Terrain-based mapping of landslide susceptibility using a geographical information system: a case study

F.C. Dai and C.F. Lee

Abstract: This paper deals with the development of a technique for mapping landslide susceptibility using a geographical information system (GIS), with particular reference to landslides on natural terrain. The method has been applied to Lantau Island, the largest outlying island within the territory of Hong Kong. Landslide susceptibility in the study area is related to a number of terrain variables, viz., lithology, slope gradient, slope aspect, elevation, land cover, and distance to drainage line. Multiple correspondence analysis (MCA) was carried out to generate the principal axes that are linear combinations of these terrain variables using occurrence data of landslides and terrain variables. A GIS is used to project the values of the principal axes, and subsequently to relate these principal axes to landslide susceptibility by logistic regression modeling. The spatial landslide susceptibility response in the study area can then be obtained by applying this logistic regression model to the study area. The results from this study indicate that such a GIS-based model is useful and suitable for the scale adopted in this study.

Key words: landslides, geographical information systems, multiple correspondence analysis, logistic regression, terrain analysis.

Résumé : Cet article traite du développement d'une technique de cartographie de la susceptibilité aux glissements utilisant un système d'information géographique (GIS), référant particulièrement aux glissements dans le terrain naturel. La méthode a été appliquée à Lantau Island, la plus grande île périphérique du territoire de Hong Kong. La susceptibilité au glissement dans la région étudiée est reliée à un certain nombre de variables de terrain telles que: la lithologie, le gradient de la pente, l'aspect de la pente, l'élévation, la couverture du terrain, et la distance à la ligne de drainage. De multiples analyses de correspondances (MCA) ont été faites pour générer les principaux axes qui sont les combinaisons linéaires de ces variables de terrain au moyen des données d'occurrence de glissements et des variables de terrain. Un GIS est utilisé pour projeter les valeurs des principaux axes, et subséquemment pour mettre en relation ces principaux axes avec la susceptibilité aux glissements par modélisation de régression logistique. La réponse spatiale de la susceptibilité aux glissements dans la région étudiée peut alors être obtenue en appliquant le modèle de régression logistique à la région étudiée. Les résultats de cette étude indiquent qu'un tel modèle basé sur un GIS est utile et convient à l'échelle adoptée dans cette étude.

Mots clés : glissements, système d'information géographique (GIS), analyse de correspondances multiples (MCA), régression logistique, analyse de terrain.

[Traduit par la Rédaction]

Introduction

Landslides in mountainous terrains often occur as a result of heavy rainfall, resulting in the loss of life and damage to the natural and (or) human environment. Sites that are prone to landslides should therefore be identified in advance to avoid such damage. In this regard, landslide-hazard mapping can provide much of the basic information essential for hazard mitigation through proper project planning and implementation.

Landslide hazard was defined by Varnes (1984) as the probability of occurrence of a potentially damaging landslide phenomenon within a specified period of time and within a given area. The factors which determine the landslide hazard of an area can be divided into two groups: (*i*) the quasi-static variables, which contribute to landslide susceptibility, such as geology, slope gradient, slope aspect (i.e., orientation of slope face), elevation, geotechnical properties, and long-term drainage patterns; and (*ii*) the dynamic variables, which tend to trigger landslides in an area of given susceptibility, such as rainfall and earthquakes (Wu and Sidle 1995; Atkinson and Massari 1998). Obviously, the probability of a landslide depends on both the quasi-static and dynamic variables. However, the dynamic variables may change over a very short time span, and are thus very difficult to estimate. The spatial distribution of the quasi-static

Received November 6, 2000. Accepted March 12, 2001. Published on the NRC Research Press Web site at http://cgj.nrc.ca on September 25, 2001.

F.C. Dai. Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences, Beijing 100101, People's Republic of China.
C.F. Lee.[1] Department of Civil Engineering, University of Hong Kong, Hong Kong.

[1]Corresponding author (e-mail: leecf@hkucc.hku.hk).

DOI: 10.1139/cgj-38-5-911

variables within a given area determines the spatial distribution of relative landslide susceptibility in that region (Carrara et al. 1995). Brabb (1984) defined landslide susceptibility as the tendency for a landslide to be generated in a specific area in the future. Up to now, most studies (e.g., Yin and Yan 1988; Carrara et al. 1991, 1995; Niemann and Howes 1991; Atkinson and Massari 1998) have focused on the indirect determination of landslide susceptibility, rather than on landslide hazard as defined by Varnes. These studies have been largely based on the general principle that "the past and the present are the keys to the future," i.e., future slope failures will more likely occur under those conditions which led to past and present landslides (Brabb 1984; Niemann and Howes 1991; Carrara et al. 1995; Atkinson and Massari 1998).

A variety of techniques have been developed to assess landslide susceptibility. They can be grouped into the inventory, heuristic, statistical, and deterministic approaches (Soeters and Van Westen 1996; Van Westen et al. 1997; Atkinson and Massari 1998). Landslide inventory mapping is the most straightforward initial approach to any study of regional landslide hazard and is the basis of most susceptibility mapping techniques (Soeters and Van Westen 1996). Landslide inventory maps can be used as an elementary form of susceptibility map because they show the location of recorded landslides. They do not, however, identify areas that may be susceptible to landslides unless landslides have already occurred in such areas in the past (Evans et al. 1997; Atkinson and Massari 1998).

Heuristic models use expert opinions to estimate landslide potential from data on quasi-static variables only. They are based on the assumption that the relationships between landslide susceptibility and the quasi-static variables are known and are specified in the models. A set of variables are entered into the model to estimate landslide susceptibility (Anbalagan 1992; Pachauri and Pant 1992; Niemann and Howes 1991; Atkinson and Massari 1998). The limitations in the heuristic models are in the reproducibility of results and in the subjectivity in assigning weightings and ratings to the variables.

Statistical models involve the statistical determination of the combinations of variables that have led to past landslides. Quantitative or semiquantitative estimates are then made for areas currently free of landslides, but where similar conditions exist. Both simple and multivariate statistical approaches have been used widely in such indirect mapping of landslide susceptibility (Yin and Yan 1988; Bernknopf et al. 1988; Gupta and Joshi 1989; Siddle et al. 1991; Carrara et al. 1991, 1995; Wang and Unwin 1992; Naranjo et al. 1994; Atkinson and Massari 1998).

Deterministic approaches are based on slope stability analyses and the limit equilibrium method. They are only applicable when the ground conditions are reasonably uniform or known across the study area and the landslide types are known and relatively easy to analyze. The infinite slope stability model has been widely used to assess landslide susceptibility in small areas (Van Westen 1993; Terlien et al. 1995; Wu and Sidle 1995). The advantage of the deterministic models is that they permit quantitative factors of safety to be calculated, and the main problem with the deterministic models is the high degree of simplification that is usually necessary for their use.

Not all methods of landslide susceptibility mapping mentioned previously are equally applicable at different scales of analysis. Deterministic techniques require very detailed input data, which can only be collected for small areas because of the required level of effort (Evans et al. 1997). Statistical techniques are generally considered the most appropriate approach for landslide susceptibility mapping at scales of 1 : 20 000 to 1 : 50 000, because at this scale it is possible to map out in detail the occurrence of past landslides and to collect sufficient information on the variables that are considered relevant to the occurrence of landslides (Naranjo et al. 1994).

Recently, the geographical information system (GIS) has become an important tool for landslide susceptibility mapping because it provides the various functions of handling, processing, analyzing, and reporting geospatial data. The overlay operation commonly applied within the GIS is useful in both the heuristic and the statistical approaches (Gupta and Joshi 1989; Carrara et al. 1991, 1995; Wang and Unwin 1992; Fernandez et al. 1999; Van Westen et al. 1997; Mark and Ellen 1995). The infinite slope stability model has also been incorporated into the GIS to calculate the spatial distribution of the factor of safety within a given region, based on the assumption that landslides generally occur along shallow failure surfaces (Terlien et al. 1995; Van Westen et al. 1997; Wu and Sidle 1995).

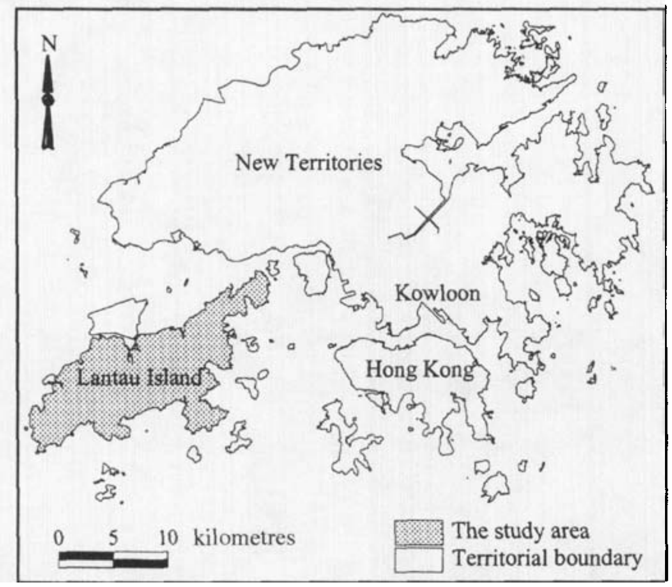
In this paper, Lantau Island, the largest outlying island of Hong Kong, is used as the study area, and a new statistical approach is presented for mapping landslide susceptibility on the island using a GIS. Landslide susceptibility in the study area is related to the terrain variables, viz., lithology, slope gradient, slope aspect, elevation, land cover, and distance to drainage line. Multiple correspondence analysis (MCA) is used to generate the principal axes, which are linear combinations of these terrain variables, using occurrence data of landslides and their terrain variables. A GIS is carried out to project the values of the principal axes across the geographic space, and subsequently to relate these principal axes to landslide susceptibility by logistic regression modeling. This methodology has been developed by using the ArcView GIS software.

Description of the study area

Lantau Island, situated in the southwestern part of Hong Kong and with a total land area of about 140 km^2, has been selected as a pilot study area (Fig. 1). This selection was made based on (*i*) the presence of numerous landslides with a high spatial concentration; (*ii*) the availability of existing datasets such as topographical maps, geological maps, land cover, and spatial distribution of landslides; (*iii*) the existence of steep terrain; and (*iv*) the presence of undeveloped terrain that is most suitable for mapping of landslide susceptibility using the GIS.

Lantau Island is virtually undeveloped and uninhabited, mainly because of the steep natural terrain. Land with slope angles >25° accounts for 44% of the total land area. The ground generally rises at about 30° from sea level everywhere on the island (Brand 1994). Elevation ranges from sea level to over 900 m above sea level and changes abruptly. The only flat land exists as occasional small coastal patches.

Fig. 1. Location of the study area.



The bedrock geology of the study area consists of volcanic rocks and a younger suite of granitic rocks. The volcanic rocks consist mainly of tuff and lava which are commonly banded. The former includes both the fine and coarse ash types. The bedrock materials, which are often heavily weathered in situ to form deep residual deposits, are sometimes overlain by deposits of younger superficial materials that are generally colluvial, alluvial, or littoral in character. The oldest rocks are the sandstones and siltstones. These sedimentary rocks occur as a small outcrops. Extensive deposits of colluvium probably blanketed the landscape as a result of numerous individual episodes of mass wasting and erosion during the Quaternary (Fig. 2). Debris-flow deposits, as part of the colluvial deposits, usually form distinct lobes within stream courses or at the mouths of drainage networks. In recent times, the alluvium and raised-beach sediments were deposited under the combined influence of higher sea levels and fluctuating climatic conditions (Geotechnical Control Office 1988*a*, 1988*b*). The area is structurally affected by two sets of faults trending northeast–north-northeast and north-northwest–northwest.

The climate is subtropical and monsoonal, with mild, dry winters and hot, humid summers. Rainfall is high and occasionally intense during the rainstorms and typhoons. Given the steep natural terrain mantled with a layer of superficial deposits and the frequent intense rainfall, it is not surprising that landslides are a common occurrence.

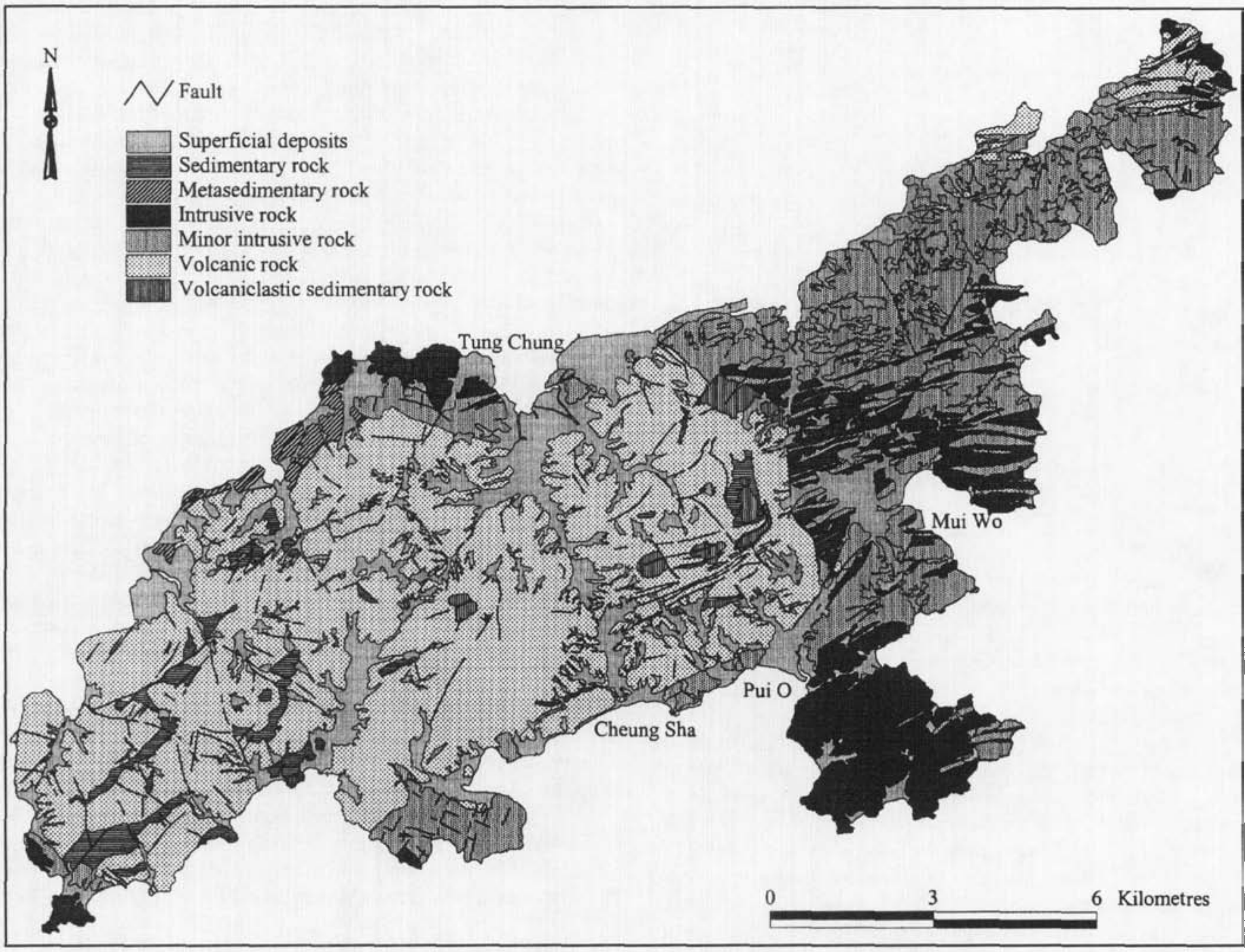
The landslide data used in this analysis were derived from the Geotechnical Engineering Office (GEO) work in which landslide locations were digitized from 23 temporal sets of 1 : 20 000 to 1 : 40 000 scale stereoscopic aerial photographs taken between 1945 and 1994 (King 1999; Evans 1998). The aerial photographs thus cover a period of 50 years and landslides up to 10 years old were visible before revegetation masked most scars. Recent landslides (Fig. 3), observed on aerial photographs as a distinctive light tone (King 1999), were extracted from the natural terrain landslide database supplied by the GEO. Thus the recent landslide data used in this analysis cover a period of about 60 years of landslide activity. The location of each identified landslide crown was recorded on the 1 : 5000 scale base map, and the centerline of any debris trail was marked with a line. Each landslide was assigned an identification number. Multiple sequential photograph sets bracketed most landslides, allowing their minimum and maximum ages to be determined. The width of each landslide scar was classified as greater or less than 20 m, and the ground slope angle across the landslide head, calculated from the distance between the steepest two adjacent contours on the 1 : 5000 scale map, was recorded. All these features have been digitized and saved in MicroStation design files (DGN) by the GEO and are available to the authors. Using the system proposed by Cruden and Varnes (1996), most of the landslides in the study area are probably debris slides, debris flows, complex debris slide–flows, or composite debris slide–flow–falls (Evans et al. 1999). The GEO carried out a systematic study of the 56 natural terrain failures in three selected areas within the study area, and factual and diagnostic reports on the investigations and observations of the landslides were given by Wong et al. (1997) and Wong et al. (1998), respectively. Field inspections of these landslides have also been carried out by the authors (Dai et al. 1999). The distributions of source (i.e., the area above the surface of rupture) length, source width, and failure depth of initial failures are shown in Fig. 4. For the landslides examined, the source lengths vary between 6 and 40 m, with a mean value of about 15 m. The source widths range from 3 to 20 m, with a mean value of about 10 m. The landslides generally have a failure depth varying between 0.5 and 2 m, with a mean value of about 1.4 m. Site inspections indicated that the failures generally occurred along the colluvium–bedrock contact and most of the landslides started as slides and were quickly converted to flows due to the abundance of water and the steep terrain below the debris sources (Wong et al. 1998; Dai et al. 1999).

Methodology

The data needed for this study were derived from existing topographic maps, superficial and bedrock geological maps, and the spatial distribution of landslides. Contour lines and drainage lines were obtained from the 1 : 20 000 scale topographic maps. Superficial and bedrock geological data were obtained from 1 : 20 000 scale geological maps developed by the GEO, Hong Kong Geological Survey. Land-cover data were derived from the Satellite Pour l'Observation de la Terra (SPOT) images using image-processing techniques. All locational, geological, and geomorphologic features provided by the different thematic maps were digitized using the GIS software Arc/Info and then transferred to ArcView for the subsequent analyses.

Statistical methods were used to relate the occurrence of a landslide to the spatial distribution of terrain variables. Spatiotemporal variations of rainfall as an indispensable dynamic variable for triggering the occurrence of a landslide were excluded from this analysis. This consideration was based on the assumption that a record of up to 60 year landslide incidences collected over many rainfall events might tend to smooth out and reduce the temporal and spatial rainfall effects, and that the recorded distribution of landslides, therefore, reflects the underlying susceptibility of the natural

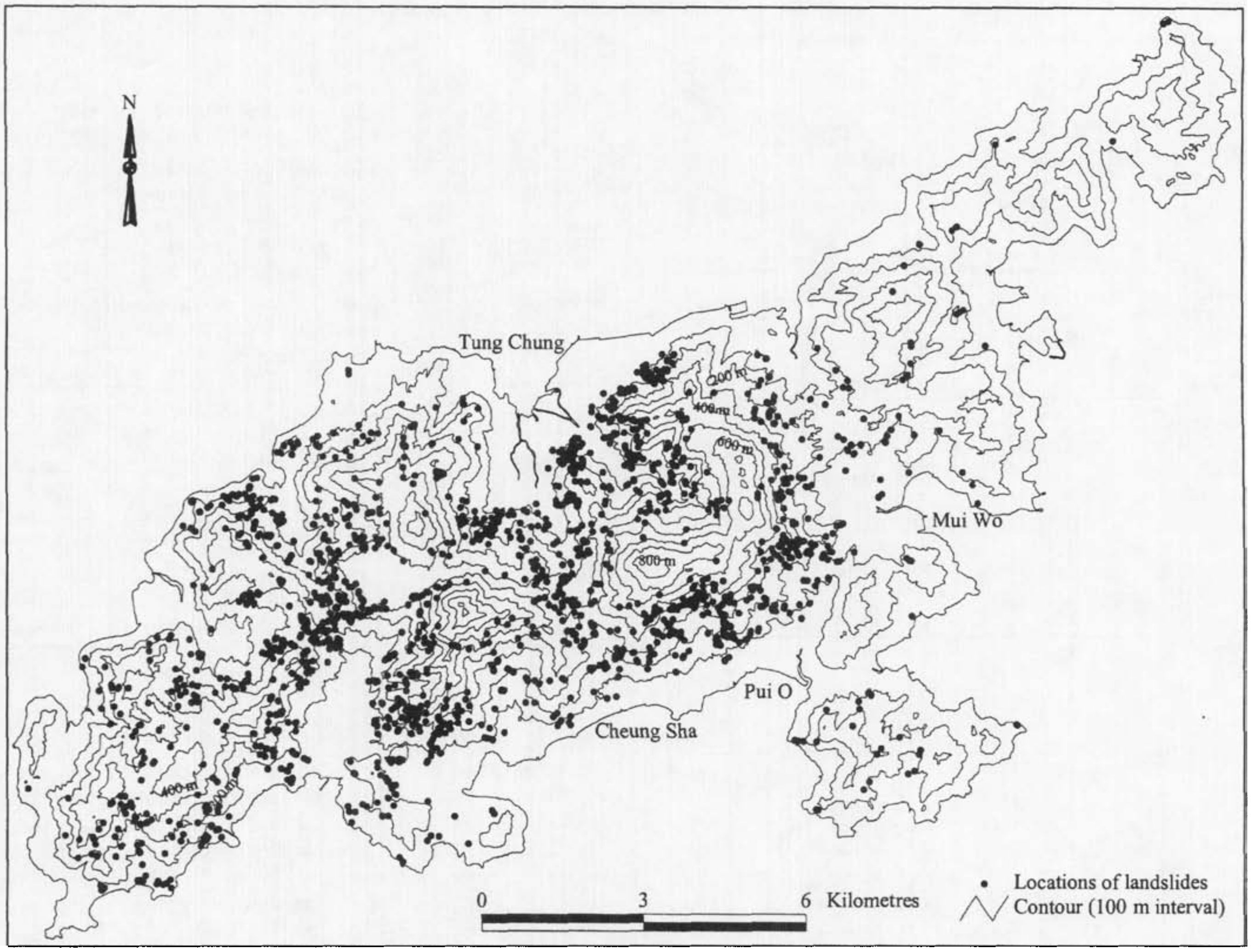
Fig. 2. Simplified geological map (data provided by the Geotechnical Engineering Office, Hong Kong).



terrain, rather than the distribution of rainstorms. Six terrain variables including lithology, slope gradient, slope aspect, elevation, distance to drainage line, and land cover were considered to have a strong influence on landslide susceptibility in the study area. Lithology exerts a fundamental control on the geomorphology of a landscape. The nature and rate of the geomorphological processes, including landslides, are partially dependent on the lithology and weathering characteristics of the underlying materials. Slope gradient is an essential component of slope stability analysis. As slope gradient increases, the level of gravitation-induced shear stress in the colluvium or residual soils increases as well. Gentle hillslopes are expected to have a low frequency of landslides because of the generally lower shear stresses associated with low gradients. The aspect of a slope has the potential to influence its physical properties and its susceptibility to failure. The processes that may be operating include exposure to sunlight, drying winds, and, possibly, rainfall (Evans et al. 1999). Figure 3 shows that the topographical variable of elevation might be associated with landslides. Intense gully erosion occurs in the study area, and field checking indicates that the proximity to drainage line (in the form of a natural gully channel) may be an important factor controlling slope failure. Land cover, especially of a woody type with strong and large root systems, helps to improve the stability of slopes by providing both hydrological and mechanical effects that generally are beneficial to stability (Gray and Leiser 1982; Greenway 1987). Franks (1998) examined natural terrain landslides on North Lantau Island and concluded that a sparsely vegetated slope is most susceptible to failure. Elevation data were obtained from the digital elevation model (DEM) derived from the 1 : 20 000 scale digital contour lines of the area. Two data layers were derived from these elevation data, namely, slope aspect and slope gradient. The data layer containing drainage lines was converted to a grid in which the cell values indicate the distance to the drainage line.

The vector datasets were then rasterized into 20 m × 20 m grid cells for subsequent analyses. This size of the grid cells was based on the scale of the topographic map used and the size of the landslides, most of which were less than 20 m in width. Each variable was divided into several categories (Table 1), and the univariate statistical method was adopted to quantify the relationship between landslide frequency and the categories of each variable. Although the conventional statistical method could give a satisfactory combination of variables,

Fig. 3. Map showing topographical features and locations of landslides (landslide data provided by the Geotechnical Engineering Office, Hong Kong).



it has a serious drawback because it uses the assumption of conditional independence (Van Westen et al. 1997). This means that different variable maps are independent with respect to landslide susceptibility. This assumption is, however, largely invalid, leading to unrealistic susceptibility values. This drawback can be avoided with the use of multiple correspondence analysis, which can detect the uncorrelated principal axes that are linear combinations of these terrain variables. Logistic regression was conducted to obtain a predictive model, and this model was then used to produce a relative susceptibility map within the GIS environment. The main steps in the procedure are shown in Fig. 5. The use of this method has the following advantages: (*i*) the results are reproducible because the mathematical operations are defined; (*ii*) the results are easy to interpret because each parameter can be evaluated separately; and (*iii*) the assumption of conditional independence usually adopted in the conventional statistical model, which generally deviates from reality, can be avoided.

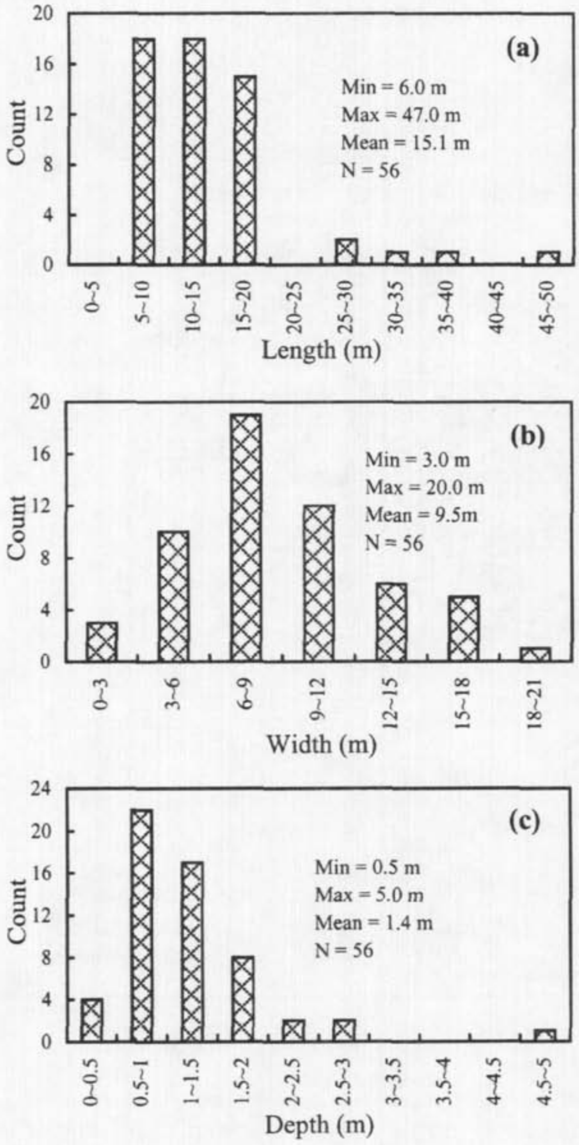
Terrain analysis and landslide distribution

Descriptive statistics for the categorized variables were compiled to investigate the relationship between the occurrence of landslides in the past and the terrain variables within the study area. Such a univariate analysis can be used to validate or verify expectations regarding the role of individual variables in the occurrence of landslides. The digital map of the distribution of recent landslides was overlain on the aforementioned raster data layers of terrain variables to automatically extract the categories of terrain variables for the landslide sites using the GIS, and the descriptive univariate statistical relations were then produced.

Lithology

The landslide frequency in each lithological category is shown in Fig. 6*a*. There are three geological categories with relatively high landslide frequency: trachydacite, dacite, and rhyolite lava (TDR); sedimentary rock (SR); and metasedimentary rock (MSR). The TDR category had the highest frequency. As noted previously, the available evidence tends to suggest that a thin surficial layer of colluvium may have played an important role in the majority of landslides. However, colluvial deposits that are less than approximately 2 m thick are not identified on the 1 : 20 000 scale geological

Fig. 4. Histograms showing characteristics of initial landslides: (*a*) source length, (*b*) source width, and (*c*) failure depth (data from Wong et al. 1997).



maps (Evans et al. 1999). Hence landslides in thin colluvium might have been recorded as occurring within the underlying geological group. This is not considered by Evans et al. (1997) to be a serious problem because the properties of the thin colluvial layers depend on the bedrock geology from which they are derived. Immediately downslope from geological group boundaries, unmapped colluvial deposits may have been partly derived from the upper geological group rather than from the underlying unit. However, the proportion of landslides affected by this situation will be very small (Evans et al. 1999).

Structural geology information is also available from the digital geological maps. However, qualitative examination of spatial distributions suggests that the correlation between landslides and mapped linear structural features at the 1 : 20 000 scale is not good, and the structural information is thus excluded from this study.

Table 1. Terrain variables and categories used for analysis.

Variable	Categories
Lithology	1, superficial deposits (Q); 2, sedimentary rock (SR); 3, metasedimentary rock (MSR); 4, intrusive rock (IR); 5, minor intrusive rock (MIR); 6, ash tuff, tuffite, tuff breccia, and eutaxite (BCT); 7, trachydacite, dacite, and rhyolite lava (TDR); 8, volcaniclastic sedimentary rock (VSR)
Slope gradient (°)	1, 0–15; 2, 15–20; 3, 20–25; 4, 25–30; 5, 30–35; 6, 35–40; 7, ≥40
Slope aspect	1, flat; 2, north; 3, northeast; 4, east; 5, southeast; 6, south; 7, southwest; 8, west; 9, northwest
Elevation (m)	1, 0–100; 2, 100–200; 3, 200–300; 4, 300–400; 5, 400–500; 6, 500–600; 7, >600
Land cover	1, developed land (DL); 2, forested land (FL); 3, shrub – forested land (SFL); 4, densely grassed land (DGL); 5, moderately grassed land (MGL); 6, sparsely grassed land (SGL)
Distance to drainage line (m)	1, <50; 2, 50–100; 3, 100–150; 4, 150–200; 5, 200–250; 6, 250–300; 7, >300

Slope gradient

Examination of the distribution of landslide frequency with the corresponding slope gradient categories, measured at the 1 : 20 000 scale, shows an increase in the frequency of landslides with an increase in the slope gradient until the maximum frequency is reached in the 35–40° category, followed by a gradual decrease in the ≥40° category (Fig. 6*b*). This is because the slope-forming material of the terrain with a gradient exceeding 40° is composed of weathered rock that is not overlain by colluvium and whose strength is much higher. In contrast, the moderately steep terrain is often covered by a thin layer of colluvium, which is more susceptible to rainfall-induced failure.

Slope aspect

To investigate the relative relationship between the occurrence of landslides and slope aspect, the DEM was used to calculate the aspect of a slope within the study area. The distribution of aspect among the mapped landslides is shown in Fig. 6*c*, which shows that landslides with south-facing aspects are relatively more common, indicating that landslides in natural terrain are more common on south-facing slopes.

Elevation

The relationship between landslide frequency and elevation is shown in Fig. 6*d*. At very high elevations there are mountain summits that usually consist of weathered rocks whose shear strength is very high. At intermediate elevations, however, slopes tend to be covered by a thin layer of colluvium, which is more prone to landsliding. At very low elevations, the landslide risk is low because the terrain itself is gentle and is covered with a thick layer of colluvium and

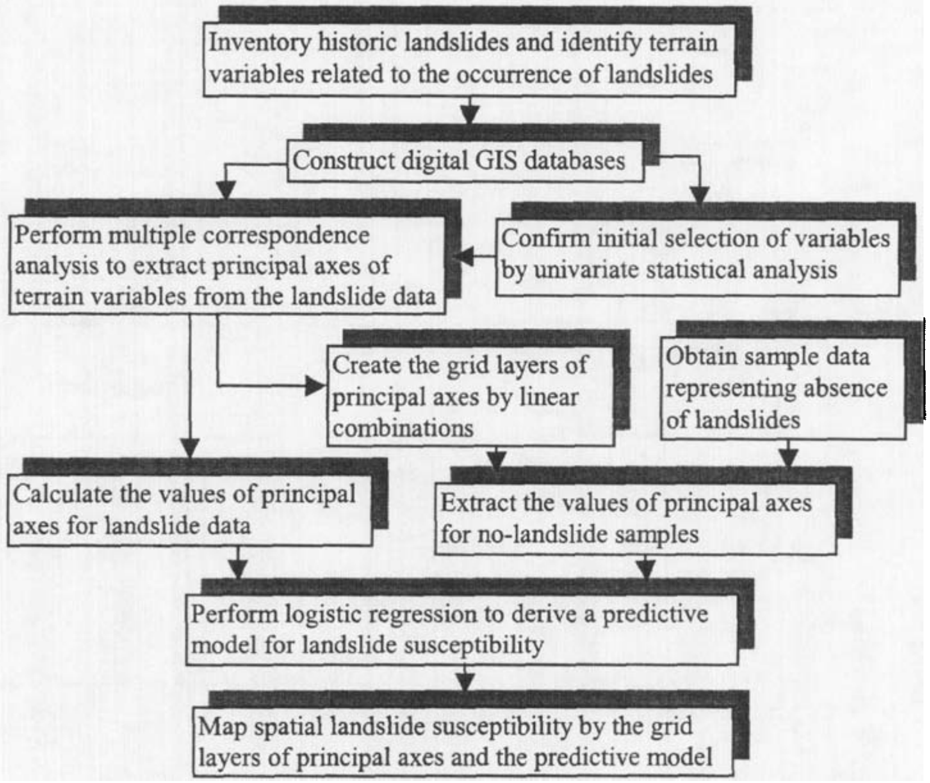

Inventory historic landslides and identify terrain variables related to the occurrence of landslides
Construct digital GIS databases
Perform multiple correspondence analysis to extract principal axes of terrain variables from the landslide data
Confirm initial selection of variables by univariate statistical analysis
Create the grid layers of principal axes by linear combinations
Obtain sample data representing absence of landslides
Calculate the values of principal axes for landslide data
Extract the values of principal axes for no-landslide samples
Perform logistic regression to derive a predictive model for landslide susceptibility
Map spatial landslide susceptibility by the grid layers of principal axes and the predictive model

Fig. 5. Procedures for mapping of landslide susceptibility adopted in this study.

(or) residual soils, and a higher perched water table will be required to initiate slope failure.

Land cover

The correlation between land cover and landslide frequency is shown in Fig. 6*e*, which shows that the landslide frequency is relatively low on "sparsely grassed land" (SGL) and highest on "densely grassed land" (DGL). This is because sparsely grassed land is composed of weathered rock not overlain by colluvium or residual soil. In contrast, the densely grassed land is often covered by a thin layer of colluvium underlain by weathered bedrock which is susceptible to landslides. It should be noted, however, that land-cover data are considered to be estimates only, because of increased development of coastal flat-lying lands with time and possible temporal change in land-cover categories over the past several decades.

Distance to drainage line

The relationship between landslide frequency and distance to drainage line is shown in Fig. 6*f*, which shows that landslide frequency decreases as the distance to drainage line increases. This can be attributed to the fact that the elevated groundwater level during storms and terrain modification caused by gully erosion may influence the initiation of landslides.

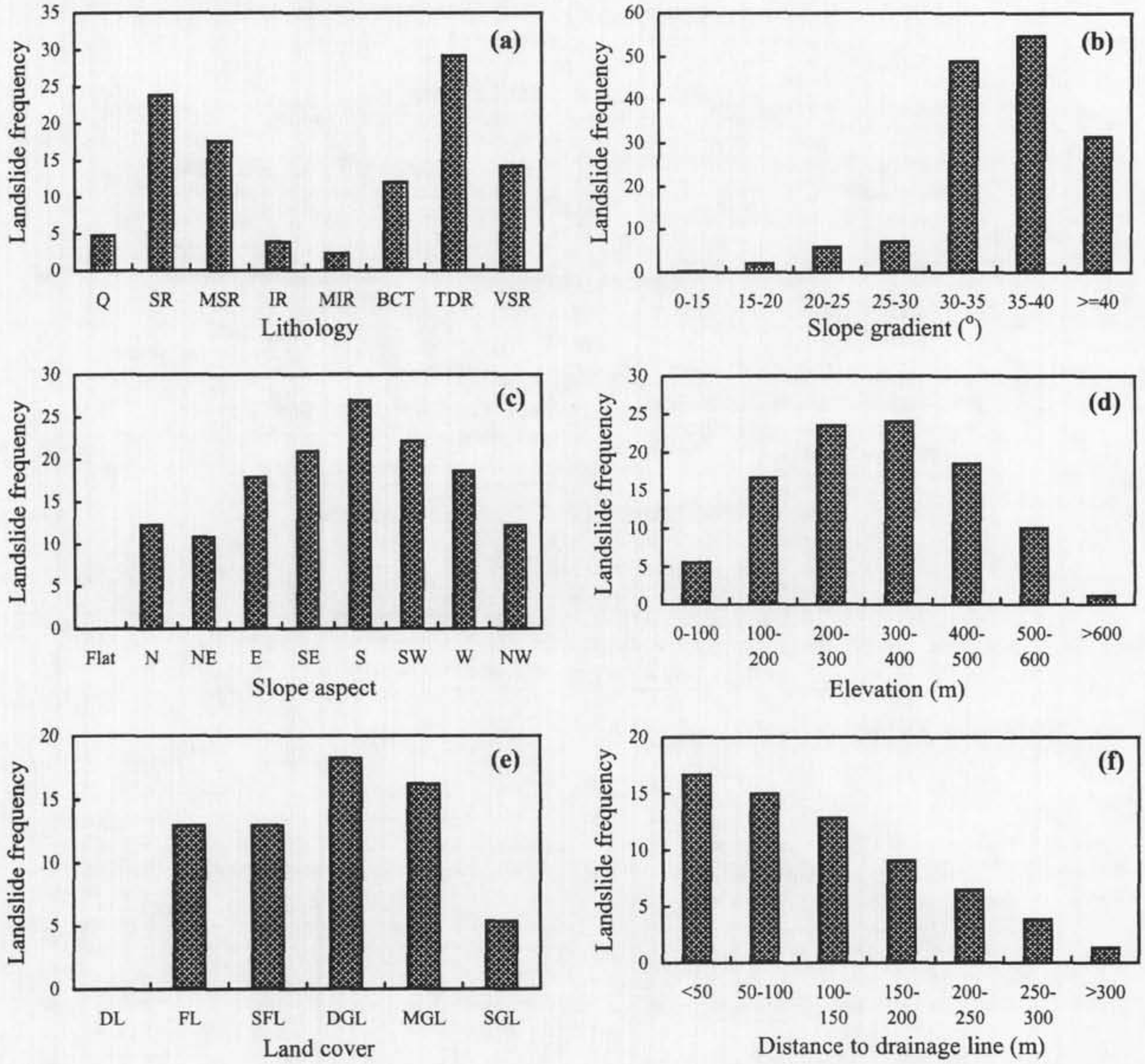
Mapping of landslide susceptibility

Logistic regression was conducted using the Statistical Analysis System (SAS), a software package for the manipulation and statistical analysis of data (Everitt and Der 1996), to predict landslide susceptibility as a function of terrain variables for 20 m × 20 m cells. However, the terrain variables used for describing landslide susceptibility were intrinsically interdependent. Thus a substantial part of the information for one or more of these variables may be redundant and the conclusions drawn from the regression analysis may be ambiguous (Glantz and Slinker 1990). Thus, multiple correspondence analysis (Greenacre 1984) was used to overcome the problem of the unavoidable intercorrelatedness among these data by extracting the principal axes. These principal axes contain the same information as the original parameters and are defined in such a way that they are mutually uncorrelated and there is no redundant information between them.

Multiple correspondence analysis (MCA)

Correspondence analysis (CA) is a weighted principal component analysis of simple two-way and multi-way tables containing some measures of correspondence between the rows and columns. It is one of the eigenvector-based techniques that assume a unimodal, rather than linear relationship among the variables. Correspondence analysis can be classified into simple and multiple correspondence analyses (MCA). Simple correspondence analysis is carried out on an indicator matrix with cases as rows and categories of variables as columns. In contrast, the MCA is not performed on an indicator matrix, which potentially may be very large if there are many cases, but rather on the inner product of this matrix, called the Burt table. A Burt table is a partitioned symmetric matrix containing all pairs of cross-tabulations among a set of categorical variables. Each diagonal partition is a diagonal matrix containing marginal frequencies (a

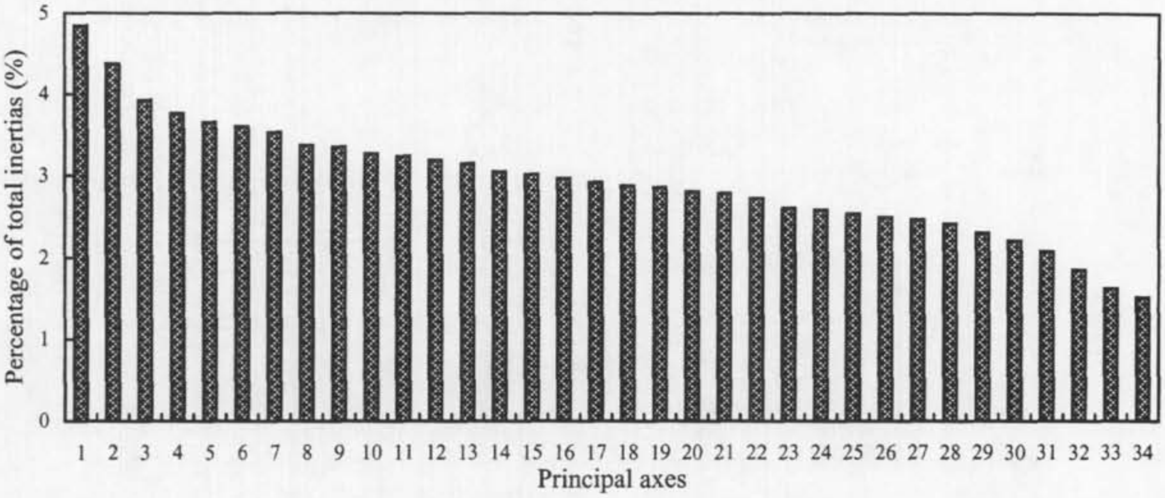
Fig. 6. Relationships of landslide frequency (number of landslides per km^2) with (*a*) lithology, (*b*) slope gradient, (*c*) slope aspect, (*d*) elevation, (*e*) land cover, and (*f*) distance to drainage line (symbols as in Table 1).



cross-tabulation of a variable with itself). Each off-diagonal partition is an ordinary contingency table. Each contingency table above the diagonal has a transposed counterpart below the diagonal (SAS Institute Inc. 1990). The results of the MCA performed on the Burt table provide information which is similar in nature to that produced by the factor analysis techniques, and they allow exploration of the structure of categorical variables in the table. The principal axes are extracted so as to give the maximum variance between variables. Thus, the extraction of principal axes is similar to the extraction of principal components in a factor analysis. In essence, correspondence analysis creates a series of orthogonal axes to identify trends that explain the data variation, with each subsequent axis explaining a decreasing amount of the variation (Benzecri 1992). A comprehensive description of this method, computational details, and its applications is given in the classic texts by Greenacre (1984) and Benzecri (1992).

As noted previously, the terrain variables considered relevant to the occurrence of landslides have been extracted for all recent landslides. This result from the overlays is then transferred to a matrix of terrain variables. The same categorization scheme as that used previously to study the relation of landslide frequency with the categories of terrain variables is adopted herein for consistency. However, for the terrain variables of land cover and slope aspect, the forested land and shrub – forested land categories and the north-facing and northwest-facing categories have been incorporated into single categories because they have nearly the same landslide frequencies. An indicator table of terrain variables for the occurrence of landslides is obtained. In this table, there are 1975 landslide cases, and at the *i*th case ($i = 1, 2, \ldots, m$) we have $A_{i1}, A_{i2}, \ldots, A_{ik}$, which represent the numeric values of the *k* terrain variables. Because A_{ij} represents a category of the *i*th landslide case in the data layer of the *j*th terrain variable, the numeric value A_{ij} cannot be used directly in the MCA. A commonly used technique is to generate a binary variable for each terrain variable category to

Fig. 7. Percentage of the total inertias for individual principal axes.



indicate the presence or absence of that category at each landslide case. Suppose that we have l categories in the jth terrain variable ($j = 1, 2, \ldots, k$). Then, at the ith case we generate l binary variables for the jth terrain variable, B_{ij1}, B_{ij2}, …, B_{ijl}, where one of B_{ij1}, B_{ij2}, …, B_{ijl} is equal to 1 and all the others are equal to 0. This procedure is repeated for all the other terrain variables, and a binary indicator table is obtained. The Burt table of this binary indicator table is then created by each binary variable being tabulated against itself and against all other binary variables.

Multiple correspondence analysis is then conducted on the indicator table using SAS (Everitt and Der 1996). The SAS has the function of creating the Burt table from an indicator table. Figure 7 shows the percentages of the total inertias for individual principal axes. The inertia is analogous to the variance in principal component analysis. The percentage of principal inertia for the first axis is very low. This is not unique to this analysis and is in agreement with the work of Micheloud (1997), which indicates that in the MCA the percentage of inertia for the first axis is very low. It is common to keep the first few principal axes which account for most of the total variability for subsequent analysis. One limitation of this approach is that the MCA only focuses on the variability in the inputs (here, terrain variables) and ignores the relationship with the output(s) (here, landslide susceptibility). A low-order principal axis which only accounts for a small proportion of the total variability in the inputs may be significant in modeling an output (Martin and Morris 1999). To avoid this problem in the analysis, all 34 principal axes obtained from the MCA are used for the subsequent logistic regression analysis.

Based on the coefficients obtained from the MCA, calculation of all principal axes from a linear combination of the original terrain variables is readily implemented in the ArcView GIS. Each principal axis constitutes a grid layer in the GIS. All these axis layers together define a new uncorrelated space.

Logistic regression

Landslide susceptibility is to be predicted using logistic regression, one of a family of generalized linear models that are well suited to analyzing a presence–absence dependent variable. Logistic regression uses a linear combination of independent variables to explain the variance in a dependent variable having only two states. Here the dependent variable was the absence or presence of a landslide, and the independent variables were the principal axes. Each sample can be represented through a binary variable Y, which indicates whether a landslide occurred ($Y = 1$) or did not occur ($Y = 0$), and n independent variables (i.e., principal axes), X_1, X_2, …, X_n, which include all the principal axes. The task is to use the available m samples ($X_{11}, X_{12}, \ldots, X_{1n}; Y_1$), …, ($X_{m1}, X_{m2}, \ldots, X_{mn}; Y_m$) to express the probability of landslide occurrence $P(Y = 1)$ as a function of $X_1, X_2, \ldots, X_n$. Since Y is an indicator variable, it follows that, for any given $X_1, X_2, \ldots, X_n$, the probability that $Y = 1$ is also the expected value of Y given $X_1, X_2, \ldots, X_n$, i.e., $P(Y = 1)$, is the regression against $X_1, X_2, \ldots, X_n$. The fact that the regression of Y against $X_1, X_2, \ldots, X_n$ has the meaning of probability implies that $P(Y = 1)$ must lie between 0 and 1. This constraint can be obtained by replacing the probability that $Y = 1$ with the odds that $Y = 1$ (e.g., Menard 1995):

$$[1] \qquad \textbf{Odds}\,(Y = 1) = \frac{P(Y = 1)}{1 - P(Y = 1)}$$

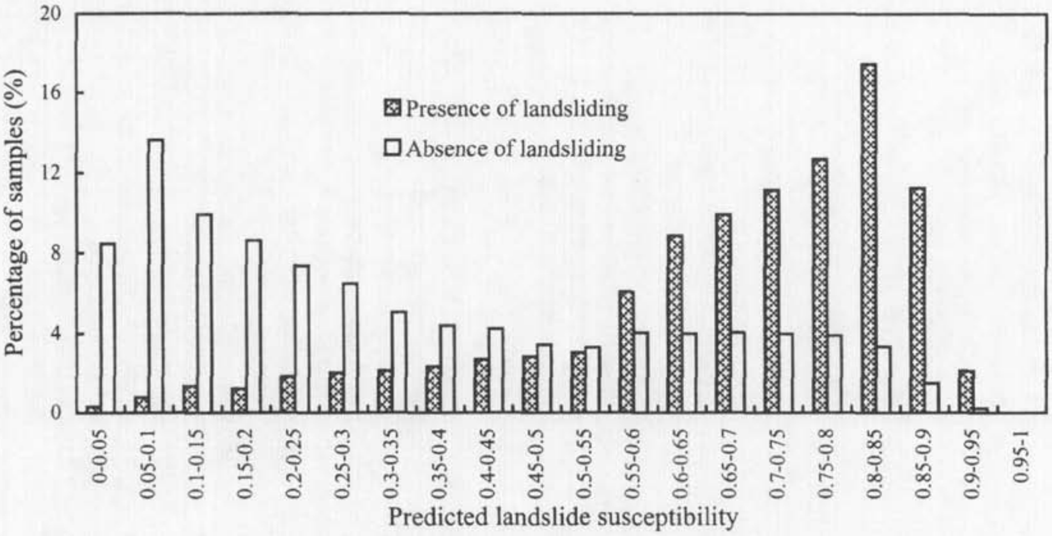
The natural logarithm of the odds, called the logit of Y, produces a variable that varies monotonically from negative infinity to positive infinity. The logit of Y, logit(Y), becomes negative and increasingly large in absolute value as the odds decrease from 1 to 0 and becomes increasingly large in the positive direction as the odds increase from 1 to infinity. If we use the natural logarithm of the odds that $Y = 1$ as the dependent variable, we no longer face the problem that the estimated probability may exceed the maximum or minimum possible values for the probability. The equation for the dependent variable and the independent variables then becomes (e.g., Menard 1995)

$$[2] \qquad \mathrm{logit}(Y) = \alpha + \beta_1 X_1 + \ldots + \beta_n X_n$$

where β_i ($i = 1,\ldots, n$) is the coefficient estimated from the sample data, and α is the intercept.

We can convert logit(Y) back to the odds and then convert the odds back to $P(Y = 1)$. This produces the following ex-

Fig. 8. Histogram of predicted landslide susceptibility.



pression for the probability of landslide occurrence in terms of the variables X_1, ..., X_n:

$$[3] \qquad P(Y=1)=\frac{1}{1+e^{-(\alpha+\beta_1 X_1+\ldots+\beta_n X_n)}}$$

In a strict sense, however, $P(Y = 1)$ is not a probability because the dynamic variables, such as rainfall, triggering landslides are not accounted for. It may be more appropriate to term it hereafter as landslide susceptibility based on the terrain variables.

The parameters of the logistic regression model are estimated using the maximum-likelihood method. In other words, those coefficients which make the observed results most "likely" are selected. Since the relationship between the independent predictor variables, i.e., the principal axes, and the landslide susceptibility is nonlinear in the logistic regression model, an iterative algorithm is necessary for parameter estimation.

Two sets of sample data representing both absence and presence of landslide must be provided to fit the logistic regression model. The way in which these data are obtained will affect both the nature of the regression relation and the nature and accuracy of the resulting estimates (Atkinson and Massari 1998). In this analysis, the dataset of landslide inventory is an indispensable data source representative of samples of landslide presence. All grid cells of the 1975 landslides studied were thus used to obtain the values of the principal axes. To eliminate bias in the sampling process, an equal number of cells were chosen from the no-landslide area as samples representing the absence of a landslide. These grid cells were obtained using systematic sampling, i.e., a spatially uniform sampling scheme, but excluding a 40 m buffer zone for all landslides. For these cells, the values of all principal axes were extracted automatically from the existing data layers of the principal axes. Each sample cell has its respective binary value on the presence or absence of landslide coded as 1 or 0, respectively, as well as information on principal axes. The training data were then used to input to the logistic regression algorithm within SAS to obtain the coefficients for the logistic regression model.

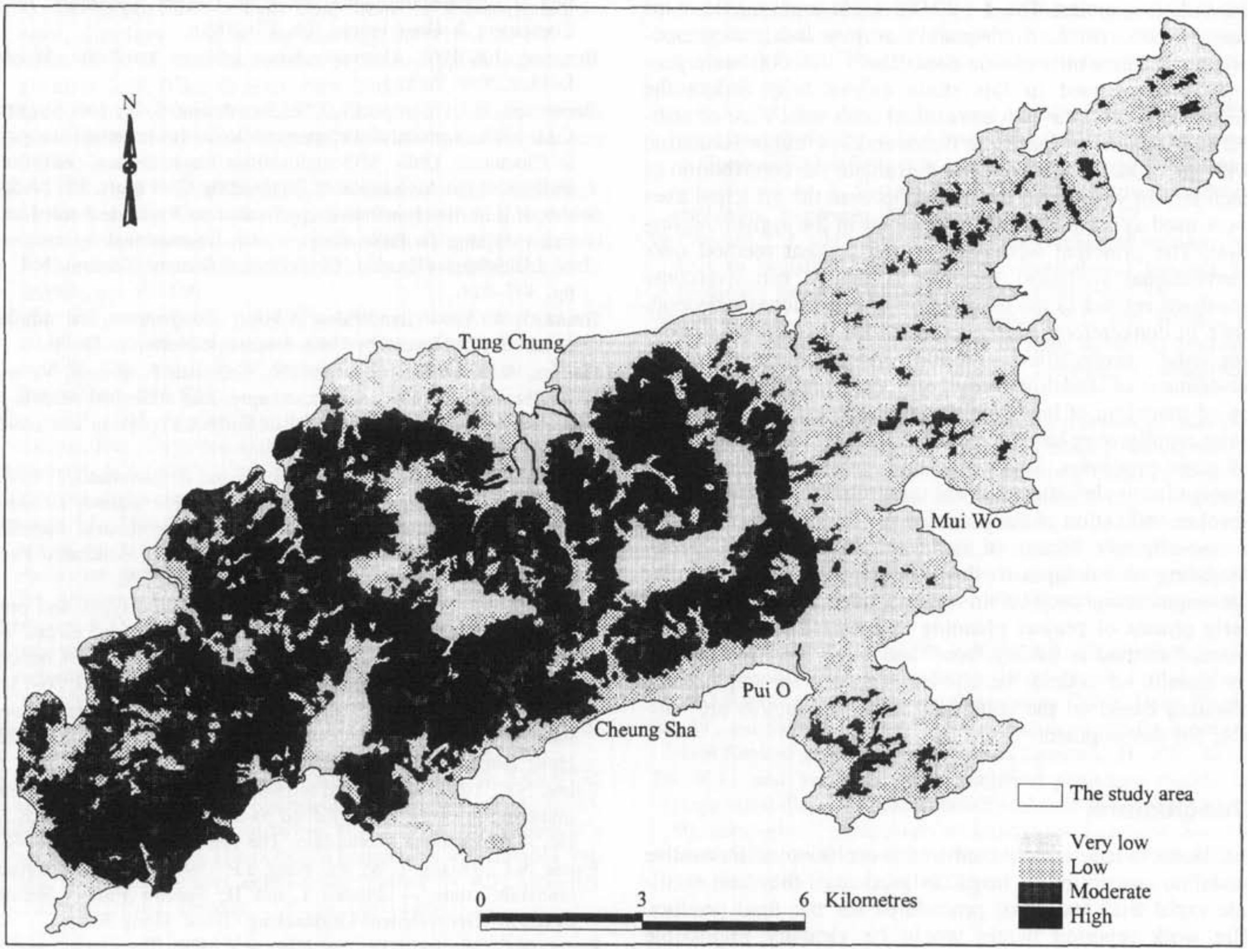
All the predictor variables, i.e., the principal axes, are then subjected to a forward stepwise procedure to generate a more parsimonious logistic regression model. At each step, principal axes are evaluated for entry into the model one by one if they contribute sufficiently to the regression equation. All the principal axes for which the attained statistical significance (p) is smaller than 0.1 were entered into the model, following the procedure available in SAS. The model developed was checked for plausibility, and variables were then entered into, or removed from, the model to improve its goodness of fit. The significance of the logistic model was ascertained from the likelihood ratio statistics by comparing the deviance for the model of interest against the deviance for the model fitted only to the intercept (Hosmer and Lemeshow 1989). Classification tables were also used for assessing the goodness of fit (Hosmer and Lemeshow 1989). Lastly, the coefficients for the logistic regression are obtained.

Figure 8 is a histogram of the predicted landslide susceptibility for the samples used in this analysis. Theoretically, if we have a model that successfully distinguishes the two groups based on a classification cutoff value of 0.5, the cases for which a landslide has occurred should be to the right of 0.5, and the cases for which a landslide has not occurred should be to the left of 0.5. The more the two groups cluster at their respective ends of the plot, the better the model is at predicting landslide susceptibility. Figure 8 shows that the model produced a concordance rate of 77.1%, and 82.5% of the actual landslides were correctly classified with the use of 0.5 as a classification cutoff value. By examining the histogram of predicted susceptibilities in Fig. 8, one can determine the classification rule that should be adopted when applying the model to each cell in the study area.

Susceptibility mapping

The logistic regression model obtained above can be readily implemented in a GIS. The model is implemented by building a single formula where each coefficient multiplies its related predictor principal axes. The result of the calculations is then subjected to inverse logistic transformation to obtain the susceptibility values between 0 and 1 at every cell

Fig. 9. Map showing relative landslide susceptibility.



of the GIS grid. The estimated susceptibilities thus obtained from the regression model are converted to the range between 0 and 1.

A general description of the spatial landslide susceptibility on the scale adopted in this analysis is to classify the range of the landslide susceptibility into several relative descriptive categories, such as high, moderate, and low. This is also convenient for the presentation of landslide susceptibility maps. In this analysis, spatial landslide susceptibility is generalized into four categories: (*i*) very low (0–0.2), (*ii*) low (0.2–0.35), (*iii*) moderate (0.35–0.55), and (*iv*) high (>0.55). These ranges of the individual categories were derived by moving the cutoff point by an increment of 0.05 along the [0, 1] susceptibility interval to allow estimates of optimal cutoff points to be made by identifying the values for which most successes are correctly classified, while minimizing the number of failures, based on the histogram of the estimated landslide susceptibility shown in Fig. 8. The final relative landslide susceptibility map is shown in Fig. 9.

Figure 9 shows that those zones classified as having "very low" susceptibility, which account for 38.5% of the study area, are satisfactorily distributed in clusters on the coastal lowland and on the top of high mountains which are characterized by relatively gentle gradients. All these sites are highly stable and are not favorable to the development of landslides. Zones of "low" susceptibility, covering 17.5% of the total land, are relatively dispersed in their spatial distribution, and hence the chance for landslides to develop is small. Zones of "moderate" susceptibility cover 13.7% of the total area, and are mainly found in lower sections of slopes and ridges. The "high" susceptibility category occupies 30.3% of the study area and exhibits a strongly clustered pattern of spatial distribution. This category is characterized by relatively high elevations and steeper terrain. Most of the locations of the identified landslides actually fall within this category. Thus, it appears that the results of this GIS-based statistical modeling are generally satisfactory.

Discrepancies in the classifications of landslides between the predictive and the actual landslides can be attributed to a variety of factors. The possible problem with the model may have been due to the limitations and assumptions inherent in the statistical techniques. The nature and details of the landslides are not necessarily well represented by the model. There may be a potential bias in the calculation of the principal axes because only those data on landslide occurrence were used in the MCA. This can be overcome by adding the sample data representing the absence of landslides. Data quality and resolution may also be factors contributing to these discrepancies. The 1 : 20 000 scale topographic maps may be too coarse to adequately portray local microtopographic features at landslide sites. The 1 : 20 000 scale geological maps used in this study cannot fully reflect the distribution of colluvium or residual soils which are of critical significance to landslide occurrence. A major limitation with the model is that we cannot evaluate the contribution of each terrain variable to the model because the principal axes were used as the independent variables in the logistic regression. The principal advantage of the present method over conventional statistical methods is that it can overcome problems related to the assumption of conditional independence in conventional statistical methods. Compared to conventional geotechnical ground investigations for the assessment of landslide hazard and slope stability, this GIS-based modeling of landslide susceptibility can be carried out quite rapidly over large areas to provide an early impression of slope processes, landforms, and stability. Although the susceptibility classifications are only relative and provide no absolute indication of the potential for failure, they represent a cost-effective means of rapid terrain appraisal, thereby providing useful input to the assessment of slope stability for engineering projects in mountainous areas during the early phases of project planning and decision-making. The zones classified as having "low" and "very low" susceptibility should, of course, be chosen for sites for engineering planning based on the condition that other factors are suitable for development.

Conclusions

GIS tools have contributed to the evolution of innovative landslide susceptibility maps. In particular, they can facilitate rapid trial and error procedures for the final product. The work reported herein would be virtually impossible without the aid of GIS. The procedures developed in this study consist of (*i*) developing a data base of landslides and terrain variables, (*ii*) relating the landslide data base to a set of terrain variables through a GIS, (*iii*) using multiple correspondence analysis to extract the principal axes, (*iv*) using the principal axes to formulate the logistic regression model, and (*v*) mapping the spatial landslide susceptibility from the logistic regression model by means of a GIS. The results obtained in this study indicate that this model is useful and suitable for the scale adopted in the study.

Acknowledgements

Funding for this research was provided by the Research Grants Council of Hong Kong and the Hong Kong Jockey Club Research and Information Centre for Landslip Prevention and Land Development, at the University of Hong Kong. The authors wish to express their sincere appreciation for the generous support received from these two organizations.

References

Anbalagan, D. 1992. Landslide hazard evaluation and zonation mapping in mountainous terrain. Engineering Geology, **32**: 269–277.

Atkinson, P.M., and Massari, R. 1998. Generalized linear modelling of landslide susceptibility in the Central Apennines, Italy. Computers & Geosciences, **24**: 373–385.

Benzecri, J.P. 1992. Correspondence analysis handbook. Marcel Dekker, New York.

Bernknopf, R.L., Campbell, R.H., Brookshire, D.S., and Shapiro, C.D. 1988. A probabilistic approach to landslide hazard mapping in Cincinnati, Ohio, with applications for economic evaluation. Bulletin of the Association of Engineering Geologists, **25**: 39–56.

Brabb, E.E. 1984. Innovative approaches to landslide hazard and risk mapping. *In* Proceedings of 4th International Symposium on Landslides, Canadian Geotechnical Society, Toronto, Vol. 1, pp. 307–324.

Brand, E.W. 1994. Landslides in Hong Kong during the rainfall event of 4–5 November 1993. Landslide News, **8**: 35–36.

Carrara, A., Cardinali, M., Detti, R., Guzzetti, F., Pasqui, V., and Reichenbach, P. 1991. GIS techniques and statistical models in evaluating landslide hazard. Earth Surface Processes and Landforms, **16**: 427–445.

Carrara, A., Cardinali, M., Guzzetti, F., and Reichenbach, P. 1995. GIS-based techniques for mapping landslide hazard. *In* Geographical information systems in assessing natural hazards. *Edited by* A. Carrara and F. Guzzetti. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 135–176.

Cruden, D.M., and Varnes, D.J. 1996. Landslide types and processes. *In* Landslides: investigation and mitigation. *Edited by* A.K. Turner and R.L. Schuster. National Research Council, Transportation Research Board, Special Report 247, pp. 36–75.

Dai, F.C., Lee, C.F., and Wang, S.J. 1999. Analysis of rainstorm-induced slide-debris flows on natural terrain of Lantau Island, Hong Kong. Engineering Geology, **51**: 279–290.

Evans, N.C. 1998. The natural terrain landslide study. *In* Slope engineering in Hong Kong. *Edited by* S. Li, J.N. Kay, and K.K.S. Ho. A.A. Balkema, Rotterdam, The Netherlands, pp. 137–144.

Evans, N.C., Huang, S.W., and King, J.P. 1997. The natural terrain landslide study — Phases I and II. Special Project Report SPR5/97, Geotechnical Engineering Office, Hong Kong.

Evans, N.C., Huang, S.W., and King, J.P. 1999. The natural terrain landslide study — Phases I and II. GEO Report 73, Geotechnical Engineering Office, Hong Kong.

Everitt, B.S., and Der, G. 1996. A handbook of statistical analyses using SAS. Chapman & Hall, London, U.K.

Fernandez, C.I., Castillo, T.F.D., Handouni, R.E., and Montero, J.C. 1999. Verification of landslide susceptibility mapping: a case study. Earth Surface Processes and Landforms, **24**: 537–544.

Franks, C.A.M. 1998. Characteristics of some rainfall-induced landslides on natural slopes, Lantau Island, Hong Kong. Quarterly Journal of Engineering Geology, **32**: 247–259.

Geotechnical Control Office. 1988*a*. Geotechnical area studies programme — North Lantau. GASP VI, Geotechnical Control Office, Hong Kong Government, Hong Kong.

Geotechnical Control Office. 1988*b*. Geotechnical area studies programme — South Lantau. GASP XI, Geotechnical Control Office, Hong Kong Government, Hong Kong.

Glantz, S.A., and Slinker, B.K. 1990. Primer of applied regression and analysis of variance. McGraw-Hill, New York.

Gray, D.H., and Leiser, A.T. 1982. Biotechnical slope protection and erosion control. Van Nostrand Reinhold, New York.

Greenacre, M.J. 1984. Theory and applications of correspondence analysis. Academic Press Inc. (London) Ltd., London, U.K.

Greenway, D.R. 1987. Vegetation and slope stability. *In* Slope stability. *Edited by* M.G. Anderson and K.S. Richards. Wiley, New York, pp. 187–230.

Gupta, R.P., and Joshi, B.C. 1989. Landslide hazard zoning using the GIS approach — a case study from the Ramganga catchment, Himalayas. Engineering Geology, **28**: 119–131.

Hosmer, D.W., Jr., and Lemeshow, S. 1989. Applied logistic regression. John Wiley & Sons, New York.

King, J.P. 1999. Natural terrain landslide study: natural terrain landslide inventory. GEO Report 74, Geotechnical Engineering Office, Hong Kong.

Mark, R.K., and Ellen, S.D. 1995. Statistical and simulation models for mapping debris flow hazard. *In* Geographical information systems in assessing natural hazards. *Edited by* A. Carrara and F. Guzzetti. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 93–106.

Martin, E.B., and Morris, A.J. 1999. Artificial neural networks and multivariate statistics. *In* Statistics and neural networks. *Edited by* J.W. Kay and D.M. Titterington. Oxford University Press, London, pp. 195–258.

Menard, S. 1995. Applied logistic regression analysis. Sage Publications, Inc., Thousand Oaks, Calif.

Micheloud, F.X. 1997. <http://www.micheloud.com/ FXM/COR/>. Correspondence analysis.

Naranjo, J.L., Van Westen, C.J., and Soeters, R. 1994. Evaluating the use of training areas in bivariate statistical landslide hazard analysis: a case study in Colombia. Journal of the International Institute for Aerospace Survey and Earth Sciences, **1994-3**: 292–300.

Niemann, K.O., and Howes, D.E. 1991. Applicability of digital terrain models for slope stability assessment. Journal of the International Institute for Aerospace Survey and Earth Sciences, **1991-3**: 127–137.

Pachauri, A.K., and Pant, M. 1992. Landslide hazard mapping based on geological attributes. Engineering Geology, **32**: 81–100.

SAS Institute Inc. 1990. The CORRESP procedure. *In* SAS/STAT user's guide. Vol. 2. SAS Institute Inc., Cary, N.C., pp. 613–675.

Siddle, H.J., Jones, D.B., and Payne, H.R. 1991. Development of a methodology for landslip potential mapping in the Rhondda Valley. *In* Slope stability engineering. *Edited by* R.J. Chandler. Thomas Telford, London, U.K., pp. 137–142.

Soeters, R., and Van Westen, C.J. 1996. Slope instability recognition, analysis, and zonation. *In* Landslides: investigation and mitigation. *Edited by* A.K. Turner and R.L. Schuster. National Research Council, Transportation Research Board, Special Report 247, pp. 129–177.

Terlien, M.T.J., Van Asch, T.W.J., and Van Westen, C.J. 1995. Deterministic modelling in GIS-based landslide hazard assessment. *In* Geographical information systems in assessing natural hazards. *Edited by* A. Carrara and F. Guzzetti. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 57–77.

Van Westen, C.J. 1993. Application of geographic information systems to landslide hazard zonation. International Institute for Aerospace Survey and Earth Sciences, Publication 15.

Van Westen, C.J., Rengers, N., Terlien, M.T.J., and Soeters, R. 1997. Prediction of the occurrence of slope instability phenomena through GIS-based hazard zonation. Geologische Rundschau, **86**: 404–414.

Varnes, D.J. 1984. Landslide hazard zonation: a review of principles and practice. Natural Hazards, No. 3. UNESCO, Paris, France.

Wang, S.Q., and Unwin, D.J. 1992. Modelling landslide distribution on loess soils in China: an investigation. International Journal of Geographical Information Systems, **6**: 391–405.

Wong, H.N., Chen, Y.M., and Lam, K.C. 1997. Factual report on the November 1993 natural terrain landslides in three study areas on Lantau Island. GEO Report 61, Geotechnical Engineering Office, Hong Kong.

Wong, H.N., Lam, K.C., and Ho, K.K.S. 1998. Diagnostic report on the November 1993 natural terrain landslides on Lantau Island. GEO Report 69, Geotechnical Engineering Office, Hong Kong.

Wu, W., and Sidle, R.C. 1995. A distributed slope stability model for steep forested basins. Water Resources Research, **31**: 2097–2110.

Yin, K.L., and Yan, T.Z. 1988. Statistical prediction models for slope instability of metamorphosed rocks. *In* Proceedings of the 5th International Symposium on Landslides, Lausanne, Switzerland. *Edited by* C. Bonnard. Vol. 2. A.A. Balkema, Rotterdam, The Netherlands, pp. 1269–1272.