LANDSLIDE HAZARD ZONING USING GENETIC PROGRAMMING

Sandra Litschert
Department of Geosciences
Colorado State University
Fort Collins, Colorado 80523

Abstract: Genetic programming (GP) is presented as a technique to induce models that can be used with GIS data to map landslide-hazard zones. GP is a data-mining technique frequently used to solve engineering and scientific problems; in recent years it has been applied to several spatial questions. This paper describes attempts to optimize the GP system, to evolve models that are then tested in different locations. GP map accuracy is evaluated by comparison to landslide-hazard maps of the same California locations created by the USGS (United States Geological Survey; Wentworth et al., 1997). GP maps are compared to maps created using linear discriminant analysis (LDA). In three out of the four study sites, GP models produce more accurate hazard maps than the LDA process. The highest accuracy attained by GP models is 84% and by LDA is 69%. Map accuracies at the training site (Yountville and Capell Valley) are higher than at other locations, highlighting the need for care when choosing a training site. Inaccuracies could be caused by using data of too coarse spatial resolution or by differences in data-processing techniques of GP and the USGS, that is, individual cell versus a regional approach. [Key words: landslides, genetic programming, hazard zoning, California.]

INTRODUCTION

Landslides are globally the most costly of the natural hazards in terms of lives lost and damage to resources and property. Accurate landslide-hazard maps are needed to aid in decision making for future development and to prevent the types of catastrophic losses seen in recent years, for example in Central and South America. Automated mapping techniques can be used to provide an initial attempt at regional landslide-hazard zoning. The variety of research in this area reflects the importance and difficulty of this problem; zoning techniques have included inventories, GIS procedures, and models.

Landslide-hazard maps are often created using a two-step method. First, inventories are formulated of areas where landslides have occurred using extensive field reconnaissance and delineation of landslides from aerial photographs (Wieczorek, 1984; Howes, 1987; Wentworth et al., 1997). Assessment of existing landslides can be confusing and subjective, particularly where multiple overlapping landslides have occurred or where there is substantial vegetative regrowth. The second step is to rate landslide hazard by spatial frequency of the landslides or some other criteria (Wentworth et al., 1997; Parise, 1999). Without further analysis, simple inventories provide little insight into causes or locations of future landslides except around known slides.

Physical Geography, 2004, **25**, 2, pp. 130–151.

Predictive models imply that explanatory factors are causative or at least associative. Such models can be based on factors such as slope, land use, and lithology. If explanatory factors in an area of interest are present or in a similar size range comparable to areas known to be susceptible to landslides, models may be able to predict landslide hazards in the area of interest. Such hazard-zoning models are relatively easy to implement with Geographic Information Systems (GIS). GIS have been used successfully in deterministic models such as weighted overlays of polygons or raster layers depicting areas of physical homogeneity (Clouatre et al., 1996; van Westen et al., 1997; Nguyen, 2000) or empirically based equations (Sakellariou and Ferentinou, 2001). Physical or statistical methods combined with GIS for landslide-hazard zonation include multivariate regression and discriminant analysis (Carrara et al., 1991; Naranjo et al., 1994; Guzzetti et al., 1999, 2000; Dhakal et al., 2000; Gritzner et al., 2001; Dai and Lee, 2003).

Relationships between enhanced processing power, data availability, and model complexity are direct and circular. Increased processing power allows use of high resolution data such as remotely sensed data which can greatly enhance studies in areas that are difficult to access (Jordan et al., 2000; Nguyen, 2000). Mathematical and physical models have been developed in recent years that successfully take advantage of such advances through intuitive and appealing graphical user interfaces, and GIS. Sinmap and Shalstab models combine easy to use software interfaces with GIS data layers to allow calculation of a spatially explicit infinite slope model resulting in stability indices (Dietrich, 1998; Pack, 1998). However such models can still be hampered by lack of available and accurate data of appropriate resolution.

Other modeling or problem-solving techniques that have been applied to geographic problems use data mining or searching paradigms. Genetic programming (GP), explained in the next section of this paper, is one such method commonly used to explore data through an evolutionary algorithm. This study explores the use of GP to discover causative or associative factors most important to landslides and to induce a model for landslide-hazard zoning using readily available data. A hazard map is also created using linear discriminant analysis and the same data. The hazard maps are compared, both quantitatively and qualitatively, to a landslide-hazard map of the same area previously created by USGS (United States Geological Survey) scientists.

GENETIC PROGRAMMING

GP has been described as a searching, data mining, and optimization technique for analytical problem solving. Since Koza invented GP, it has become a widely used method for solving a variety of questions in science and engineering (Koza, 1992, 1994). In a geographic setting, Whigham (2000) and McKay (2001) examined natural variability of marsupial populations using raster data. GP has also been used in image processing for classification and feature extraction (Daida, 1995, 1996; Jan, 1997). An advantage of GP over similar methods is that model results are visible so that it is possible to determine which factors are used and to evaluate how

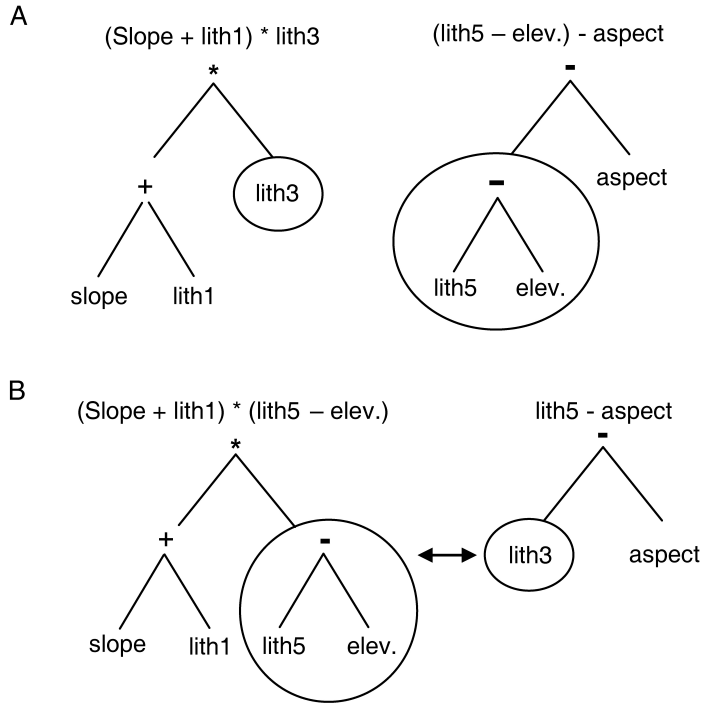

Fig. 1. Crossover operation, (A) before and (B) after.

factors are used (Diplock, 1998). However, Turton et al. (1996) found that GP induced such complex models that they were hard to interpret.

GP is based on the Darwinian theory of evolution. The evolutionary aspect of GP stems from its iterative processing of generations of individual models. Each model is designed to solve a problem, in this case, to predict and map landslide-hazard zones.

A model can be represented by a tree structure with terminals being user-defined variables such as slope or lithology. Interior nodes of the tree represent user provided functions with which to combine the variables into a predictive model. The set of functions may include mathematical, conditional, and logical operators.

The first generation of models is created randomly from the variables and functions, and is evaluated for fitness by comparison to a training set of data provided by the user. The fittest models are carried over into the next generation using randomly assigned breeding operators: reproduction, crossover, and mutation. During reproduction, an exact replica of a model is copied to the next generation. The crossover operation involves an exchange of subtrees between two fit models (Fig. 1). This operation is an exchange of genetic material and results in two new models within the next generation. Mutation causes replacement of a subtree of a randomly selected model.

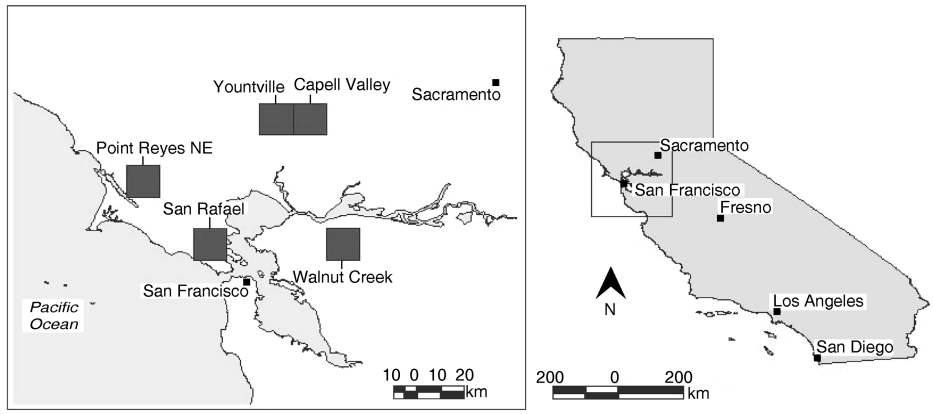

Fig. 2. California study sites.

GP iterates through several generations until it reaches a solution model of 100% accuracy or, more likely, it attains a certain number of generations or reaches a time limit. The result is then the fittest predictive model. More detailed explanations of GP are given in Koza (1992, 1994), Fogel (1999), and Langdon and Poli (2002).

LINEAR DISCRIMINANT ANALYSIS

LDA is a parametric maximum likelihood statistical procedure that classifies observations into one of n classes by using a training sample set to define a solution space for each class. Each dimension of the space corresponds to a different independent variable. Test points are then classified into the class whose solution space centroid is closest to the point itself. Distances in this analysis are generally computed using the Mahalanobis index:

$$D^2 = (\chi_i - \bar{\chi}_j) S^{-1} (\chi_i - \bar{\chi}_j)$$

where x_i is the vector of independent variables for point I, x_j is the mean vector for the jth class and S is the pooled covariance matrix (Marcoulides and Hershberger, 1997).

STUDY SITES

Five 7.5-min. U.S. Geological Survey quadrangle map areas from the San Francisco Bay area of California were selected for use in this study based on data availability, physical diversity and susceptibility to landslides: Yountville, Capell Valley, Point Reyes NE, San Rafael, and Walnut Creek (Fig. 2). Yountville and Capell Valley are contiguous training areas and are considered as one site.

The study sites are physically diverse. They provide a range of lithologies of sedimentary, igneous and metamorphic origin. Point Reyes NE and San Rafael are dominated by sandstone and shale. Yountville and Capell Valley are composed of mafic and some felsic volcanic rocks, serpentinite, sandstone/mudstone/shale, and surficial deposits from alluvial sources and landslides. Walnut Creek is typified by sandstone, conglomerate, mudstone, shale, and some surficial deposits (Wentworth, 1997). This variety of parent material ensures various soil types and therefore a wide range of soil physical properties. The study sites have varied terrain that includes the steeply sloping hills of the coastal ranges and flat alluvial valleys. The terrain suggests a range of landslide-hazard susceptibilties. Although there is a wide variety of slope angles, about one third of the terrain has slopes between 18–35% with the mean slope for the areas ranging between 20% (Walnut Creek) to 28% (San Rafael; Soil Conservation Service, 1977, 1985, 1978).

METHODS

This study compared two techniques for creating landslide-hazard maps, GP and LDA, against a USGS landslide-hazard map. The comparison was evaluated quantitatively using an overall accuracy percentage and the kappa statistic.

Data

All of the GIS data layers for the study were downloaded from Internet sites during 2000–2001. URLs are given in the Appendix. A GIS data layer of landslide-hazard classes developed by USGS scientists was used to evaluate model accuracies (Wentworth et al., 1997). Landslide hazard was rated low (few to no landslides or presence of surficial deposits), medium (landslides greater than 455 m apart) or high (landslides closer than 455 m). Medium and high hazard classes were buffered by an unknown amount.

A list of factors important in the formation of landslides was gleaned from the literature (Howes, 1987; Tsukamoto and Minematsu, 1987; Ritter et al., 1995; Mantovani et al., 1996; California Department of Conservation, 1997; van Westen et al., 1997; Larsen and Torres-Sanchez, 1998; Pack et al., 1998). Based on this list, the following layers were found or derived: aspect, elevation, slope, slope curvature, flow accumulation, soil physical properties (bulk density, rock depth, available water capacity, liquid limit, plasticity index, particle size), lithology, geologic age, land use/land cover (LULC), and distances to streams and roads. Using ArcInfo GIS software (ESRI), the data were standardized to UTM coordinates, zone 10, NAD 27 datum, resampled to 30 m cells and clipped to the study site.

In order to use GP and LDA, the data were processed further into two subsets, a training set for model building and a model-testing set. The training set, from the Yountville and Capell Valley area, was a stratified random sample that for each of the three hazard ratings numbered 50% of the points within the smallest class of the hazard rating layer. This set totaled almost 70,000 points. Data were extracted from every raster layer, including the USGS hazard layer, at each of the points for use with GP and LDA. The testing data set comprised the remaining data set. This

included about 260,000 points from Yountville and Capell Valley, 63,000 from Point Reyes NE, 76,000 from San Rafael, and 48,000 from Walnut Creek. The different numbers of points for each area reflect the presence of water bodies or of no data values. Both GP and LDA were extremely sensitive to missing data values.

GP Methods

This study used Lilgp genetic programming software run on the Sun Solaris operating system (Punch, 1998). Eight unique combinations of independent variables and initial functions were evaluated using GP analysis (Table 1). Initially, each independent variable was used in the GP analysis in its original units of measure. This approach seemed reasonable since, according to Koza (1992), GP is robust to different data units.

Using the premise that a model should begin simply and only increase in complexity as necessary, the first GP configuration contained only four independent variables and two initial functions. This resulted in trivial solution models that identified only constant hazard ratings regardless of the values of the independent variables. In reaction to the trivia, the second GP configuration included 65 independent variables and 14 initial functions. Initial functions included the mathematical, trigonometric, relational, and logarithmic functions. This "try everything" approach resulted in two findings. First, a size restriction on the trees resulting from the GP analysis is necessary to maintain relatively simple trees that can be interpreted and transcribed into the GIS programs to produce landslide hazard maps. As Turton et al. (1996) found, the complex models that can be produced by GP often gave no insight into the data or processes involved. The second approach resulted in a huge search space that was time consuming for the GP program to explore effectively and accuracies were very low (<40%).

In order to reduce the size of the search space, the function and data sets were reduced in size and the data sets were radically simplified. Of the large function set in configuration 2, only the addition, subtraction, multiplication, division and if/then/else functions were retained and tested (Table 1). The land-use/land-cover layer had much lower spatial resolution than the remainder of the data (1:100,000 scale and a minimum mapping unit of 12.5 ha) and contained unclear data categories, so it was dropped completely from the remainder of the analysis. Other layers were categorized into two values (1 or 0) depending upon the perception of their contribution to landslide hazard. For example, the aspect layer was simplified from a 0–360° scale by assigning 0 to values 0–180° and 1 to values 180–360°. This simplification represented differences between windward and leeward hill slopes that might indicate differences in moisture content of the soil and vegetation cover. Two slope layers were created based on thresholds of 10° or 20° (Ellen et al., 1997; Wentworth et al., 1997). If there was no obvious threshold, a cutoff value was computed by using the mean value for each independent variable. The resulting simplified data are described in Tables 2 and 3, and configurations 3–8 use data sets with 8 or 27 of these variables (Table 1). Relative and absolute means were originally calculated. That is, the absolute means were calculated from the Yountville-Capell

Table 1. Genetic Programming Run Configurations[a]

	Configuration							
	1	2	3	4	5	6	7	8
Function								
	+	+	+	+	+	+	+	+
	-	-	-	-	-	-	-	-
		x		x			x	
		/		/			/	
		1/x						
		sqrt						
		pow						
		trig						
		log						
		abc[b]			c[b]			c[b]
Variable								
Aspect		x	x	x	x	x	x	x
Elevation	x	x	x	x	x	x	x	x
Slope (10° or 20°)	x	x	x	x	x	x	x	x
Flow accumulation	x	x	x	x	x	x	x	x
Slope curvature		x	x	x	x	x	x	x
Soil rock depth	x	x	x	x	x	x	x	x
Soil bulk density		x	x	x	x	x	x	x
Soil water capacity		x				x	x	x
Soil liquid limit		x				x	x	x
Soil plasticity index		x				x	x	x
Soil particle sizes		x				x	x	x
Soil clay content		x						
Soil permeability		x						
Presence or absence of geologic formations (16)		x						
Geologic age		x						
Presence or absence of lithologic types (13)		x				x	x	x
Presence or absence of land-use or cover types (13)		x						
Distance to roads		x				x	x	x
Distance to streams		x				x	x	x
Constants: configuration 2 (1,2,3) and configurations 6–8 (1,0)		x				x	x	x
Ephemeral random constant		x						

[a]Configurations 1 and 2 use raw data or presence/absence data; Configurations 3–8 use transformed data or presence/absence data.

[b]"a" represents if...≥...then...else; "b" represents if...<...then...else; "c" represents if...then...else.

Table 2. Data Layer Simplification by Threshold Value

Original data	Original range	Simple range	Rationale for selected cutoff
Aspect	0–360°	0 = 0–180°, 1 = 180–360°	Given prevailing winds, leeward slope = 1 and windward slope = 0.
Slope 10°	0–56°	0 = 0–10°, 1 > 10°	Because of the complexity of the landslide process, a threshold value to indicate slope cutoff could not be picked with absolute accuracy. We used two values from the documentation of the USGS hazard layer: 20° (Ellen, 1997) and 10° (Wentworth, 1997).
Slope 20°	0–56°	0 = 0–20°, 1 > 20°	Because of the complexity of the landslide process, a threshold value to indicate slope cutoff could not be picked with absolute accuracy. We used two values from the documentation of the USGS hazard layer: 20° (Ellen, 1997) and 10° (Wentworth, 1997).
Slope curvature	-1.1–0.8	$0 \leq 0$, 1 > 0	Negative values indicate upwardly concave surface; positive values indicate upwardly convex surface. Not used outside of Yountville and Capell Valley.
Presence or absence of lithologic types (13)	n.a.	n.a.	
Constants: 1 and 0	n.a.	n.a.	

Valley data and the relative means were calculated from each study site. The results were so similar that they are not reported here (Litschert, 2002).

Linear Discriminant Analysis Methods

Linear discriminant analysis (LDA) using SAS (1999) was applied using the training data sets from Yountville and Capell Valley. These data sets consisted of 8 and 27 variables, and 70,000 points: they were exactly the same as those data sets used to train the GP. The resulting models were then applied to all areas using the same 8 and 27 variable testing data sets used in GP analysis.

Overall and Kappa Accuracies

The models produced by GP and LDA were used with the test data to create raster maps. The raster maps were compared cell by cell against the landslide-hazard maps created by the USGS. Overall percentage accuracy was calculated by dividing the total number of correct cells by the total number of cells for each map. The kappa statistic was used to calculate a more telling accuracy percentage by using the numbers of incorrect cells in each predicted hazard class as well as correct cells (Jensen, 1996). Kappa of 100% indicates complete agreement between two hazard maps and 0% indicates no agreement beyond what would be expected by chance.

Table 3. Data Layer Simplification by Mean Value

	Original range				Means[a]			
Original data	Yountville and Capell Valley	Point Reyes NE	San Rafael	Walnut Creek	Yountville and Capell Valley	Point Reyes NE	San Rafael	Walnut Creek
Soil bulk density	0–1.55	0–1.6	0–1.45	1.27–1.7	1.39	0.68	0.64	1.40
Soil water capacity	0–0.185	0–0.180	0–0.180	0.02–0.2	0.13	0.07	0.06	0.16
Soil liquid limit	0–6	0–55	0–53	18–53	26.3	18.8	14.8	38.3
Soil particle size (no. 4 sieve)	0–100	Not used	Not used	Not used	74.1	Not used	Not used	Not used
Soil plasticity index	0–35	0–30	0–25	0–30	8.70	8.09	5.08	18.70
Rock depth (in)	0–60	15–50	0–50	15–50	20.5	32.4	22.1	26.0
Distance to roads (m)	0–1,510	Not used	Not used	Not used	286.0	Not used	Not used	Not used
Distance to streams (m)	0–2,058	Not used	Not used	Not used	422.0	Not used	Not used	Not used
Elevation (m)	12–804	0–422	0–785	0–1,173	316.5	141.7	158.3	165.0
Flow accumulation	0–457,761	0–9,050	0–8,513	0–60,719	578	20.0	32	93

[a]Values below the mean = 0. Values above the mean = 1.

RESULTS AND DISCUSSION

Training and Testing GP at Yountville and Capell Valley

In seeking to optimize the use of GP, it was important to find a workable choice of inputs, including functions and cartographic variables, in order to evolve appropriate models. Of the first two configurations shown in Table 1, the first was too small and produced trivial results; and the second configuration was too large and unwieldy for GP to search and to discover useful models. Configurations 3–8, using the simplified inputs, produced better initial results so they were tested more rigorously and results are reported here.

The GP procedure was trained on the Yountville and Capell Valley site using data sets of both 8 and 27 variables with three different combinations of functions (Table 1). The six GP configurations were each run 10 times to 20 generations, resulting in 60 hazard models. The models were then tested by creating landslide hazard maps using the remaining data sets of 8 and 27 variables for the same location. Models and maps were scored for accuracy by comparison to the USGS hazard layer and the accuracies are shown in Table 4. The highest test accuracy was 80.7% which was attained using the larger data set with two different function sets: {+, -} and {+, -, *, /}. It was apparent from the success of the larger data set that more rather than fewer variables were needed to capture the variability inherent in the landscape in order to predict landslide hazard.

Many of the initial GP runs did not evolve to their highest fitness levels until they reached their 20th generation; these configurations were rerun to 60 generations to test whether this would improve model accuracies (Table 5). The 60 generation runs produced the highest test accuracy of 83.7% using the 27 variable data set and the function set {+, -,*,/}. Figure 3 shows the tree structure of the model that scored the highest accuracy. Lith*x* are lithology data layers (see Table 6 for a description); sl20 is the slope layer where 20 was selected as the data threshold; awc is soil available water capacity; flowacc is flow accumulation, a measure of topographic convergence; and LL is the soil liquid limit. The accuracy of this model was only a 3% improvement over the best accuracy obtained from the 20 generation runs. It is debatable whether increasing the number of generations further would have produced any substantial improvement beyond this 3% since most of the GP runs attained their maximum accuracy before the 60 generation limit was reached (Table 5). Accuracy may have been improved by an increase in the size of each generation; however it was decided not to pursue this since there were already exact replicas of some trees despite different initial generations. In addition to input choices and number of generations, there are several other parameters that the user can vary to improve the performance of GP including tree size, generation size, number of runs, and probability of each breeding operation as described in the section entitled Genetic Programming.

Table 4. Genetic Programming Run Results for 20 Generations

	8-variable data set			27-variable data set		
Run	Training accuracy	Test accuracy	Generation of best algorithm	Training accuracy	Test accuracy	Generation of best algorithm
Function set +, -						
1	65.34	23.6	8	79.03	64.6	18
2	65.34	23.6	11	75.01	49.4	20
3	65.20	42.4	7	77.81	49.1	20
4	65.15	36.9	4	77.23	56.2	20
5	65.15	36.9	2	77.89	80.7	15
6	65.22	36.6	7	76.94	54.8	14
7	65.22	36.6	10	79.39	57.7	20
8	65.22	36.6	9	78.89	73.9	19
9	65.20	26.0	2	78.30	63.4	20
10	65.34	23.6	10	79.68	77.9	19
Highest	65.34	42.4		79.68	80.7	
Average	65.24	32.3		78.02	62.8	
Lowest	65.15	23.6		75.01	49.1	
Function set +, -, *, /						
1	69.74	23.3	20	79.68	78.3	17
2	69.37	64.6	19	80.08	44.9	20
3	68.80	28.3	19	79.88	34.2	20
4	68.49	39.1	20	79.12	21.5	20
5	68.96	55.7	17	79.66	80.1	20
6	68.30	33.5	20	79.63	68.0	19
7	68.94	8.3	20	77.89	80.7	14
8	68.34	45.8	16	80.73	57.3	20
9	68.47	49.8	17	80.71	22.9	20
10	68.52	38.1	14	81.10	68.0	20
Highest	69.74	64.6		81.10	80.7	
Average	68.79	38.7		79.85	55.6	
Lowest	68.30	8.3		77.89	21.5	
Function set +, -, *if*						
1	65.21	9.1	20	76.52	35.9	15
2	65.20	34.4	4	77.68	8.9	18
3	65.34	30.5	2	78.01	22.3	20
4	65.38	18.3	14	77.81	53.6	13
5	65.34	22.5	20	77.20	53.1	17
6	65.34	26.4	7	78.00	52.5	20
7	65.15	30.5	3	77.29	22.3	20
8	65.15	27.3	3	79.68	75.5	17
9	65.15	44.8	3	77.89	37.3	16
10	65.15	34.9	3	77.89	30.7	20
Highest	65.38	44.8		79.68	75.5	
Average	65.24	27.9		77.80	39.2	
Lowest	65.15	9.1		76.52	8.9	

Table 5. Genetic Programming Run Results for 60 Generations

	8-variable data set			27-variable data set		
Run	Training accuracy	Test accuracy	Generation of best algorithm	Training accuracy	Test accuracy	Generation of best algorithm
Function set +, -[a]						
1				79.68	77.9	25
2				79.68	27.7	49
3				78.00	80.7	29
4				79.68	77.9	27
5				78.00	80.7	21
6				78.97	56.1	41
7				79.68	77.9	34
8				79.68	77.9	24
9				79.68	77.9	36
10				79.68	77.9	19
Highest				79.68	80.7	
Average				79.28	71.3	
Lowest				78.00	27.7	
Function set +, -, *, /						
1	69.74	31.7	20	79.92	27.7	50
2	69.75	25.9	49	82.46	8.2	54
3	70.38	33.2	59	83.33	8.2	57
4	69.37	38.3	56	80.71	19.3	59
5	69.39	61.0	29	81.04	59.0	55
6	68.87	30.8	58	80.40	68.4	59
7	70.33	31.1	55	81.01	69.8	47
8	70.23	33.1	50	83.29	23.5	58
9	70.69	63.5	58	82.71	83.7	59
10	68.56	21.3	52	83.27	12.7	60
Highest	70.69	63.5		83.33	83.7	
Average	69.73	37.0		81.81	38.1	
Lowest	68.56	21.3		79.92	8.2	

[a]This set was not run at 60 generations because it attained highest scores before 20 generations.

Testing GP at Other Locations

The 14 models that scored the highest accuracies in Yountville and Capell Valley were selected from the 20 and 60 generation GP runs. It is possible that using models with the highest testing accuracies may have biased results toward GP (Dr. Charles Anderson, professor of computer science, Colorado State University, pers. comm., February 2002). However, in this case several of these models also had high

Table 6. Independent Variables in the Genetic Programing Best 14 Models

Independent variable	Number of uses	Lithx[a]	Yountville/ Capell Valley	Point Reyes NE	San Rafael	Walnut Creek
Sand, gravel, silt, and mud[b]	17	8	11.6%	3.0%	5.2%	3.6%
Clay, silt, sand, gravel[b]	16	2	2.2%	0.3%	1.7%	5.8%
Landslide deposits[b]	16	4	4.0%	0.0%	2.1%	0.0%
Low-grade metasandstone and shale[b]	14	7	5.7%	9.1%	10.5%	0.0%
Serpentinite[b]	14	11	6.1%	0.4%	2.9%	0.0%
Mudstone and shale, some sandstone[b]	12	5	9.2%	0.0%	0.0%	13.1%
Sandstone and conglomerate, some mudstone or shale[b]	11	12	0.2%	7.4%	0.0%	17.4%
Slope 10° or 20°[c]	8		23.7%	26.2%	27.6%	20.4%
Depth to bedrock (in)[c]	5		20.5	32.4	22.1	26.0
Elevation (m)[c]	5		316.5	141.7	158.3	165.0
Sheared sandstone and shale (melange)[b]	4	6	0.5%	79.5%	76.8%	0.0%
Liquid limit index	4		26.3	18.8	14.8	38.3
Tuff, tuffaceous sandstone, some sandstone, volcanic rock[b]	3	13	1.5%	0.0%	0.0%	0.0%
Flow accumulation (cells)[c]	3		578	20	32	93
Mafic volcanic[b]	2	1	29.4%	0.0%	0.0%	0.0%
Aspect (°)[c]	2		169.0	179	164	165.0
Soil bulk density (g cm-3)[c]	2		1.39	0.68	0.64	1.40
Available water capacity (cm cm-1 horizon)	2	0.13	0.07	0.06	0.16	
Felsic volcanic[b]	1	3	2.2%	0.0%	0.0%	0.0%
Sandstone or mudstone or shale[b]	1		27.3%	0.0%	0.0%	58.8%
Plasticity index	1		8.70	8.09	5.08	18.70
Porcelaneous or siliceous mudstone and shale; chert[b]	0	9	0.1%	0.3%	0.9%	1.4%
Slope curvature[c]	0		-0.0001	-0.1960	0.0076	-0.1660
Distance to roads (m)	0		286.0		not used	
Distance to streams (m)	0		422.0		not used	
Particle size (no. 4 sieve)	0		74.1%		not used	

[a]Lithx numbers denote variable names used in Figure 3.

[b]Lithologic values are % data area. Other values are means of data set.

[c]Designates one of the eight terminals in the small data set.

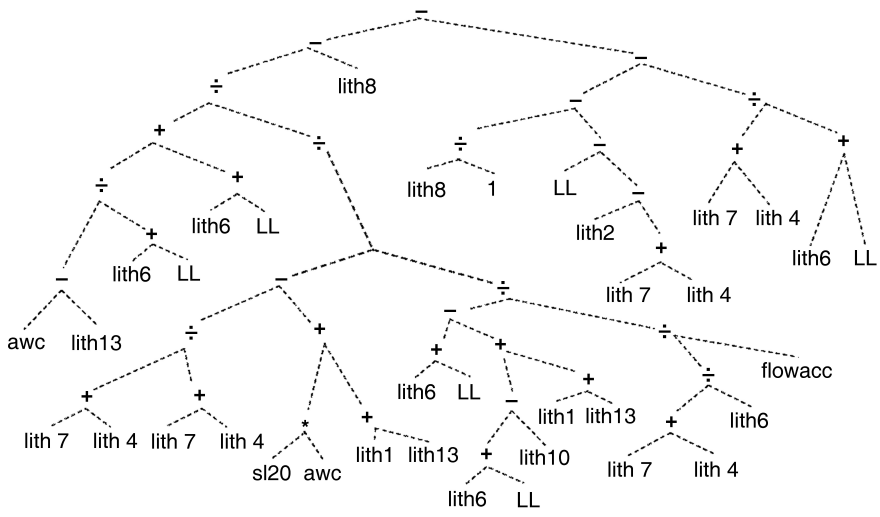

Fig. 3. Tree representing highest scoring genetic programming model (84%).

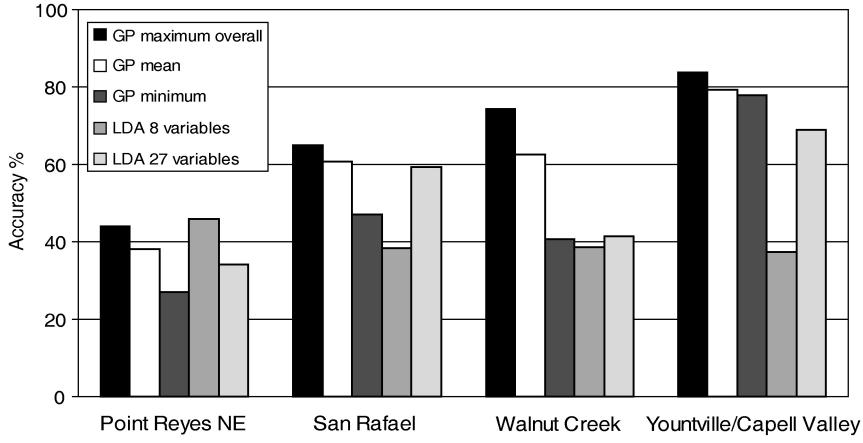

Fig. 4. Accuracy comparison for genetic programming (GP) and linear discriminant analysis (LDA).

training accuracies and would have been chosen for further testing under this different criterion.

The 14 best models were applied to the other three study sites, Point Reyes NE, San Rafael and Walnut Creek, to see if the models would be accurate at other locations. The GP models using the larger data set score highest at Walnut Creek (maximum = 74%, mean = 63%), and lowest in Point Reyes NE (maximum = 44%, mean = 38%; Fig. 4). The physical characteristics of Walnut Creek are quite similar to Yountville and Capell Valley, but Walnut Creek also had the lowest number of

data points, which may have biased results in favor of this area (Table 6). The higher scores at Yountville and Capell Valley probably indicate the importance of local training of the models. All but one of the models uses the larger data set, indicating that the smaller data set may not be adequate to characterize landslide hazard in these areas.

The models were complex and not obvious as representations of the physical reality or as representations of a logical, hazard-rating process. However, some parts of the models are clear, for example the subtraction of the "sand, gravel, silt and mud" variable. This variable would have the value of one in low slope areas that encouraged deposition and would probably not be prone to landslides. Subtraction of this variable in a model would cause the hazard rating to decrease appropriately if the variable is present (i.e., value = 1) or have no effect if the variable is absent (i.e., value = 0).

Table 6 shows the number of times each independent variable was used in the GPs14 best models and the area at the site occupied by each variable. The table demonstrates how GP discovered the importance of some of the lithological variables by the large number of times that they are used. The two variables "clay, silt, sand, gravel" and "sand, gravel, silt, mud" are used the most frequently. These variables define locations of alluvial deposits and are typically found in areas of low slope with little potential for landslide hazard. Such variables were used to define partly the low landslide hazard class on the USGS map.

Further examination of Table 6 reveals an interesting lack of use by GP of some layers. The two slope layers were only used in seven out of the 14 best models: similarly, the soil physical properties were rarely used. At least two possibilities exist here. The data simplifications performed on each layer may have used inappropriate threshold values. Secondly, as Gritzner et al. (2001) also suggested, the resolution of the spatial data at 30 m cell size may have been too coarse, resulting in reduced variability of data values.

Comparison of GP and LDA

GP performs better than LDA at all locations except Point Reyes NE where the accuracies are also the closest: the LDA result for Point Reyes NE is 46% compared to a GP accuracy of 44% (Fig. 4). Point Reyes seems to be an anomaly. It is the only site where (1) the small dataset scored higher than the large dataset (using LDA by 12%) and (2) LDA scored higher than GP (by 2%). As mentioned earlier, Point Reyes has quite different physical characteristics from the other study sites. Whereas these differences do not necessarily explain these accuracy scores, they might be a contributing factor.

LDA scores for the large data set reflect a similar pattern to GP models, with Yountville and Capell Valley scoring the highest (69%) and Point Reyes NE scoring the lowest (34%) (Fig. 4). LDA results for the smaller data set are confusing since Point Reyes NE is the highest at 46% and Yountville and Capell Valley is the lowest at 37%. The low scores, which are not much higher than what might be expected from a random classification of the landscape, may indicate, as with the GP results, that the smaller data set was inadequate for landslide-hazard zoning.

Table 7. Error matrices for best Genetic Programming (GP) and Linear Discriminant Analysis (LDA)

Functions: +, -,*, /, Generations: 60	Low	Medium	High	Row total	Omission	Producer
			GP: 27-variable data set			
Low	17,715	14	3,502	21,231	3,516	83.4%
Medium	3,674	188,116	11,150	202,940	14,824	92.7%
High	274	23,859	12,012	36,145	24,133	33.2%
Column total	21,663	211,989	26,664	260,316		
Commission	3,948	23,873	14,652			**Overall = 83.7%**
User	81.78%	88.74%	45.05%			**Kappa = 52.6%**
			LDA: 27-variable data set			
Low	21,216	118	1	21,335	119	99.4%
Medium	8	128,919	6,811	135,738	6,819	95.0%
High	7	73,903	29,333	103,243	73,910	28.4%
Column total	21,231	202,940	36,145	260,316		
Commission	15	74,021	6,812			**Overall = 68.9%**
User	99.93%	63.53%	81.15%			**Kappa = 41.6%**

Lower LDA scores might be attributed to the unfulfilled needs of LDA. LDA requires data to be normally distributed: similar to many geographic data, the data layers used in this study did not have normal distributions. LDA may require data to be in their usual wider physical range of values in order to classify data points distinctly and not be simplified to only two values.

Kappa Statistic of Yountville and Capell Valley

The kappa statistic was run for the best GP and LDA overall results at Yountville and Capell Valley. Error matrices are shown in Table 7: as expected GP (53%) performed better than LDA (42%). It was apparent from the error matrices that the low hazard class was more distinct than the medium and high classes. This was probably partly due to the use of data layers describing surficial deposits to outline the low hazard class. Medium and high hazard classes were delineated using an entirely different process: actual landslide polygons were buffered and rated depending on proximity. Both LDA and GP examined physical data layers on an individual cell basis, so it was unlikely that these techniques could determine any such spatial relationships. Given the different ways of tackling this issue, it was not surprising that neither GP nor LDA attained close to 100% accuracy when compared to the USGS hazard layer.

Qualitative Comparison of GP and LDA at Yountville and Capell Valley

Best GP, best LDA, and USGS hazard layers for Yountville and Capell Valley are shown in Figure 5. Both GP and LDA mapped the low hazard class well,

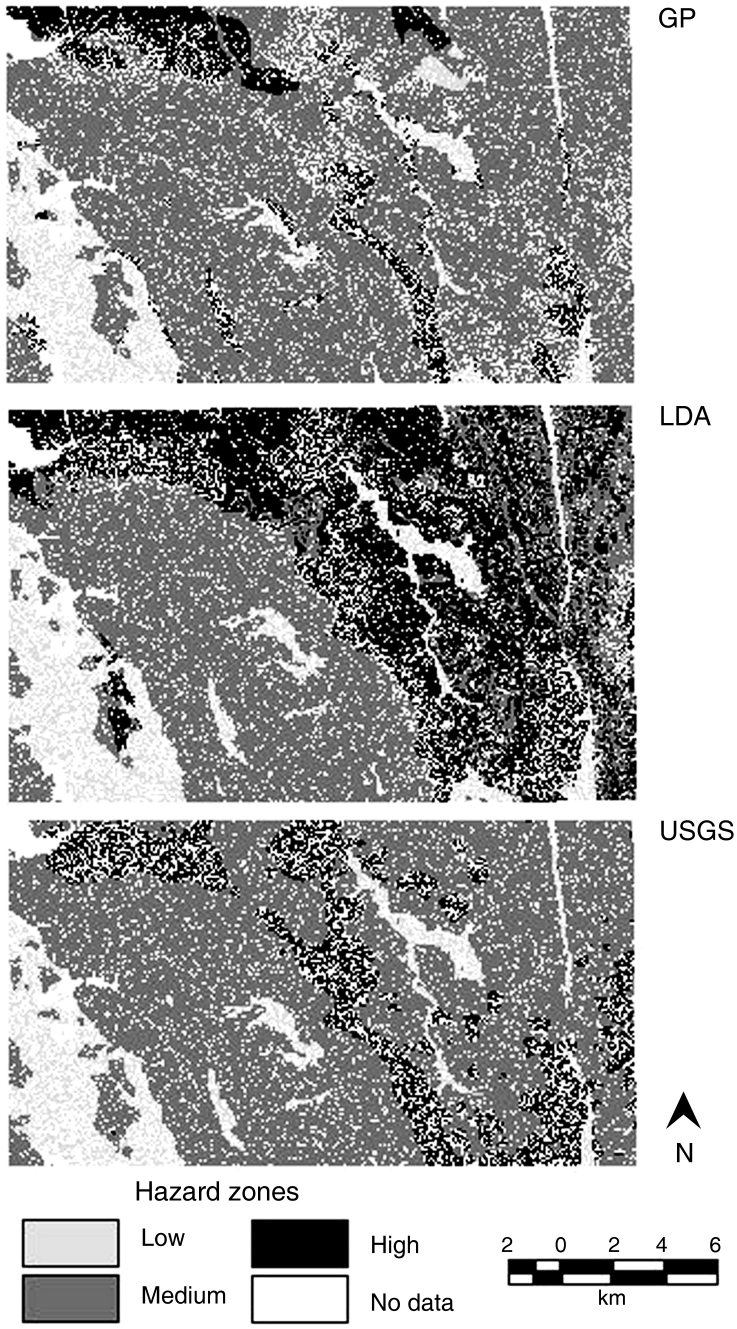

Fig. 5. Hazard maps for Yountville and Capell Valley. Speckled appearance is a result of absence of training data.

particularly in Napa Valley that dominates the west side of Yountville and Capell Valley. GP delineates quite accurately the medium hazard class throughout the central and east sides whereas on the LDA map the high hazard class erroneously dominates the eastern and northern portions.

CONCLUSION

This study has described the comparison of landslide hazard maps created using GP and LDA to a USGS hazard map. GP proved eventually to be a very flexible tool but it required a large effort initially to hone it for this study. Eventually GP was moderately successful, attaining at best 84% accuracy and performing better than LDA 3 out of 4 times. The larger data sets were required to capture more landscape variability to produce more accurate models. For this study, local training data and similarity of study sites proved to be important factors in model accuracy. Despite the proximity of Point Reyes NE to the Yountville and Capell Valley training sites, the Point Reyes NE maps showed the lowest accuracy when compared the USGS hazard maps.

Possibilities are available to increase GP performance further, for example by reducing sensitivity to areas of no data; by extracting data from surrounding cells where appropriate to the application; or by using data of finer spatial resolution. With further work, GP may be a useful technique to create preliminary landslide hazard maps at the regional scale that can then be refined as necessary by more detailed site specific work.

Acknowledgments: I thank Denis Dean of Colorado State University for his contributions to this study, and Chuck Anderson and Lee MacDonald, also of CSU, for some enlightening conversations.

REFERENCES

California Department of Conservation, Division of Mines and Geology (1997) *DMG Note 50: Factors affecting landslides in forested terrain*. Available from the California Department of Conservation Web site http://www.conserv.ca.gov/dmg/pubs/notes/50/index.htm

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*, Vol. 16, 427–445.

Clouatre, E., Dubois, J. M., and Poulin, A. (1996) The geographic information system and regional delimitation of zones at risk for landslides: Hull-Gatineau region, Quebec. *Canadien Geographer-Geographe Canadien*, Vol. 40, 367–386.

Dai, F. C. and Lee, C. F. (2003) A spatio-temporal probabilistic modelling of storm-induced shallow landsliding using aerial photographs and logistic regression. *Earth Surface Processes and Landforms*, Vol. 28, 527–545.

Daida, J. M., Hommes, J. D., Ross, S. J., and Vesecky J. F. (1995) Extracting curvilinear features from synthetic aperture radar images of arctic ice: Algorithm discovery using the genetic programming paradigm. *Proceedings of the 1995*

International Geoscience and Remote Sensing Symposium: Quantitative Remote Sensing for Science and Applications, July 10–14, 1995, Firenze, Italy.

Daida, J. M., Bersano-Begey, T. F., Ross, S. J., and Vesecky, J. F. (1996) Computer aided design of image classification algorithms: Dynamic and static fitness evaluations in a scaffolded genetic programming environment. In J. Koza, D. Goldberg, D. Fogel, and R. Riolo, eds., *Genetic Programming 1996: Proceedings of the first Annual Conference*. Stanford University, California. MIT Press, Cambridge, MA.

Dhakal, A. S., Amada, T., and Aniya, M. (2000) Landslide hazard mapping and its evaluation using GIS: An investigation of sampling schemes for a grid based quantitative method. *Photogrammetric Engineering and Remote Sensing*, Vol. 66, 981–989.

Dietrich, W. and Montgomery, D. (1998) *SHALSTAB: A Digital Terrain Model for Mapping Shallow Landslide Potential*. Available from the Socrates Web site http://socrates.berkeley.edu/~geomorph/shalstab/index.htm

Diplock, G. (1998) Building new spatial interaction models by using genetic programming and a supercomputer. *Environment and Planning A*, Vol. 30, 1893–1904.

Ellen, S. D., Mark, R., Wieczorek, G., Wentworth, C., Ramsey, D., and May, T. (1997) *Open-File Report 97-745E, San Francisco Bay Region Landslide Folio Part E, Map of Debris-Flow Source Areas in the San Francisco Bay Region, California*. Available from the Western Region Geologic Information Web site http://wrgis.wr.usgs.gov/open-file/of97-745e.html

Environmental Systems Research Institute (ESRI) Arc 8.2. (2002) *ESRI Arc 8.2*. Redlands, CA: Author.

Fogel, D. (1999) *Evolutionary Computing: Toward a New Philosophy of Machine Intelligence*. New York, NY: IEEE Press.

Gritzner, M. L., Marcus, W. A., Aspinall, R., and Custer, S. G. (2001) Assessing landslide potential using GIS, soil wetness modeling and topographic attributes, Payette River, Idaho. *Geomorphology*, Vol. 37, 149–165.

Guzzetti, F., Carrara, A., Cardinali, M., and Reichenbach, P. (1999) Landslide hazard evaluation: A review of current techniques and their application in a multi-scale study, central Italy. *Geomorphology*, Vol. 31, 181–216.

Guzzetti, F., Carrara, A., Cardinali, M., and Reichenbach, P. (2000) Comparing landslide hazard maps: A case study in the upper Tiber river basin, central Italy. *Environmental Management*, Vol. 35, 247–263.

Howes, D. (1987) A method for predicting terrain susceptible to landslides following forest harvesting: A case study from the southern Coast Mountains of British Columbia. *Forest Hydrology and Watershed Management. (Proceedings of the Vancouver Symposium, August 1987)*. Wallingford, UK: IAHS, *IAHS-AISH Publication 167*.

Jan, J. (1997) Classification of Remotely Sensed Data Using Adaptive Machine Learning Methods. Unpublished dissertation, Colorado State University, Fort Collins, CO.

Jensen, J. (1996) *Introductory Digital Image Processing*. Upper Saddle River, NJ: Prentice-Hall.

Jordan, C. J., O'Connor, E. A., Merchant, A. P., Northmore, K. J., Greenbaum, D., McDonald, A. J., Kovacik, M., and Ahmed, R. (2000) Rapid landslide susceptibility mapping using remote sensing and GIS modelling. *Proceedings of the 14th International Conference on Applied Geologic Remote Sensing*, Vol. 14, 113–120. Ann Arbor, MI: Veridam ERIM International.

Koza, J. (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press.

Koza, J. (1994) *Genetic Programming II: Automatic Discovery of Reusable Programs*. Cambridge, MA: MIT Press.

Langdon, W. B. and Poli, R. (2002) *Foundations of Genetic Programming*. Berlin, Germany: Springer-Verlag.

Larsen, M. and Torres-Sanchez, A. J. (1998) The frequency and distribution of recent landslides in three montane tropical regions of Puerto Rico. *Geomorphology*, Vol. 24, 309–331.

Litschert, S. (2002) A Comparison of Genetic Programming and Linear Discriminant Analysis Techniques to Identify Regions at Risk for Landslides. Unpublished master's thesis, Colorado State University, Fort Collins, CO.

McKay, R. I. (2001) Variants of genetic programming for species distribution modeling—Fitness sharing, partial functions, population evaluation. *Ecological Modelling*, Vol. 146, 231–241.

Mantovani, F., Soeters, R., and van Westen, C. (1996) Remote sensing techniques for landslide studies and hazard zonation in Europe. *Geomorphology*, Vol. 15, 213–225.

Marcoulides, G. and Hershberger, S. (1997) *Multivariate Statistical Methods: A First Course*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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. *ITC Journal*, Vol. 3, 292–300.

Nguyen, T. B. T. (2000) Application of remote sensing data and GIS techniques to land hazard zonation mapping in Vanchan area (northwest part of Vietnam). *Technical Bulletin—Coordinating Committee for Coastal and Offshore Geosciences Programmes in Est and Southeast Asia*, Vol. 29, 39–44.

Pack, R. T., Tarboton, D. G., and Goodwin, C. N. (1998) *The SINMAP Approach to Terrain Stability Mapping*. Paper submitted to 8th Congress of the International Association of Engineering Geology, Vancouver, British Columbia, Canada. Available from the SINMAP Web site http://moose.cee.usu.edu/sinmap/sinmap.htm

Paris, M. and Wasowski, J. (1999) Landslide activity maps for landslide hazard evaluation: Three case studies from southern Italy. *Natural Hazards*, Vol. 20, 159–183.

Punch, B. (1998) Lilgp software and documentation. Available from the Michigan State University Genetic Algorithms Research and Applications Group Web site http://garage.cps.msu.edu/software/lil-gp/lilgp-index.html

Ritter, D. F., Kochel, R., and Miller, J. R. (1995) *Process Geomorphology*. Dubuque, IA: Wm. C. Brown.

Sakellariou, M. G. and Ferentinou, M. D. (2001) GIS based estimation of slope stability. *Natural Hazards Review*, Vol. 2, 12–21.

SAS. (1999) *SAS Statistical Software 8.0*. Cary, NC: SAS Institute.

Soil Conservation Service (SCS) U.S. Department of Agriculture. (1977) *Soil Survey of Contra Costa County, California*. Davis, CA: Natural Resources Conservation Service.

Soil Conservation Service (SCS) U.S. Department of Agriculture. (1978) *Soil Survey of Napa County, California*. Available from the National Resources Conservation Service Web site http://www.ca.nrcs.usda.gov/mlra/NapaSS/napass.html

Soil Conservation Service (SCS) U.S. Department of Agriculture. (1985) *Soil Survey of Marin County, California*. Davis, CA: Natural Resources Conservation Service.

Swan, A. and Sandilands, M. (1995) *Introduction to Geologic Data Analysis*. Malden, MA: Blackwell Science.

Turton, I., Openshaw, S., and Diplock, G. (1996) Some geographical applications of genetic programming on the Cray T3D supercomputer. C. R. Jessope and A. V. Shafarenko, eds., *Proceedings of UKPAR '96*, Berlin, Germany: Springer, 135–150.

Tsukamoto, Y. and Minematsu, H. (1987) Evaluation of the effect of deforestation on slope stability and its application to watershed management. *Forest Hydrology and Watershed Management. (Proceedings of the Vancouver Symposium, August 1987)*. Wallingford, UK: IAHS, *IAHS-AISH Publication 167*.

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

Wentworth, C. M. (1997) *General Distribution of Geologic Materials in the San Francisco Bay Region, California: A Digital Map Database*. Denver, CO: United States Geological Survey, *Open-File Report 97-774*.

Wentworth, C., Graham, S., Pike, R., Beukelman, G., Ramsey, D., and Barron, A. (1997) *Summary Distribution of Slides and Earth Flows in Napa County, California*. Denver, CO: United States Geological Survey, *Open-File Report 97-745C*.

Whigham, P. A. (2000) Induction of a marsupial density model using genetic programming and spatial relationships. *Ecological Modelling*, Vol. 131(2–3), 299–317.

Wieczorek, G. F. (1984) Preparing a detailed landslide-inventory map for hazard evaluation and reduction. *Bulletin of the Association of Engineering Geologists*, Vol. 21, 337–342.

APPENDIX

Data URLs (Accessed September 2000–March 2001)

Napa County Soil Survey: http://www.ca.nrcs.usda.gov/mlra/NapaSS/napass.html

Other Soils—Contra Costa, Marin Counties: http://www.statlab.iastate.edu/soils/nssc/

Land Use: http://www.consrv.ca.gov/dlrp/fmmp

DEMs (digital elevation models): http://bard.wr.usgs.gov/

DLGs (digital line graphs): http://www.usgs.gov/

Geology: http://wrgis.wr.usgs.gov/open-file/of97-744

Landslides: http://bard.wr.usgs.gov/