Available online at www.sciencedirect.com
ScienceDirect
Engineering Geology 94 (2007) 19–26
ENGINEERING GEOLOGY
www.elsevier.com/locate/enggeo

New entropy-based method for variables selection and its application to the debris-flow hazard assessment

Chien-chih Chen [a,*], Chih-Yuan Tseng [b], Jia-Jyun Dong [c]

[a] Institute of Geophysics, National Central University, Jhongli 320, Taiwan
[b] Institute of Physics, National Central University, Jhongli 320, Taiwan
[c] Institute of Applied Geology, National Central University, Jhongli 320, Taiwan

Received 21 December 2006; received in revised form 4 June 2007; accepted 21 June 2007
Available online 7 July 2007

Abstract

We propose a new data analyzing scheme, the method of minimum entropy analysis (MEA). New MEA method provides a quantitative criterion for selecting relevant variables to model the studied physical system. This method can be easily extended to the analyses of various geophysical/geological data, where many relevant or irrelevant available measurements may obscure the understanding of the highly complicated physical system like the occurrence of debris-flows. After demonstrating and testing the MEA method, we apply this method to a dataset of debris-flow occurrences in Taiwan and successfully identify three relevant variables, i.e. the hydrological form factor, landslide area, and number of landslides, to the occurrence of observed debris-flow events by the 1996 Typhoon Herb.

Keywords: Debris flow; Data analysis; Minimum entropy analysis

1. Introduction

Most geophysical/geological problems, e.g. the occurrence of debris-flows or earthquakes, are so complicated that many observed and/or unobserved variables have the obscure contributions to the geophysical/geological systems (Rundle et al., 2000). From the viewpoint of practical experiment setting, scientists often encounter the problem of variables selection when choosing relevant measurement to the studied physical systems. For instance, three categories of variables describing three aspects of topography, geology, and hydrology, are usually used in the geographic information system (GIS) to assess the hazard potential of debris-flows (e.g. Lin et al., 2002; Rupert et al., 2003). Although some consensus could be reached for the issue of debris-flow hazard assessment, the observed variables could be much different among different research groups (Wieczorek and Naeser, 2000). Therefore, when lots of measurement could be probably made and available to use, we are forced to face a fundamental question of *which variables are relevant to describe a highly complex physical system* like the debris-flow system.

To answer the abovementioned question, we introduce a new data analyzing scheme, i.e. the minimum entropy criterion (Tseng, 2006), to the problem of selecting the variables which dominate the debris-flow occurrence. We

* Corresponding author. Tel.: +886 3 422 7151 65653; fax: +886 3 422 2044.
E-mail address: chencc@ncu.edu.tw (C. Chen).

0013-7952/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.enggeo.2007.06.004

first present the principle of minimum entropy analysis (MEA) and verify its result when applied to a geological example extracted from the textbook of Davis (2002). Then, in Section 3, we demonstrate the application of MEA to an observed debris-flow dataset (Lin et al., 2000), consisting of the binary outcome (the response) of debris-flow occurrences and measurement (the covariates) of some topographic, geologic, and hydrologic variables. Conclusions will be given at the end of this paper.

2. Minimum entropy rule for variables selection

2.1. Principle of minimum entropy analysis

Models selection in data processing is usually accomplished by ranking models according to some kind of measure of *preference*. Several methods such as *P*-value, Bayesian approach, and Kullback–Leibler distance, etc., could be used to provide the selection criteria (Weiss, 1995; Raftery et al., 1997; Forbes and Peyrard, 2003; Dupuis and Robert, 2003; Tseng, 2006). Tseng (2006) reviews those methods and suggests an entropy-based criterion as the selecting preference of models.

The principle of maximum entropy proposed by Jaynes (1957a,b, 2003) is recognized as a tool for assigning a probability distribution to a system. Inspired by Shannon's axiomatic approach, Jaynes' tool involves the use of a unique functional form of entropy $S[P] = -\sum_x P(x)\ln P(x)$, where x denotes the states of the model and P is the probability density function. It was further extended to become an inductive inference tool in data processing for updating the probability distribution of a system according to available information (Skilling, 1989; Caticha, 2004). When a reference density function $\mu(x)$ is available, it can be shown that the approach of updating $P(x)$ involves another unique functional form of the *relative* entropy $S[P|\mu] = -\sum_x P(x)\ln(P(x)/\mu(x))$ (Tseng, 2006).

Furthermore, Tseng (2006) has also shown that the relative entropy uniquely determines the preference for models selection. Suppose a family of models are given by the probability distributions $\{P^m(x)\}$, where m labels the model. The preference is given by the scalar relative entropy to a reference distribution $\mu(x)$ of a model $P^m(x)$,

$$S[P^m|\mu] = -\sum_x P^m(x)\ln\frac{P^m(x)}{\mu(x)}. \quad (1)$$

This relative entropy measures the difference between the model $P^m(x)$ and the reference distribution $\mu(x)$ (Tseng, 2006). The larger the relative entropy $S[P^m|\mu]$ is, the closer the gap between $P^m(x)$ and $\mu(x)$ is.

If the *real* distribution function $P_{\text{real}}(x)$ being able to correctly interpret the system is chosen as the reference distribution, a model $P^m(x)$ with the maximum $S[P^m|P_{\text{real}}]$ will be the most preferable. Unfortunately, the real distribution is usually difficult to be practically determined. Tseng (2006) thus proposes ranking all the candidate models by means of their values of the relative entropy to a uniform distribution function P_{uni}, i.e. $S[P^m|P_{\text{uni}}]$. A model $P^m(x)$ with a maximum $S[P^m|P_{\text{uni}}]$ carries little information about the studied system since a uniform distribution does not carry any useful information and a maximum of the relative entropy indicates two distribution functions are identical. On the other hand, when a model $P^m(x)$ is *codified* with more information, $P^m(x)$ more differs from P_{uni} and a smaller relative entropy of $S[P^m|P_{\text{uni}}]$ would be expected. Consequently, sorting $S[P^m|P_{\text{uni}}]$ of candidate models in ascending order would give the preference to those models with larger values of $S[P^m|P_{\text{real}}]$.

In the case of variables selection, let's suppose that a regression model $P(\vec{x})$ associated with N variables $\vec{x} = \{x_1, x_2,\ldots, x_N\}$ is used to reveal the behavior of an unknown system by experimental measurements. For example, the logit model is often used to investigate the binary outcome of some systems (Johnson and Albert, 1999; Dupuis and Robert, 2003; Rupert et al., 2003). Note that the variables x_i are usually estimated by experiment, and they may or may not be crucial characteristics of the system interested. They may be also correlated to each other. Then our question is that, after modeling the system with various combinations of variables, which ones play more important roles allowing the model to best interpret the studied system. Namely, what are the preferences of these variables? This is basically the same question addressed in the course of models selection (Tseng, 2006).

Suppose that a *full* model defined by $P_f(\vec{x}) = P(\vec{x})$ contains all the N variables available from experiment. For a set of N variables there will be $(2^N - 2)$ combinations (subsets) of variables $\vec{x}_{s_i} \in \vec{x}$. Each subset of variables forms a *sub*-model with a distribution $P_s(\vec{x}_{s_i}) = P(\vec{x}_{s_i})$. Replacing $P^m(x)$ and $\mu(x)$ in Eq. (1) by $P_s(\vec{x}_{s_i})$ and P_{uni}, respectively, the preference of the sub-model is thus given by decreasing the relative entropy

$$S[P_s|P_{\text{uni}}] = -\sum_{\vec{x}_{s_i}\in\vec{x}} P_s(\vec{x}_{s_i})\ln\frac{P_s(\vec{x}_{s_i})}{P_{\text{uni}}} = S[P_s] + \ln P_{\text{uni}}, \quad (2)$$

where the sub-model $P_s(\vec{x}_{s_i})$ contains n_i variables and $S[P_s] = -\sum_{\vec{x}_{s_i}\in\vec{x}} P_s(\vec{x}_{s_i})\ln P_s(\vec{x}_{s_i})$. Since $\ln P_{\text{uni}}$ is constant,

Table 1
Chemical analyses of brines (in ppm) recovered from drillstem tests of three carbonate rock units (Ellenburger Dolomite, Grayburg Dolomite = Unit G, Viola Limestone) in Texas and Oklahoma

Unit G	HCO_3	SO_4	Cl	Ca	Mg	Na
N	10.4	30	967.1	95.9	53.7	857.7
N	6.2	29.6	1174.9	111.7	43.9	1054.7
N	2.1	11.4	2387.1	348.3	119.3	1932.4
N	8.5	22.5	2186.1	339.6	73.6	1803.4
N	6.7	32.8	2015.5	287.6	75.1	1691.8
N	3.8	18.9	2175.8	340.4	63.8	1793.9
N	1.5	16.5	2367	412	95.8	1872.5
Y	25.6	0	134.7	12.7	7.1	134.7
Y	12	104.6	3163.8	95.6	90.1	3093.9
Y	9	104	1342.6	104.9	160.2	1190.1
Y	13.7	103.3	2151.6	103.7	70	2054.6
Y	16.6	92.3	905.1	91.5	50.9	871.4
Y	14.1	80.1	554.8	118.9	62.3	472.4
N	1.3	10.4	3399.5	532.3	235.6	2642.5
N	3.6	5.2	974.5	147.5	69	768.1
N	0.8	9.8	1430.2	295.7	118.4	1027.1
N	1.8	25.6	183.2	35.4	13.5	161.5
N	8.8	3.4	289.9	32.8	22.4	225.2
N	6.3	16.7	360.9	41.9	24	318.1

Adapted from Davis (2002).

ranking order given by decreasing $S[P_s|P_{\text{uni}}]$ is identical to ranking by decreasing $S[P_s]$. The selection of variables, then, can be made from the analysis of the preference of sub-models.

Before illustrating the detailed process in the following, a remark has to be made. The proposed MEA method ranks variables according to the preferences of corresponding sub-models $P_s(\vec{x}_{s_i})$. It thus suggests that the MEA can always give correct ranking as long as the variables can be correctly codified into a sub-model $P_s(\vec{x}_{s_i})$ even though some of the variables may be redundant.

2.2. Demonstration and verification of minimum entropy analysis

We test our data processing procedure of the MEA method with an example of samples classification extracted from the book of Davis (2002). Table 1 contains the results of brine analyses for oil-field waters from three groups of carbonate units in Texas and Oklahoma. Brines recovered during drillstem tests of wells may have remanent compositional characteristics that provide clues to the origin or depositional environment of their source rocks. The first column in Table 1 denotes brine samples belonging or not belonging to some specific carbonate unit, i.e. Grayburg Dolomite (briefly in "Unit G" here), while the rest are the percentages of six chemical ions. Davis (2002) applied the discriminant function analysis (DFA) to these six multivariate measurements for finding a projection, i.e. a linear combinations of measurements, allowing distinguishing various categories of samples. The first discriminant function thus calculated is $(-0.3765, -0.0468, 0.0112, -0.0148, -0.0174, -0.0110)\cdot(HCO_3, SO_4, Cl, Ca, Mg, Na)^T$, which can clearly separate samples of Unit G from other units. Note that the weighting factors in the first discriminant function of the variables of HCO_3 and SO_4, i.e. -0.3765 and -0.0468, represent the first two largest factors in magnitude among six, indicating these two variables play the most dominant effect in classification.

By means of our entropy-based procedure, can we identify the relevant variables to the problem of determining the category of samples in Table 1?

Let's consider the response to be the binary outcome belonging ("Yes" or "1") or not belonging ("No" or "0") to Unit G and the covariates these percentages of six chemical ions in Table 1. We can apply the logit model (Johnson and Albert, 1999; Dupuis and Robert, 2003)

$$R(\vec{x}) = \frac{\exp\left(\sum_{i=1}^{N} \beta_i x_i\right)}{1 + \exp\left(\sum_{i=1}^{N} \beta_i x_i\right)} \tag{3}$$

to relate the response to the covariates. After normalizing Eq. (3), the probability distribution of the response associated with a given subset of all the six variables is

$$P(\vec{x}) = R(\vec{x})/Z = \frac{1}{Z}\frac{\exp\left(\sum_{i=1}^{N} \beta_i x_i\right)}{1 + \exp\left(\sum_{i=1}^{N} \beta_i x_i\right)}, \tag{4}$$

where $Z\left(=\sum_{\vec{x}} \frac{\exp(\sum_{i=1}^{N} \beta_i x_i)}{1+\exp(\sum_{i=1}^{N} \beta_i x_i)}\right)$ is a normalization constant. Note that the coefficients β_i could be determined through fitting the logit model to experimental measurements by the maximum likelihood estimation (Johnson and Albert, 1999). Thus the entropy of $P(\vec{x}_{s_i})$ related with a subset of variables $\vec{x}_{s_i} \in \vec{x}$ could be calculated and gives the ranking order of a sub-model $P(\vec{x}_{s_i})$ defined by Eq. (4).

In the example of brine data there are 62 sub-models. As shown in Table 2, we found that 16 sub-models out of 62 have the minimum entropy value of ~ 1.7918 while the rest of the sub-models (not shown in Table 2) have the entropy larger than 2. The MEA suggests that these 16 sub-models are the most preferable. Yet, due to intrinsically limited precision of measured data, we can not distinguish the preferences for these 16 sub-models further. For tackling the resolution issue of the entropy resulted from intrinsic data precision, there are many

Table 2
Entropy (S) for sixteen sub-models with different combinations of six variables in Table 1 (A = HCO_3, B = SO_4, C = Cl, D = Ca, E = Mg, and F = Na)

A	B	C	D	E	F	S
1	1	1	0	1	1	1.79183823
1	1	1	1	0	1	1.79183829
1	1	0	1	1	1	1.79183836
1	1	1	1	1	0	1.79183836
1	1	0	1	1	0	1.79184075
1	1	0	0	1	1	1.79184177
1	1	1	0	1	0	1.79184215
1	1	1	0	0	1	1.79184241
1	1	0	0	1	0	1.79184396
1	1	0	1	0	1	1.79184653
1	1	1	1	0	0	1.79184701
1	0	1	1	1	1	1.79184888
1	1	0	1	0	0	1.79184968
1	1	1	0	0	0	1.79185471
1	1	0	0	0	1	1.79185668
1	1	0	0	0	0	1.79185738

"1" or "0" denotes the variable selected or not selected in each sub-model.

possible ways (e.g. Dupuis and Robert, 2003) to determine the most dominant variables in this example. Here we simply count the frequencies of six variables appearing in these 16 sub-models. It turns out that the frequencies for two variables of HCO_3 and SO_4 are 16 and 15, respectively, and 8 for the rest of variables. This result suggests the ability interpreting the measurement in the logit model is strongly dominated by involving simultaneously two variables of HCO_3 and SO_4 in the model. Therefore the MEA result is quite consistent with the DFA result. Comparing both results of the DFA and MEA procedures helps the understanding of our entropy-based technique and enhances confidence in this MEA tool.

3. Application of minimum entropy analysis to the debris-flow hazard assessment

Locating at an active convergent plate boundary, the Island of Taiwan is rugged in relief and severe in erosion by the weather. After the heavy rainfalls brought by typhoons, the occurrence of the debris-flow often results in enormous damage of lives and buildings (Chen et al., 1997; Lin and Jeng, 2000; Lin et al., 2002, 2003; Jan and Chen, 2005). There are absolutely many factors affecting intricately the occurrence of the debris-flow and various measurement in the field has been conducted to assess the occurrence potential of the debris-flow in Taiwan (Chen et al., 2000; Lin et al., 2000; Wieczorek and Naeser, 2000; Chen and Su, 2001; Lin et al., 2002, 2003; Jan and Chen, 2005). Can we figure out the observations relevant to the occurrence of the debris-flow by means of our MEA procedure?

To preliminarily apply the MEA method we have used a relatively small dataset (Table 3) published in Lin et al. (2000), documenting the occurrence of the debris-flows in the Hsinyi area of Nantou County, Central Taiwan, during the 1996 Typhoon Herb. As shown in Fig. 1 is a geological map (CGS, 2000) and drainage system in the Hsinyi area. Eocene–Oligocene metamorphic rocks are thrusting over the Miocene rocks along the Chenyeolanchi Fault. The Chenyeolan River is the major river flowing through the Hsinyi area which is closely related to the Chenyeolanchi Fault. Gullies selected by Lin et al. (2000) in the Shinshan (No. 1–12) and Shenmu (No. A–G) areas are also shown in Fig. 1. In Shinshan area, the outcropped Oligocene metamorphic strata in the catchments of Gullies 1–6 are the Shuichangliu Formation (OS) mainly composed of thick-bedded argillites and slates with thin-bedded meta-sandstone. The outcropped Miocene sedimentary strata in the catchments of Gullies 7–12 are the Nankang Formation (M2) composed of thick- to thin-bedded fine-

Table 3
Debris-flow occurrences of 22 creeks during the 1996 Typhoon Herb in Hsin-Yi area of the Nantou County, Central Taiwan, together with their corresponding characteristics including gully lengths (Le), areas of drainage basin with slope >15° (Ad), form factor ($Ff = Ad/Le^2$) and numbers (Nl) and areas (Al) of landslides

Gully No.	Debris-flow occurrence	Le [m]	Ad [km^2]	Ff	Nl	Al [km^2]
1	No	1505	0.86	0.3797	0	0
2	Yes	1876	1.27	0.3609	3	10.9
3	Yes	1640	0.35	0.1301	2	4.5
4	Yes	1560	0.57	0.2342	3	3.4
5a	Yes	2158	1.82	0.3908	5	3.7
5b	Yes	1035	1.82	1.6990	1	7.8
6a	Yes	582	3.4	10.0377	5	3.7
6b	Yes	2445	3.4	0.5687	4	4.4
7	Yes	2685	3.5	0.4855	9	8.4
8	No	2350	1.97	0.3567	0	0
9	No	142	1.15	57.0323	4	2.7
10	No	1349	1.55	0.8517	4	2.9
11	No	1337	0.74	0.4140	4	1.2
12	No	911	0.72	0.8676	3	0.8
A	Yes	2048	0.78	0.1860	5	0.086
B	Yes	2960	2.18	0.2488	6	0.226
C	No	2010	1.65	0.4084	6	0.061
D1	Yes	675	0.58	1.2730	1	0.033
D2	Yes	4947	2.24	0.0915	13	0.362
E	No	3185	4.05	0.3992	4	0.045
F	Yes	4209	6.63	0.3742	7	0.084
G	Yes	4444	6.93	0.3509	21	0.371

Adapted from Lin et al. (2000).

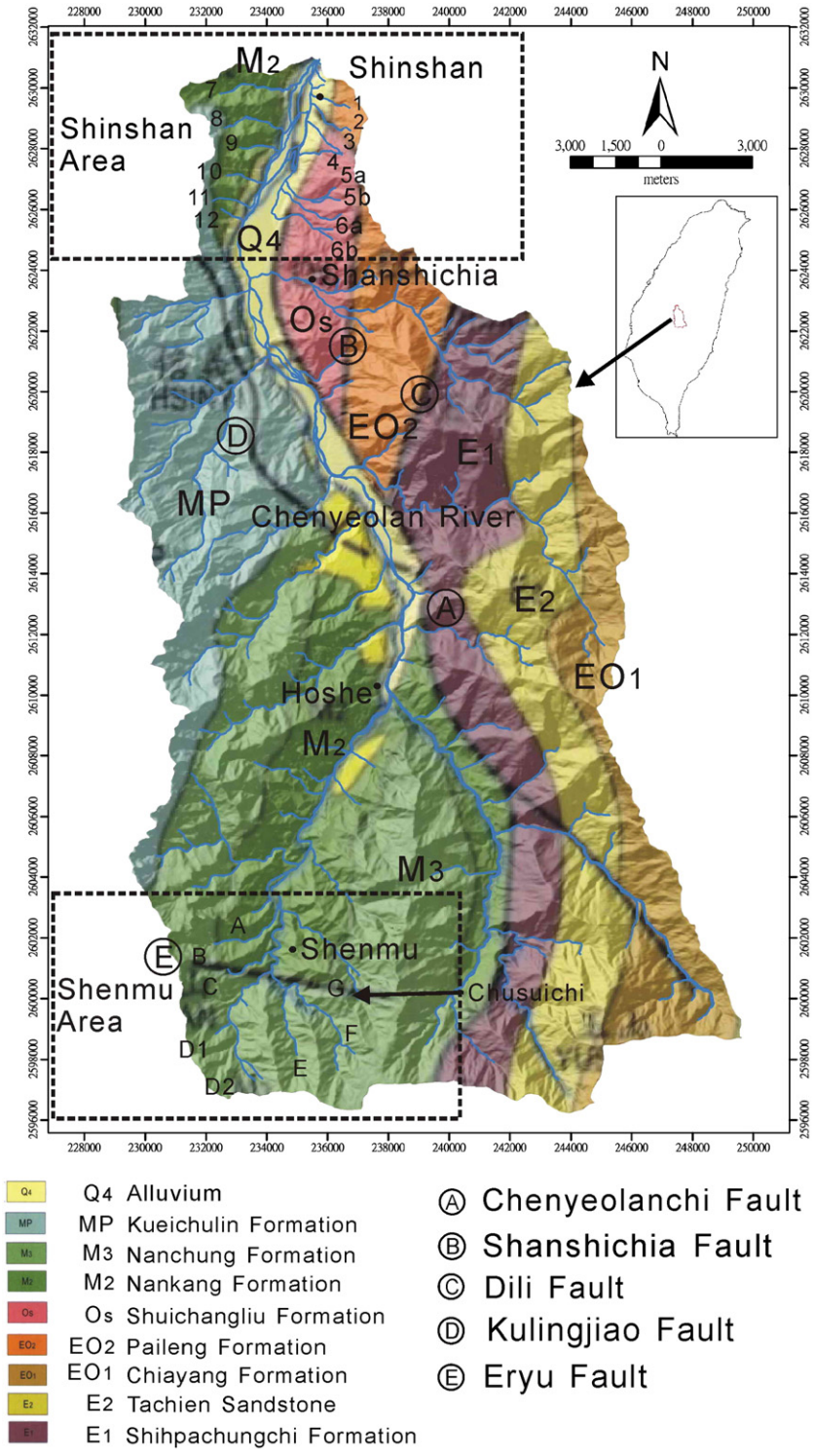

Fig. 1. Geological map (CGS, 2000) and drainage system in the Hsinyi area of Nantou County, Central Taiwan. The Chenyeolan River is the major river flowing through the Hsinyi area which is closely related to the Chenyeolanchi Fault. In Shinshan area, the Oligocene metamorphic strata outcropped in the catchments of Gullies 1–6 and the Miocene sedimentary strata in the catchments of Gullies 7–12. In Shenmu area, the Miocene sedimentary strata outcropped in the catchments of Gullies A–G.

grained calcareous sandstone, dark gray shale and siltstone. In Shenmu area, the outcropped Miocene sedimentary strata in the catchments of Gullies A–G are the Nankang Formation (M2) and the Nanchung Formation (M3). The Nanchung Formation is mainly composed of thick-bedded to massive sandstone interlayered with thick- to thin-bedded shale.

Five variables (Table 3) presented in Lin et al. (2000) for the gullies include (1) the gully lengths (Le), (2) areas of drainage basin with the slope >15° (Ad), (3) form factor ($Ff = Ad/Le^2$), (4) numbers (Nl) and (5) areas (Al) of landslides, which implicitly reflect the topographic, geologic, and hydrologic characteristics of the gullies. Takahashi (1978) suggested that the debris deposited on the gully with a slope angle >15° contributes the material to the occurrence of debris-flow. Consequently, the factor of areas of drainage basins with the slope angle >15° was selected in Lin et al. (2000). The landslides and debris-flow occurrences were identified through the field investigation and interpretation of aerial photos with a scale of about 1:15,000 to 1:22,000 taken before and after the Typhoon Herb. The other variables were derived from the topographic map with a scale of 1:25,000. For the detailed description of those observations, please refer to the paper of Lin et al. (2000).

Rupert et al. (2003) utilized a logistic regression to predict the probability of the debris-flow occurrence, and their results show that the logistic regression is useful to the debris-flow prediction. Therefore, we also utilize the logit model to relate the binary outcome of the debris-flows with those five covariates in Table 3 (the last five columns). Following the MEA procedure demonstrated in Section 2.2, we find 3 out of 30 sub-models, i.e. Sub-models 1, 2 and 3, having the minimum entropy of ~2.93 as shown in Table 4. The problem of the entropy resolution also appears in this case. Two strategies are useful in determining the relevant variables through the entropy calculations of these sub-models. We have listed in Table 4 all the calculated entropy of 30 sub-models for the debris-flow data. The entropy value ranges between 2.9346 and 3.0726, and the difference in entropy is about 0.138. When the resolution level in entropy is assigned 10%, which could be related to the precision in measurement, the first 3 sub-models can be thus discriminated from the rest of sub-models with the entropy values >2.95. The entropy difference between Sub-model 4 and the first 3 sub-models is larger than 10% of 0.138. On the other hand, according to the debris-flow data shown in Table 3, it seems fairly conservative to consider the measurement precision to be three significant figures. Note that, in Table 3, there are four significant figures for the factor of Le, three for Ad and four for Al. Therefore, the significant figure in entropy is truncated to two decimal places and, again, we can conclude the first 3 sub-models with the minimum entropy of ~2.93 are the most preferable.

Three variables of the form factor, numbers and areas of landslides (Columns C, D and E in Table 4) are involved into all the first 3 sub-models with the minimum entropy of ~2.93, meaning that these three variables are important to the debris-flow hazard assessment in the study area. We have noticed that the same watershed in the Central Taiwan was independently studied in the papers of Lin et al. (2000, 2002), and the observation spans of the debris-flow occurrences are same after the 1996 Typhoon Herb. It is therefore interesting to compare our MEA result with the assessing variables used by Lin et al. (2002). In the paper of Lin et al. (2002), an overall debris-flow hazard

Table 4
Entropy (S) for total thirty sub-models with different combinations of five variables in Table 3 (A = Le, B = Ad, C = Ff, D = Nl, and E = Al)

Sub-model No.	A	B	C	D	E	S
1	0	1	1	1	1	2.9346
2	1	0	1	1	1	2.9351
3	0	0	1	1	1	2.9355
4	1	1	1	0	1	2.9538
5	1	1	0	1	1	2.9542
6	1	0	1	0	1	2.9542
7	1	0	0	1	1	2.9557
8	1	0	0	0	1	2.9649
9	1	1	0	0	1	2.9653
10	0	1	0	1	1	2.9725
11	0	0	0	1	1	2.9738
12	0	1	1	0	1	2.9747
13	0	0	1	0	1	3.0065
14	0	1	0	0	1	3.0117
15	1	1	1	1	0	3.0240
16	1	0	1	1	0	3.0279
17	0	1	1	1	0	3.0289
18	0	0	1	1	0	3.0296
19	1	1	1	0	0	3.0446
20	0	1	1	0	0	3.0447
21	0	0	0	0	1	3.0447
22	1	0	1	0	0	3.0498
23	1	1	0	1	0	3.0550
24	1	0	0	1	0	3.0550
25	0	1	0	1	0	3.0560
26	0	0	0	1	0	3.0570
27	0	0	1	0	0	3.0607
28	1	1	0	0	0	3.0632
29	1	0	0	0	0	3.0638
30	0	1	0	0	0	3.0726

"1" or "0" denotes the variable selected or not selected in each sub-model.

index is derived from the sophisticated GIS analysis of nine factors, i.e. the rock formation, fault length, landslide area, slope angle, slope aspect, stream slope, watershed area, form factor and C factor. For the detailed explanation of those factors, please refer to the paper of Lin et al. (2002). Although the data presented in Lin et al. (2000), as mentioned above, is only a relatively small dataset, two variables of the form factor and landslide area are in agreement selected to be the important factors for assessing the debris-flow occurrence in both our MEA result and the GIS analysis in Lin et al. (2002). This indicates a fairly good performance of our new MEA procedure. One important fact is that, while the reason for the selection of those nine factors in Lin et al. (2002) is quite subjective, as they mentioned in their paper, our MEA procedure straightforwardly provides a quantitative criterion in the variables selection for the debris-flow hazard assessment.

We have further compared the predicting performance of the full model with all five observed factors and two sub-models, i.e. 3 and 28 in Table 4, by the Relative Operating Characteristic (ROC) diagram (Fig. 2). The ROC diagram is a well-established way to examine the performance of model predictor (e.g. Chen et al., 2005, 2006). A ROC graph is a plot with the *false alarm rate* (FAR) on the horizontal axis and the *hit rate* (HR) on the vertical axis. The HR is the fraction of positive occurrences of debris-flows that were correctly predicted to be occurred, while the FAR is the fraction of non-occurrence cases that were incorrectly predicted as the positive occurrence. The point (0, 1), which means the FAR is 0 and the HR is 1, on an ROC graph is the perfect predictor. It predicts all occurrences and non-occurrences of events correctly. In many cases, a model predictor has a probability threshold that can be adjusted to increase HR at the cost of an increased FAR or decrease FAR at the cost of a decrease in HR. For example, for the full model, the pair of (0, ~0.65) could be obtained by using a probability threshold larger than ~0.76. Lowering the threshold to ~0.25, we obtained another pair of (~0.63, ~0.93). Each probability threshold setting provides a (*FAR*, *HR*) pair and a series of such pairs can be used to plot an ROC curve. In the ROC diagram, a larger area under the ROC curve indicates higher skill of the model predictor. Therefore, based on the ROC curves of the full model and sub-models 3 and 28 (Fig. 2), it can be seen that the prediction performance of sub-model 3 is significantly better than sub-model 28 and slightly over the full model. However, note that only three preferred factors selected by the above MEA procedure is used in sub-model 3 while all five measurements are involved in the full model.

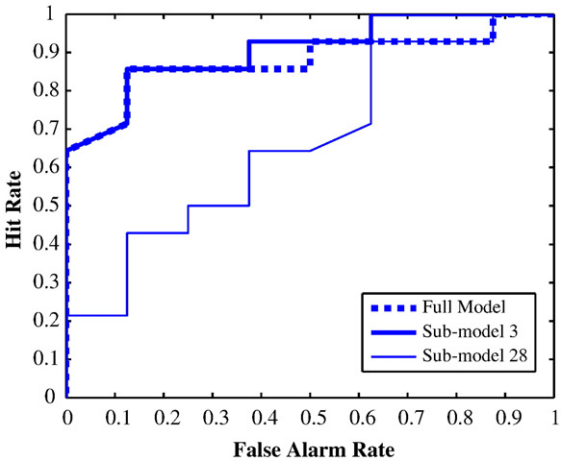

Fig. 2. ROC curves of the full model and sub-models 3 and 28 in Table 4. A larger area under the ROC curve indicates higher skill of the model predictor. For the details please see the text.

Regarding the variable of landslide number, our MEA procedure then raises an open issue about whether it really does matter to the occurrence of the debris-flows. We postpone to future work the examination of this issue.

4. Concluding remark

In the course of data analyses two fundamental questions are commonly queried. What is the *pertinent* model to interpret best the experimental measurement for understanding the investigated physical system? And, what are the important variables that should be involved in the model? One may be able to reveal the nature and characteristics of the investigated system through answering these two questions. For addressing the first question, unfortunately, there are no systematical methods. It is usually resolved through the ways of trials and errors, doing empirical regression, and giving some intuitive assumptions etc. Therefore we focus on addressing the second question in this paper. Our proposed MEA procedure represents a systematical scheme to tackle the second issue. We establish, demonstrate, and test our MEA procedure by studying two geo-scientific examples in this paper. The MEA procedure, then, can satisfactorily provide a quantitative criterion for selecting the relevant variables in both examples.

The MEA procedure seems simple and straightforward. It is thus expected that the MEA procedure could be easily extended and applied to various geophysical/

geological data analyses, where many relevant or irrelevant observations could be possibly made and obscure the understanding to a highly complicated physical system. The occurrence of debris-flows is such an example. Here we would also like to emphasize that the MEA procedure only provides an honest way to extract the most important information from dazzling data available. The entropy resolution of the MEA method is restricted to the precision and the correctness of measurement, which means the data themselves could be incorrect and the observations could be conducted in ill condition.

Acknowledgements

The authors thank two anonymous reviewers and the editor for their helpful suggestions to improve our paper. CCC is grateful for research support from both the National Science Council (ROC) and the Institute of Geophysics (NCU, ROC). Research by CYT was funded by grant from the NSC, ROC. Thanks are also extended to Owen Huang at Indiana University (USA) for correcting the English usage and to Shawn Huang at NCU for producing Fig. 1.

References

Caticha, A., 2004. Relative Entropy and Inductive Inference. In: Erickson, G., Zhai, Y. (Eds.), Bayesian Inference and Maximum Entropy Methods in Science and Engineer, vol. 707. AIP, New York.

CGS (Central Geological Survey), 2000. Geologic Map of Taiwan 1:500,000, Central Geological Survey, MOEA, Taiwan.

Chen, H., Su, D.Y., 2001. Geological factors for hazardous debris flows in Hoser, central Taiwan. Environmental Geology 40, 1114–1124.

Chen, J.D., Wu, H.L., Chen, L.J., 1997. A comprehensive debris flow hazard mitigateion program in Taiwan. In: Chen, C.L. (Ed.), Proceedings of 1st International Conference on Debris-flow Hazards Mitigation: Mechanics, Prediction, and Assessment. Am. Soc. Civil Engineers, New York, pp. 93–102.

Chen, J.D., Su, R.R., Wu, H.L., 2000. Hydrometeorological and site factors contributing to disastrous debris flows in Taiwan. In: Wieczorek, G.F., Naeser, N.D. (Eds.), Debris-flow Hazards Mitigation: Mechanics, Prediction, and Assessment. A.A. Balkema, Rotterdam, Netherlands, pp. 583–592.

Chen, C.C., Rundle, J.B., Holliday, J.R., Nanjo, K.Z., Turcotte, D.L., Li, S.C., Tiampo, K.F., 2005. The 1999 Chi-Chi, Taiwan, earthquake as a typical example of seismic activation and quiescence. Geophys. Res. Lett. 32 (22), L22315. doi:10.1029/2005GL023991.

Chen, C.C., Rundle, J.B., Li, H.C., Holliday, J.R., Nanjo, K.Z., Turcotte, D.L., Tiampo, K.F., 2006. From tornadoes to earthquakes: forecast verification for binary events applied to the 1999 Chi-Chi, Taiwan, earthquake. Terr. Atmos. Ocean. Sci. 17, 503–516.

Davis, J.C., 2002. Statistics and Data Analysis in Geology, 3rd. Ed. Wiley, New York.

Dupuis, J.A., Robert, C.P., 2003. Variable selection in qualitative models via an entropic explanatory power. J. Stat. Plan. Inference 111, 77–94.

Forbes, F., Peyrard, N., 2003. Hidden Markov random field model selection criteria based on mean field-like approximations. IEEE Trans. Pattern Anal. Mach. Intell. 25 (9), 1089–1101.

Jan, C.D., Chen, C.L., 2005. Debris flows caused by Typhoon Herb in Taiwan. In: Jakob, M., Hungr, O. (Eds.), Debris-flow Hazards and Related Phenomena. Springer, New York, pp. 539–563.

Jaynes, E.T., 1957a. Information theory and statistical mechanics. Phys. Rev. 106, 620–630.

Jaynes, E.T., 1957b. Information theory and statistical mechanics II. Phys. Rev. 108, 171–190.

Jaynes, E.T., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK.

Johnson, V.E., Albert, J.H., 1999. Ordinal Data Modeling. Sringer, New York.

Lin, M.L., Jeng, F.S., 2000. Characteristics of hazards induced by extremely heavy rainfall in Central Taiwan: Typhoon Herb. Eng. Geol. 58, 191–207.

Lin, C.W., Wu, M.C., Shieh, C.L., Shieh, Y.C., 2000. Influence of geology on debris-flows: examples from Hsin-Yi, Nantou County, Taiwan. In: Wieczorek, G.F., Naeser, N.D. (Eds.), Debris-flow Hazards Mitigation: Mechanics, Prediction, and Assessment. A.A. Balkema, Rotterdam, Netherlands, pp. 169–176.

Lin, P.S., Lin, J.Y., Hung, J.C., Yang, M.D., 2002. Assessing debris-flow hazard in a watershed in Taiwan. Eng. Geol. 66, 295–313.

Lin, C.W., Shieh, C.L., Yuan, B.D., Shieh, Y.C., Liu, S.H., Lee, S.Y., 2003. Impact of Chi-Chi earthquake on the occurrence of landslides and debris flows: example from the Chenyulan River watershed, Nantou, Taiwan. Eng. Geol. 71, 49–61.

Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for regression models. J. Am. Stat. Assoc. 92, 179–191.

Rundle, J.B., Klein, W., Tiampo, K.F., Gross, S.J., 2000. Linear pattern dynamics in nonlinear threshold systems. Phys. Rev., E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics 61 (3), 2418–2431.

Rupert, M.G., Cannon, S.H., Gartner, J.E., 2003. Using Logistic Regression to Predict the Probability of Debris Flows Occurring in Areas Recently Burned by Wildland Fires. Open-file Report OF 03-500US Geological Survey.

Skilling, J., 1989. Maximum Entropy and Bayesian Methods. Kluwer, Dordrecht.

Takahashi, T., 1978. Mechanical characteristics of debris flow. J. Hydraul. Div. 104 (HY8), 1153–1169.

Tseng, C.Y., 2006. Entropic criterion for model selection. Physica, A 370, 530–538.

Weiss, R.E., 1995. The influence of variable selection: a Bayesian diagnostic perspective. J. Am. Stat. Assoc. 90, 619–625.

Wieczorek, G.F., Naeser, N.D., 2000. Debris-flow Hazards Mitigation: Mechanics, Prediction, and Assessment. A.A. Balkema, Rotterdam, Netherlands.