Journal of Zhejiang University SCIENCE A
ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

GIS-based logistic regression method for landslide susceptibility mapping in regional scale*

ZHU Lei[1,3], HUANG Jing-feng[†‡2]

([1]Department of Natural Science, Zhejiang University, Hangzhou 310029, China)
([2]Institute of Agricultural Remote Sensing and Information Technology Application, Zhejiang University, Hangzhou 310029, China)
([3]Shandong Agriculture Administrators College, Jinan 250100, China)
[†]E-mail: hjf@zju.edu.cn

Received Mar. 20, 2006; revision accepted Sept. 21, 2006

Abstract: Landslide susceptibility map is one of the study fields portraying the spatial distribution of future slope failure susceptibility. This paper deals with past methods for producing landslide susceptibility map and divides these methods into 3 types. The logistic linear regression approach is further elaborated on by crosstabs method, which is used to analyze the relationship between the categorical or binary response variable and one or more continuous or categorical or binary explanatory variables derived from samples. It is an objective assignment of coefficients serving as weights of various factors under considerations while expert opinions make great difference in heuristic approaches. Different from deterministic approach, it is very applicable to regional scale. In this study, double logistic regression is applied in the study area. The entire study area is first analyzed. The logistic regression equation showed that elevation, proximity to road, river and residential area are main factors triggering landslide occurrence in this area. The prediction accuracy of the first landslide susceptibility map was showed to be 80%. Along the road and residential area, almost all areas are in high landslide susceptibility zone. Some non-landslide areas are incorrectly divided into high and medium landslide susceptibility zone. In order to improve the status, a second logistic regression was done in high landslide susceptibility zone using landslide cells and non-landslide sample cells in this area. In the second logistic regression analysis, only engineering and geological conditions are important in these areas and are entered in the new logistic regression equation indicating that only areas with unstable engineering and geological conditions are prone to landslide during large scale engineering activity. Taking these two logistic regression results into account yields a new landslide susceptibility map. Double logistic regression analysis improved the non-landslide prediction accuracy. During calculation of parameters for logistic regression, landslide density is used to transform nominal variable to numeric variable and this avoids the creation of an excessively high number of dummy variables.

Key words: Landslide, Susceptibility, Logistic regression, GIS, Spatial analysis
doi:10.1631/jzus.2006.A2007 **Document code:** A **CLC number:** P642.2

INTRODUCTION

Landslide is one of the most serious geological hazards in mountain areas. Globally, they cause hundreds of billions of dollars in damage, and hundreds of thousands of deaths and injuries each year (Aleotti and Chowdhury, 1999). Over the past few decades, scientists have shown an ever increasing interest in this natural hazard. One of the study fields is to produce landslide susceptibility map, i.e. a map portraying the spatial distribution of the future susceptibility of slope failure, based on GIS. Among past studies are three main types of GIS-based methods to assess future landslide susceptibility: deterministic, statistical and heuristic approaches.

Heuristic approaches also called expert-driven approach in which expert opinions make great difference during assessing of the type and degree of landslide hazard. To create susceptibility map, the

‡ Corresponding author
* Project supported by the Natural Science Foundation of Zhejiang Province (No. 30295) and the Key Project of Zhejiang Province (No. 011103192), China

direct mapping method is to identify regions susceptible to slope failure by comparing detailed geological and geomorphological properties with those of landslide sites while indirect mapping method integrates many factors and weighs the importance of different variable using subjective decision rules, based on the experience of geoscientists involved. There are a variety of subjective decision rules and most commonly used one being the analytic hierarchy process (AHP) of Saaty (1980) and weighted linear combination (WLC) (Barredo *et al*., 2000; Ayalew and Yamagishi, 2004; Ayalew *et al*., 2004).

Deterministic and statistical methods are quantitative approaches based on numerical expressions of the relationship between controlling factors and landslides. Deterministic approach, based on stability models, is also called geotechnical process model. This method is very useful for mapping hazard in regional scale, for instance for some construction purposes (Barredo *et al*., 2000). In view of the need for detailed and exhaustive geotechnical and groundwater data, this method is often effective for creating a small map in small areas.

Another quantitative approach, statistical method is more objective for mixing a little expert opinion. Significant factors associated with landslide occurrence are analyzed and ranked by statistics or theory such as chi-square test, bayesin, cumulative frequency (Gritzner *et al*., 2001) and other model. Then landslide susceptibility mapping, one way of landslide prediction, can be obtained based on GIS technology. Landslide susceptibility mapping uses either multivariate or bivariate statistical approaches to analyze the historical link between landslide-controlling factors and the distribution of landslides (Guzzetti *et al*., 1999). Bivariate statistical analyses (BSA) involve the idea of comparing a landslide inventory map with maps of landslide influencing parameters in order to rank the corresponding classes according to their role in landslide formation (Ayalew and Yamagishi, 2005). The most used statistical techniques are clearly the multivariate ones, especially linear logistic regression, and discriminate analysis (Clerici *et al*., 2002). Logistic linear regression is further elaborated on by crosstabs method, which is used to analyze categorical data. It describes the relationship between the categorical or binary response variable and one or more continuous or categorical or binary explanatory variables derived from samples and yields the coefficient for each variable. These coefficients serve as weights in an algorithm which can be used in the GIS database to produce a map depicting the probability of landslide occurrence (Dai and Lee, 2002). The independent variables in this model can have value of 0 and 1 representing the absence and presence of landslide cells or sites, model outcome between 0 and 1 shows the susceptibility of landslide. This method is very suitable for landslide prediction and papers using GIS based logistic linear regression method to study landslide problem and assess future landslide probability have been published (Dai and Lee, 2002; Tasser *et al*., 2003; Ohlmacher and Davis, 2003; Yesilnacar and Topal, 2005; Ayalew and Yamagishi, 2005; Beguería, 2006).

Another statistical method, ANNS, has been applied to the study of landslides, with reference to the indirect determination of the triggering parameters (Yesilnacar and Topal, 2005; Gómez and Kavzoglu, 2005) and also to landslide susceptibility mapping, with physical terrain factors (Lee *et al*., 2001). This method is introduced for two reasons. First, it allows a black box model which starts from the database containing the variables of that particular process and assuming that the user does not have any statistical background to find out the multicollinearity problem. The second assumption is that superiority of intelligent systems increases as the dimensionality and/or nonlinearity of the problem increases, which is when traditional regression techniques often fail to produce accurate approximation (Yesilnacar and Topal, 2005).

This paper aims at using GIS database to create logistic multiple regression equations and obtain realistic susceptibility map of potential landslide hazards.

STUDY AREA

Chun'an County located in the west of Zhejiang Province in China (29°11′~30°02′ N, 118°21′~119°20′ E), covers an area of 442717 ha. The Xin'an River, the upper reach of the Qiantang River, flows through this county and its neighbor county, Jiande. It is blocked by a big dam within Jiande County which is the dam for the Xin'an River reservoir and power station. The reservoir boasts 1078 islets, hence the name 1000-Islet Lake. This 573-km^2 lake, over 100 m in depth, is 109 times the size of West Lake in Hangzhou. Nowadays, with numerous tourist attractions, the scenery in 1000-Islet Lake is especially enchanting.

Fig.1 The location of study area in Zhejiang Province of China

Geography summary

In this area, most hills are about 1000 m high. The highest is 1523 m and the extended foothill inclining from circumjacent to inter space is below 400 m. With long-term exogenic force of tectonic movement, tight complex folds and fractures are formed inclining from north-east to south-west and take on a series of stripped mountains and foothills in physiognomy. The tectonic pack is of Yang meta-platform Qiantang fold zone of Yang meta-platform, stratum developed well and exposed completely. Tectonic pack is from Pt_3s of Sinian Period to Quaternary Period of Cenozoic Era except for middle and lower Devonian series, upper Permian series, Triassic series, lower Jurassic series, Cretaceous period and Tertiary Period, especially for lower Paleozoic Era. Rolling country is constituted of flag mudstone, siliceous stone of Ordovician Period and in some part of pelitic limestone of Cambrian Period. The tectonic, rock crushed and weathering resistance is very weak. This is one causative reason for the geological hazard.

This area is in the north of subtropical monsoon climate. It is warm and humid, four seasons distinctly account for the obvious climate difference. Average annual rainfall is 1720 mm. March and April rains occur at high frequency but low intensity. In the flood season from May to July, rainfall is concentrated at high intensity, especially in June. Rainstorm and super rainstorm are other causative reasons for the geological hazard mainly of landslide in Chun'an County.

Landslides

Geological hazard, especially landslide distributed all over the county. According to the geological survey of Land Management Bureau of Zhejiang Province in 2002, there were 200 landslides and 146 latent landslides. The work was accomplished during all-around field survey in the area seriously affected by landslide. In other areas, the landslide was found by inquiry and validation on the spot. Latent landslides here refer to those areas of mass movement and show obvious evidence of landslide. For example, there is camber tension crack at the top of these areas and sink or rise in the central and lower section. Trees become crooked and most walls are ruptured. The geological setting of this area is unstable but landslide has not occurred. There is great latent threat and landslide is prone to happen when the area is subject to rainstorm or other engineering action.

Among 200 landslides that occurred in past half century, more than 150 landslides occurred in latest ten years. Almost all the landslides in Chun'an are earth slides. Only 6 landslides is rock slide, with gliding mass made up of mudstone, siltstone and bedrock debris. But the area of these rock slides is small. According to the geological survey work in 2002, the length of 70% landslides and the width of 80% landslides recorded were less than 30 m. So each landslide can be assumed to be a single 30 m pixel in the grid layer. The pixel in the center of the landslide site recorded is taken into account to collect factors influencing landslide.

DATA AND METHODS

In order to assess differential incidence of every predisposing factor to landslide and assess the landslide susceptibility, maps of the various factors contributing to the initiation of landslides were constructed in ARC/INFO 9. The factors considered were elevation, slope, aspect, pro-curvature (the curvature of the surface in the direction of slope), plan-curvature (the curvature of the surface perpendicular to the slope direction, referred to as the plan-form curvature), distance to river, fault, drainge line, residential area, road and digging, land use type, soil type, engineering and geological conditions.

Geomorphological factor

Surface topography by controlling flow sources, flow direction and soil moisture concentration, is an important factor limiting the density and spatial extent of landslides (Ayalew and Yamagishi, 2005). So it is necessary to use DEM to get some geomorphological information such as elevation, slope, aspect, plan-curvature and pro-curvature. DEM is used to get the 1:10000 topographical map with 25 m contour interval by spatial interpolation. The cell size of DEM is 30 m.

Engineering and geological conditions

Engineering and geological conditions are other important causative factors for landslide when there is external force such as slope undercutting. Here, petrofabric is used to represent different engineering and geological conditions (Fig.2). Considering different geology, 19 petrofabrics were divided including three petrofabrics in river valley plain region and 16 petrofabrics in hilly and rolling area (Table 1).

Table 1 Petrofabrics representing different engineering and geological condition

Petrofabrics codes	Petrofabrics types	Stratum included (stratum code)
ST	Petrofabrics mainly of incompact sandy soil and sub sandy soil	Q_hy^{al}, Q_{hy}, Q_3, Q_1
LT	Petrofabrics mainly of incompact gravel, gravelly and sandy soil	Q_4y^{al}, Q_3s^{pal}, $Q_4{}^{al}$
Sfn	Petrofabrics mainly interlaced with soft and rigid medium-bedded and thin-bedded mudstone, siltstone, sandy stone	K_1h, J_2m, J_2y, S_2k, S_1d (S_1h), S_1a (S_1x), O_3w (O_3y), O_3c (O_3yq)
Sfy	Petrofabrics mainly interlaced with soft and rigid medium-bedded and thin-bedded mudstone, siltstone, shale and siliceous shale	P_2l, P_1d, ϵ_2y, ϵ_1h
Sfg	Petrofabrics mainly of semi-compact medium-bedded and thin-bedded calcareous mudstone and mudstone	O_1n, O_1y
Scc	Petrofabrics mainly of semi-compact medium-bedded and thick-bedded quartzose sandstone, carbonaceous shale, mudstone	C_1y, D_3z, P_1x, P_1l
Tch	Petrofabrics mainly of semi-compact medium-bedded and thick-bedded crystallinoclastic rock, pelitic limestone	ϵ_3x, ϵ_3h
Sff	Petrofabrics mainly of semi-compact medium-bedded and thick-bedded sandy mudstone, post stone tuffaceous siltstone	Z_1n (Z_1l), Z_1x (Z_1z)
Tcb	Petrofabrics mainly of compact medium-bedded and thick-bedded limestone, dolomitic limestone and dolomite	C_3c, C_2h, C_2l, ϵ_1d
Sfx	Petrofabrics mainly of compact and semi-compact medium-bedded and thick-bedded calcareous mudstone and mudstone	O_3s, O_3h, O_2y, O_2h
Scn	Petrofabrics mainly of compact and semi-compact block conglomerate, silt mudstone, siltstone, fine sand with pebble	K_2j
Scs	Petrofabrics mainly of compact and semi-compact medium-bedded and thick-bedded quartzose sandstone, siltstone, sandy dolomite and silicalite	Z_2dy, Z_2p, Z_2b, Z_2d
Byy	Petrofabrics mainly of compact and semi-compact medium-bedded and thick-bedded compact and semi-compact epizonal sand rock, epizonal rhyolite, epizonal tuff	Pt_3s, Pt_3h, Pt_3l
Scl	Petrofabrics mainly of compact medium-bedded and thick-bedded block quartzose sandstone and sand rock	D_3x, S_3t
Hin	Petrofabrics mainly of compact medium-bedded and thick-bedded block rhyolite and tuff	J_3h, J_3l, J_3s
Qg	Eugeogenous acidite rock petrofabrics mainly of compact block granite	–
Qd	Intermediate rock petrofabrics mainly of compact block diorite	–
Rr	Acidite petrofabrics mainly of compact block rhyolite	–
Rb	Basic rock petrofabrics mainly of compact block basalt	–

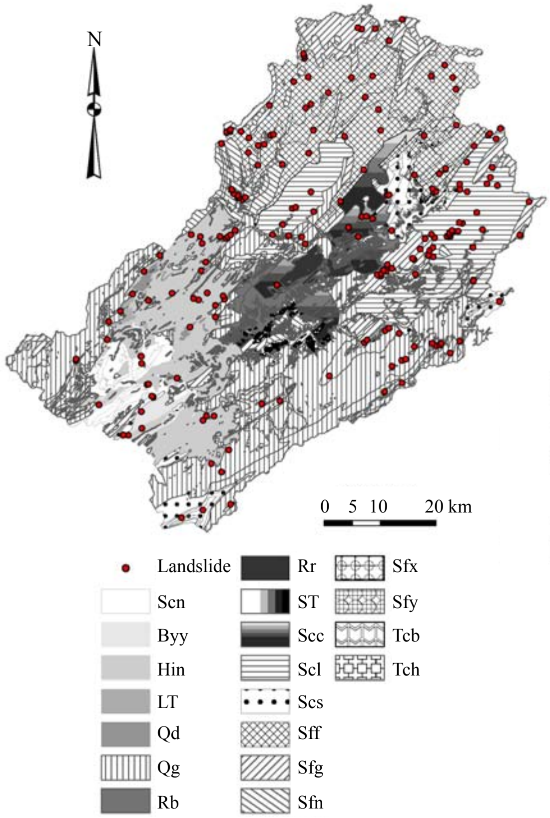

Fig.2 Engineering and geological map of study area

Land use and soil type

Extensive investigations showed that land-use cover or vegetation cover, especially of a woody type with strong and large root systems, helps to improve the stability of slopes (Gray and Leiser, 1982; Greenway, 1987). A 1:10000 land use map for the study area is used for analysis. It is the fruit of land use detailed field survey based on the interpretation of SPOT images with verification of field checking by the Information Center of Land Management Bureau in Zhejiang. This work lasted for two years from 2003 to 2005. The land use map and data were checked and accepted by Land Management Bureau of Chun'an County and Zhejiang Province and the precision of this map can meet this study. According to this map, there are more than 30 land use types in this area. For the purpose of this study, they are simplified into 8 categories (Fig.3): (1) farm land; (2) forestland with vegetation cover of more than 50%; (3) developed land including buildings, roads and irrigation establishment; (4) sparsely grassed land on rock outcrop-dominated areas; (5) bare land; (6) water such as river, pond, bottomland and reservoirs; (7) land with vegetation cover of less than 50%; (8) other undeveloped land. For the increased development of Chun'an economy, land use types were changed from time to time in the past several decades. So it should be noted that land use type is considered to be only estimates.

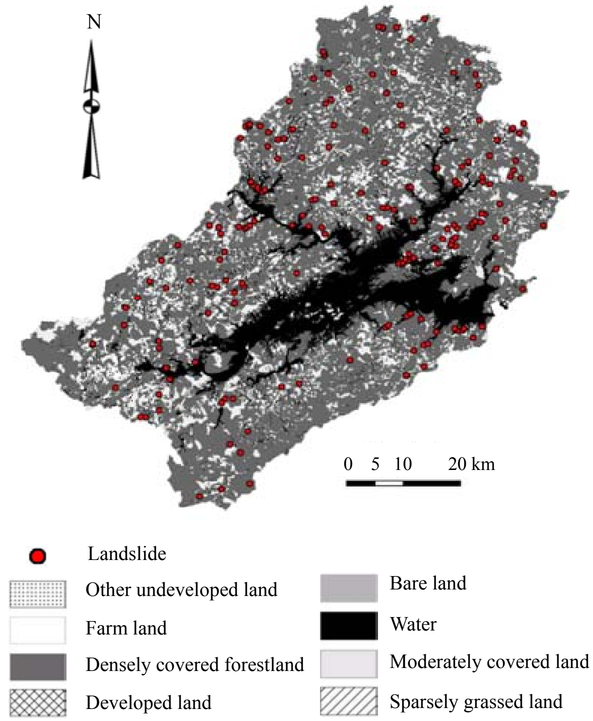

Fig.3 Land use map of study area

Most landslides in Chun'an are earth slides. Soil depth, texture and moisture may be other factors contributing to the initiation of landslides. The soil survey was completed in 1985 and the scale was 1:50000. Here, 14 soil types were taken as analysis data (Fig.4): (1) dry calcium purple sand; (2) dry red earth; (3) dry yellow-red earth; (4) dry eroded red earth; (5) dry limestone soil; (6) yellow earth; (7) mountain calcium purple sand; (8) mountain red earth; (9) mountain yellow-red earth; (10) mountain eroded red earth; (11) mountain limestone soil; (12) gleyed paddy soil; (13) percogenic paddy soil; (14) hydragric paddy soil.

All these datasets in vector format (coverage of ARC/INFO 9) mentioned above were converted into raster format (grid of ARC/INFO 9).

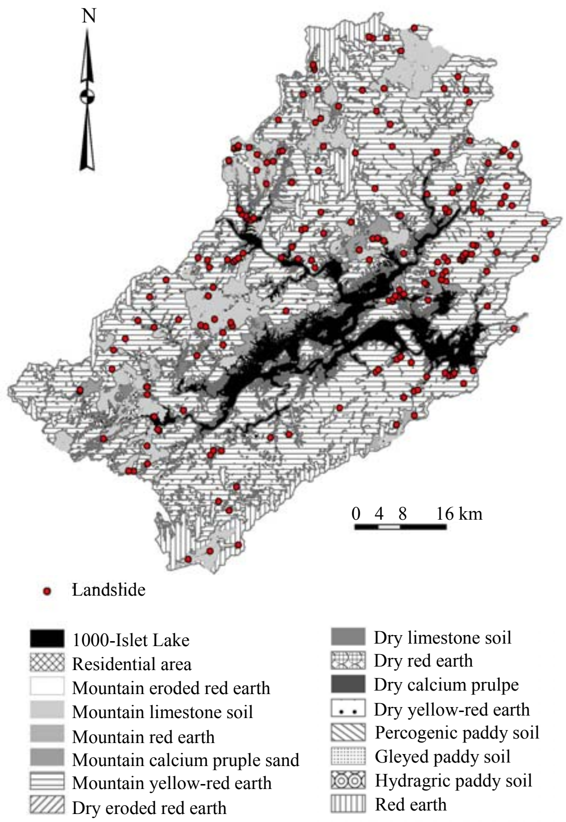

Fig.4 Soil map of study area

The distance to river or fault

It was assumed that the intermittent flow regime of the streams and gullies in the basin involved remarkable erosive processes, which, in turn, are the cause of intense, superficial mass wasting phenomena in areas adjacent to drainage channels (Barredo *et al.*, 2000). Besides, river or streams may induce failure of the banks due to slope undercutting. In this study, river map is offered by Water Conservancy Department of Chun'an County and primary rivers and rivulets are included.

Distance from direct faults and the thrusts faults are also main causative reasons for landslides in that the presence of these tectonic structures breaks the rock mass reducing its strength (Donati and Turrini, 2002). Fault lines were derived from 1:50000 scale geology maps.

Some anthropogenic factors

Road-cuts are usually sites of anthropologically induced instability. It is the same as slope undercutting by house. Fig.5 shows that most landslides are near to road and residential area in this study area. What is more, analysis was carried out to assess the influence of drainge lines on landslide occurrence (Dai and Lee, 2002). For these reasons, proximity to road, house and drainage line are also considered as parameter in the logistic regression model. The road, residential area and drainage line are all extracted from land use map, the resulting map of land use detailed survey in 2004.

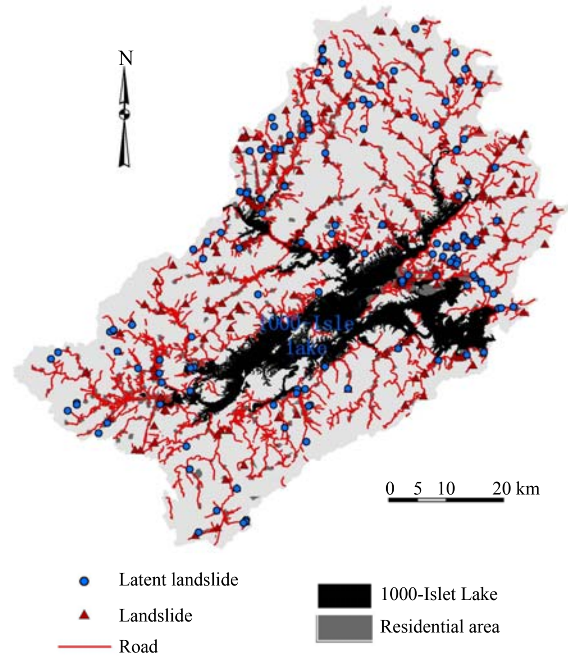

Fig.5 Road, residential area and landslide distribution

Mining can also undercut the slope and reduce the strength of rock mass. There are 129 diggings in this area with the location recorded by mining survey. So the distance of landslides to diggings is also calculated to be involved in the model. When calculating distance, search radius is 10000 m.

Logistic regression model

As the dependent variable is dichotomous and the relationship between the dependent variable and independent variables is nonlinear, logistic multiple regression model was constructed based on the physical parameters defined above.

When using logistic regression model, an issue is how many samples should appropriately be taken to create dependent variables. Literature showed that there are mainly 3 types in practice. First one is using data from all over the study area, which undoubtedly leads to unequal proportions of landslide and non-landslide pixels (Ohlmacher and Davis, 2003; Guzzetti *et al*., 1999). Large volume of data is included in this method. Second method is using all the landslide pixels and equal non-landslide pixels. This may decrease data number and eliminate bias in the sampling process. But the model constructed and validated using the same landslide data may reduce the reliability. Yesilnacar and Topal (2005) used the total number of landslide pixel and randomly selected cells from landslide free areas, with these data divided into six different training datasets with the accuracy percentage being used to determine which datasets will work well in the logistic model. The most reasonable method is to divide landslide pixels into two parts: training data and test data. There are also two cases of this method. For example, Atkinson and Massari (1998) used unequal pixels while Dai and Lee (2002) used equal proportion of landslide and non-landslide pixels.

As for this area, the number of landslide pixels is comparatively little for the area of each landslide is small. All the pixels should be taken into account to increase the accuracy. In order to decrease the effect of unequal proportion of landslide and non-landslide pixels, equal number of non-landslide pixels is randomly selected from free landslide area.

Another issue in this study is the disposal of nominal variables such as land use type, soil type and engineering and geological type. The conversion of parameters from nominal to numeric can be done through the creation of dummy variables or by coding and ranking the classes based on the relative percentage of the area affected by landslides (Yesilnacar and Topal, 2005). Here, the landslide density is used to transform land use, soil and engineering and geological nominal variable to numeric variable. It avoids the creation of an excessively high number of dummy variables and allows consideration of the so-called "previous knowledge" of landslide susceptibility (Carrara, 1983). This is the formula to calculate the densities:

$$Densiti = (B_i / A_i) \Big/ \sum_{i=1}^{N} (B_i / A_i), \tag{1}$$

where A_i is the area of ith type of a certain parameter; B_i is landslide area of ith type of a certain parameter; N is the type number of a certain parameter.

Then a certain parameter map was compared with the landslide inventory map to calculate landslide densities based on Eq.(1). Then the parameters of landslide and non-landslide sample points are calculated; Table 2 shows the descriptive statistics of

Table 2 Descriptive statistics of parameters of landslides and non-landslides

Parameters (pixels)	Range		Minimum		Maximum		Mean		Median		Std. deviation		Variance		Skewness		Kurtosis	
	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
Dem	680	1029	116	111	796	1140	261	412	203	359	145	221	21108	48658	1.5	0.9	2.0	0.7
Slope	43.3	77.5	0.0	0.0	43.3	77.5	16.4	23.6	17.3	23.9	11.2	12.0	125.1	144.7	0.1	0.1	−0.9	1.1
Aspect	359	361	−1.0	−1.0	358	360	164	167	160	159	104	104	10823	10735	0.0	0.2	−1.1	−1.0
Pro-curvature	5.8	9.6	−2.9	−6.6	2.9	3.1	0.2	0.0	0.0	0.0	0.8	1.0	0.7	1.0	0.1	−1.6	3.2	11.7
Plan-curvature	4.1	22.3	−1.5	−2.3	2.6	20.0	0.1	0.1	0.0	0.0	0.6	1.6	0.3	2.7	1.2	9.1	4.3	109
Landslide density of petrofabrics types	0.5	0.5	0	0	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	4.2	6.4	17.4	51.6
Landslide density of soil type	0.3	0.2	0.0	0.0	0.3	0.2	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	1.1	−0.7	3.4	2.5
Landslide density of land use type	0.3	0.3	0.0	0.0	0.3	0.3	0.2	0.3	0.3	0.3	0.2	0.1	0.0	0.0	−0.3	−1.2	−1.8	−0.6
Distance to road	2436	4232	0.1	3.4	2436	4235	205	704	98.4	572	311	626	96898	391792	4.0	2.3	21.6	8.3
Distance to river	5227	4268	0.8	1.5	5228	4269	495	935	234	798	723	706	522109	499043	3.1	1.4	12.5	3.2
Distance to residential area	1373	2489	0.2	4.8	1373	2494	129	668	66.4	621	191	453	36346	205126	3.4	1.2	14.2	1.8
Distance to fault	5744	4887	4.1	1.5	5749	4888	906	1088	512	801	1023	1007	1045402	1013691	2.1	1.4	5.4	2.0
Distance to digging	9875	9951	126	49	10000	10000	4627	5129	3582	4659	3320	2947	11019947	8686526	0.5	0.3	−1.3	−1.1
Distance to drainge line	2207	2501	2.5	4.0	2210	2505	317	396	157	305	400.0	386	159957	149112	2.5	2.4	7.2	7.8

1 refers to landslides and 0 refers to non-landslides

every parameter with the One-sample Kolmogorov-Smirnov test indicating that none of data list except slope and aspect is close to normal distribution.

So it is evident that both the multiple regression method and discriminant analysis do not work well in this case. Also, test for two independent samples shows the difference of landslide and non-landslide data of every parameter. Significant of aspect, pro-curvature and plan-curvature is more than 0.05 showing that these three parameters of landslide and non-landslide data list show little difference. So the value of independent variables besides these three parameters and its respective binary value on landslide and non-landslide sample points are exported as ASSII file from GIS software and analyzed in SPSS to construct best logistic regression equation.

RESULTS AND DISCUSSION

Forward stepwise logistic regression was used to analyze data. This method begins with no stepwise terms in the model. At each step, the most significant term is added to the model until none of the stepwise terms left out of the model would have a statistically significant contribution if added to the model. After the forward stepwise logistic regression analysis, variables such as distance to road, distance to river, distance to residential area and elevation were selected for being statistically significant. The overall model statistics of the regression conducted in this study using SPSS is shown in Table 3.

Table 3 The overall statistics of the logistic regression model involved 4 variables

Hosmer-Lemeshow test			−2log likelihood	Cox and Snell R^2	Nagelkerke R^2
chi-square	df	sig			
12.581	8	0.13	405.38	0.309	0.413

Hosmer-Lemeshow test showed that the goodness of fit of the equation can be accepted because the significance of chi-square is larger than 0.05. The value of Cox and Snell R^2 and Nagelkerke R^2 showed that the independent variables can explain the dependent variables in a way. Besides, the predicted accuracy for landslide is 88% and for non-landslide is 80%; the overall predicted accuracy is 84%.

Though overall statistics of the model is good, variables such as slope, distance to fault, engineering and geological conditions that are generally accepted as important triggering factors were not included in the model. There may be many reasons. Collinearity is one possible reason. Tolerance and the variance inflation factor are two important indexes for multicollinearity diagnosis. Tolerance smaller than 0.2 is one indicator for multicollinearity and smaller than 0.1 means that there is serious multicollinearity between independent variables (Menard, 1995). The smallest tolerance being larger than 0.5 in this study (Table 4) showed that there is little multicollinearity between independent factors. Variance inflation factor (VIF) is another index the reciprocal of tolerance index.

Table 4 The multicollinearity diagnosis indexes for variables

Independent variables	Tolerance	*VIF*
Slope	0.7348	1.3610
Aspect	0.9499	1.0528
Pro-curvature	0.7019	1.4247
Plan-curvature	0.6599	1.5153
Distance to road	0.5486	1.8228
Distance to river	0.8505	1.1757
Distance to residential area	0.5490	1.8214
Distance to fault	0.9336	1.0711
Distance to drainge line	0.8460	1.1820
Distance to digging	0.9367	1.0675
Landslide density of petrofabrics types	0.9518	1.0507
Landslide density of soil type	0.8947	1.1176
Landslide density of land use type	0.9019	1.1088

As the geological survey work is not from image interpretation, some of the landslides in the remote mountains and far from road or residential area may be missed. This may lead to deviation of training data in a way and is one kind of input noise. It may be one causative reason for those factors such as slope, distance to fault being not involved in the equation. So landslide survey should be developed using remote sensing technology in other area and in later work in this area.

In order to improve the prediction veracity and as best as we can to decrease the effect of possible survey mistakes, 7 significant variables are used to construct the logistic regression equation using forced entry terms. Besides 4 parameters statistically significant, 3 parameters generally considered as important are also included in the equation. Using this equation, the predicted accuracy for landslide is 89.9% and for non-landslide is 81%. The prediction accuracy is increased from 84% to 85.5% although goodness of fit is decreased. Table 5 is some statistic of the new equation and the equation is as follows:

$$f(x)=\ln(p/(1-p))=2.694199+0.004957S -0.000916d_{\mathrm{ro}}-0.000551d_{\mathrm{ri}}-0.005799d_{\mathrm{ra}} -0.000112d_{\mathrm{f}}+2.862699C-0.001134E, \quad (2)$$

where p is the landslide probability; S is the slope; d_{ro} is the distance to road; d_{ri} is the distance to river; d_{ra} is the distance to residential area; d_{f} is the distance to fault; C is the engineering and geological condition; E is the elevation.

Table 5 The overall statistics of the logistic regression model involved 7 variables

Hosmer-Lemeshow test			−2log likelihood	Cox and Snell R^2	Nagelkerke R^2
chi-square	df	sig			
33.4487	8	0.0840	303.89	0.46	0.52

According to Eq.(2) obtained from the training data, these 7 parameter layers of the entire study area are calculated to generate landslide susceptibility map. The value of each cell in the map represents landslide susceptibility with the range of landslide susceptibility being classified into 4 categories for general purpose. These 4 categories are very low (0~25%), low (25%~50%), medium (50%~75%) and high (75%~100%). Comparing landslides and latent landslides map with the landslide susceptibility map, 89% of landslides fall into high and medium landslide susceptibility zone while 83% of latent landslides fall in high and medium landslide susceptibility zone. 26.48% of the study area falls in high susceptibility zone.

The precision accuracy for landslide can be accepted. But these areas are like buffer zone of road and residential area and almost all the area along road falls in high susceptibility zone. There must be many lands with low landslide susceptibility incorrectly separated into high landslide zone. It indicates that the prediction accuracy for non-landslide is not high and that it may be a conservative prediction. More conservative thresholds would result in a smaller number of negative errors (landslides not recognized) at the expense of a higher rate of positive errors (safe areas classified as hazardous). This conservative approach should be used in territorial planning or in preliminary phases of risk analysis, in which the focus must be concentrated on predicting possible hazards and revealing the location of key areas (Ermini *et al.*, 2005). In this study, in order to decrease parts incorrectly divided into high landslide zone near the road and residential area and improve the prediction for non-landslide, high landslide susceptibility area are separated from total study area and are calculated to get landslide probability again with the same method as that for the total study area. Two hundred non-landslide cells were randomly selected in this area. Fourteen parameters mentioned above for these 200 cells were calculated; these data and parameters of original 200 landslide cells were again analyzed in SPSS to select important factors. In the stepwise logistic regression, only engineering and geological conditions were included in the model as important factors. Then the conclusion can be drawn that only areas with special engineering and geological conditions will become unstable and be prone to landslide when large engineering activity occurs. By some calculation of the engineering and geological condition layers using the new logistic regression equation, the landslide susceptibility map of these areas can be created. Then it was divided into two parts according to the probability. Those areas where the probability is more than 0.47 are considered to be highly susceptible to landslide in the total study area and less than 0.47 as medium susceptibility to landslide. The low and very low landslide susceptibility area of the study area is obtained from the original landslide susceptibility map with the cutoff point being 0.35. Incorporation of these two maps yields a new landslide susceptibility map of the total study area (Fig.6). Table 6 shows some statistics of the first and second landslide susceptibility map. Although the landslide prediction decreases a little, the area of high and medium landslide susceptibility decreases more. So it indicates that there is little non-landslide incorrectly divided into landslide zone in the new landslide susceptibility map. Overall, the second landslide susceptibility map can be accepted.

Table 6 Statistics of first and second landslide susceptibility map

Susceptibility levels	First landslide susceptibility map					Second landslide susceptibility map				
	PTA (%)	*PLC* (%)	*PLLC* (%)	Ratio		*PTA* (%)	*PLC* (%)	*PLLC* (%)	Ratio	
				PLC/*PTA*	*PLLC*/*PTA*				*PLC*/*PTA*	*PLLC*/*PTA*
Very low	59.84	6.00	8.84	10.03	14.77	65.75	7.00	10.88	10.65	16.55
Low	13.68	4.50	8.84	32.89	64.62	21.12	22.50	23.13	106.53	109.52
Medium	13.35	19.50	17.01	146.07	127.42	4.78	7.50	14.29	156.90	298.95
high	13.13	70.00	65.31	533.13	497.41	8.32	63.00	51.70	757.21	621.39

PTA: percent in total area; *PLC*: percent in landslide cells; *PLLC*: percent in latent landslide cells

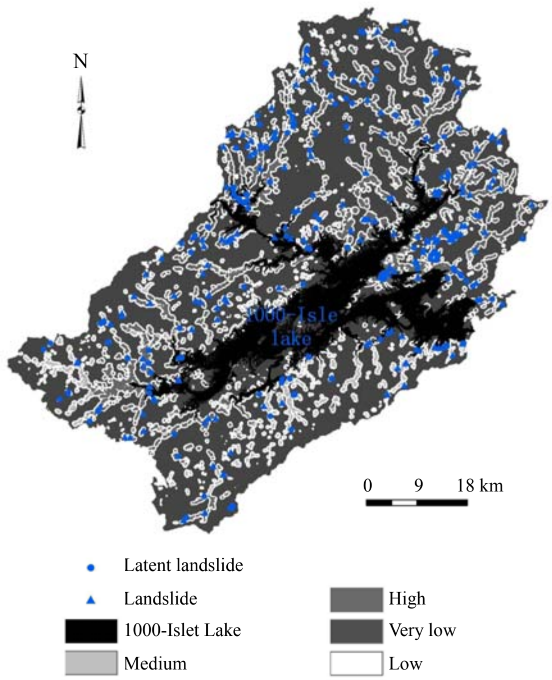

Fig.6 Improved landslide susceptibility map of the study area

CONCLUSION

In hazard management, landslide susceptibility map can help to prevent and manage hazard effectively. Nowadays, GIS tools exist for possible production of innovative landslide susceptibility maps. Many qualitative and quantitative techniques are useful for analyzing the relationship between landslides and their influencing parameters. In this study, logistic model regression model is used twice. First analyze the whole study area then high landslide susceptibility zone. The influencing parameters considered include elevation, aspect, slope, pro-curvature, plan-curvature, distance to road, distance to river, distance to residential area, distance to digging, distance to fault, distance to drainage line, engineering and geological conditions, soil type and land use type. Landslide densities were used to transform nominal variables to numeric variables.

Model statistics and coefficients were two results of logistic regression, which were useful for assessing the accuracy of the regression function and the role of parameters on landslide occurrence. The susceptibility map was another outcome of the regression process. In first logistic regression, elevation, the distance to road, residential area and river are main important factors for presence or absence of landslides of the total area. The first landslide susceptibility map was divided into 4 zones of landslide susceptibility, namely very low, low, medium and high. Eighty-nine percent landslide cells and 83% latent landslide cells were included in the high and medium susceptibility zone. But it is just a conservative method and the high prediction accuracy of landslide is at the expense of a higher rate of positive errors (safe areas classified as hazardous). So the second landslide probability prediction is calculated in high landslide probability zone in the first result map by logistic regression. In this logistic regression equation of high landslide susceptibility zone of first result map, only engineering and geological conditions were not removed. It indicates only land with unstable engineering and geological conditions is prone to landslide.

Then take these two logistic regression results into account to obtain a new landslide susceptibility map which greatly improves the result that large non-landslide area is incorrectly divided into high landslide susceptibility area.

On the whole, the logistic regression model based on GIS can effectively determine which parameter is most important to landslide and predict landslide susceptibility. But when the logistic regression model is used to predict landslide susceptibility, attention should be paid to the multicollinearity. When there is multicollinearity, the coefficient is very sensitive to samples and model setting. Little change of model and deletion or addition of one sample may lead to large change of coefficient estimate. So diagnoses and elimination of multicollinearity are very important though independent variables show little multicollinearity in this study. Factor analysis is one method to eliminate multicollinearity effect when using logistic regression model and ANNs may be a substitute for logistic regression model when the researcher has no statistical background.

There are still some issues that need further study. The rainfall factor is not available in this study but it should be included in the analysis. Select appropriate rainfall index with temporal and spatial scale being issues in further study. Besides, landslide inventory should be perfected and validated through interpretation of image based on RS technology.

References

Aleotti, P., Chowdhury, R., 1999. Landslide hazard assessment: summary review and new perspectives. *Bulletin of Engineering Geology and the Environment*, **58**(1):21-44. [doi:10.1007/s100640050066]

Atkinson, P.M., Massari, R., 1998. Generalized linear modelling of landslide susceptibility in the Central Apennines. *Computers and Geosciences*, **24**(4):373-385. [doi:10.1016/S0098-3004(97)00117-9]

Ayalew, L., Yamagishi, H., 2004. Slope movements in the Blue Nile basin, as seen from landscape evolution perspective. *Geomorphology*, **57**(1-2):95-116. [doi:10.1016/S0169-555X(03)00085-0]

Ayalew, L., Yamagishi, H., 2005. The application of GIS-based logistic regression for susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology*, **65**(1-2):15-31. [doi:10.1016/j.geomorph.2004.06.010]

Ayalew, L., Yamagishi, H., Ugawa, N., 2004. Landslide susceptibility mapping using GIS-based weighted linear combination, the case in Tsugawa area of Agano River, Niigata Prefecture, Japan. *Landslides*, **1**(1):73-81. [doi:10.1007/s10346-003-0006-9]

Barredo, J.I., Benavidesz, A., Herhl, J., van Westen, C.J., 2000. Comparing heuristic landslide hazard assessment techniques using GIS in the Tirajana basin, Gran Canaria Island, Spain. *International Journal of Applied Earth Observation and Geoinformation*, **2**(1):9-23. [doi:10.1016/S0303-2434(00)85022-9]

Beguería, S., 2006. Changes in land cover and shallow landslide activity: A case study in the Spanish Pyrenees. *Geomorphology*, **74**(1-4):196-206. [doi:10.1016/j.geomorph.2005.07.018]

Carrara, A., 1983. Uncertainty in Evaluating Landslide Hazard and Risk. *In*: Nemec, J., Nigs, J.M., Siccardi, F. (Eds.), Prediction and Perception of Natural Hazards. Kluwer, Dordrecht, The Netherlands, p.101-111.

Clerici, A., Perego, S., Tellini, C., Vescovi, P., 2002. A procedure for landslide susceptibility zonation by the conditional analysis method. *Geomorphology*, **48**(4):349-364. [doi:10.1016/S0169-555X(02)00079-X]

Dai, F.C., Lee, C.F., 2002. Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. *Geomorphology*, **42**(3-4):213-228. [doi:10.1016/S0169-555X(01)00087-3]

Donati, L., Turrini, M.C., 2002. An objective method to rank the importance of the factors predisposing to landslides with the GIS methodology: application to an area of the Apennines (Valnerina; Perugia, Italy). *Engineering Geology*, **63**(3-4):277-289. [doi:10.1016/S0013-7952(01)00087-4]

Ermini, L., Catani, F., Casagli, N., 2005. Artificial Neural Networks applied to landslide susceptibility assessment. *Geomorphology*, **66**(1-4):327-343. [doi:10.1016/j.geomorph.2004.09.025]

Gómez, H., Kavzoglu, T., 2005. Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. *Engineering Geology*, **78**(1-2):11-27. [doi:10.1016/j.enggeo.2004.10.004]

Gray, D.H., Leiser, A.T., 1982. Biotechnical Slope Protection and Erosion Control. Van Nostrand-Reinhold, New York, p.271.

Greenway, D.R., 1987. Vegetation and Slope Stability. *In*: Anderson, M.G., Richards, K.S. (Eds.), Slope Stability. Wiley, New York, p.187-230.

Gritzner, M.L., Marcus, W.A., Aspinall, R., Custer, S.G., 2001. Assessing landslide potential using GIS, soil wetness modeling and topographic attributes, Payette River, Idaho. *Geomorphology*, **37**(1-2):149-165. [doi:10.1016/S0169-555X(00)00068-4]

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, central Italy. *Geomorphology*, **31**(1-4):181-216. [doi:10.1016/S0169-555X(99)00078-1]

Lee, S., Ryu, J., Min, K., Won, J., 2001. Development of Two Artificial Neural Network Methods for Landslide Susceptibility Analysis. Proceedings of the Geoscience and Remote Sensing Symposium, IGARSS'01, IEEE 2001 International, **5**:2364-2366.

Menard, S.W., 1995. Applied Logistic Regression Analysis. SAGE Publication, Inc., Thousand Oaks, CA.

Ohlmacher, G.C., Davis, J.C., 2003. Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. *Engineering Geology*, **69**(3-4):331-343. [doi:10.1016/S0013-7952(03)00069-3]

Saaty, T.L., 1980. The Analytical Hierarchy Process. McGraw Hill, New York, p.350.

Tasser, E., Mader, M., Tappeiner, U., 2003. Effects of land use in alpine grasslands on the probability of landslides. *Basic Appl. Ecol.*, **4**(3):271-280. [doi:10.1078/1439-1791-00153]

Yesilnacar, E., Topal, T., 2005. Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). *Engineering Geology*, **79**(3-4):251-266. [doi:10.1016/j.enggeo.2005.02.002]