Environ Earth Sci (2011) 63:981–997
DOI 10.1007/s12665-010-0775-0

ORIGINAL ARTICLE

The knowledge expression on debris flow potential analysis through PCA + LDA and rough sets theory: a case study of Chen-Yu-Lan watershed, Nantou, Taiwan

Tsu-Chiang Lei · Shiuan Wan · Tien-Yin Chou · Hung-Chieh Pai

Received: 22 May 2009 / Accepted: 28 September 2010 / Published online: 10 November 2010

Abstract Debris flow is often performed through identifying and analyzing the soil condition, hydraulic, geomorphological factors and vegetation conditions. In the present study, a spatial information analysis system is combined with a linear statistical method (principle components analysis with linear discriminant analysis, PCA + LDA) and an advanced data mining technique (discrete rough sets, DRS) to investigate the debris flow occurrence based on geomorphological and vegetation conditions factors. The analyzed data sources include (1) digital elevation model: to investigate the variation in the landscape, and (2) remote sensing data: to analyze the vegetation and plant conditions on the ground surface. The objective of this research is to define a method with the ability to forecast the level of debris flow susceptibility through the parallel study of statistical outcomes (PCA + LDA) and data mining results (DRS). The outcomes from PCA + LDA are inadequate due to the thresholds of the influenced variables not being examined. In this study, the DRS approach not only showed satisfactory results for the thresholds of influenced variables in the study area, but also the occurrence rules of debris flow are generated. Finally, the results show superior classification accuracy (70.8% for debris flow occurrence) for the DRS method over those of PCA + LDA analysis (54.2% for debris flow occurrence) for the analysis of debris flow occurrence. Therefore, this is an encouraging preliminary approach in the hazard assessment of debris flow.

Keywords Debris flow · Discrete rough set · Linear discriminant analysis · Principle components analysis · Occurrence rule

Introduction

Taiwan is an island with one-third of its area located in mountainous zones. The scarcity of usable land resulted in many housing units and farmhouses being built into the hillsides. In addition, earthquakes and typhoons frequently occur because Taiwan is located on the Circum-Pacific Earthquakes Belt and Western-Pacific Typhoon Path. First, landslides and severe erosion over the years on steep hills consisting of relatively erosive geological materials have resulted in abundant colluvial accumulation, especially after the 921 Chi-Chi Earthquake. Second, the average annual rainfall is more than 2,500 mm with a significant contribution from typhoons. The heavy rainfalls result in fragile geological materials and colluvium. The site conditions of steep landform and heavy rainfall will easily produce debris flow. Consequently, the casualties, property loss, and structure damage caused by debris flows have dramatically increased in recent years.

T.-C. Lei (✉)
Department of Urban Planning and Spatial Information,
Feng Chia University, Taichung 407, Taiwan, ROC
e-mail: tclei@fcu.edu.tw

S. Wan
Department of Information Management, Ling Tung University,
Taichung, Taiwan, ROC
e-mail: shiuan123@mail.ltu.edu.tw

T.-Y. Chou
Department of GIS Research Center, Feng Chia University,
Taichung 407, Taiwan, ROC
e-mail: jimmy@gis.tw

H.-C. Pai
Department of Environmental and Spatial Information Science
and Technology, Feng Chia University, Taichung, Taiwan, ROC

In reality, the mechanism of debris flows is quite difficult to analyze. Efforts to develop adequate theoretical models for debris flow have been limited by the lack of understanding of how such flows occur with regard to the given site conditions such as (1) geomorphology, (2) geology, (3) hydraulic properties, and (4) soil conditions (Pierson 1994; Floris et al. 2004). However, since the debris flows have led to many disasters, the literatures have to be reviewed for a possible resolution. Based on the previous literature, the analysis processes of debris flow can be classified into one of three categories:

1. *Indicators of factors*. Chang and Hsieh (1997) and Chang (1998) investigated the debris flow events in several potential regions in Taiwan. Hsieh et al. (1995) used critical precipitation line for predicting the occurrence or non-occurrence of a debris flow problem. In addition, the occurrence predictions of debris flow with instant rainfall intensity could not be estimated precisely, i.e., rainfall intensity was not the only dominant factor causing debris flows. As a matter of fact, the obstacle for analyzing the debris flow involved many environmental factors which require a good strategy to extract their importance (Pachauri and Pant 1992; Donati and Turrini 2002; Carrara et al. 2003). Floris et al. (2004) reported the core triggering factors of debris flow in Northern Apennines (Italy). The slope, which consists of mainly clayey and clayey-marly terrains, has been affected by landslide triggering phenomena. However, the development of an analysis model on prediction can be of help to create a warning system for landslide risks mitigation.
2. *Statistical factor analysis*. Johnson and Rodine (1984) presented the slope stability is an important factor causing debris flow. Besides these two factors of rainfall and slope, soil weight and water level are also important factors (Wang 1994). Recently, Melelli and Taramelli (2004) started from an inventory of hillslope hollows from air photographs and fieldwork for mapping debris-flow events down-slope from the initiation sites. Their analysis of the morphogenetic factors influencing slope instability processes was used to define a representative elementary area (REA) and to examine causal relations between the factors and debris-flow events. Some of the GIS-based methods used for landslide susceptibility mapping draw a great attention to scientists and engineers; for example, the weights of evidence method (Van Westen et al. 2003), the multi-variable statistical analysis with logistic regression (Dai et al. 2001; Dai and Lee 2003; Wang and Sassa 2005), frequency ratio method (Carrara et al. 1999; Lee and Sambath 2006) and the discriminant analysis (Baeza and Corominas 2001; Santacana et al. 2003) are most widely used.
3. *Artificial intelligent analysis approaches*. Lee and Chang (1995) presented a fuzzy model for the prediction of debris flow, but did not consider the cumulative rainfall as an important factor. Moreover, this fuzzy model after defuzzification methods cannot provide a boundary of clear ranges which can be used for the indication of occurrence of a debris flow. Chang and Lee (1997) analyzed the instant rainfall intensity in the areas of debris flow using the group method of data handling (GMDH; developed by Ivakhnenko (1970)) with artificial neural networks (ANN).The ANN model with physical terrain factors was applied to the study of landslides, in particular, indirect determination of the triggering parameters and landslide susceptibility mapping (Mayoraz et al. 1996; Fernàndez-Steeger et al. 2002; Ermini et al. 2005). ANN provides a quick and effective way to estimate the occurrence in a given zone; however, it adopts a black-box model, and the trained network (weights and thresholds) cannot be reused.

With the progress of spatial data survey techniques in geosciences, massive amounts of data or information can be easily collected and monitored. Thus, the analysis of influenced variables of debris flow/landslides becomes complicated. There are many possible techniques for data classification. However, those variables of depictive data in a given watershed have many uncertainties which may request a preprocessing analysis for enhancing their accuracy. Further, those variables also need to be sieved out their characteristics and features. Basically, common concepts for the dimension reduction approaches are to reduce the dimensionality but, in the process, irreversibly transform the descriptive dataset features. These methods include. (1) Hard dimension reduction problems: typical methods include principle component analysis (PCA) (Devijver and Kittler 1982) and rough set analysis (Nguyen and Skowron 1995; Chouchoulas and Shen 2001). (2) Soft dimension reduction problems: the typical method is factor analysis (Friedman and Tukey 1974). (3) Visualization problems: typical methods include projection pursuit (Mardia, et al. 1979) and multidimensional scaling (Torgerson 1952). However, these data reduction methods are widely applied in many fields, especially when the data contain (1) reduced chaotic information in the data set, (2) redundant depictive factors, and (3) incomplete measurements in the factor analysis. Accordingly, a possible mining technique (dimension reduction) has to be developed to extract the influencing factor(s) on debris occurrence. In the mean time, once the factor analysis can be successfully produced, a preliminary analysis for collecting the influencing factors will be greatly reduced, i.e., a valuable concept and knowledge may become a crucial idea to tackle debris flow hazards.

As part of this study, a well-known statistical method, combining PCA and linear discriminant analysis (LDA), is used to study the data classification and dimension reduction problem involved in debris flow. The PCA is a linear combination of variables (attributes) to access a compromised output with the purpose of reducing with data dimensions. To sum up, PCA is a dimensionality reduction tool in common use, perhaps due to its conceptual simplicity and the existence of relatively efficient algorithms for computation. PCA aims to find a new set of dimensions (attributes) that better captures the variability in multiple dimensions. However, the drawback of PCA is that all attributes influence the output decision. PCA aims to find a new set of dimensions (attributes) that better captures the variability in multiple dimensions, i.e., the first dimension is chosen to capture as much of the variability as possible. It requests the reduction of the dimensions of data (Tian et al. 2005; Mundt et al. 2005). Therefore, data mining could be a possible solution to tackle the tedious computational work in reducing dimensionality.

Data mining (Lei et al. 2008; Wan et al. 2008, 2009) had become a brand new approach in analyzing landslides and geosciences. This research used the discrete rough sets (DRS; Nguyen and Skowron 1995; Nguyen and Nguyen 1998a, b) to tackle the uncertainties arising from the materials and parameters involved in an observed landslide. The concept of DRS arises from conventional rough sets. The conventional rough set can only resolve data that are pre-classified into certain levels of groups. As a matter of fact, natural or environmental data are distributed either uniformly or normally. However, the revolution of DRS turns those continuous data into appropriate levels of groups mathematically. In other words, the separate point of DRS can successfully break the real world data into several levels and transform them to the Information Table. It (1) extracts numerous debris flow description factors for core influencing factor(s), (2) searches the segmentation points (thresholds) with regard to the core influencing factor(s), (3) establishes debris flow occurrence of the knowledge description (interpretation rule). In accordance with the previous steps, the accuracy of the debris flow can be greatly improved.

The study has been divided into four parts. In the first part, the development of the study area, geomorphology and land-cover factors for database, is discussed. In the second part, combined LDA + PCA methods (LDA + PCA) are introduced. The third part will briefly introduce the DRS method. The fourth part shows the results of a parallel analysis of the landslide problem through (a) LDA + PCA method and (b) DRS method. The data analysis is carried out by the DRS method and rational results are obtained.

The geomorphology, land-cover factors and vegetation index for the debris flow problem

Varnes (1978) defined a debris avalanche as a rapid flow of predominantly coarse debris consisting of soil and/or weathered bedrock. Debris flow originates when poorly sorted debris (rock, soil, woody debris, etc.) is mobilized from hill slopes and channels by sufficient moisture in the soil. Cruden and Varnes (1996) proposed a classification process: landslide events can be classified as rotational–translational movements with respect to earth slides and earth flows. They are primarily affected by their formations, often highly deformed, that widely outcrop in the mountain chain. Lin et al. (1993) presented the characteristics of debris flow in gravelly deposits as stream slope, rainfall, rainfall intensity, geological condition, grain size distribution, void ratio, shear strength, vegetation condition, and channeled topography. Lin et al. (1998) also discussed the contributing factors of debris flow events for the application of spatial information techniques (Remote Sensing and GIS). To sum up, from the past instances, it is quite difficult to determine a possible solution in finding the induced-factors or core factors with regard to the mechanism of the debris flow events.

Accordingly, the observed influenced factors from debris flow should be discussed rationally. The most likely contributing factors toward debris flow are topography, geology, watershed geometry factors and remote sensing data for vegetation condition (Lee and Choi 2004; Lin et al. 2007; Tian et al. 2005; Wan et al. 2008). However, the influenced factors can be organized into three major areas: (1) the geomorphology of surrounding watershed, (2) the geomorphology of surrounding stream, and (3) the land-cover by vegetation. In this research, the potential debris-flow streams distribution from the Water Conservation Bureau (WCB) database was categorized as the sub-watershed data of Chen-Yu-Lan stream (WCB Website 2008a). Further, in this study, a GIS database to describe debris flow in this study area is generated that can be further divided using 18 factors, including (1) watershed area, (2) watershed perimeter, (3) watershed of average elevation, (4) watershed of average slope, (5) watershed of primary length, (6) stream length, (7) geology index, (8) watershed width, (9) form factor, (10) stream density, (11) stream sinuosity, (12) average slope of stream, (13) total length of stream, (14) NDVI, (15) cover and management factor, (16) bare-soil land area, (17) bare-soil land evaluation rate, and (18) bare-soil land geology index. The index and symbol definition are given in Table 1. Some of the definitions of the symbols are quite straightforward, but some require further explanation as described in the following:

Table 1 The environmental factors of study site

Class	Factors	Symb.	Unit	Class	Factors	Symb.	Unit
Geomorphology of watershed	Watershed area	A	km^2	Geomorphology of stream	Stream density (*6)	SD	%
	Watershed perimeter	P	km		Stream sinuosity (*7)	T	–
	Watershed of average elevation	WD	m		Average slope of stream	SS	%
	Watershed of average slope	WS	deg		Total length of stream	L	–
	Watershed of primary length	L_0	km	Land-cover	NDVI (*2)	N	–
	Stream length	L_T	km		Cover and management factor (*3)	C	–
	Geology index	E	km		Bare-soil land area	D	km^2
	Watershed width	WW	–		Bare-soil land evaluation rate (*5)	DR	%
	Form factor (*1)	F	–		Bare-soil land geology index (*4)	DE	–

Note the "*" is referred to the indicators of influenced factors from study site

1. *Form factor*. Form factor (also known as shape factor) is the ratio of the minor axis to the major axis of the watershed area (Pareschia et al. 2002). Shrestha (2001) proposed a study on the restoration of vegetation for the conservation of the dilapidated mountainous regions of Nepal. He found that the slender-shaped (form factor is 0.14) watershed is useful in investigating the watershed characteristics. Herein, it was shown that the debris flow is highly correlated to the shape of the watershed area and his study selected this attribute to evaluate the occurrence of debris flow.
2. *Normalized difference vegetation index* (*NDVI*). To determine the density of vegetation on a patch of land, researchers must observe the distinct colors (wavelengths) of visible and near-infrared sunlight reflected by the plants (Lin et al. 2006a). As can be seen through a prism, many different wavelengths make up the spectrum of sunlight. Nearly all satellite vegetation indices employ this difference formula to quantify the density of plant growth on the earth—near-infrared radiation (NIR) minus red radiation (R) divided by near-infrared radiation plus red radiation (Bannari et al. 1995). The result of this formula is called the normalized difference vegetation index (NDVI). The values for NDVI are obtained from SPOT image. The range of this value is [−1, 1].
3. *Cover and management factor*. Cover and management factor (C value) is taken from the plant-cover condition of the universal soil loss equation (USLE) (WCB Website 2008b). The C value is in the 0–1 range. When the land is bared, C value is assigned as one. On the contrary, when the land has good vegetation condition, C value is approaching to zero. The C value varies with the vegetation type, season change, and the percentage of covered land (Lin et al. 2002a, b, 2006b; Özhan et al. 2005).

$$C = \frac{1 - \text{NDVI}}{2}. \tag{1}$$

4. *Bare land geology index* (*DE*). The weight factor of geology index was reported by CGS (CGS 2005). The Bare land geology index was computed by the following equation:

$$\text{DE} = \frac{\sum E_i \times A_i}{A}, \tag{2}$$

where E_i is the value with regard to the associated soil type, A_i is the observed sub-area of bare land, and A is the total watershed area. For instance, $E = 10$ for Schist or Slate type soil, $E = 6$ for old tertiary sedimentary rock, and $E = 3$ for New Tertiary Pleistocene (Lin et al. 2006b).

5. *Bare-soil land evaluation rate*. The bare-soil land evaluation rate is a ratio defined as the new landslide area divided by the water basin area. If this ratio is high, the watershed area becomes fragile. In this study, these areas of landslide change are collected and then identified based on two different periods of SPOT data (1999, 2001).This can be a detection process using them as a material to monitor the land-cover change area of debris flow occurrence.
6. *Stream density*. This is defined as the total length of all the streams and rivers in a watershed divided by the total area. Considering a certain level of stream density, the factor of soil permeability and the underlying rock type affect the runoff in a watershed which become a dominant factor on debris flow occurrence. Stream density can also affect the shape of a river's hydrograph during a rain storm.
7. *Stream sinuosity*. This is defined as the extent to which a river meanders within its valley, calculated by dividing the total stream length by the valley length. A high value of stream sinuosity is most likely to be flooded. Therefore, this value can be attained from GIS and DEM data will be helpful to understand the occurrence of debris flow.

Study area and material

The environmental features of debris flow events

The watershed of Chen-Yu-Lan River, located in the central part of Taiwan, was selected to be the study site as shown in

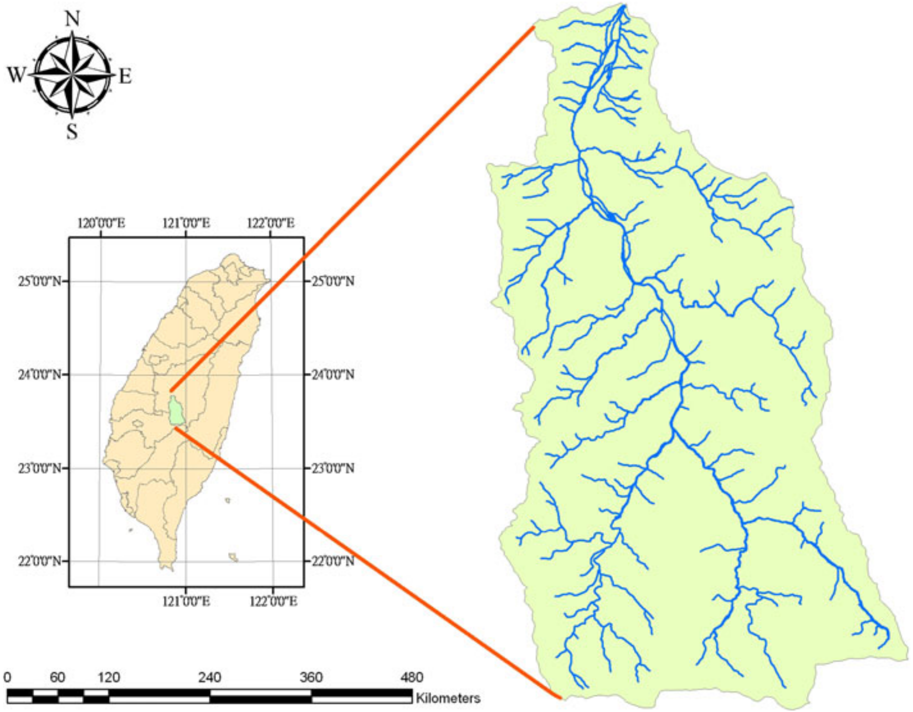
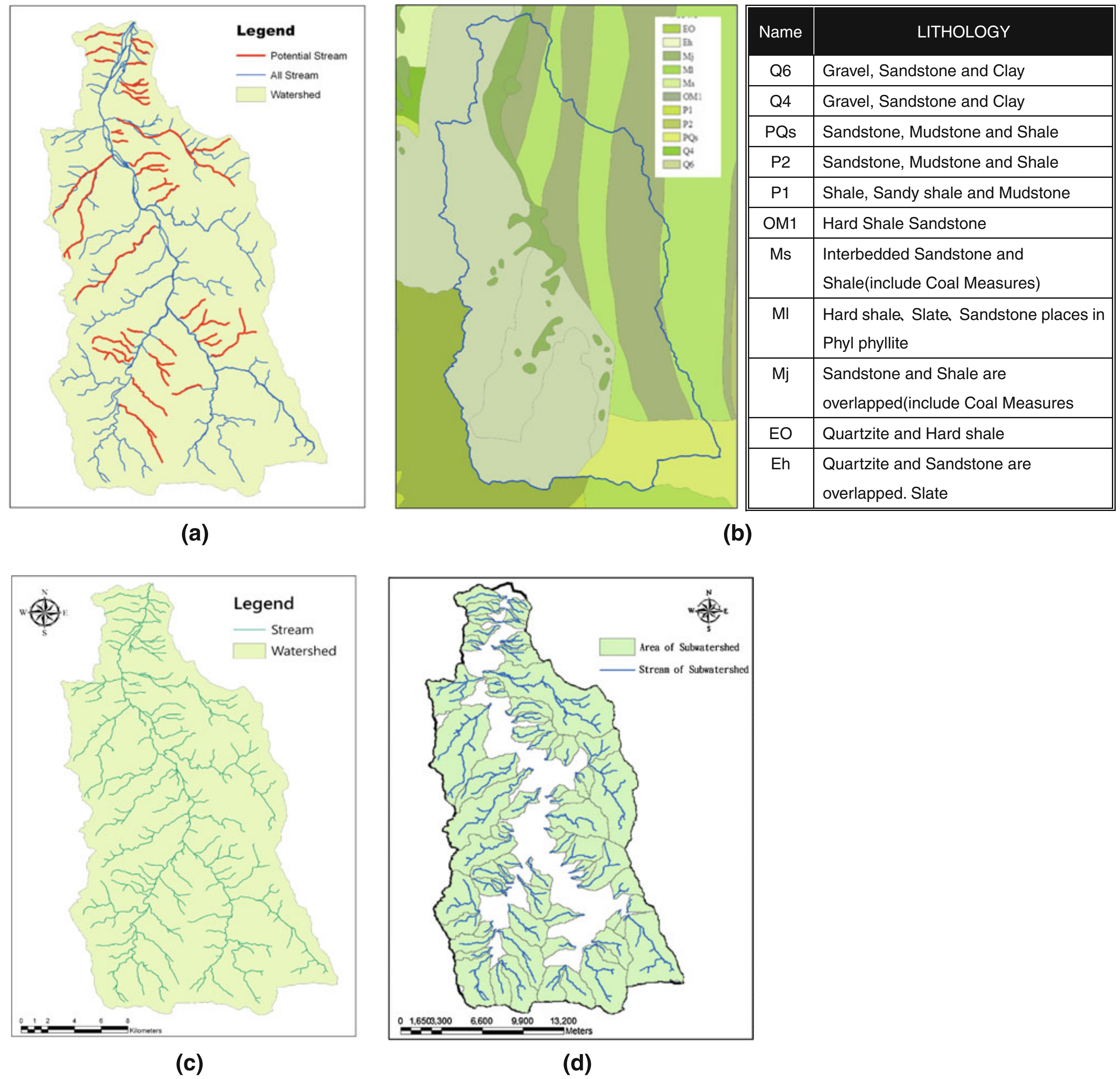


Fig. 1 The study area of Chen-Yu-Lan Stream

Fig. 1. The Chen-Yu-Lan River originates from the north peak of Yu Mountain with an elevation of 3,910 m. Chen-Yu-Lan River is one of the upper rivers of the Zhuoshui River system, which is the largest river system in Taiwan. Furthermore, Chen-Yu-Lan River has a length of 42.4 km with an average declination slope of 5%, and its watershed area is about 45,000 ha. From 31 July through 1 August (1996), the heavy rainfall brought by Typhoon Herb which induced 34 debris flows in the watershed of the Chen-Yu-Lan River (see Fig. 2a). As aforementioned, this area was already very fragile from the strong ground motion of Chi-Chi earthquake. Afterwards, a large precipitation of about 1,291 mm (peak discharge 195 m^3/s of 73 mm in 1 h) brought into the Chen-Yu-Lan River.

In this study, the research data consists of two formats: (1) vector and (2) raster data. The vector data includes (a) potential stream of debris flow, (b) geology, (c) river system, and (d) boundary line of sub-watershed (see Fig. 2). The potential stream of debris flow (see Fig. 2a) has ever taken place the debris disaster and the disaster may occur in the near future. However, this result from WCB (2008a) and Central Geological Survey (2005) and has developed a series evolution process of the environment factors. The scale of the geology diagram (Fig. 2b) is 1/250,000 displaying the geology distribution of the study region from CGS (2005). In fact, geology and morphology conditions can affect the occurrence of landslides and debris flow problem. In this map, different colors represent the different geological conditions of the Chen-Yu-Lan River. For example, the symbol of Q6 represents the Holocene epoch and the soil that is mainly constituted of gravel and sand. Other geological conditions are listed in Fig. 2a. Basically, these regions having complicated geology and fault crossovers can be good materials for fracture geology and the debris flow problem. Figure 2c demonstrates the river system and Fig. 2d illustrates the boundary line of the sub-watershed in the study area. On the other hand, the raster data consists of digital elevation model (DEM) data and remote sensing (SPOT4) data. To handle the geomorphology characteristics of the study area, a well-developed DEM data are generated. These data will be used to construct a knowledge rule for landslides. The DEM data were extracted from the aerial photos which adopted the HEC-Geo HMS module with the DEM data (40 m × 40 m resolution). The geomorphology factors of aspect, evaluation, slope and river system maps are then extracted from the DEM data. The SPOT Image of Chen-Yu-Lan River is monitored from upstream to downstream, as shown in Fig. 3. These data render a good evaluation on the overall range of these study samples of debris flow. In Taiwan, the SPOT image resolution cell was 12.5 m × 12.5 m (Center for Space and Remote Sensing Research of National Central University in Taiwan, CSRSR 2008). It was decided to reduce to 12.5 m × 12.5 m to meet the standard grid size. Additionally, the size of debris flow area was generated through artificial diagnosis (Arc-GIS file). Meanwhile, the boundary was generated. One of the advantages of using these areas is easily identified in the SPOT satellite image data by means of their spectral characteristics and regular field geometry. In addition, these areas of landslide change are identified based on two periods of SPOT data (1999, 2001). These images were collected after a typhoon or



Name	LITHOLOGY
Q6	Gravel, Sandstone and Clay
Q4	Gravel, Sandstone and Clay
PQs	Sandstone, Mudstone and Shale
P2	Sandstone, Mudstone and Shale
P1	Shale, Sandy shale and Mudstone
OM1	Hard Shale Sandstone
Ms	Interbedded Sandstone and Shale(include Coal Measures)
Ml	Hard shale、Slate、Sandstone places in Phyl phyllite
Mj	Sandstone and Shale are overlapped(include Coal Measures
EO	Quartzite and Hard shale
Eh	Quartzite and Sandstone are overlapped. Slate

(a) (b)

(c) (d)

Fig. 2 The potential stream and geology distributions of Chen-Yu-Lan Stream. **a** Potential stream of debris flow. **b** Geological map. **c** River system. **d** Bounder line of sub-watershed

heavy rains with cumulative precipitations of over 100 mm, which was taken in the aftermath of Bilis Typhoon (21/08/2000), Toraji Typhoon (28/07/2001) and Nari Typhoon (16/09/2001). When the Toraji Typhoon struck the Central Part of Taiwan, it triggered extensive landslides. The typhoon delivered very heavy rain to the Chen-Yu-Lan River (1,217 mm over 3 days). The area was already very disturbed following the strong ground motion of Chi-Chi earthquake and this led to a major calamity. Thus, this can be a detection process for using them as a material by monitoring the land-cover change area of debris flow occurrence.

Land-cover data extraction from the SPOT image

In this work, a two-stage study is designed to interpret the landslide pattern of satellite images. In the first stage, the traditional supervised classification method of Maximum Likelihood Classification (MLC) is used to obtain four major land-cover categories of (1) water, (2) forest, (3) landslide and (4) bare-soil area (almost river round). The detailed classification results are shown in Fig. 3. In addition, the use of the above measurements enables the monitoring of the land-cover change area of debris flow occurrence. To obtain a complete

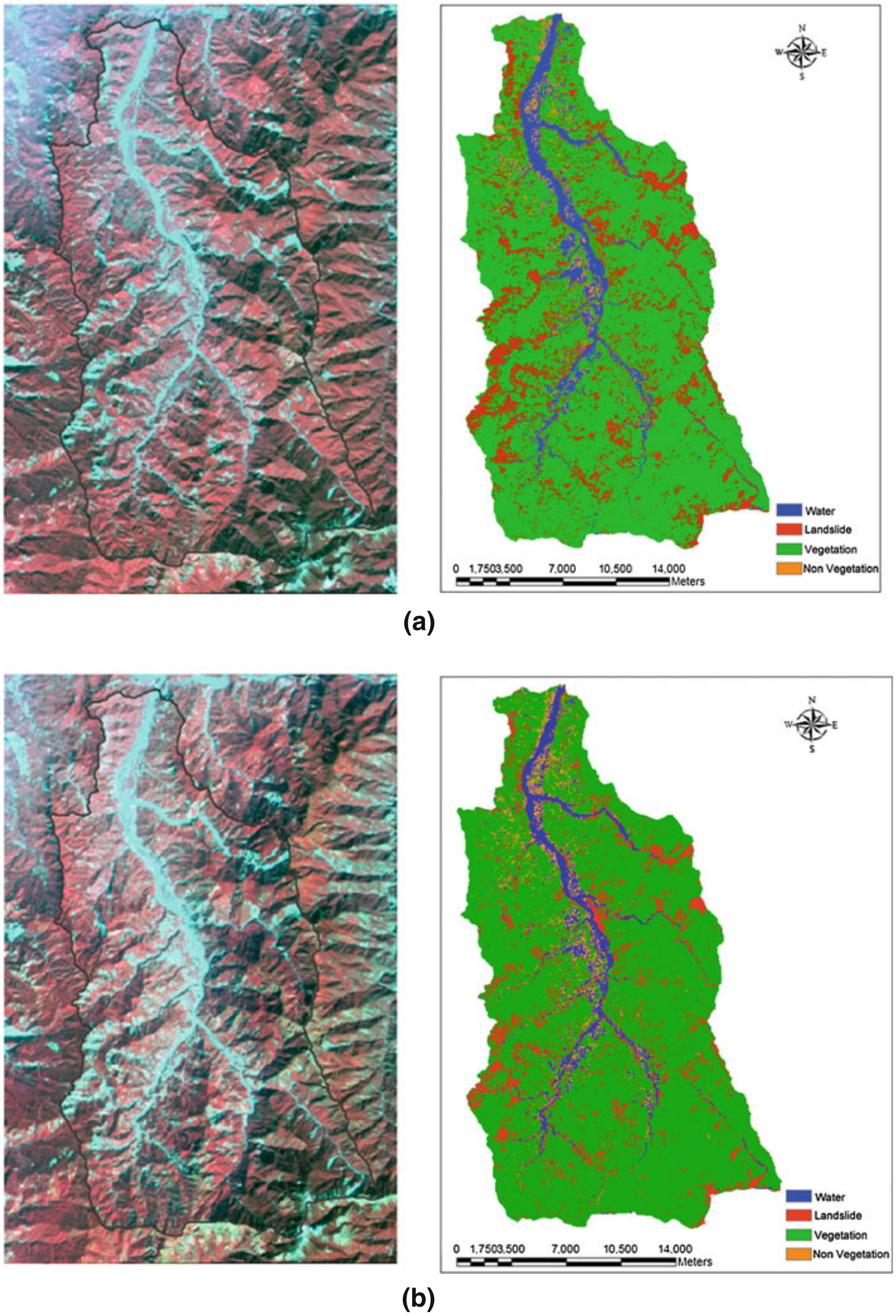


Fig. 3 The origin image and classification result of SPOT satellite data. **a** Origin image and classification result on 31/10/1999. **b** Origin image and classification result on 05/03/2001

land change distribution in the study site, a series of aerial photos are used to identify the classification process from SPOT images. These ground truth map is a raw data from the Aerial Survey Office, Provincial Department of Agriculture and Forestry of Taiwan (ASO 2008). Table 2 presents the accuracy of the image classification results (or the so-called error matrix) of sampling. It displays the accuracy outcomes from those four categories of satellite images. There are 400 samples randomly selected through ERDAS that are reliable enough to present the classification accuracy for each category. However, the overall accuracy is 92.2% (1999) and 91.75% (2001), respectively. The results and extractions are satisfactory for the generation of landslide data to the spatial information database. To observe the variation of vegetation on the land surface, the NDVI and cover and management factor are required. Lin et al. (2002a, b, 2006b) study some similar researches about debris flow occurrence indices. However, the study may involve redundant and surplus variables (Hsieh et al. 1995; Lin et al. 1998). Hence, the techniques for feature extraction and feature selection of core factors from debris flow events should be applied and discussed. The authors have a good experience in using feature

Table 2 Classification results of error matrix evaluation

	Reference data					User's accuracy (%)
	Landslide	Bare soil	Forest	Water	Total no. of rows	
1999 image						
Classification results						
Landslide	91	5	0	4	100	91.00
Bare soil	3	89	2	6	100	89.00
Forest	0	5	95	0	100	95.00
Water	3	3	0	94	100	94.00
Total no. of columns	97	102	97	104	400	–
Producer's accuracy (%)	93.81	87.25	97.94	90.38	–	–
2001 image						
Classification results						
Landslide	89	6	0	5	100	89.00
Bare soil	5	87	2	6	100	87.00
Forest	0	4	96	0	100	96.00
Water	2	3	0	95	100	95.00
Total no. of columns	96	100	98	106	400	–
Producer's accuracy (%)	92.71	87.00	97.96	89.62	–	–

extraction in paddy rice image classification (Lei et al. 2008) and feature selection debris flow classification analysis (Wan et al. 2008), and landslide susceptibility map (Wan et al. 2009). Applying this concept, feature extraction and feature selection are proposed in the following section.

Basic principle of classification method

Feature extraction process (PCA + LDA)

PCA

PCA is a well-known multivariate analysis technique for reducing data dimensions. The use of PCA allows a smaller number of variables in a multivariate data set (Fukunaga 1990; Mundt et al. 2005; Tian et al. 2005). Mathematically, PCA is a process that decomposes the covariance matrix of a matrix into two parts: eigenvalues and column eigenvectors. The reduction process is achieved by taking p variables X_1, X_2, …, X_p which are then combined to produce principal components (PCs) PC_1, PC_2, …, PC_p, that are uncorrelated. These PCs are also termed eigenvectors. The lack of correlation is a useful property as it means that the PCs are measuring different "dimensions" in the data. Nevertheless, PCs are ordered so that PC_1 exhibits the largest amount of variation, PC_2 exhibits the second largest amount of variation, PC_3 exhibits the third largest amount of variation, and so on. When using PCA, it is hoped that the eigenvalues of most of the PCs will be low so that they are virtually ignorable. Accordingly, sieving small amounts of variables in the original number of variables (X variables) can be described using the smaller number of PCs (Fukunaga 1990).

LDA

LDA is a classical statistical approach for classifying samples of unknown classes, based on training samples with known classes. LDA-related Fisher's linear discriminant (Fukunaga 1990) and machine learning to find the linear combination of features which best separate two or more classes of objects or events. Discriminant analysis differs from factor analysis in that it is not an interdependent technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made. The detailed processes can be found in Fukunaga (1990).

PCA + LDA

PCA aims to project the data in the direction of maximal variance. LDA is supervised and is used as the project axes. Among these extensions, PCA + LDA, a two-stage method, received relatively more attention in handling decision science. For instance, face recognition and analysis are extracted successfully from the features of the face patterns (Sahoolizadeh et al. 2008). Moreover, the LDA + PCA is the main trend in feature extraction has been representing the data in a lower dimensional space computed through a linear transformation satisfying certain properties (Yang and Yang 2003; Sahoolizadeh et al. 2008). Few of the studies used this method to solve the environmental problems in Geosciences. In this study, the evaluation of the debris flow (Baeza and Corominas 2001; Santacana et al. 2003) on Chen-Yu-Lan River through the PCA and LDA method which presents a similar concept of face recognition and analysis. Both methods have advantages and disadvantages for debris flow recognition. However, this study compared the performance of the two methods: (1) PCA + LDA method and (2) DRS for debris flow factors recognition, and the steps for the construction of the DRS theory are introduced in the following section.

Feature selection process (DRS)

This section introduces the progress of DRS. Unfortunately, the conventional rough set can only resolve data that are pre-classified into certain levels of groups. As a matter of fact, the actual environmental data are distributed uniformly. Hence, DRS is employed as an appropriate tool to evaluate them.

Quantization problems (Nguyen and Skowron 1995)

If $\mathbf{A} = (U, A \cup \{d\})$ is a decision table with a large number of values of objects from U for some $a \in A$, then there is a very small probability that a new object will be recognized by matching its attribute value vector with the rows of this table. Hence, for decision table with real value attributes, some discretization strategies are built to achieve a higher quality of classification.

Let $\mathbf{A} = (U, A \cup \{d\})$ be a decision table where $U = \{x_1, x_2, \ldots, x_n\}$. It is assumed that $V_a = [l_a, r_a) \subset \Re$ for any $a \in \mathbf{A}$, where $\Re$ is the set of real numbers. $\mathbf{A}$ is assumed to be a consistent decision table. Let P_a be a partition on V_a (for $a \in \mathbf{A}$) into subintervals i.e.

$$P_a = \left\{ \left[C_0^a, C_1^a\right], \left[C_1^a, C_2^a\right], \ldots, \left[C_k^a, C_{k+1}^a\right) \right\} \tag{3}$$

for some integer k, where $l_a = C_0^a < C_1^a < C_2^a < \cdots < C_k^a < C_{k+1}^a = r_a$ and,

$$V_a = \left[C_0^a, C_1^a\right) \cup \left[C_1^a, C_2^a\right) \cup \cdots \cup \left[C_k^a, C_{k+1}^a\right). \tag{4}$$

Any P_a is uniquely defined by the set $C_a = \{C_0^a, C_1^a, C_2^a, \ldots, C_k^a, C_{k+1}^a\}$ called the set of cuts on V_a [the set of cuts is empty if *card* $(P_a) = 1$]. In the sequel, one identify P_a with the set of cuts on V_a defined by C_a. Any family $\{P_a : a \in A\}$, where P_a is a partition on V_a called a partition on $\mathbf{A}$. Then any family $P = \{P_a : a \in A\}$ of partitions can be represented by $P = \cup \{a\} \times C_a$. Any pair $(a, c) \in P$ will be called a cut on V_a.

Any family $P = \{P_a : a \in A\}$ of partitions on $\mathbf{A}$ defines from $A = (U, A \cup \{d\})$ a new decision table $A^P = (U, A^P \cup \{d\})$, where $A^P = (a^P : a \in A)$ and $a^P(x) = i \Leftrightarrow a(x) \in [C_i^a, C_{i+1}^a)$ for any $x \in U$ and $i \in \{0, \ldots, k\}$. The table A^P is called $\mathbf{P}$-quantization of $\mathbf{A}$.

Two families of partitions $\boldsymbol{P}'$, $\boldsymbol{P}$ on $\mathbf{A}$ are equivalent, i.e. $P' \equiv {}_A P$, if and only if $A^P = A^{P'}$. The equivalence relation '$\equiv_A$' has a finite number of equivalence classes. In the sequel, it is not being distinguished between equivalent families of partitions.

The quantization problems of real value attributes of A can be described as decision problem:

Complexity of quantization problems (Nguyen and Skowron 1995)

Let $A = (U, A \cup \{d\})$ be a decision table where $U = \{x_1, x_2, \ldots, x_n\}$. An arbitrary attribute $a \in A$ defines a sequence $V_1^a < V_2^a < \cdots < V_{na}^a$, where $\{V_1^a, V_2^a, \ldots, V_{na}^a\} = \{a(x) : x \in U\}$ and $n_a \le n$

Let P_k^a be a prepositional variable corresponding to the interval $[v_k^a; v_{k+1}^a)$ for any $k \in \{1, \ldots, n_a - 1\}$and $a \in A$. By $BV(A)$, it is denoted as the set of all prepositional variables of the above form.

Any partition $P \subseteq \cup_{a \in A} \{a\} \times V_a$ defines a valuation val-p of prepositional variables P_k^a by valp$(P_k^a) = \mathbf{true}$ iff there exists a cut $(a, c_a) \in \boldsymbol{P}$ satisfying $v_k^a \le c_a < v_{k+1}^a$. Instead of valp$(P_k^a) = \mathbf{true}$, it can be also written as $P| = P_k^a$.

By $\varphi\{a, i, j\}$ it is denoted as a disjunction of all Boolean variables from the set:

$$P_k^a : \left[v_k^a, v_{k+1}^a\right) \subseteq \left[\min(a(x_i), a(x_j)); \max(a(x_i), a(x_j))\right) \tag{5}$$

Hence valp$\{\varphi(a, i, j)\} = \mathbf{true,}$ iff there is a cut in $\mathbf{P}$ on V_a between $a(x_i)$ and $a(x_j')$.

By $\Re^k(i, j)$, it is denoted as a disjunction of all $\varphi\{a, i, j\}$,where $a \in A$ and $a(x_i) \neq a(x_j)$. Formula $\Re^k(i, j)$ is called the discernibility formula for objects $x_{i,}$ x_j (it is assumed that the disjunction of the empty set of variables to be equivalent to **true**).

The discernibility Boolean prepositional formula of $\mathbf{A}$ is denned by:

$$\Phi^A = \Lambda\left\{\psi(i,j) : d(x_i) \neq d(x_j)\right\} \tag{6}$$

Any non-empty set $S = \{P_{k1}^{a1}, P_{k2}^{a2}, \ldots, P_{kr}^{ar}\}$of Boolean propositional variables from BV(A) defines a family of partition $P(S)$ as follows:

$$P(S) = \left\{ \left(a_{1,} \frac{v_{k1}^{a1} + v_{k1+1}^{a1}}{2}\right), \left(a_{2,} \frac{v_{k1}^{a1} + v_{k2+1}^{a2}}{2}\right), \ldots, \left(a_1 \frac{v_{kr}^{ar} + v_{kr+1}^{ar}}{2}\right) \right\}. \tag{7}$$

To make the theory more easily understandable, an example based upon the above is demonstrated.

Stages of DRS

There are four stages in DRS analysis. In the first stage, the development of the "Information Table" is required for describing the characteristic attributes. In this table, the relation in the multi-attribute set is displayed. In the information table, each row represents a new case (or object). Each of the columns represents the respective variables (or condition attributes). In this study, the variables can be the site condition such as the geomorphology, land-cover, and river density. The outcome (also called the concept or decision attribute) of each object is either 1 or 0, indicating whether the particular case of debris-flow has occurred.

In the second stage, all the attributes must be clustered into *appropriate classes* to construct a "Decision Attribute". The crucial aspect is to find the *appropriate classes*. In other words, if the separate points can be found, then the appropriate classes of attributes can be determined. Please refer to Nguyen and Skowron (1995) or Eq. 7 for the detailed process of how the separate points are calculated. The rough set provides a possible solution in discretizing the chaotic information. In this study, a new concept is proposed to deal with the uncertainty of classification in the debris flow problem.

The third stage is to attain the cores and reducts of the data attributes. There are two fundamental concepts related to attribute reduction. The minimal subsets of attributes that discern all equivalent classes of the relation, which is discernable by the entire set of attributes, are called reducts. The core is the common part of all reducts.

The last stage is the most important application of rough sets which is the generation of decision rules for a given Information Table to predict the classes for new objects that are beyond visual observation. Using a reduced Information Table, the rules could be found through determining the decision attributes value based on the condition attributes values. Therefore, the rules are presented in an "IF condition(s) THEN decision(s)" format. If the condition(s) in the IF part matches with the given fact(s), the decision(s) in the THEN part will be performed.

Results

The results of this study are divided into five parts: (1) strategy for selecting effective samples, (2) application of multi-variable analysis, (3) results of PCA + LDA, (4) results of DRS, and (5) comparison of results for (3) and (4).

Strategy for selecting effective samples

One thousand and four hundred and twenty (1,420) creeks in Taiwan are classified as hazardous debris-flow creeks according to the maps published by the Council of Agriculture. They are referred to the report of the Council of Agriculture (Lin et al. 2002a, b, 2003). This research found there are 73 potential streams (including 146 recorded data from 2000 and 2001) in the study region. The maps are generated by the government of Taiwan and are generally recognized as a relatively accurate source of resource data (WCB Website 2008a). They are the raw data (study material) collected for the spatial information database. In this study, 18 selected typical streams [a total of 36 samples including 9 occurrences and 9 non-occurrences of debris flow from the WCB Website (2008a)] were judged to be debris-flow hazards based on the evaluation of the aforementioned factors (see Table 1) reported by Lin et al. (2002a, b, 2003), Jan and Chen (2005), CGS (2005) and WCB Website (2008a). In this study, it is decided to select the most vulnerable catchments as the training dataset for watershed areas of different sizes suggested by Lin et al. (2002a, b, 2003) and CGS (2005). The rest of the data (testing data for verification) were used from Lin et al. (2002a, b), which have been evaluated as the most fragile debris flow area. The attributes represent the in situ conditions and the decisions represent the occurrence ($d = 1$) or non-occurrence ($d = 0$) of the debris flow (refer to Table 3). These selected data are recognized as the most representative data of the training data. On the other hand, the rest of the data (55 streams for 110 testing data) are used for verification. During the heavy rainfall of Toraji Typhoon (28/07/2001), the geological materials and colluvium are easily weakened, which often leads to a debris flow. In this study, related data concerning the Chen-Yu-Lan River were collected on November 2001, through the spatial database and site investigation of a typhoon that occurred on 22 May 2002.

Application of multi-variable analysis

Initially, multi-variable analysis is used to solve the debris flow training data. In the study cases, PCA is used to compute its data reduction and the feature extraction method of the training data. Table 4 shows the total variance outcomes of the dataset. Among those components, after the fourth component is selected, it can be observed that the trend of the accumulated given value becomes smoother. That is, the components after fifth do not significantly influence the decision. Consequently, the first to fourth components govern 84.87% of the variation in outcomes. Finally, this case selected the first four factors (PCA1–PCA4) to present the major factors of the study site. On the other hand, in reviewing the past literatures, Pratsinis et al. (1988) announced that when the eigenvector is greater than 0.7, then the related factors can be considered as prominent factors of datasets for the debris flow

Table 3 The 36 testing samples of debris flow decision table from study site

No.	*A*	*P*	WD	WS	L_0	L_T	*L*	WW	*F*	SD	*T*	SS	*E*	*N*	*C*	*D*	DE	DR	Decision
41	0.0060	0.0681	0.1452	0.3808	0.1055	0.1182	0.0490	0.0363	0.0207	0.7216	0.2406	0.5393	0.9871	0.6591	0.3409	0.0000	0.0000	0.0000	1
43	0.0506	0.1798	0.1170	0.4480	0.2416	0.2557	0.1123	0.0961	0.0051	0.4705	0.3827	0.2375	0.1512	0.2673	0.7327	0.0611	0.6000	0.3294	1
45	0.0222	0.1134	0.1015	0.5459	0.1496	0.1740	0.0695	0.0729	0.0181	0.5489	0.1700	0.1411	0.1512	0.1138	0.8862	0.0703	0.6000	0.6310	1
46	0.0334	0.0970	0.2087	0.5278	0.1126	0.1289	0.0933	0.1805	0.0682	0.5527	0.2014	0.3354	1.0000	0.6581	0.3419	0.0212	1.0000	0.1507	1
47	0.0108	0.0862	0.1598	0.4814	0.1301	0.1340	0.0605	0.0383	0.0133	0.7100	0.4168	0.2924	1.0000	0.4446	0.5554	0.0292	1.0000	0.3566	1
53	0.0907	0.2198	0.2846	0.4530	0.2410	0.2036	0.1289	0.2171	0.0281	0.2904	0.9550	0.4258	1.0000	0.6005	0.3995	0.0379	1.0000	0.1307	1
66	0.0601	0.1575	0.4801	0.6950	0.2337	0.2532	0.1295	0.1315	0.0131	0.4693	0.3276	0.4495	0.1311	0.5633	0.4367	0.1200	0.5849	0.5708	1
68	0.2047	0.2953	0.5648	0.7768	0.3231	0.3346	0.1950	0.4067	0.0416	0.1581	0.4441	0.2549	0.1397	0.7676	0.2324	0.0744	0.5992	0.1266	1
69	0.2603	0.4378	0.6834	0.6907	0.4428	0.4647	0.2359	0.3743	0.0191	0.1434	0.4185	0.2181	0.1512	0.6235	0.3765	0.1897	0.6000	0.2590	1
114	0.0060	0.0681	0.1452	0.3808	0.1055	0.1182	0.0490	0.0363	0.0207	0.7216	0.2406	0.5393	0.9871	0.6488	0.3512	0.0129	0.9792	0.1864	1
116	0.0506	0.1798	0.1170	0.4480	0.2416	0.2557	0.1123	0.0961	0.0051	0.4705	0.3827	0.2375	0.1512	0.5934	0.4066	0.0746	0.6000	0.4025	1
118	0.0222	0.1134	0.1015	0.5459	0.1496	0.1740	0.0695	0.0729	0.0181	0.5489	0.1700	0.1411	0.1512	0.6482	0.3518	0.0577	0.6000	0.5175	1
119	0.0334	0.0970	0.2087	0.5278	0.1126	0.1289	0.0933	0.1805	0.0682	0.5527	0.2014	0.3354	1.0000	0.7486	0.2514	0.0162	1.0000	0.1149	1
120	0.0108	0.0862	0.1598	0.4814	0.1301	0.1340	0.0605	0.0383	0.0133	0.7100	0.4168	0.2924	1.0000	0.4164	0.5836	0.0545	1.0000	0.6664	1
126	0.0907	0.2198	0.2846	0.4530	0.2410	0.2036	0.1289	0.2171	0.0281	0.2904	0.9550	0.4258	1.0000	0.5441	0.4559	0.0584	1.0000	0.2012	1
139	0.0601	0.1575	0.4801	0.6950	0.2337	0.2532	0.1295	0.1315	0.0131	0.4693	0.3276	0.4495	0.1311	0.6920	0.3080	0.0976	0.5891	0.4641	1
141	0.2047	0.2953	0.5648	0.7768	0.3231	0.3346	0.1950	0.4067	0.0416	0.1581	0.4441	0.2549	0.1397	0.7187	0.2813	0.0858	0.5963	0.1460	1
142	0.2603	0.4378	0.6834	0.6907	0.4428	0.4647	0.2359	0.3743	0.0191	0.1434	0.4185	0.2181	0.1512	0.5414	0.4586	0.2975	0.6000	0.4062	1
13	0.0379	0.1095	0.4797	0.8103	0.0338	0.0313	0.0157	0.5456	0.4381	0.0337	0.4419	0.3196	1.0000	0.1682	0.8318	0.0019	1.0000	0.0125	0
14	0.0189	0.0895	0.6244	0.9151	0.0355	0.0391	0.0165	0.3207	0.2707	0.1401	0.2409	0.7385	1.0000	0.1207	0.8792	0.0030	1.0000	0.0296	0
19	0.0171	0.0587	0.3462	0.7598	0.0249	0.0281	0.0116	0.3681	0.3586	0.1094	0.2075	0.3763	0.7410	0.7638	0.2363	0.0048	1.0000	0.0492	0
25	0.0131	0.0678	0.8007	0.8726	0.0113	0.0098	0.0052	0.4297	0.5245	0.0776	0.3337	0.8680	1.0000	0.5510	0.4490	0.0044	1.0000	0.0507	0
26	0.0230	0.0670	0.4378	0.5324	0.0700	0.0608	0.0325	0.2151	0.1246	0.2397	0.6650	0.1704	0.1512	0.7997	0.2003	0.0002	0.6000	0.0015	0
29	0.1418	0.2892	0.9186	0.8883	0.2432	0.2146	0.1241	0.3665	0.0556	0.1341	0.8397	0.5799	1.0000	0.6086	0.3914	0.0186	1.0000	0.0439	0
31	0.0280	0.0887	0.8774	0.9727	0.0537	0.0568	0.0249	0.3174	0.2099	0.1424	0.3134	0.9411	1.0000	0.1519	0.8481	0.0013	1.0000	0.0106	0
38	0.1596	0.2709	0.9228	0.9344	0.2369	0.2440	0.1101	0.4316	0.0701	0.0767	0.4466	0.4178	0.8054	0.3078	0.6922	0.0096	0.6000	0.0205	0
39	0.0081	0.0775	0.1059	0.4283	0.0860	0.1076	0.0400	0.0742	0.0452	0.5440	0.0453	0.5673	0.8773	0.5712	0.4288	0.0017	1.0000	0.0221	0
86	0.0379	0.1095	0.4797	0.8103	0.0338	0.0313	0.0157	0.5456	0.4381	0.0337	0.4419	0.3196	1.0000	0.3938	0.6062	0.0032	1.0000	0.0213	0
87	0.0189	0.0895	0.6244	0.9151	0.0355	0.0391	0.0165	0.3207	0.2707	0.1401	0.2409	0.7385	1.0000	0.3667	0.6334	0.0057	1.0000	0.0551	0
92	0.0171	0.0587	0.3462	0.7598	0.0249	0.0281	0.0116	0.3681	0.3586	0.1094	0.2075	0.3763	0.7410	0.6315	0.3685	0.0103	0.9346	0.1054	0
98	0.0131	0.0678	0.8007	0.8726	0.0113	0.0098	0.0052	0.4297	0.5245	0.0776	0.3337	0.8680	1.0000	0.7384	0.2616	0.0053	1.0000	0.0605	0
99	0.0230	0.0670	0.4378	0.5324	0.0700	0.0608	0.0325	0.2151	0.1246	0.2397	0.6650	0.1704	0.1512	0.7553	0.2447	0.0233	0.6000	0.2057	0
102	0.1418	0.2892	0.9186	0.8883	0.2432	0.2146	0.1241	0.3665	0.0556	0.1341	0.8397	0.5799	1.0000	0.6245	0.3755	0.0050	1.0000	0.0118	0
104	0.0280	0.0887	0.8774	0.9727	0.0537	0.0568	0.0249	0.3174	0.2099	0.1424	0.3134	0.9411	1.0000	0.3255	0.6745	0.0020	1.0000	0.0159	0
111	0.1596	0.2709	0.9228	0.9344	0.2369	0.2440	0.1101	0.4316	0.0701	0.0767	0.4466	0.4178	0.8054	0.4217	0.5783	0.0354	0.8294	0.0754	0
112	0.0081	0.0775	0.1059	0.4283	0.0860	0.1076	0.0400	0.0742	0.0452	0.5440	0.0453	0.5673	0.8773	0.5913	0.4087	0.0218	0.8458	0.2916	0

The decision column record 1 was happened debris flow events and the decision column record 0 was not happened debris flow events in study site of 2000 and 2001

Table 4 Total variance outcomes

PCA	Eigenvalue	% of variance	Cumulative %
1	*7.037*	*39.095*	*39.095*
2	*4.944*	*27.465*	*66.560*
3	*2.023*	*11.238*	*77.799*
4	*1.272*	*7.067*	*84.866*
5	0.857	4.763	89.629
6	0.710	3.947	93.576
7	0.511	2.838	96.413
8	0.281	1.558	97.972
9	0.174	0.966	98.938
10	0.085	0.474	99.412

Note: In this study, PCA1 to PCA4 component was selected to presented the major factors of the study site and the 1st to 4th componets can govern 84.87% of the variation of the outcomes

problem. Applying this concept, Table 5 shows the results of the influencing factors and their contribution to the occurrence of debris flow. Meanwhile, the influencing factors of this case reduced the dimension of environmental factors from 18 to 14. Finally, the (1) stream sinuosity, (2) average slope of stream, (3) bare-soil land evaluation rate, and (4) bare-soil land geology index are removed from the dataset by means of data reduction process.

Results of PCA + LDA (feature extraction)

Four prominent factors (PCA1–PCA4) and LDA are used to generate the discrimination function, i.e., it can provide information on 14 major factors influencing debris flow occurrence (refer to Fig. 4a). It has to be pointed out while using the discrimination, the cover and management factor is not utilized in the analysis process. In fact, the cover and management factor is derived from the NDVI, thus we decided to use the factor NDVI as a substitute for the cover and management factor (see Eq. 1). Equation 8 shows the results of PCA + LDA:

$$\begin{aligned} Z = {} & 630.2x_1 - 453.2x_2 + 480.7x_3 - 337.6x_4 + 34.1x_5 \\ & + 140.6x_6 + 106.9x_7 - 4.4x_8 - 28.6x_9 + 104.8x_{10} \\ & + 9.1x_{11} + 34.6x_{12} - 25.6x_{13} - 64.8, \end{aligned} \tag{8}$$

where Z is the decision function of debris flow occurrence or non-occurrence, x_1 the primary length of the watershed, x_2 the stream length, x_3 the total stream length, x_4 the watershed perimeter, x_5 the bare-soil land area, x_6 the watershed area, x_7 the form factor, x_8 the geology index, x_9 the watershed of average elevation, x_{10} the stream density, x_{11} the watershed width, x_{12} the watershed of average slope, and x_{13} is the NDVI.

Using Eq. 8, the dataset outcomes can be divided into two categories (occurrence and non-occurrence) and the accuracy is 100% for the training dataset when using those 36 training samples. There are 18 occurrence samples and 18 non-occurrence samples for the study site in 2000 and 2001, respectively. The discrimination function is Z generated as:

IF $Z \geq 0$ *Then* Decision as debris flow is classified as an occurrence sample

IF $Z < 0$ *Then* Decision as debris flow is classified as a non-occurrence sample

From the foregoing statements, the discrimination function is also applied to the testing dataset (110 testing samples) for attaining the accuracy (see Table 6). There are 24 occurrences and 86 non-occurrences. In the occurrence sample, there are 13 accurate samples verified by PCA + LDA with an accuracy rate of 54.2%. Also, in the non-occurrence sample, there are 42 accurate samples verified by PCA + LDA with an accuracy rate of 48.8%. LDA is one of the well-known linear projection techniques for feature extraction in classification problems. The basic concept is to use a process of generalized eigenvalue decomposition. The major drawback of applying LDA is often degraded the "Small Sample Size" (SSS) problem (Lu et al. 2003). Generally, one popular solution to the SSS problem is to combine with a PCA method. However, this

Table 5 Component matrix

Attributes	PCA				Attributes	PCA			
	1	2	3	4		1	2	3	4
L_T	**0.961**	0.181	0.113	0.071	*DE*	−0.574	0.233	0.015	0.443
L_0	**0.95**	0.238	0.072	0.155	*WD*	−0.081	**0.916**	0.009	−0.015
L	**0.939**	0.24	0	0.147	*SD*	0.162	**−0.916**	0.098	0.229
D	**0.834**	0.103	0.2	−0.194	*WW*	−0.165	**0.914**	−0.11	−0.21
P	**0.802**	0.553	0.063	0.154	*WS*	−0.286	**0.86**	0.147	−0.208
A	**0.761**	0.616	−0.024	0.06	*DR*	0.51	−0.532	0.387	−0.139
F	**−0.735**	0.407	−0.12	−0.372	*N*	0.295	−0.169	**−0.917**	−0.068
E	**−0.704**	0.138	−0.01	0.579	*C*	−0.295	0.169	**0.917**	0.068
SS	−0.632	0.332	0.066	0.209	*T*	0.264	0.4	−0.256	0.523

Note: In this table, the major influencing factors (eigenvector is greater than 0.7) are the main contribution of the occurrence of debris flow are shown in bold

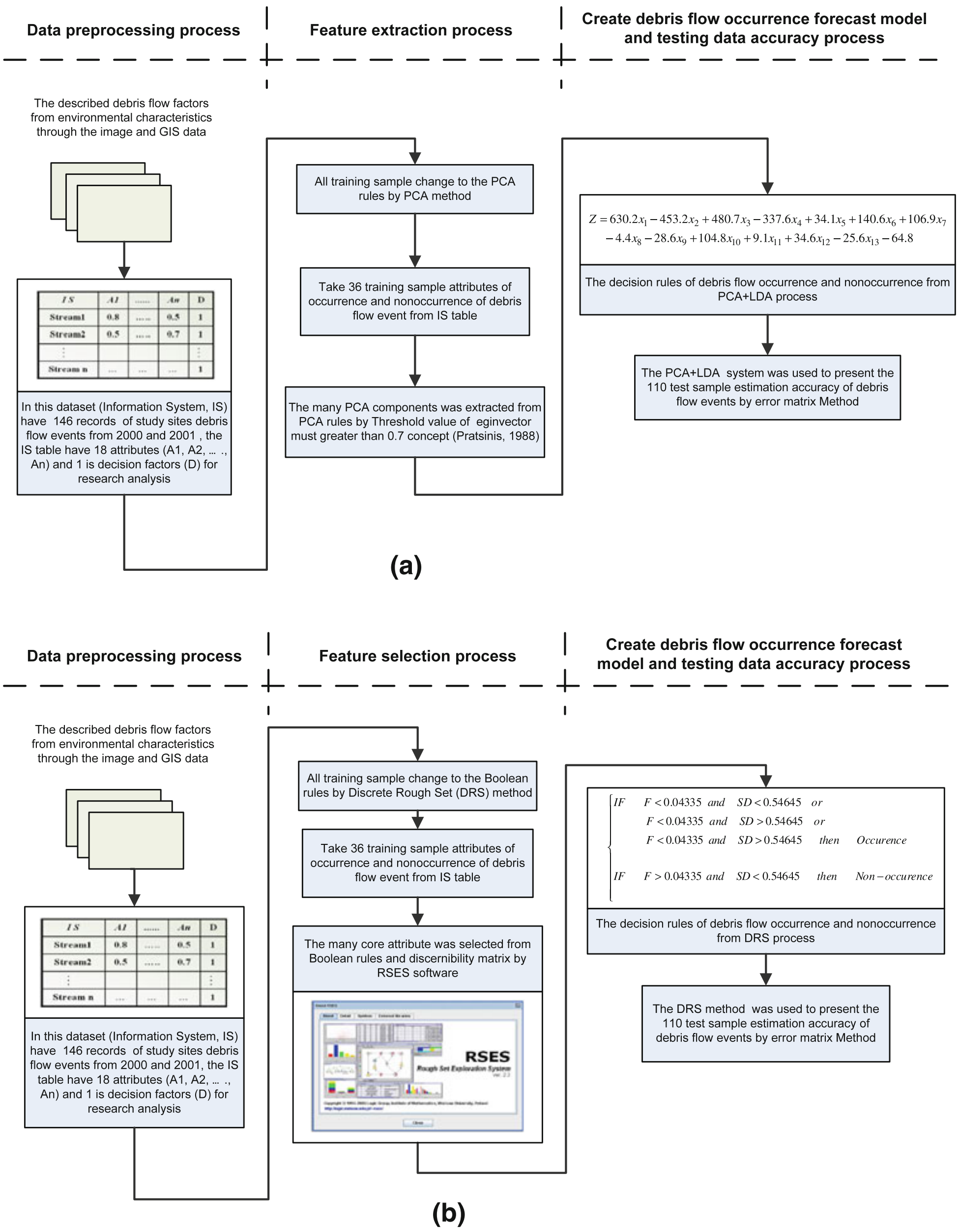

Data preprocessing process
Feature extraction process
Create debris flow occurrence forecast model and testing data accuracy process
The described debris flow factors from environmental characteristics through the image and GIS data
IS
A1
……
An
D
Stream1
0.8
0.5
1
Stream2
0.5
0.7
1
Stream n
In this dataset (Information System, IS) have 146 records of study sites debris flow events from 2000 and 2001 , the IS table have 18 attributes (A1, A2, … ., An) and 1 is decision factors (D) for research analysis
All training sample change to the PCA rules by PCA method
Take 36 training sample attributes of occurrence and nonoccurrence of debris flow event from IS table
The many PCA components was extracted from PCA rules by Threshold value of eginvector must greater than 0.7 concept (Pratsinis, 1988)
$Z = 630.2x_1 - 453.2x_2 + 480.7x_3 - 337.6x_4 + 34.1x_5 + 140.6x_6 + 106.9x_7 - 4.4x_8 - 28.6x_9 + 104.8x_{10} + 9.1x_{11} + 34.6x_{12} - 25.6x_{13} - 64.8$
The decision rules of debris flow occurrence and nonoccurrence from PCA+LDA process
The PCA+LDA system was used to present the 110 test sample estimation accuracy of debris flow events by error matrix Method
(a)
Data preprocessing process
Feature selection process
Create debris flow occurrence forecast model and testing data accuracy process
The described debris flow factors from environmental characteristics through the image and GIS data
In this dataset (Information System, IS) have 146 records of study sites debris flow events from 2000 and 2001, the IS table have 18 attributes (A1, A2, … ., An) and 1 is decision factors (D) for research analysis
All training sample change to the Boolean rules by Discrete Rough Set (DRS) method
Take 36 training sample attributes of occurrence and nonoccurrence of debris flow event from IS table
The many core attribute was selected from Boolean rules and discernibility matrix by RSES software
RSES
Rough Set Exploration System
$IF \quad F < 0.04335 \ and \ SD < 0.54645 \ or$
$F < 0.04335 \ and \ SD > 0.54645 \ or$
$F < 0.04335 \ and \ SD > 0.54645 \ then \ Occurence$
$IF \quad F > 0.04335 \ and \ SD < 0.54645 \ then \ Non-occurence$
The decision rules of debris flow occurrence and nonoccurrence from DRS process
The DRS method was used to present the 110 test sample estimation accuracy of debris flow events by error matrix Method
(b)

Fig. 4 Dimensional reduction approach to described the debris flow factors. **a** PCA + LDA, **b** DRS

procedure cannot effectively eliminate the uncertain information in the debris flow data set (Information Table). On the other hand, rough set analysis can provide an effective feature selection procedure to keep useful information in the data analysis process. Thus, as a part of this study, the DRS method is used to extract a better outcome and then the performance is compared.

Results of DRS (feature selection)

This study also used DRS as a parallel study for comparison. The first step is to create the Information Table. Then, through the Boolean operation, the discernibility matrix is generated. The second step is to calculate the core factors of the most influenced factors to the decisions. The third step is to calculate the separate points with respect to the core factors (refer to Fig. 4b). The results of DRS can be stated as:

1. The core factors are form factor (F) and stream density (SD).
2. The cutting points for form factor and stream density are 0.04335 (the real value from the dataset is 0.3244) and 0.54645 (the real value from the dataset is 0.0018), respectively. In general, the alarm values of form factor usually display differently for various study areas over the world, i.e., the debris flow occurrences are governed by the environmental conditions such as

Table 6 The 110 testing samples classification result by PCA + LDA method

	Reference data (debris flow record)	
	Occurrence	Non-occurrence
Classification result		
Occurrence	13	44
Non-occurrence	11	42
Total	24	86
Classification accuracy	54.2%	48.8%

Table 7 Debris flow decision rule recreated by DRS method

Class	No.	*F*	SD	Class	No.	*F*	SD
Occurrence	41	1	2	Non-occurrence	13	2	1
	43	1	1		14	2	1
	45	1	2		19	2	1
	46	2	2		25	2	1
	47	1	2		26	2	1
	53	1	1		29	2	1
	66	1	1		31	2	1
	68	1	1		38	2	1
	69	1	1		39	2	1
	114	1	2		86	2	1
	116	1	1		87	2	1
	118	1	2		92	2	1
	119	2	2		98	2	1
	120	1	2		99	2	1
	126	1	1		102	2	1
	139	1	1		104	2	1
	141	1	1		111	2	1
	142	1	1		112	2	1

Table 8 The 110 testing Samples classification result by DRS method

	Reference data (debris flow record)	
	Occurrence	Non-occurrence
Classification result		
Occurrence	17	35
Non-occurrence	7	51
Total	24	86
Classification accuracy	70.8%	59.3%

geological factors, hydraulic situations and vegetation terms. These variables will totally affect the form factor of a watershed in this case. Specifically, the form factors are subjectively or statistically observed in the range of 0.1–0.6 (Jan and Chen 2005) and 0.2–0.5 (Chen et al. 2004). This threshold is more rational than some previous studies (Wu 1999; Chen et al. 2004; Jan and Chen 2005; Chen and Jan 2008) in the central part of Taiwan. However, these ranges require a systematic manner to analyze them. Fortunately, DRS renders a great help in solving them. Stream density is the total length of all the streams or rivers in a watershed divided by the total area of this region. In addition, stream density can effect the erosion of soil during a rainstorm. From another viewpoint, physically, high stream density will correlate to poor permeable soil properties because the water runoff is quite large in this area.

3. The Information Table is transferred to the Boolean Table (see Table 3; the data must be preprocessed by the normalization process then plugged into the RSES program). If the attributes values are greater than the cutting point, the Boolean value will be assigned as 2; otherwise, the Boolean value will be assigned as 1. For instance, if the form factor is greater than 0.04335, the Boolean value will be assigned as 2 or the Boolean value will be assigned as 1 (refer to columns *F* and SD in Table 7). The rules will be created as:

$$\left\{\begin{array}{lllll} IF & F < 0.04335 & \text{and} & SD < 0.54645 & \text{or} \\ & F < 0.04335 & \text{and} & SD > 0.54645 & \text{or} \\ & F < 0.04335 & \text{and} & SD > 0.54645 & \text{then occurrence} \\ & & & & \\ IF & F > 0.04335 & \text{and} & SD < 0.54645 & \text{then non-occurrence} \end{array}\right. \tag{9}$$

Also refer to the new decision information system in Table 7. From Eq. 9, if the form factor is greater than 0.04335 (the real value from the dataset is 0.3244), then it may induce debris flow. In other words, if the shape of the watershed area appears circular, it may have a higher capability to retain water. In addition, if the stream density is lower than 0.54645 (the real value from the dataset is 0.0018), it may show the rainfall has a low probability of infiltrating the zone between the soil-layer and laccolith. If the stream density is lower than the threshold value, it seems infiltration will be higher than in the usual case. Physically, higher infiltration will induce debris occurrence.

Table 8 shows the classification results of debris flow events using the DRS method. There are 24 occurrences and 86 non-occurrences. In the occurrence sample, there are 17 accurate samples verified by the DRS method with an accuracy 70.8%. Also, in the non-occurrence sample, there are 51 accurate samples verified by DRS with an accuracy rate of 59.3%. The outcomes of verification accuracy are much higher than PCA + LDA. The reason for better performance is the core factor(s) and separate point(s) are successfully attained and then the redundant spatial data are eliminated. Through this process, the evaluation of the spatial data requested a mining technique to diminish the uncertainties and useless attributes.

How to apply DRS to creat hazard levels in debris flow

In the past, Lin et al. (2002a, b, 2003); Auer and Shakoor (1993) and Lin et al. (2006a, b) used different observations and flowcharts to distinguish various levels of hazards of debris flow. Despite the results from their approaches being attained by statistical analysis, this study proposes a new concept for classifying three hazard levels and determining the DRS results. There are four types in three levels (see Table 9; Fig. 5) and the detailed outcomes can be categorized as:

Type A. It is classified as level 1 (red line). The observed data points were from two given typhoon events. The in situ conditions (geomorphology and land-cover) are relatively fragile and sensitive to the debris occurrence. It is requested to install monitoring devices since there are dangerous areas.

Table 9 Classification on decision matrix by DRS method

	Reference data (debris flow record)	
	Occurrence	Non-occurrence
Classification result		
Occurrence	Case A: 17	Case B: 35
Non-occurrence	Case C: 7	Case D: 51

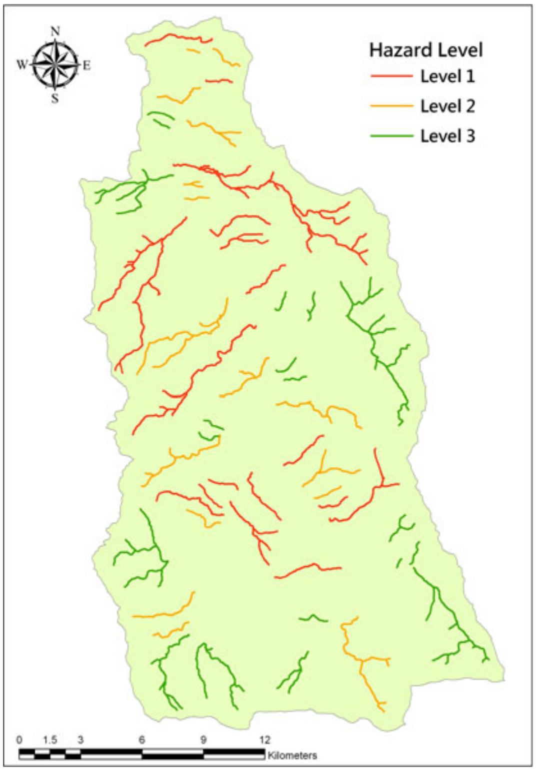


Fig. 5 The reclassification 3 levels of debris flow distribution by DRS method from WCB data and 2000, 2001 records

Type B. It is classified as level 2 (yellow line). Using the knowledge rules from the DRS method, the conditions (geomorphology and land-cover) are identical to Type A but the outcomes are non-occurrence. This is the same as the term "an error of commission". Thus, in the next typhoon event, this area could be dangerous for human beings.

Type C. It is classified as level 2 (yellow line). Using the knowledge rules from the DRS method, the conditions (geomorphology and land-cover) are in terms of non-occurrence, but it had a debris flow occurrence in Toraji typhoon. This is same as the term "an error of omission". Thus, in the next typhoon event, this area might also be dangerous.

To sum up, in the level 2 of this study, it is found that nine streams are the potentially occurred debris flow and seven of them are not. It can be concluded seven of these high potential streams should be monitored in the next storm because there are many uncertainties in this analysis. Applying this method, the original numbers of the potential stream should be modified.

Type D. It is classified as level 3 (green line). The samples extracted from the case excluded Types A–C. They are relatively safe regions. In Fig. 5, the hazard levels of the entire river system can be plotted and the vulnerable watersheds are rationally found.

Summary and conclusions

Although various useful methods have been applied to debris flow events over the whole world, it is important to choose an effective and quick approach in advance to understand the debris flow problem. In particular, one must comprehend the landslide mechanisms spatially, especially in regions frequently affected by earthquakes and rainfalls. From the literature review, the factors influencing debris flow are quite difficult to understand and analyze their occurrence. In this study, the statistical outcomes from PCA + LDA are insufficient, owing to the thresholds of the variables not being evaluated. Thus, an advanced data mining (DRS) approach is used to attain the thresholds. This approach not only showed satisfactory results for the thresholds of influenced variables of the debris flow, but the occurrence rules were also successfully generated. However, in this study, the authors encountered two major problems in their debris flow investigations. First, most of the debris flow occurred in inaccessible places, thus making the measurement of site data difficult or impossible. The geology, geomorphology, water system and vegetation conditions were attained from GIS and remote sensing techniques. Second, conventional statistical methods such as PCA + LDA are very difficult to use, while some spatial acquisition data is surplus. An effective classifier is required to diminish the amount of useless in situ information. DRS can offer a positive knowledge description of the debris flow problems. In other words, following this reduction process, the new knowledge eliminates a lot of noise and chaotic information substitution, and this process improves the classification result. In this study, form factor and stream density are the dominant factors that affect the occurrence of debris flow. The threshold values are 0.3244 and 0.0018, respectively. The DRS can sieve out useless information of measures. Thus, the performance on classification accuracy of DRS is higher than PCA + LDA by about 15%. Also, four different types are classified to illustrate the level of hazards in the debris flow. Clearly, applying this concept, the susceptibility (potential) maps are generated to visualize the overall distribution of debris flow area. This could be of help to the decision-makers to evacuate the affected population away from the disaster zone. Thus, improvements are made and levels of hazards are rationally clarified for various in situ conditions.

Acknowledgments The authors would like to express their gratitude to the GIS Research Center, Feng Chia University, for providing all the data relevant to the debris flow on Chen-Yu-Lan River. The authors also would like to thank National Science Council also supported this work (97-2625-M-275-001).

References

Aerial Survey Office (2008) Aerial photo resource (accessed 15 May 2008). http://www.afasi.gov.tw/

Auer K, Shakoor A (1993) A statistical approach to evaluate debris avalanche activity in central Viginia. Eng Geol 33:305–321

Baeza C, Corominas J (2001) Assessment of shallow landslide susceptibility by means of multivariate statistical techniques. Earth Surf Proc Landf 26:1251–1263

Bannari A, Morin D, Bonn F, Huete AR (1995) A review of vegetation indices. Rem Sens Rev 13:95–120

Carrara A, Guzzetti F, Cardinali M, Reichenbach P (1999) Use of GIS technology in the prediction and monitoring of landslide hazard. Nat Hazards 20:117–135

Carrara A, Crosta G, Frattini P (2003) Geomorphological and historical data in assessing landslide hazard. Earth Surf Proc Landf 28:1125–1142

Center for Space and Remote Sensing Research (2008) SPOT image resource (accessed 10 May 2008). http://www.csrsr.ncu.edu.tw/08CSRWeb/EngVer/index.php

Central Geological Survey (2005) Study on geology factors of debris flow in Taiwan (accessed 16 June 2008). http://www.moeacgs.gov.tw/plan/view.jsp?plan=41

Chang TC (1998) Field investigation and analysis of potential debris flow in Northern Taiwan. J Chin Agric Eng 44:51–63 (in Chinese)

Chang TC, Hsieh CL (1997) Field investigation and analysis of debris flow in Central Taiwan. J Chin Agric Eng 43:31–46 (in Chinese)

Chang FJ, Lee SP (1997) A study of the intelligent control theory for the debris flow warning system. Proceedings of the 1st debris flows conference, pp 109–123

Chen JC, Jan CD (2008) Probabilistic analysis of landslide potential of an inclined uniform soil layer of infinite length: application. Environ Geol 54:1175–1183

Chen JC, Shieh CL, Lin CW (2004) Topographic properties of debris flow in Central Taiwan. J Chin Soil Water Conserv 35:25–34 (in Chinese)

Chouchoulas A, Shen Q (2001) Rough set-aided keyword reduction for text categorization. Appl Artif Intell 15:843–873

Cruden DM, Varnes DJ (1996) Landslide types and processes. In: Turner AK, Schuster RL (eds) Landslides: investigation and mitigation. Special-Report 247, Transportation Research Board, National Research Council. National Academy Press, Washington, DC, pp 36–75

Dai FC, Lee CF (2003) A spatiotemporal probabilistic modeling of storm-induced shallow landsliding using aerial photographs and logistic regression. Earth Surf Proc Landf 28:527–545

Dai FC, Lee CF, Li J, Xu ZW (2001) Assessment of landslide susceptibility on the natural terrain of Lantau Island, Hong Kong. Environ Geol 40:381–391

Devijver P, Kittler J (1982) Pattern recognition: a statistical approach. Prentice Hall, Englewood Cliffs

Donati L, Turrini MC (2002) An objective method to rank the importance of the factors predisposing to landslides with the GIS methodology: application to an area of the Apennines (Valnerina; Perugia, Italy). Eng Geol 63:277–289

Ermini L, Catani F, Casagli N (2005) Artificial neural networks applied to landslide susceptibility assessment. Geomorphology 66:327–343

Fernàndez-Steeger TM, Rohn J, Czurda K (2002) Identification of landslide areas with neural nets for hazard analysis. Proceedings of IECL, Balkema, Netherlands, pp 163–168

Floris M, Mari M, Romeo RW, Gori U (2004) Modeling of landslide-triggering factors: a case study in the northern Apennines, Italy. Lect Notes Earth Sci 04:745–753

Friedman JH, Tukey JW (1974) A projection pursuit algorithm for exploratory data analysis. IEEE Trans Comput C 23:881–890

Fukunaga K (1990) Introduction to statistical pattern recognition. Academic Press, San Diego

Hsieh CL, Lu YC, You BS, Chen LR (1995) Methodology for critical precipitation line of debris flow occurrence. Chin Water Soil Reserve Rep 26:167–172 (in Chinese)

Ivakhnenko AG (1970) Heuristic self-organization in problem of engineering cybernetics. Automatica 6:207–219

Jan CD, Chen CL (2005) Debris flow caused by typhoon herb in Taiwan, Chap 21. In: Book of debris-flow hazards and related phenomena. Springer, Praxis Publishing Ltd, Berlin, pp 539–563

Johnson AM, Rodine JR (1984) Debris flow in slope instability. Wiley, New York, pp 257–361

Lee SP, Chang FJ (1995) Study of fuzzy control theorem for debris flow warning system. J Chin Soil Water Conserv 26:145–154 (in Chinese)

Lee S, Choi J (2004) Landslide susceptibility mapping using GIS and the weight-of-evidence model. Int J Geogr Inf Sci 18:789–814

Lee S, Sambath T (2006) Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. Environ Geol 50:847–855

Lei TC, Wan S, Chou TY (2008) The comparison of PCA and discrete rough set for feature extraction of remote sensing image classification - a case study on rice classification, Taiwan. Comp Geosci 12:1–14

Lin PS, Feng TY, Lee CM (1993) A study on the initiation characteristics of debris flow in gravelly deposits. J Chin Soil Water Conserv 24:55–64 (in Chinese)

Lin JY, Yang MD, Lin PS (1998) An evaluation on the applications of remote sensing and GIS in the estimation system on potential debris flow. 1998 Annual Conference of Chinese Geogr Inf Society (CD edition)

Lin CY, Lin WT, Chou WC (2002a) Soil erosion prediction and sediment yield estimation: the Taiwan experience. Soil Tillage Res 68:143–152

Lin PS, Lin JY, Hung JC, Yang MD (2002b) Assessing debris-flow hazard in a watershed in Taiwan. Eng Geol 66:295–313

Lin CW, Shieh CL, Yuan BD, Shieh YC, Huang ML, Lee SY (2003) Impact of Chi-Chi earthquake on the occurrence of landslides and debris flows: example from the Chen-Yu-Lan River watershed, Nantou, Taiwan. Eng Geol 71:49–61

Lin PS, Lin JY, Lin SY, Lai JR (2006a) Hazard assessment of debris flows by statistical analysis and GIS in Central Taiwan. Int J Appl Sci Eng 4:165–187

Lin WT, Lin CY, Chou WC (2006b) Assessment of vegetation recovery and soil erosion at landslides caused by a catastrophic earthquake: a case study in Central Taiwan. Ecol Eng 28:79–89

Lin WT, Chou WC, Lin CY, Huang PH, Tsai JS (2007) Win-Basin: using improved algorithms and GIS technique for automated watershed modeling analysis from digital elevation models. Int J Geogr Inf Sci 22:47–69

Lu J, Plataniotis KN, Venetsanopoulos AN (2003) Regularized discriminant analysis for the small sample size problem in face recognition. Pattern Recogn Lett 24:3079–3087

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis probability and mathematical statistics series. Academic Press, New York

Mayoraz F, Cornu T, Vuillet L (1996) Using neural networks to predict slope movements. Proceedings of VII international symposium on landslides, Trondheim, Balkema, Rotterdam, pp 295–300

Melelli L, Taramelli A (2004) An example of debris-flows hazard modeling using GIS. Nat Hazards Earth Syst Sci 4:347–358

Mundt JT, Glenn NF, Weber KT, Prather TS, Lass LW, Pettingill J (2005) Discrimination of hoary cress and determination of its detection limits via hyperspectral image processing and accuracy assessment techniques. Rem Sens Environ 96:509–517

Nguyen SH, Nguyen HS (1998a) Pattern extraction from data. Fundam Inform 34:129–144

Nguyen SH, Nguyen HS (1998b) Pattern extraction from data. In: Proceedings of conference on inf. proc. management of uncertainty in knowledge-based systems IPMU'98, pp 1346–1353

Nguyen HS, Skowron A (1995) Quantization of real values attributes, rough set and Boolean reasoning approaches. In: Proceedings of 2nd Conference on Inf. Sci., Wrightsville Beach, NC, pp 34–37

Özhan S, Balcı NA, Özyuvaci N, Hızal A, Gökbulak F, Serengil Y (2005) Cover and management factors for the Universal Soil-Loss Equation for forest ecosystems in the Marmara region, Turkey. For Ecol Manag 214:118–123

Pachauri AK, Pant M (1992) Landslide hazard mapping based on geological attributes. Eng Geol 32:81–100

Pareschia MT, Santacroceb R, Sulpiziob R, Zanchetta G (2002) Volcaniclastic debris flows in the Clanio Valley (Campania, Italy): insights for the assessment of hazard potential. Geomorphol 43(3–4)

Pierson TC (1994) Flow characteristics of large eruption-triggered debris flow at snow-clad volcanoes: constraints for debris-flow models. J Volcanol Geotherm Res 66:283–294

Pratsinis SE, Zeldin MD, Ellis EC (1988) Source resolution of the fine carbonaceous aerosol by principal component-stepwise regression analysis. Environ Sci Technol 22:212–216

Sahoolizadeh AH, Heidari BZ, Dehghani CH (2008) A new face recognition method using PCA, LDA and neural network. Int J Comp Sci Eng 2:218–223

Santacana N, De Paz A, Baeza B, Corominas J, Marturi J (2003) A GIS based multivariate statistical analysis for shallow landslide susceptibility mapping in La pobla de Lillet area (Eastern Pyrenees, Spain). Nat Hazards 30:281–295

Shrestha MB (2001) Study on restoration of vegetation for conservation of the dilapidated mountainous regions of Nepal. Dissertation. University of Gifu

Tian Y, Guo P, Lyu MR (2005) Comparative studies on feature extraction methods for multispectral remote sensing image classification. IEEE Int Conf Syst Man Cyber 2:1275–1279

Torgerson WS (1952) Multidimensional scaling: I. Theory and method. Psychometrika 17:401–419

Van Westen CJ, Rengers N, Soeters R (2003) Use of geomorphological information in indirect landslide susceptibility assessment. Nat Hazards 30:399–419

Varnes DJ (1978) Slope movement types and processes. Landslide Anal Control Natl Acad Sci Transp Res Board Spec Rep 176:11–13

Wan S, Lei TC, Huang PC, Chou TY (2008) Knowledge rules of debris flow event: a case study for investigation ChenYu Lan River, Taiwan. Eng Geol 98:102–114

Wan S, Lei TC, Chou TY (2009) A novel data mining technique of analysis and classification for landslide problems. Nat Hazards. doi:10.1007/s11069-009-9366-3

Wang DS (1994) Study of mechanism of debris flow occurrence. Dissertation. University of Taiwan (in Chinese)

Wang HB, Sassa K (2005) Comparative evaluation of landslides susceptibility in Minamata area, Japan. Environ Geol 47:956–966

Water Conservation Bureau web site (2008a) Potential debris-flow streams distribution of Taiwan (accessed 20 July 2008). http://246.swcb.gov.tw/DebrisPage/distribution/taiwan.asp

Water Conservation Bureau web site (2008b) Handbook of soil and water conservation, Taiwan (accessed 12 June 2008). http://www.swcb.gov.tw/class2.asp?ct=dlbook

Wu YR (1999) Debris-Flow potential analysis and its application in Tainan Count. Dissertation. University of Cheng Kung (in Chinese)

Yang J, Yang JY (2003) Why can LDA be performed in PCA transformed space? Pattern Recogn 36:563–566