VERIFICATION AND RISK ASSESSMENT FOR LANDSLIDES IN THE SHIMEN RESERVOIR WATERSHED OF TAIWAN USING SPATIAL ANALYSIS AND DATA MINING

J-S. Lai [a, *], F. Tsai [a, b]

[a] Department of Civil Engineering, National Central University, No. 300, Zhong-Da Rd., Jhongli, Taoyuan, Taiwan 32001 – jslai0726@gmail.com
[b] Center for Space and Remote Sensing Research, National Central University, No. 300, Zhong-Da Rd., Jhongli, Taoyuan, Taiwan 32001 - ftsai@csrsr.ncu.edu.tw

Commission II /3

KEY WORDS: Hazard assessment, Landslides, Disaster Mitigation, Data Mining

ABSTRACT:

Spatial information technologies and data can be used effectively to investigate and monitor natural disasters contiguously and to support policy- and decision-making for hazard prevention, mitigation and reconstruction. However, in addition to the vastly growing data volume, various spatial data usually come from different sources and with different formats and characteristics. Therefore, it is necessary to find useful and valuable information that may not be obvious in the original data sets from numerous collections. This paper presents the preliminary results of a research in the validation and risk assessment of landslide events induced by heavy torrential rains in the Shimen reservoir watershed of Taiwan using spatial analysis and data mining algorithms. In this study, eleven factors were considered, including elevation (Digital Elevation Model, DEM), slope, aspect, curvature, NDVI (Normalized Difference Vegetation Index), fault, geology, soil, land use, river and road. The experimental results indicate that overall accuracy and kappa coefficient in verification can reach 98.1 % and 0.8829, respectively. However, the DT model after training is too over-fitting to carry prediction. To address this issue, a mechanism was developed to filter uncertain data by standard deviation of data distribution. Experimental results demonstrated that after filtering the uncertain data, the kappa coefficient in prediction substantially increased 29.5%.The results indicate that spatial analysis and data mining algorithm combining the mechanism developed in this study can produce more reliable results for verification and forecast of landslides in the study site.

1. INTRODUCTION

Taiwan has complicated geological conditions, high density of population and other potential factors making it vulnerable to natural hazards. The geological structures have become very fracture after the 1999 Chichi earthquake. Moreover, typhoons and other extreme weathers also frequently happen in this region. The heavy rainfall often triggers serious landslides and debris flows, and then causing human casualties and property damages. Therefore, the World Bank listed Taiwan as one of the countries that is most vulnerable to natural disasters in the world in terms of lands and population exposing to the danger. Thus, to prevent and mitigate natural hazards such as landslides has become an important issue in Taiwan.

Spatial information technologies and data such as remotely sensed images, LIDAR point clouds and GIS datasets can be used effectively to investigate and monitor natural disasters contiguously and to support policy- and decision-making for hazard prevention, mitigation and reconstruction. In addition, previous studies (e.g. Sakar & Kanungo, 2004; Metternicht et al., 2005; Nichol & Wong, 2005; Tsai & Chen, 2007; Peduzzi, 2010) have demonstrated that geo-informatics techniques can perform investigation successfully in LULC (Land Use/ Land Cover) and natural disaster applications. However, in addition to the vastly growing data volume, these spatial data usually come from different sources and with different formats and characteristics. Therefore, it is necessary to develop effective algorithms to extract useful and valuable information that may not be obvious in the original datasets from complicated datasets for efficient analysis.

Data Mining (DM) is an important and effective technique in the field of Knowledge Discovery (KD) that extracts knowledge from vast data, database or data warehouse as the primary objective. Therefore, it may be a viable solution to fulfill the demand of identifying possible landslide factors from heterogeneous spatial datasets. In addition, Spatial Analysis (SA) can supply DM with advanced information with overlay, buffer and other GIS processes. Decision Tree (DT) algorithm is a classical, universal and comprehensible method in the DM domain. The outcomes of DT are constituted by “If and Then” rules, and these sequences are helpful in realizing the reasons and interactions between causative factors in the landslide records. Based on these spatial analysis components, this research has adopted and developed DT and SA algorithms on the validation and forecast (risk assessment) of landslide events induced by heavy rainfall in a regional scale.

2. STUDY SITE AND MATERIALS

The Shimen reservoir watershed (see Figure 1) that covers a region of about 763.4 km^2 in Taiwan was selected as the study

[*] Corresponding author.

site of this research The elevation in the study site ranges between 250m to 3,500m. The primary land-cover is forest, but there are limited agricultural activities. Landslides are commonly induced by heavy rainfall in the area and the debris flows are flushed into the reservoir, causing various problems in water supply and resource management. Previous studies related to landslide (e.g. Sidle et al., 1985; Wu and Sidle, 1995; Zhou et al., 2002; Dahal et al., 2008) divided the causative factors of landslides into latent and triggering catalogs. This study does not focus on different types of landslides. Instead, it explores the knowledge of landslides induced by heavy torrential rains using data mining and spatial analysis. Therefore, eleven latent factors were considered, including elevation (Digital Elevation Model, DEM), slope, aspect, curvature, NDVI (Normalized Difference Vegetation Index), fault, geology, soil, land use, river and road. Besides, landslide inventories since 2004 to 2008 also adopted for extracting each causative factors. Detail information of the selected factors is listed in Table 1.

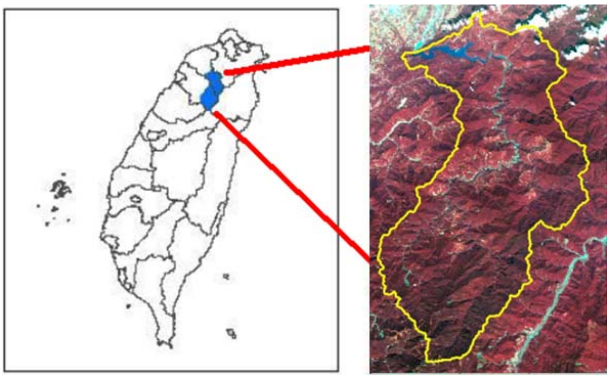

Figure 1. Study site

Data Type	Original Data	Derived Data	Notations
Raster	DEM	Elevation	40 m x 40 m
		Slope	
		Aspect	
		Curvature	
	SPOT Images	NDVI	10 m x 10 m
Vector	River	Distance of each pixel to the nearest river	-
	Road	Distance of each pixel to the nearest road	Large scale topographic map
	Fault	Distance of each pixel to the nearest fault	1/50,000
	Landuse	-	-
	Soil	-	1/25,000

Table 1. All materials in this research

Among the selected factors, NDVI data were derived from original satellite images that contain different radiometric and atmospheric conditions. Consequently, there may be biases if these NDVI values are compared or analyzed directly. Pseudo Invariant Features (PIFs) normalization is one of relative radiometric correction methods, which performs linear stretch or histogram matching based on PIFs (Schott et al., 1988). It is very convenient and suitable for analyzing multi-temporal NDVIs. Therefore, all the NDVI images were normalized using PIFs in this study.

3. PROCEDURE AND METHODOLOGY

The objective of this research is to perform validation and risk assessment of landslide events in the study site using SA and DT algorithms. There are four primary steps, i.e. data pre-processing and integration, analytic strategies, kernel computation and results as illustrated in Figure 2. In the data pre-processing and integration step, eleven factors were considered, including vector and raster data. Because the algorithm is record- or grid-based, vector data need to be rasterized. Furthermore, all data were pre-processed to remove null value or noise, and resampled to the same cell size. In our case, 10 m by 10 m pixel size was used. Subsequently, some factors that can provide advanced information after SA, such as aspect, curvature and slope were derived from DEM; NDVI was produced from original satellite images and normalized with PIFs; distance information about each pixel to the nearest target was generated from GIS poly-lines of rivers and roads. Finally the pre-processed data were integrated for subsequent analysis.

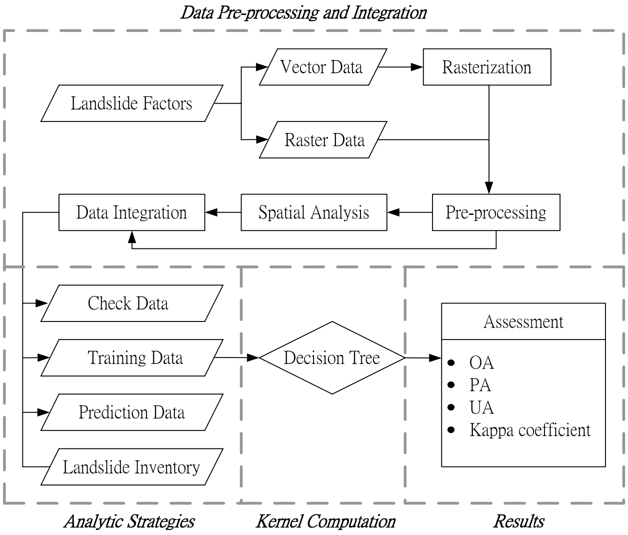

Figure 2. The procedure of this study

For the analytic strategies, this study used 2/3 of collected landslide inventory from 2004 to 2007 as the training dataset; the remainder in the same period were used as check-data. The models of data mining results were used to predict landslide events in 2008. The forecast was then verified and evaluated using real landslides in 2008 supplied by the watershed authority as reference data. On the other hand, all non-landslide records were randomly sampled approximately tenfold of landslide records. Landslide records of training, check and prediction are listed in Table 2. For the kernel computation, DT algorithm was utilized. Finally, the results of OA (Overall Accuracy), PA (Producer's Accuracy), UA (User's Accuracy) and kappa coefficient were computed from error or confusion matrix for assessment, for both check and prediction data.

Data	Training	Check	Prediction
Records	25193	12596	1839

Table 2. Landslide records of analytical data

DT algorithm is a supervised classification, and is a classical, universal and comprehensible method in the DM domain. The outcomes of DT are constituted by “If and Then” rules, and these sequences are helpful in realizing the reasons and interactions between causative factors in the landslide records. It usually forms a tree structure built with training data. The construction is composed by root (i.e. one of condition attributes), internal nodes (i.e. condition attributes) and leaf nodes (i.e. decision attribute). Each path from root to a leaf node represents a rule.

There are two important steps in the DT operator. The first is to develop the tree, i.e. branches are separated by computing and comparing degree of impurity of each condition attribute. The second is to prune the tree. The purpose of pruning is to avoid DT model becoming too over-fitting to carry validation and prediction. However, the arithmetic degree of impurity is different between nominal (or discrete) and quantitative (or contiguous) data because of their data structures. In general, information gain and Gini index are the major quantification for degree of impurity. The former is defined as Eq (1), the latter is defined as Eq (2). This study utilized the J48 algorithm in WEKA software (http://www.cs.waikato.ac.nz/ml/weka/) to build the decision tree because it can handle nominal and quantitative data at the same time.

$$
\begin{aligned}
&I(p,\ n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}\\
&E(A) = \sum_{i=1}^{v}\frac{p_i+n_i}{p+n}I(p_i,\ n_i)\\
&Gain(A) = I(p,\ n) - E(A)
\end{aligned}
\qquad (1)
$$

where p = number of positive in the decision attribute
n = number of negative in the decision attribute
I(p, n) = entropy of all condition attributes
A = one of the condition attributes
v = number of different contents in a specific attribute
E(A) = entropy of specific condition attribute
Gain(A) = information gain of specific condition attribute

$$
\begin{aligned}
&Gini(\mathrm{A} \le D\ or\ \mathrm{A}{>}D) = 1 - \sum_{j=1}^{2} f_j^2\\
&Gini_A(D) = \frac{N_1}{N}Gini(\mathrm{A} \le D) + \frac{N_2}{N}Gini(\mathrm{A}{>}D)
\end{aligned}
\qquad (2)
$$

where D = segmented point to divide contiguous data into two parts
A = one of the condition attributes
f = frequency of the positive and negative in A≦D or A>D range
N = total records of training data
N1, N2 = total numbers of A≦D and A>D domain respectively

4. RESULTS AND DISCUSSIONS

4.1 Preliminary results

The evaluations of check and prediction are shown in Table 3 (a) and (c), respectively. It is clear that the check result is good enough. However, the omission (100%-PA) and commission (100%-UA) of landslides are too high in the prediction (risk assessment) phase to obtain acceptable results (low kappa values). It may be caused by interaction between uncertain and multi-temporal problems in different data sets obtained from multi-sources. Therefore, the DT model after training becomes too over-fitting to carry prediction. To address this issue, a mechanism was developed to filter out uncertain data by standard deviation of data distribution (see next section).

	Non-landslide		Landslide		OA (%)	Kappa
	PA (%)	UA (%)	PA (%)	UA (%)		
(a)	99.1	98.9	88.5	90.2	98.1	0.8829
(b)	95.5	92.8	95.5	97.3	95.5	0.9053
(c)	93.1	98.8	65.3	23.9	92.2	0.3182
(d)	75.1	92.4	89.8	68.9	80.7	0.6134

Table 3. Accuracy assessment

4.2 The mechanism to filter out uncertain data

Because real world is a complex, varied and non-linear system, the measures are difficult to fit it. Thus each spatial data set probably contains uncertainty in the location, attribute, topology, process and visualization. Moreover, some biases may occur in multi-source and multi-temporal analysis. It is assumed that each data set of continuous type is a normal distribution. Every condition attribute of continuous data has a standard deviation (σ). Smaller standard deviation can get more pure data but fewer records. In the filtering process, if a record with all attributes conforming to a threshold (e.g. 5σ), it will be reserved. Table 4 shows the original records and the results after filtering. It is noted that smaller than 5σ is too strict to obtain enough records, but larger than 5σ, it may loss the meaning of filtering. Therefore, this paper utilized 5σ to filter uncertain data.

Records	Landslide		Non-landslide	
	Training & Check	Prediction	Training & Check	Prediction
Original	37790	1839	329786	19184
3σ	260	0	0	0
4σ	260	117	0	158
5σ	34624	1336	21993	1360

Table 4. Records of test data

4.3 Results after filtering

The results of landslide validation and prediction after filtering are shown in Table 3 (b) and (d). The check (validation) result remains excellent after filtering. In the prediction results, the OA decreased because the PA of non-landslide reduced. Extracting landslides knowledge is usually complex than non-landslides, but the samples in non-landslide are more than landslide, which may dominate the “overall” accuracy (OA) and reach high precision. However, the records of non-landslide in prediction are reduced by the filtering process (see Table 4); this is the reason that OA and PA of non-landslides decreased. In addition, the PA and UA of landslide substantially increased 24.5 % and 45 %, so kappa coefficient has improved significantly. In other words, this test indicates that the mechanism to filter out uncertain data can produce more reliable verification and risk assessment for landslides.

5. CONCLUSIONS

The decision tree algorithm and spatial analysis were utilized to extract landslide knowledge for validation and risk assessment of landslides in a watershed in Taiwan. Moreover, this paper presents a mechanism for filtering uncertain data from the data sets to improve the reliability of landslide predictions. The experimental results indicate that OA and kappa coefficient in verification can reach 98.1 % and 0.8829, respectively. However, the DT models after training are too over-fitting to carry prediction. After filtering uncertain data, PA, UA of landslides and kappa coefficient in the prediction task substantially increased at least 20 %. In conclusion, the spatial analysis and data mining algorithms combining the mechanism of filtering uncertainty data can perform verification and forecast of landslides with more reliable results in the study site.

References:

[1] Dahal, R. K., S. Hasegawa, A. Nonomura, M. Yamanaka, S. Dhakal, and P. Paudyal, 2008. Predictive modelling of rainfall-induced landslide hazard in the lesser himalaya of nepal based on weights-of-evidence, *Geomorphology*, 102(3-4), pp.496-510.

[2] Mettemicht, G., L. Hurni, and R. Gogu, 2005. Remote sensing of landslides: An analysis of the potential contribution to geo-spatial systems for hazard assessment in mountainous environments, *Remote Sensing of Environment*, Vol. 98, pp. 284-303.

[3] Nichol, J. and M. S. Wong, 2005. Satellite remote sensing for detailed landslide inventories using change detection and image fusion, *International Journal of Remote Sensing*, 26(9), pp. 1913-1926.

[4] Peduzzi, P., 2010. Landslide and vegetation cover in the 2005 Northern Parkistan earthquake: a GIS and statistical quantitative approach, *Natural Hazards and Earth System Sciences*, Vol. 10, pp. 623-640.

[5] Sakar, S. and D. P. Kanungo, 2004. An integrated approach for landslide susceptibility mapping using remote sensing and GIS, *Photogrammetric Engineering & Remote Sensing*, Vol. 70, pp. 614-625.

[6] Schott, J. R., Salvaggio, C. and Volchok, W. J., 1988. Radiometric Scene Normalization using Pseudoinvariant features, *Remote Sensing of Environment*, Vol. 26, pp. 1-16.

[7] Sidle, R.C., A. J. Pearce, and C. L. O'Loughlin, 1985. Hillslope stability and land use, *Water Resources Monograph*, Vol. 11, pp.140-141.

[8] Tsai, F. and L. C. Chen, 2007. Long-term landcover monitoring and disaster assessment in the Shiman Reservoir Watershed using satellite images, *in: Proc. 13th CeRES International Symposium on Remote Sensing*, Chiba, Japan.

[9] Wu, W., and R.C. Sidle, 1995. A distributed slope stability model for steep forested basins, *Water Resource Research*, Vol. 31, pp.2097-2110.

[10] Zhou, C., C. Lee, J. Li and Z. Xu, 2002. On the spatial relationship between landslides and causative factors on Lantau Island, Hong Kong, *Geomorphology*, Vol. 43, pp. 197-207.