Support vector machine modeling of landslide susceptibility using a GIS: A case study

X. YAO[1] & F.C. DAI[2]

[1] Institute of Geology and Geophysics, Chinese Academy of Sciences. (e-mail: yaoxinphd@yahoo.com)
[2] Institute of Geology and Geophysics, Chinese Academy of Sciences. (e-mail: daifc@mail.igcas.ac.cn)

Abstract: Support vector machine (SVM) is used as a universal constructive learning procedure based on the statistical learning theory. It can provide non-linear solutions to regression and classification problems by transforming the input variables in a large-dimension space, whose inner product is given by positive definite kernel functions. In this paper, SVM was used to model landslide susceptibility using a Geographic Information System (GIS). The study area was selected from the natural terrain of Hong Kong. Both datasets representing occurrence of landslides and non-occurrence of landslides are needed to develop a SVM model. The Natural Terrain Landslide Inventory (NTLI) interpreted from aerial photographs was used to represent the occurrence of landslides. The dataset representing non-occurrence of landslides was randomly sampled with a buffer zone from the locations of landslides. The factors influencing the occurrence of landslides, such as slope gradient, slope aspect, elevation, slope form, topographic index, vegetation cover, lithology, were automatically extracted from relevant grid themes using a GIS. Both datasets representing occurrence of landslides and non-occurrence of landslides and corresponding influencing factors were then divided into three sub-datasets used for training, validating and testing. A SVM model was trained and validated repeatedly by the training and validating sub-dataset respectively to obtain the optimum kernel function, and the performance of the model was tested by the testing sub-dataset. The SVM model developed was finally used to model landslide susceptibility of the study area, and the modeling result was compared with that obtained using Artificial Neural Network (ANN). The advantages and disadvantages of SVM model were also discussed.

Résumé: La machine de vecteur de soutien (SVM) est utilisée comme procédé d'étude constructif universel basé sur la théorie d'étude statistique. Il peut fournir les solutions non linéaires aux problèmes de régression et de classification en transformant les variables d'entrée dans un espace de grand-dimension, à qui produit intérieur est donné par des fonctions définies positives de grain. En cet article, SVM a été employé pour modeler la susceptibilité d'éboulement enutilisant un système d'information géographique (GIs). Le secteur d'étude a été choisi parmi le terrain normal de Hong Kong. Les ensembles de données représentant l'occurrence des éboulements et l'non-occurrence des éboulements sont nécessaires pour développer un modèle de SVM. Le inventaire normal d'éboulement de terrain (NTLI) a interprété des photographies aériennes a été employé pour représenter l'occurrence des éboulements. L'ensemble de données représentant l'non-occurrence des éboulements a été aléatoirement prélevé avec une zone-tampon des endroits des éboulements. Les facteurs influençant l'occurrence des éboulements, comme le gradient de pente, aspect de pente, altitude, forme de pente, index topographique, couverture de végétation, lithologie, ont été automatiquement extraits à partir d'employer approprié de thèmes de grille des GIs. Des ensembles de données représentant l'occurrence des éboulements et l'non-occurrence des éboulements et correspondant influençant des facteurs ont été alors divisés en trois secondaire-ensembles de données utilisés pour la formation, valider et examiner. Un modèle de SVM a été formé et validé à plusieurs reprises par la formation et de valider secondaire-ensemble de données respectivement pour obtenir la fonction optima de grain, et l'exécution du modèle a été examinée par l'secondaire-ensemble de données d'essai. Les SVM modèlent développé ont été finalement employés pour modeler la susceptibilité d'éboulement du secteur d'étude, et le résultat modelant a été comparé à cela obtenu en utilisant le réseau neurologique artificiel (ANN). Les avantages et les inconvénients du modèle de SVM ont été également discutés.

Keywords: landslides, safety, environmental geology, environmental urban geotechnics, data analysis, numerical models, SVM , Hong Kong

INTRODUCTION

In Hong Kong , landslides on natural terrain are a common occurrence due to natural slopes mantled with a layer of colluvium or residual soil and frequent intense rainstorms. Some of the landslides on natural terrain can develop into large mobile, channelized debris flows. Such debris flows can travel over a great distance and may cause significant damage to developed areas adjacent to steep slopes. In addition, the demand for land in Hong Kong is such that there will a continuing trend to develop areas close to steep, nature hill slopes. Therefore, stability of natural slopes has become a significant concern in land-use planning and hazard mitigation. Decisions need to be made regarding existing and proposed development in areas that are vulnerable to landslides originating in natural slopes.

The need to assess landslide hazard for the purpose of land-use planning and hazard mitigation has led to the development of many statistical and process-based models, with increasing emphasis on the use of GIS. Statistical modelling approaches have generally taken the form of multivariate statistical analyses of physical characteristics associated with past landsliding (Carrara et al. 1991; Mark & Ellen 1995; Chung & Fabbri 1999; Guzzetti et al. 1999)

or weighted hazard ratings based on environmental attributes related to landsliding (Gupta & Joshi 1989), whereas the process based models have generally used infinite slope stability theory coupled with geomorphological, hydrological, geological and vegetation data to estimate slope, strength parameters and pore pressure distribution and then to model the factor of safety of the slopes (Montgomery and Dietrich 1994; Wu and sidle 1995). In general, GIS-based statistical models have been able to indicate general regions of landslide susceptibility, whereas GIS-based process models specifically include the physical processes involved in landsliding and therefore can often better describe the physical processes of mass movement. Theoretically, process models commonly use site-specific data and should produce a more realistic result with slope instability modelling than most statistical approaches. However, data requirements on soil properties and thickness of colluvial or residual soil for process based GIS landslide models can be prohibitive for a relatively large area, and frequently it is impossible to acquire the input data necessary to use the models.

Natural hazard has been defined as the probability of occurrence of a potentially damaging phenomenon in a given area and in a given period of time (Varnes 1984). In this sense, landslide hazard modelling should provide information on the probability of landsliding in terms of both space and time in a certain area. Susceptibility mapping of landslides combined with storm events is the trend in landslide forecasting, and the former is a condition of landslide, the latter is the trigger. Storm events involve meteorological parameters, which are difficult to assess; furthermore it has little effect on land-use planning which only requires information on the probability of landsliding in terms of space in a certain area. So, studies concerning landsliding susceptibility have been carried out by all kinds of methods: such as, GIS overlay model (Gupta & Joshi 1989), statistical models combined with GIS technologies (Carrara et al. 1991), statistical and simulation models (Mark & Ellen 1995), generalized linear models (Atkinson & Massari 1998), logistic regression models (Dai et al. 2004), analytical hierarchy process (Lulseged et al.2005) and artificial neural networks (Leonardo et al.2005, Go'meza & Kavzoglub 2005). However, they mostly promised better performance of those models, gaps between application and models still remain, and new mathematic tools of prediction are continuing to be invented, so studies about this work should be boosted incessantly.

Support vector machine (SVM) is one of new mathematic tools which is used as a universal constructive learning procedure based on the statistical learning theory developed by V. Vapnik (Vapnik, 1995). It provides non-linear solutions to regression and classification problems by transforming the input variables in a large-dimension space, whose inner product is given by positive definite kernel functions. SVM is trained using dual optimisation techniques with constraints. Recently several research groups about engineering have shown excellent performance of SVM on different problems of regression and classification.

In general, this study will concentrate on a representative part of Hong Kong (Figure 1), reaching the purposes as follows:

- To develop a SVM method based on ArcGIS and Matlab for determining the probability of landsliding in term of space.
- To apply the model in the study area to map the probability of landsliding.
- To compare performances of SVM model with the counterparts of Artificial Neural Network model applying to the same study area.

It is hoped that this paper can be an introduction of SVM used in the field of environmental geology. It also provides valuable new information for landslide assessment and hazard mitigation planning, and may serve as a prototype for further work on the natural terrain of the whole territory of Hong Kong and contribute to landslide hazard warning strategies.

SUPPORT VECTOR MACHINE

In recent years, with the advance of computational efficiency combined with sophisticated statistical methods, machine learning methods have been increasingly used and shown as powerful tools in a wide variety of science disciplines including planetary science, computer science, bioinformatics, and environmental science (Mjolsness & DeCoste 2001). Among many machine-learning methods, SVM, originally developed by Vapnik (1995), are considered to be a new generation of learning algorithms. SVM have several appealing characteristics for modellers, including: they are statistically based models rather than loose analogies with natural learning systems, and they theoretically guarantee performance (Cristianini & Scholkopf 2002). SVM have been applied successfully to text categorization, handwriting recognition, gene-function prediction, remote sensing classification and ecology (Guo Q. H. et. al. 2005) demonstrating the utility of the method across disciplines, and proving that SVM produce very competitive results with the best available classification methods, and require just a minimum amount of model tuning (Joachims 1998, Brown et al., 2000, Cristianini & Scholkopf 2002, Decoste & Scholkopf 2002, Huang et al. 2002). Typically, SVM are designed for two-class problems where both positive and negative objects exist. For these classification problems, two-class SVM seek to find a hyperplane in the feature space that maximally separates the two target classes.

Consider a set of training points x_i (i = 1, 2, . . .,l) which are assigned to one of two classes with corresponding labels $y_i = \pm 1$. The goal of the two-class SVMs is to find an optimal separating hyperplane with the maximal margin between the training points for class –1 and class +1. Define a discriminant function:

$$g(x) = (w \cdot x) + b \tag{1}$$

where w = (w1, . . ., wn) is a vector of n elements, n is the dimension of the feature space; b is a scalar. $(w \cdot x)$ represents the inner product between w and x.

The classification rule is:

$$f(x) = \text{Sign}((w \cdot x) + b) \tag{2}$$

$$f(x) > 0 \Rightarrow x \in \text{class } y_i = +1 \tag{3}$$

$$f(x) < 0 \Rightarrow x \in \text{class } y_i = -1$$

Hence, the optimization problem can be formulated as:

$$\text{Minimize } \frac{1}{2}\|w\|^2 \tag{4}$$

Subject to:

$$y_i((w \cdot x_i) + b) \geq 1 \tag{5}$$

The problem can be solved by the Lagrangian:

$$L = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{l} a_i(y_i((\omega \cdot x_i) + b) - 1) \tag{6}$$

where a_i: $i = 1, \ldots, l$; $a_i \geq 0$ are the Lagrange multipliers.

Taken the derivative with respect to w and b, and

$$w = \sum_{i=1}^{l} a_i y_i x_i \tag{7}$$

$$\sum_{i=1}^{l} a_i y_i = 0 \tag{8}$$

Substituting (7) and (8) into (9) gives the dual form of the Lagrangian:

$$L = \sum_{i=1}^{l} a_i - \frac{1}{2} l \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j (x_i \cdot x_j) \tag{9}$$

Subject to:

$$a_i \geq 0, \ \sum_{i=1}^{l} a_i y_i = 0 \tag{10}$$

So far, what we have discussed above is suitable for linearly separable cases. We now turn to more general cases:

One case, for linearly non-separable cases (e.g. some classes overlap in the feature space), a slack variable ξ_i (i= 1,. . . , l) is introduced into the constraints to give:

$$y_i((w \cdot x) + b \geq 1 - \xi_i \tag{11}$$

$$\xi_i \geq 0$$

An additional term is introduced to the cost function by replacing (4) by:

$$\text{Minimize} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i \tag{12}$$

where C is a parameter to measure the amount of penalty for misclassification.

The other, for the non-linear decision boundary, similar to one-class SVM, kernel functions are introduced.

The optimization problem of Eq.(9) becomes:

$$L = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j K(x_i, x_j) \tag{13}$$

where:

$$K(x_i, x_j) = \phi(x_i)^T \cdot \phi(x_j) \tag{14}$$

is called the kernel function. The $\phi(x)$ is called the feature function that mapped training vectors x_i into a higher (maybe infinite) dimensional space.

And the decision rule can be expressed as:

$$f(x) = \mathbf{Sign}\left(\sum_{i=1}^{Ns} a_i y_i K(x_i, x_j) + b\right) \tag{15}$$

where *N*s is the number of support vectors. More detailed mathematical description about two class SVM can be found in Hastie et al. (2001) and Webb (2002).

The kernel is key to SVM model. Though new kernels are being proposed by researchers, the betters are the following four basic kernels:

linear: $K(x_i, x_j) = x_i^T x_j$ (16)

polynomial: $K(x_i, x_j) = (\gamma\, x_i^T x_j + r)^d, \gamma > 0$ (17)

radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2) \quad \gamma > 0$ (18)

sigmoid: $K(x_i, x_j) = \tanh(\gamma\, x_i^T x_j + r)$ (19)

Here γ, r and d are kernel parameters

In reality, we often do not have negative data, and thus commonly have only a one-class dataset. One-class SVM models also be developed by many study groups, but their theories are not reach perfection and their performances are less excellence than tow-class SVM (Guo Q. H. et. al. 2005, Deng & Tian 2005), so this study based on tow-class SVM. A more detailed mathematical derivation of one-class SVM can be found in Scholkopf et al. (1999), Deng & Tian (2005), and Tax & Duin (1999).

MATERIALS

Study area

The area of land in Hong Kong is 1099.5 km^2 (Figure 1), but the natural terrain covering 681.3 km^2 is as high as 62 % of total (Figure 2), and which is generally steep and regions with slope gradients exceeding 30° account for 29%. The climate of it is subtropical and monsoonal, characterized by hot, humid summers and mild, dry winters. Rainfall is heavy and occasionally intense during the rainstorms and typhoons. Average annual precipitation varied from 2,000 to 2,400 mm over the period 1961-1991 (Lam & Leung 1994). The hill slopes are often deeply gullied as a result of erosion caused by ephemeral streams. The relatively high permeability of the colluvium, when compared with the underlying saprolite or weathered rock, allows the development of transient perched water tables at the interface during, or following, periods of intense rainfall. So landslides are common, and a total of 29,229 landslides were recorded in Natural Terrain Landslide Inventory (NTLI) from 1945 to 1997 (GEO 1997).

The landslides observed in the area are typically small movements involving mainly the colluvium and regolith. The scar is usually small, with about 80% of the total slides, as interpreted from the available aerial photographs, being less than 20m in source width. Most of the landslides begin as a simple translational or in some cases rotational landslip and transform into a hill slope, or even channelized debris flow (Dai & Lee 1999, Dai & Lee 2002). So most of the landslides identified can therefore be characterized primarily as debris slides/flows using the terminology of Cruden & Varnes (1996). Based on field observations in some typical areas (Franks 1996, Wong et al. 1998, Dai & Lee 1999), the vast majority of the landslides, with an average failure depth of 1.4 m, occurred along the interface between the overlying thin layer of colluvium or residual soil weathered bedrock. (Figure 3)

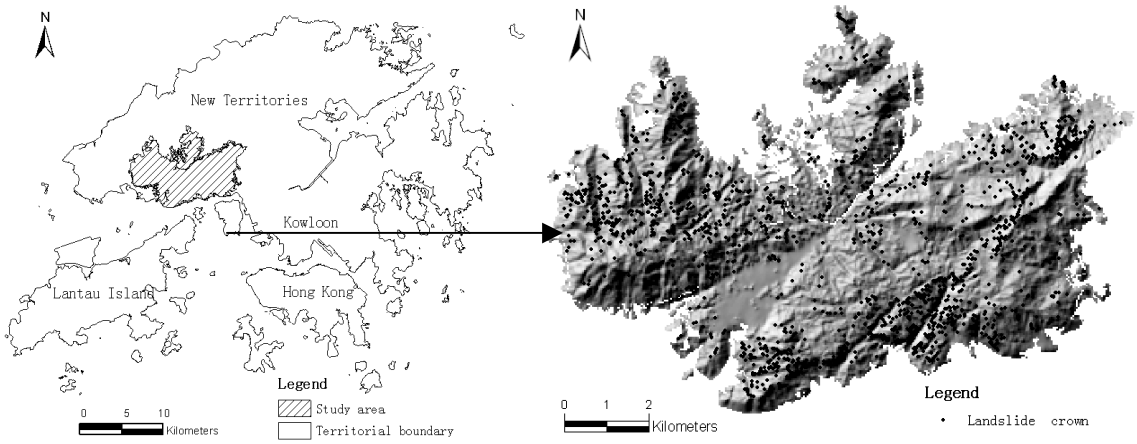

Figure 1. Study area located on Hong Kong, and its shade-map with 1680 landslides recorded in the NTLI.

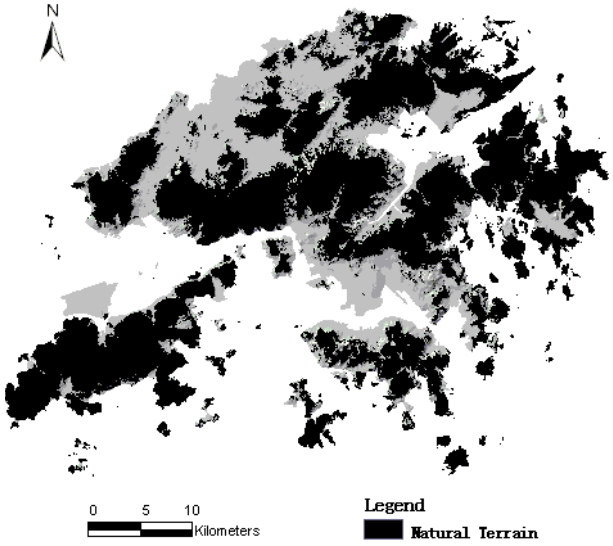

Figure 2. Natural terrain in Hong Kong

Figure 3. Shallow landslides in Hong Kong's natural terrain
*The whiter strips are landslide crowns and their debris flows.

The study area covers 142.5 km^2 (north-south: 10,000m, east-west: 14,200m) is located on the South of New Territories of Hong Kong; and it is total natural terrain (Figure 1) which was defined as: terrain that has not been modified substantially by human activity but including areas where grazing, hill fires and deforestation may have occurred (GEO, 1997). We abstract this region as study area because it is can represent of the natural terrain feature of Hong Kong, including both the inner and the coast, both the steep and the plainness, and 1680 landslides recorded in the NTLI (Figure 1) whose environmental parameters, such as slope gradient, slope aspect etc., nearly cove all numerical range in the NTLI .

Data

The goal was to predict regions where in presently have not been found out landslides are potentially likely to be risk area of landslides, thus made the assumption that if the area share similar condition features (both topographic, geology, and vegetable etc.) with landsliding sites, they are more likely to be potential targets for Landslide. The condition features used in this study were educed from the follow the base data:

- 1:5,000 scale digital topographic maps compiled by the Land Information Centre, Hong Kong
- 1:20,000 scale digital surficial and bedrock geological maps with accompanying memoirs compiled by the Geotechnical Engineering Office (GEO), Hong Kong. The Geological groups and their lithological description covering the natural terrain
- 1:20,000 scale digital Natural Terrain Landslide Inventory (NTLI) maps compiled by the GEO, based on the interpretation of high-level aerial photographs taken in 1945, 1964 and annually from 1972 to1997. (GEO 1997). The features recorded in the NTLI include both fresh scars of landslides that occurred during, or shortly before, the period covered by the aerial photography (1945-1997) and overgrown scars originally formed by earlier landslides. These two types of features are referred to as recent and relict landslides, respectively. According to the year of first observation, the relict landslides generally occurred before 1972 and nearly all the recent landslides occurred after 1945.

- 1:20,000 scale digital vegetation cover maps compiled by World Wide Fund for Nature (Hong Kong), based on the December 1989 aerial photographs and field check.

The SVM model demand two kinds of information: (1) sites of landslide and sites where are supposed stable permanently, and their condition features, (2) condition features of predicted regions. The former is samples that used to train two-class SVM model, the latter was inputted in the model to predict risk of region including them. Data sets represent a categorical attribute must be converted to numeric code in order to compute in SVM model, and the common method is rasterizing. A 20 m×20 m grid size was considered optimal for this study, as the scales of base maps are equal or greater than 1:20,000, so total 355,000 (500 rows×710 lines) grids ought to be generated in study area. Landslides recorded by NTLI is 1,680 in study area, which can covert to 1,587 grids, because some landslides, spacing less 20m, share one gird. The two-class SVM requires both absence and presence data to train the model. Since there is no absence data for presence-only data, one solution is to generate 'pseudo' absence data. The 'pseudo' absence points were generated regularly by 200m interval in both north and south direction, If the regular 'pseudo' absence points 'fall' in the 80 m-buffer of presence, it will be deleted, then 1,658 grids (nearly equal number of presence data) representing stable region features be abstract from study data. So total 3,245grids, the presence with '1' is 1,587, the 'pseudo' absence with '0' is 1,658, compose labels of sample set (Figure 4).

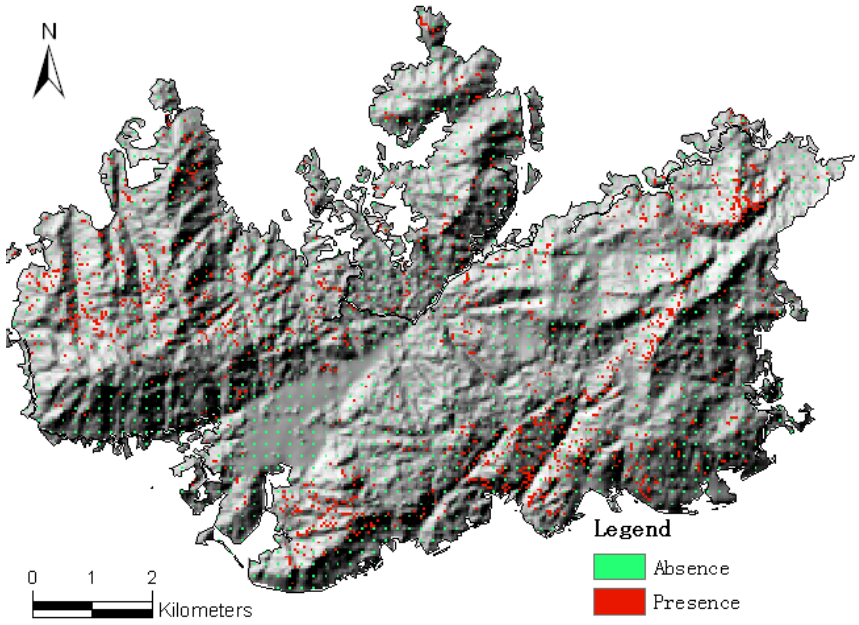

Figure 4. Presence/absence grids distribution.

Eight environmental variables were used to train the model and to predict the potential distribution for landslides. The variables included: (1) slope gradient, (2) slope aspect, (3) elevation, (4) plan of slope, (5) profile of slope, (6) curvature of the slope, (7) lithology, (8) vegetation cover. The above selections were made based on the authors' knowledge of the physical environment and landsliding in the study area (Wong et al. 1998, Dai & Lee 1999, Dai & Lee 2002). A principal component analysis (PCA) does not show apparently significant redundancy in the variables, as we can also retain eight principal components, which account for 99.9%, and seven principal components, which account for 99%, of the variation in the data set. The variables, slope gradient, slope aspect, elevation, plan of slope, profile of slope and curve of slope can be derived from digital elevation model (DEM). Vegetation cover was derived from 1:20,000 scale digital vegetation cover maps. For ease of analysis, the 1:20,000 scale superficial and solid geology map covering the study area was divided into 14 groups (Table 2) based on lithological characteristics and engineering properties as indicated in Table 1. Other seven environmental variables were each divided into seven or eight classes by standard divarication except slope gradient, which is first broken at 15°because very fewer landsides were probable on shallower slopes.

Table 1. Lithological groups.

Category	Group	Description
Volcanic rocks	TB	Fine-coarse ash tuff, tuffite and tuff breccia
	CT	Coarse ash crystal tuff
	TDR	Trachydacite, dacite and rhyolite lava
	E	Extaxite
Volcaniclastic sedimentary rocks	LCC	Coarse ash tuff, rhyolite lave, mudstone and siltstone-Lai Chi Chong Formation
	LVS	Lavas, volcaniclastic sediments and sediments
Intrusive rocks minor intrusive rocks	G	Granodiorite, granite aplite, trachyte
	MI	Aplite, trachyte, monzonite, syenite, latite, lamprophyre, basalt and dolerite
	RF	Feldsparphyric rhyolite and microgranite
Sedimentary rock	CBS	Conglomerate, breccia, sandstone and siltstone
	MS	Musdstone and siltstone
Metasedimentary rock	CS	Carboniferous siltstone, sandstone, graphite schist, quartzite, chart and quartz
Superficial deposits	DF	Debris flow deposits, talus
	ATB	Alluvial, terrace and beach deposits

Table2. Data sets and their classes for this study.

Data set	Classes
Slope aspect	(1)N; (2)NE; (3) E; (4) SE; (5)S; (6)SW; (7)W; (8)NW
Elevation	(1)< 60; (2) 60－120 ; (3) 120－180; (4) 180－240; (5) 240－300; (6) 300－360; (7) 360－420; (8) 420－570(max)
Slope gradient	(1)< 15; (2)15－20; (3) 20－25; (4)25－30; (5) 30－35; (6)35－40; (7) 40－46(max);
Profile of slope	(1) <-1.6; (2) -1.6－-0.8; (3) -0.8－-0.3; (4) -0.3－-0.3; (5) -0.3－-0.8; (6) 0.8－1.5; (7) 1.5－4.3 (max)
Plan of slope,	(1) <-1.8; (2) -1.8－-1.2; (3) -1.2－-0.5; (4) -0.5－0.6; (5) -0.6－– 1.2; (6) 1.2－1.8; (7) 1.8－3.7(max);
Curve of slope	(1)<-2.8; (2)-2.8－-1.8; (3) -1.8－-0.9; (4)-0.9－0.9; (5) 0.9－1.9; (6) 1.9－2.8; (7) 2.8－5.9(max);
Vegetation cover	(1)Bare rock or soil ; (2) Grassland; (3) Low shrubs with grass; (4); Low shrubs; (5) Tall shrubs with grass; (6)Tall shrubs; (7) Plantation woodland; (8) Woodland
Lithology	(1) TB; (2) CT; (3) TDR; (4) E; (5) LCC; (6) LVS; (7) G; (8) MI; (9) RF; (10) CBS; (11) MS; (12) CS; (13) DF; (14) ATB

Another import geological factor, faulting, was not been involved because they have little effect on the shallow landslides according to the experiences of GEO. The data sets above were all rasterized by a spatial size of 20 m×20 m to build eight rows × 355,000 columns matrix that was used to prediction, from which an eight rows × 3,245 columns matrix, matching the presence/absence labels, was abstracted.

APPLICATION OF DATA IN SVM MODEL

Software

The graphic operation and computation of model in this study were done in ArcGIS8.3 and Matlab7.01 respectively. Matlab was chose for computing because it not only possesses good performance and face-friendly but also provide tools of Artificial Neural Network (ANN), the ANN method was used to prediction the susceptibility map (this will described in later) in order to evaluating the performance of SVM. However, there are no SVM tools in Matlab, many free SVM programs can be downloaded from the internet, providing all kinds of interfaces to other software. LibSVM (Chang & Lin 2005) and SVM- Light (Joachims 2004) are two widely used SVM software library, which all provide the interface to Matlab. Data converted to Matlab standard can be utilized repeatedly in SVM and ANN. The process of data rasterizing was based on the TopoGrid, Space Analysis and 3D module in the ArcGIS8.3 software. ArcGIS's Grid to ASCII module can transfer grid to text format, but which cannot be recognized by Matlab

directly, so a little program compiled by author interfaced them. The reverse process, data transferred from Matlab to ArcGIS, can be made utilizing the little program and ASCII to Grid module in ArcGIS.

Procedure of computing

Procedure of SVM computing by following steps:

- • Transform data to the format of libSVM software
- • Conduct scaling on the data
- • Consider the RBF kernel $K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$
- • Use cross-validation to find the best parameter C and γ
- • Use the best parameter C and γ to train the whole training set
- • Test (Wei et al. 2005)

Scaling them before applying SVM is very important. (Sarle 1997, Part 2 of Neural Networks FAQ) explains why we scale data while using Neural Networks, and most of considerations also apply to SVM. The main advantage is to avoid attributes in greater numeric ranges dominate those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. Because kernel values usually depend on the inner products of feature vectors, e.g. the linear kernel and the polynomial kernel, large attribute values might cause numerical problems. We recommend linearly scaling each attribute to the range [–1, +1] or [0, 1]. Of course we have to use the same method to scale testing data before testing.

Though there are only four common kernels mentioned in section 1, we must decide which one to try first. Then the penalty parameter C and kernel parameters are chosen. In general, the RBF is a reasonable first choice. The RBF kernel nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. Furthermore, the linear kernel is a special case of RBF as (Keerthi & Lin 2003) shows that the linear kernel with a penalty parameter ˜ C has the same performance as the RBF kernel with some parameters (C,). In addition, the sigmoid kernel behaves like RBF for certain parameters (Lin & Lin 2003). The second reason is the number of hyperparameters which influences the complexity of model selection. The polynomial kernel has more hyperparameters than the RBF kernel. Finally, the RBF kernel has less numerical difficulties. One key point is $0 < K_{ij} \leq 1$ in contrast to polynomial kernels of which kernel values may go to infinity $(\gamma x_i^T x_j + r > 1)$ or $(\gamma x_i^T x_j + r < 1)$ while the degree is large. Moreover, we must note that the sigmoid kernel is not valid (i.e. not the inner product of two vectors) under some parameters (Vapnik 1995). The more details be discuss in the Wei et. al. 2005

By above straining tactics, sample sets were then divided into three sub-datasets used for training, validating and testing by scale 2: 1:1. A SVM model was trained and validated repeatedly by the training and validating sub-dataset respectively. Changing the C and γ a two-dimension accuracy graph was gain, from which the optimum parameters for this study, C with 2.1 and γ with 0.9, were selected. And the performance of the model was tested by the testing sub-dataset. Then, the optimum function was trained by all 3,245 presence/absence samples, which subsequently computed 355,000 grids.

In general, ANN model regard as a perfect learning machine. Matlab has developed dozens of successful ANN models, and provided perfect artifice in its help document. Based on the back-propagation (BP) model in ANN toolbox and following its ‘cookbook’, an optimum network including 2 inner layers, each inner layer has 10 ‘tansig’ nodes, was selected. The same data sets were applied to BP model combine with Matlab’s optimum operating method. Finally, outputs of SVM and BP were all converted to ArcGIS to analyse the performance.

Susceptibility mapping

Accuracy is a standard to distinguish performance of predict models. However, “more higher, more performance” is not exact, because train set can train model to over fit it. If the over fit model is used to test train set, it can reach a higher accuracy. In this study, over fit accuracy can reach 99%, but it is not more than 55% to validate set. A model with 94% accuracy to its train data and 69% validation accuracy is more powerful than the over fit model. i.g. we focus on the power of generalization capacity. We can get the fact that accuracy of self-validate is 81% in the optimum parameter(C=2.1, γ =0.9).

The realization of prediction of SVM and ANN led to the productions of two landslide susceptibility maps. These maps portray the location of slopes that are prone to shallow and channelized slides. For comparison purposes, the two susceptibility maps were classified to four classes: very stable, stable, susceptible and very susceptible. Both SVM (don’t use decision rule in Eq.15) and ANN produce sequential outputs which scaled to range from -1 to 1, and thresholds of division must be reasonable decided because thresholds marked affect the classified accuracies. A simple criterion, equal amount of grids to four classes, was used in this paper, i.g. the divided areas of four classes is equal. Figure 5 shows the susceptibility map produced by SVM. Similarly, Figure 6 presents the susceptibility map created by ANN.

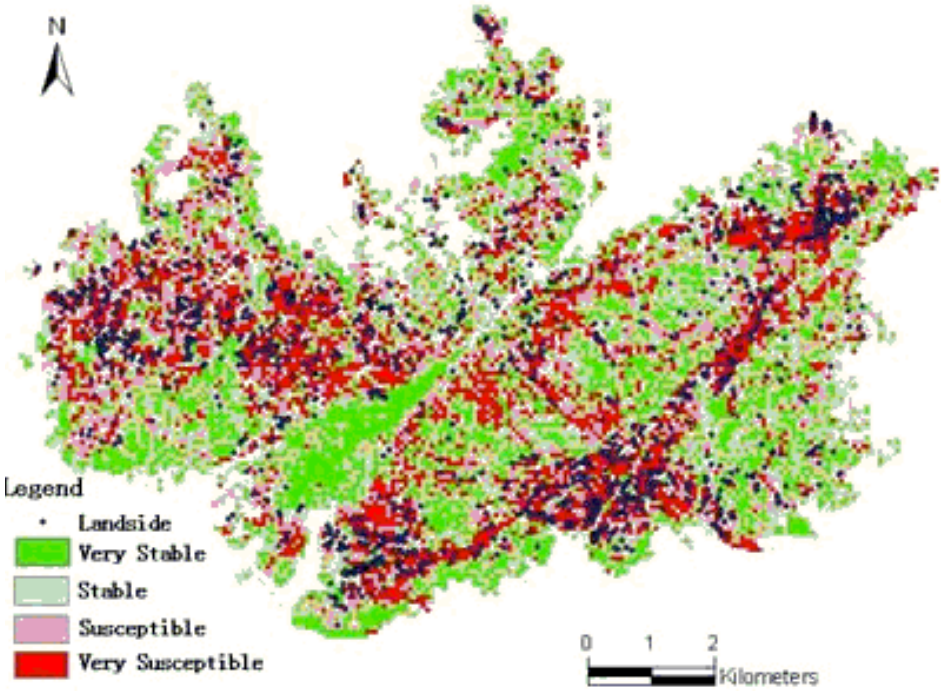

Figure 5. Susceptibility map of landslides predicted by SVM

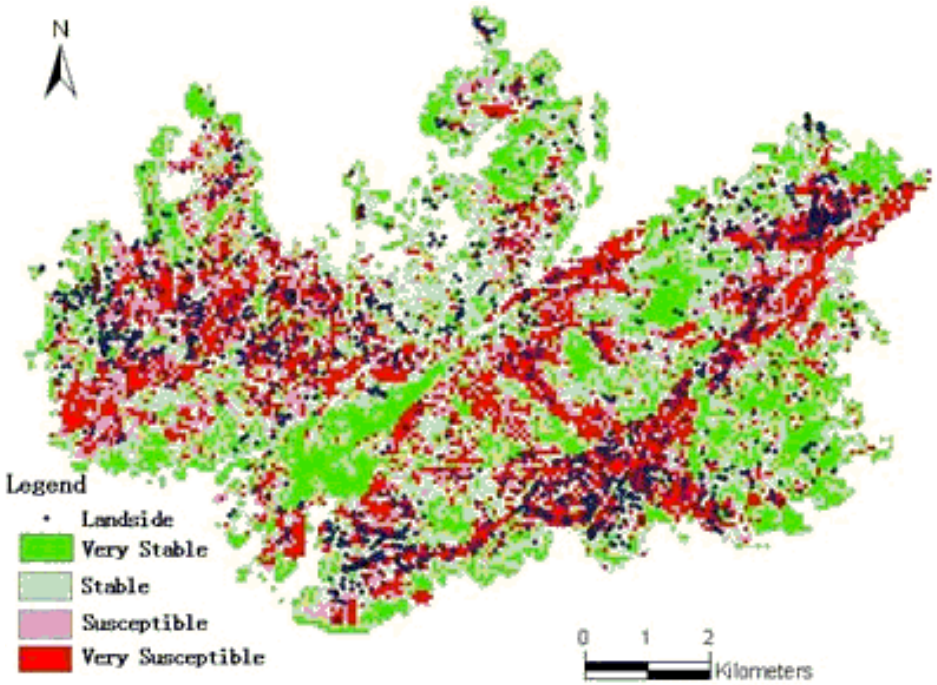

Figure 6. Susceptibility map of landslides predicted by ANN.

Looking at the two maps, there are places where differences are subtle and there are also areas with obvious dissimilarities. Both maps don't hold the distinct hallmarks of any environmental factors and present a synthetical result, as coincides with the result of PCA (described in section 3). Besides, such classification roughly reduces a high susceptible northwestern, medium and southeastern susceptible central, and between which low susceptible parts disperse. However, there are more convergent red colors on the susceptibility map produced by ANN, as result in some present point fall into the green region, in contrast, the map of SVM avoid this kind of mistake. This means that the susceptibility map of SVM is more discrete and more flexible than ANN's, and SVM can gain exacter trend than ANN. As shown in Table 4, distribution of presence/absence samples in the four classes produced by SVM, have better performances than the counterparts of ANN. However we also ought to pay attention to the fact that distributions of 'pseudo' absences are not good, as shows the method generating absence samples has something and should be promote.

Table 4. The distribution of presences/absences in four predicted regions.

Classes	SVM prediction		BP network prediction	
	Amount of presences	Amount of absences	Amount of presences	Amount of absences
In very stable terrain	46	979	73	486
In stable terrain	49	172	270	469
In susceptible terrain	411	356	470	407
In very susceptible terrain	1081	151	774	296

VERIFICATIONS

To verify the practicality of the results, a comparison was made between the two susceptibility maps. The verification process has started in such a way that each of the two susceptibility maps firstly unite four classes into tow classes•stable and susceptible. Second, the amounts of the presence/absence in the destination were surveyed for two maps respectively, and accuracies were computed. Furthermore, a landsliding map of 255 landslides occurring from 1998 to 2000, delineated by GEO, was employed and done the same investigation as the 3245 samples, and the 225 landslides mainly distributed in the northwestern and southeastern study region (Figure 7). Results of verification are listed in Table 5. Accuracies of SVM are all higher than counterparts of ANN, and the results are also higher than the accuracies predicted by the logistic regression model in Hong Kong natural region, with the similar environmental variables (Dai et al. 2004). Especially, accuracy using the data from 1998 to 2000 is as high as 85%, which present SVM possesses powerful generalization capacity in this paper's study method.

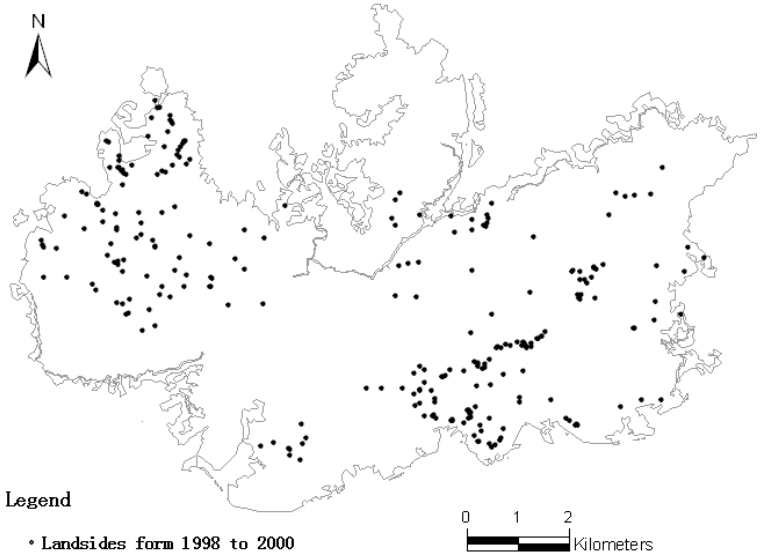

Figure 7. 225 landslides occurred from 1998 to 2000.

Table 5. The accuracy of classification.

Observed	amount	SVM prediction		ANN network prediction	
		amount in destination	Accuracy	amount in destination	Accuracy
Presence of slide	1587	1492	94%	1244	78%
Absence of slide	1658	1151	69%	955	58%
Total	3245	2643	81%	2199	68%
Verifications (1998—2000)	255	218	85%	198	78%

DISCUSSION

The accuracies of many similar studies present by others are higher than this paper's, e.g. a similar study by ANN (Leonardo et al. 2005, GO´meza & Kavzoglub 2005) more than 90%. A biological study, Support Vector Machines

for Predicting Distribution of Sudden Oak Death in California (Guo et al. 2005), is as high as 92%. However, this study does not reach those levels; why? Accuracy is relative, we will discuss by follow reasons:

First, data used to prediction have different precision in different study. The scales of all maps are more than 1: 20,000, and the landslide records are rang about 50 years in this study. In contrast, the 'Oak Death' study (Guo et al. 2005) with 1:100,000 scale maps and records were collected in two years. Furthermore, the conditions of Leonardo et al. (2005) and GO´meza & Kavzoglub (2005) studies are more simple than this study's, so it is certain that accuracies are different in the same predicting model. The factor of second is that the simple samples sets can get a better performance, and the rate of access to reality is low, but the complex samples get inverse results. Thirdly, sampling method has a significant affect on the result. In this study ,grid spacial size changing to 40m×40m, we can only get a total accuracy of 75%, but magnifying the buffer of presence will increase the accuracy. Especially the 'pseudo' absence has fault because some factors couldn't be take into account, which may depress the performance. Finally, experience of authors is another important factor. Not as the logistic regression model and bayes classifier, which can get the optimal result by auto computation, initial parameters of SVM and ANN must be manual selected. This can induces different accuracy in the same condition. Above all, accuracy of prediction are determined by many factors, it can't veridical present the performance of a model, so we must synthetically consider the results.

Though the accuracy of SVM isn't very wonderful in this study, it is higher than the ANN's, and better than similar study in Hong Kong, which can prove that it is powerful and potential in landslide susceptibility mapping, after all it is only ten years old and developing rapidly. This paper provides only an introduction to SVM used in environmental geology; more detail need more authors to continue.

CONCLUDES

In this study, we used two-class SVM to predict potential distribution of landslides in natural terrain located on New territories of Hong Kong. SVM and ANN were also compared used the same presence/absence samples. SVM has better accuracy than ANN when evaluated at both cross validation, total self validation and true future data verification. ArcGIS combined with Matlab and SVM software, can successful predict hundred thousands of grids, and the data can be effectively used by other predict tools in Matlab. It is also an efficient method to deal with similar geological study. Two models show that landslides will occur in the half of natural terrain, and the region SVM showed is more discrete and more flexible (Figures 5 and 6). We believe that SVM, while not used commonly in geology, is a useful addition to environmental geology modeling. When coupled with geographic information systems, SVM will be a useful method in geological analysis. We plan to further investigate the differences between the models in regions, to refine our understanding of the complex interaction between the environmental variables in areas, such as Lantau Island of Hong Kong, and compare the results of this work with other modeling approaches by previous research (Dai & Lee 2002). We also plan to expand this modeling to deal with presence-only data and promote its accuracy.

Acknowledgements: Supported By Open Research Fund Program of the Geomatics and Applications Laboratory, Liaoning Technical University . The data used in this paper was provided by the Geotechnical Engineering Office of Hong Kong. The authors wish to express their sincere appreciation for the generous support.

Corresponding author: Mr. X. Yao, Institute of Geology and Geophysics, Chinese Academy of Sciences, P.O. Box 9825, Beijing 100029. Tel: +86 010 62007746. Email: yaoxinphd@yahoo.com

REFERENCES

ATKINSON P.M. & MASSARI R. 1998. Generalized linear modeling of landslide susceptibility in the Central Apennines, Italy, *Computing Geoscience,* **24**:373–385.

BROWN, M.P.S., GRUNDY, W.N., LIN, D., CRISTIANINI, N., SUGNET, C.W., FUREY, T.S., ARES, M. & HAUSSLER, D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *In: Proceedings of the National Academy of Sciences of the United States of America,* ***97***, 262–267.

BURGES. J.C. 1997.A Tutorial on Support Vector Machines for Pattern Recognition. *Bell Laboratories*, Lucent Technologies.

CARRARA A., CARDINALI M., DETTI R., GUZZETTI F., PASQUI V. & REICHENBACH P. 1991. GIS techniques and statistical models in evaluating landslide hazard .*Earth Surf Process Landforms* **16**:427-445

CRUDEN D.M. & VARNES, D.J. 1996. *Landslides type and processes*. In: Schuster RL, Turner AK eds Landslides, investigation and mitigation. Transportation Research Board Special Report 247, National Academy Press, Washington, DC, 36-75.

CHANG, C. & LIN, C. 2005. LIBSVM: A Library for Support Vector Machines. *Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.*

CRISTIANINI, N. & SCHOLKOPF, B. 2002. Support vector machines and kernel methods—the new generation of learning machines. *Ai Magazine,* **23**, 31–41.

DAI, F.C., LEE, C.F. & SIJING, W. 1999. Analysis of rainstorm-induced slide-debris flows on natural terrain of Lantau Island, Hong Kong. *Engineering Geology,* **51**, 279-290.

DAI, F.C., LEE, C.F. 2002. Landslide characteristics and slope instability modelling using GIS, Lantau Island, Hong Kong. *Geomorphology,* **42,** 3-228.

DAI, F.C., LEE, C.F., THAM, L.G., NG, K.C. & SHUM, W.L. 2004. Logistic regression modelling of storm-induced shallw landsliding in time and space on natural terrain of Landtau Island, Hong Kong. *Bulletin of Engineering Geology and the Environment*, **63,** 315-327.

DECOSTE, D. & SCHOLKOPF, B., 2002. Training invariant support vector machines. *Machine Learn*ing **46,** 161–190.

DENG, N.Y. & TIAN, Y.J. 2005. *Support Vector Mechine: A new method of data mining.* Science Press of China, **3**:164~223 (in Chinese).

FRANKS C.A.M. 1996. Study of rainfall induced landslide on natural slopes in the vicinity of Tung Chung New Town, Lantau Island. *Special Project Report SPR* **4**/96, Geotechnical Engineering Office, Hong Kong.

GEO. ab. Geotechnical Engineering Office. 1997. *Natural Terrain Landslide Study Report* Technical Note TN 10/97 Geotechnical Engineering Office of Hong Kong Special Administrative Region

GUZZETTI, F., CARRARA, A., CARDINALI, M. & REICHENBACH, P. 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, center Italy. *Geomorphology,* **31**, 181-216.

GUPTA, R.P. & JOSHI, B.C. 1998. Landslide hazard zoning using the GIS approach: a case study from the Ramganga catchment, Himalayas. *Engineering Geology,* **28,** 119-131.

GO´MEZA H, KAVZOGLUB H, 2005 Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela *Engineering Geology* **78**: 11 –27

GUO, Q.H., KELLY, M. & GRAHAM, C.H. 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling,* **182** 2005 ,75–90

HUANG, C., DAVIS, L.S., TOWNSHEND, J.R.G., 2002. An assessment of support vector machines for land cover classification. *International Journal of Remote Sens*ing, **23**, 725–749.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2001. The Elements of Statistical Learning: Data Mining Inference and Prediction. Springer, New York.

JOACHIMS, T. 1998. Text categorization with support vector machines:learning with many relevant features. *In: Proceedings of European Conference on Machine Learning*. Springer-Verlag, Berlin, 137–142.

JOACHIMS, T. 2004. SVM-Light: a library for support vector machines, Version: 6.01 software available at http://svmlight.joachims.org/

KEERTHI, S.S. & C.-J. LIN. 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation,* **157**, 1667–1689.

LIN, H.-T. & C.-J. LIN. 2003. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Technical report, Department of Computer Science and Information Engineering, National Taiwan University. *Available at http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf.*

AYALEW, L., YAMAGISHI, H., MARUI, H. & KANNO, T. 2005. Landslides in Sado Island of Japan: Part II. GIS-based susceptibility mapping with comparisons of results from two methods and verifications. *Engineering Geology,* **84,** 432-445.

LAM, C.C.•LEUNG, Y.K. 1994. *Extreme rainfall statistics and design rainstorm profiles at slected locations in Hong Kong*, Technical Note No. 70, Royal Observatory of Hong Kong, Hong Kong.

LEONARDO ERMINI, FILIPPO CATANI, NICOLA CASAGLI, 2005 Artificial Neural Networks applied to landslidesusceptibility assessment *Geomorphology,* **66,** 327–343.

MJOLSNESS, E. & DECOSTE, D. 2001. Machine learning for science: state of the art and future prospects. *Science* **293**, 2051–2055.

MARK, R.K. & ELLEN, S.D. 1995. Statistical and simulation models for mapping debris flow hazard. *In CARRARA A, GUZZETTI F eds Geographical information systems in assessing natural hazards*. Kluwer, Dordrecht, 93-106.

MULIER, F. 1999. Vapnik-ChervonenkisVC Learning Theory and Its Applications. *IEEE Transactions on Neural Networks*. **105**.

MONTGOMERY DR , DIETRICH WE 1994 A physically based model for the topographic control on shallow landsliding. *Water Resour* **30**:1153-1171

SARLE, W.S. 1997. Neural Network FAQ. Periodic posting to the Usenet newsgroup comp.ai.neural-nets.

TAX, D. & DUIN, E. 1999. Support vector domain description. *Pattern Recognition Letters*, **20**, 1191–1199.

VAPNIK, V. 1995. *Nature of statistical learning theory.* John Wiley and Sons Inc., New York (in preparation).

VARNES, D.J. 1984. Landslide hazard zonation: a review of principles and pactice. *Natural Hazards* No.3, UNESCO, Paris.

WEBB, A., 2002. *Statistical Pattern Recognition*. JohnWiley & Sons,New York.

WEI CHIH HSU, CHIH-CHUNG CHANG & CHIH-JEN LIN. 2005. A Practical Guide to Support Vector Classification. Department of Computer Science and Information Engineering *Available at http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf.*

WONG, H.N., LAM, K.C. & HO, K.K.S. 1998. *Diagnostic report on the November 1993 natural terrain landslides on Lantau Island* .GEO Report No. 69, Geotechnical Engineering Office, Hong Kong

WU, W. & SLIDE, R.C. 1995. A distributed slope stability model for steep forested basin. *Water Resources Research,* **31**, 2097–2110.