979.
[6] Dai, F.C. and C.F. Lee, "Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong," Geomorphology. vol 3-4: pp. 213-228, 2002.
[7] Lee, S., "Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data," International Journal of Remote Sensing. vol 7: pp. 1477-1491, 2005.
[8] van Den Eeckhaut, M., T. Vanwalleghem, J. Poesen, G. Govers, G. Verstraeten, and L. Vandekerckhove, "Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium)," Geomorphology. vol 3-4: pp. 392-410, 2006.
[9] Yesilnacar, E. and T. Topal, "Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey)," Engineering Geology. vol 3-4: pp. 251-266, 2005.
[10] Chang, K.T., S.H. Chiang, and M.L. Hsu, "Modeling typhoon-and earthquake-induced landslides in a mountainous watershed using logistic regression," Geomorphology. vol 3-4: pp. 335-347, 2007.
[11] Melchiorre, C., M. Matteucci, A. Azzoni, and A. Zanchi, "Artificial neural networks and cluster analysis in landslide susceptibility zonation," Geomorphology. vol 3-4: pp. 379-400, 2008.
[12] Ermini, L., F. Catani, and N. Casagli, "Artificial Neural Networks applied to landslide susceptibility assessment," Geomorphology. vol 1-4: pp. 327-343, 2005.
[13] van Asch, T.W.J., J.P. Malet, L.P.H. van Beek, and D. Amitrano, "Techniques, advances, problems and issues in numerical modelling of landslide hazard," Bulletin de la Societe Geologique de France. vol 2: pp. 65-88, 2007.

[14] van Westen, C.J., A.C. Seijmonsbergen, and F. Mantovani, "Comparing Landslide Hazard Maps," Natural Hazards. vol 2: pp. 137-158, 1999.
[15] Lee, S., J.H. Ryu, M.J. Lee, and J.S. Won, "The Application of Artificial Neural Networks to Landslide Susceptibility Mapping at Janghung, Korea," Mathematical Geology. vol 2: pp. 199-220, 2006.
[16] Neaupane, K.M. and S.H. Achet, "Use of backpropagation neural network for landslide monitoring: a case study in the higher Himalaya," Engineering Geology. vol 3-4: pp. 213-226, 2004.
[17] Paliwal, M. and U. Kumar, "Neural networks and statistical techniques: A review of applications," Expert Syst Appl. vol 1: pp. 2-17, 2009.
[18] Brieman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone, "Classification and Regression Trees," Wadsworth Inc. vol, 1984.
[19] Camp, N.J. and M.L. Slattery, "Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States)," Cancer Cause Control. vol 9: pp. 813-823, 2002.
[20] De'ath, G. and K.E. Fabricius, "Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis," Ecology. vol 11: pp. 3178-3192, 2000.
[21] Pal, M. and P.M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," Remote Sens Environ. vol 4: pp. 554-565, 2003.
[22] Friedl, M.A. and C.E. Brodley, "Decision tree classification of land cover from remotely sensed data," Remote Sens Environ. vol 3: pp. 399-409, 1997.
[23] Scull, P., J. Franklin, and O.A. Chadwick, "The application of classification tree analysis to soil type prediction in a desert landscape," Ecological Modelling. vol 1: pp. 1-15, 2005.
[24] Lagacherie, P. and S. Holmes, "Addressing geographical data errors in a classification tree for soil unit prediction," International Journal of Geographical Information Science. vol 2: pp. 183-198, 1997.
[25] Wu, J., et al., "Exploratory spatial data analysis for the identification of risk factors to birth defects," BMC Public Health. vol 1: pp. 23, 2004.
[26] Getis, A. and J. Ord, "The analysis of spatial association by use of distance statistics," Geographical analysis. vol 3: pp. 189-206, 1992.
[27] Ord, J. and A. Getis, "Local spatial autocorrelation statistics: distributional issues and an application," Geographical analysis. vol 4: pp. 286-306, 1995.