

IDENTIFYING REGIONS AT RISK FOR LANDSLIDES USING COMBINED GIS AND GENETIC ALGORITHM PROCEDURES

Sam Litschert
Master of Science Candidate
Geomatics Program
Department of Forest Sciences
Colorado State University

Denis Dean
Associate Professor
Geomatics Program
Department of Forest Sciences
Colorado State University

ABSTRACT

Evaluating the susceptibility of land to various natural hazards (such as landslides, wildfire, flooding, and so on) is an obvious factor in evaluating land use suitability. In a GIS environment, such hazard susceptibility evaluations are typically accomplished by using standard statistical techniques to find relationships between expertly-appraised hazard rating levels and a variety of cartographic variables (such as elevation, soil types and conditions, proximity to other important physiographic features, etc.). Using standard GIS techniques, these cartographic variables can then be derived for regions that have not been hazard-rated by experts, and estimated hazard ratings for these regions can be obtained using these cartographic variables and the statistical relationship derived earlier.

There are a number of limitations to procedures such as this, not the least of which is the inherent limitations of most statistical procedures when applied to spatial data. This study will evaluate the alternative of using genetic algorithms instead of statistical techniques to derive relationships between cartographic variables and hazard ratings. Using a database describing landslide hazard ratings as an example, both statistical and genetic algorithm techniques will be used to identify relationships between hazard ratings and cartographic variables. The absolute and relative accuracies of both techniques' predictions will be compared, and the strengths and weaknesses of both techniques will be identified.

INTRODUCTION

Increased global population has led to increased urban development. Some of this development has encroached on locations prone to geologic hazards, and in some cases, development may have even increased the risk of these hazards. For example, deforestation and road construction in unstable areas have greatly increased the incidence of landslides (Tsukamoto, 1987; Larsen, 1997, quoted in Larsen 1998). In recent years landslides have become particularly prevalent and disastrous, especially in more vulnerable, developing nations. For example, intense rainfall in Venezuela during December 1999 caused devastating flash floods and landslides on steep slopes, particularly in areas where the landscape was denuded by development. According to U.S. Agency for International Development, these events resulted in about 30,000 deaths and 65,000 houses destroyed (USAID, 2000). Events of this magnitude are fortunately rare, but less disastrous but still destructive and dangerous landslides are becoming increasingly common in populated areas.

The process and geologic principles at work within landslides have been extensively studied, and a highly specialized terminology has been developed. However, for the purposes of this study, a general definition will suffice. For our purposes, a landslide can be defined as “a movement of a mass of rock, earth or debris down a slope” (Cruden 1991, quoted in Dikau). This definition encompasses what geologists refer to as slides, flows, and creeps.

It is usually a combination of several factors that cause a landslide. These causative factors can be divided into natural and anthropogenic categories. Common natural factors include rainfall, topography, lithology, bedrock structure and faulting, soil type, soil permeability, soil porosity, soil cohesion, clay content, shrink-swell characteristics, and soil depth (Dikau 1996; Kalvoda 1998; Dunne 1978; Tsukamoto 1987). Anthropogenic causes are any human activities that increase the mass of material resting on landslide-prone slopes, decrease the resistive forces holding landslide-prone materials in place, or alter the hydrologic conditions present in a landslide-prone area.

Given the increasing prevalence of development in landslide-prone areas, many experts have constructed landslide hazard maps to assist planners and political leaders. Traditional methods of landslide hazard mapping have been based on extensive fieldwork by expert geologists in potentially dangerous areas. This is a slow, expensive and very labor intensive operation, and as such can not be widely applied. With the increasing availability of high-resolution spatial data sets, GIS, and computers with large and fast processing capacity, it is becoming possible to partially automate the landslide hazard mapping process and minimize fieldwork (Mejia-Navarro 1993). However, due to the difficulties inherent in generalizing and quantifying landslide causative factors, no practical hazard zoning technique that eliminates the need for extensive expert input has been developed. Several studies have used GIS and statistics to evaluate landslide hazards, but none of these studies have successfully produced hazard maps without site-specific expert input (Howes 1987; van Westen 1997).

The purpose of this study is to investigate a new approach to developing partial landslide hazard maps that eliminates the need for extensive expert input. We intend to evaluate the use of GIS combined with genetic algorithms (GAs) in determining the relationship between geographic variables and slope instability, and hence predict areas that may be highly susceptible to landslides. Note that predicting slope instability is only part of predicting landslide hazards.

According to van Westen et al. (1997), landslide hazard ratings consists of both an assessment of slope instability and a determination of the probability that landslide triggering conditions will occur. This study focuses exclusively on the first step.

STUDY SITE

The study area for this project consists of two 7.5-min quad sheets (Yountville and Capell Valley) in Napa County, California. These were chosen because they provide a contiguous area with a large range of cartographic variables and a variety of landslide potential. The lithology of Napa County varies with several different formations of igneous, sedimentary and metamorphic origin. The valleys are filled with alluvial and colluvial deposits from adjacent bedrock. Several fault lines are apparent. Rocks are of different ages and have different resistances to physical and chemical weathering processes. The variety of parent material for soils ensures heterogeneous soil types. Soil formation occurs through rapid chemical weathering processes in the spring and early summer due to warm, moist conditions.

The study site has varied terrain – steeply sloping hills and flat alluvial valleys. The elevation ranges from 130ft to 8041 ft. Although there is a wide variety of slope angles, about one third of the terrain has slopes between 18 – 35%. Population centers are mainly located in the valleys with some roads traversing the hills. The area has a Mediterranean climate with warm, dry summers and cool wet winters. Snow can occur at higher elevations.

DATA SET

Landslide hazard ratings for Napa County were developed and mapped in of a series of studies conducted by Western Region USGS geologists. These hazard maps cover the Yountville and Capell Valley quads that constitute the current study area. The hazard ratings from these maps will be used as the “true” ratings to which our model will be compared.

Ideally, we plan to use the following variables to predict hazard ratings:

- ?Slope angle and length
- ?Aspect
- ?Elevation
- ?Distance to roads and railroads
- ?Distance to streams
- ?Drainage configuration
- ?Geology – lithology and age, structure, distance to faults
- ?Soils – type, depth, permeability, clay content, shrink/swell characteristics, unit weight
- ?Land use / land cover.

Efforts are currently underway to locate data sources for all of these variables. The Internet has proved to be an invaluable source for the digital data needed for this study. Data has been downloaded from several web sites administered by the U.S. Geological Survey, the Natural Resources Conservation Service, University of California at Santa Barbara, and the U.S. Census Bureau. Most, but not all, of the downloaded data was in raster formats based on 30m.-by-30m. cells. According to Mantovani (1996), 30m. cell size should provide enough detail for a medium scale analysis to predict moderately sized landslides. Unfortunately, not all of the downloaded data was available at 30m. resolution; some was available only at 100m.-by-100m cell sizes and had to be resampled to 30m. At 30m cell size, the study area provides about 343,000 (468 rows by 732 columns) pixels for analysis.

Downloaded data needs considerable preprocessing before it is suitable for analysis. One or more of the following processing steps were applied to each of the downloaded data files. Most of the data was projected in UTM Zone 10 coordinated based on the NAD 27 datum. Data not in projected this way were reprojected to this system. All vector data was converted to raster format. Large data sets were clipped to fit the boundary of the study site. In several cases it is necessary to relate or join scalar data tables to spatial coverages.

Not all of the required data has yet been found, but we are confident that we will be able to secure all of the data listed previously.

GENETIC PROGRAMMING

Once the required data set has been constructed, preprocessed and put into appropriate formats (all of these steps will be accomplished using standard GIS software and existing spatial data analysis and processing techniques), genetic algorithms (GAs) will be used to predict (on a cell-by-cell basis) landslide hazard ratings from the independent variables discussed in the previous section. Genetic algorithms have been applied to subjects as diverse as timber harvesting, crop management and aerospace engineering. GAs can be used wherever there is a complex interrelationship between a number of variables that describe or predict a phenomenon. Unlike classical classification techniques based on statistical approaches, GAs make no assumptions about the statistical distributions of the variables involved in the analysis nor the mathematical form of the relationship between the dependent and independent variables. In addition, they lend themselves to the development of hierarchical prediction rules that mimic the type of decision making usually used by human experts. However, unlike expert system analysis, GA analysis derives these rules from available data; it does not depend on a knowledge engineer and one or more experts to formulate these rules *a priori*. According to Jan (1997), the GA approach to problem solving incorporates an important learning ability that is not available with other methods. It is hoped that the GA method will allow us to succeed where more traditional predictive modeling methods have failed and accurately predict landslide hazard ratings from the available data.

The genetic algorithm approach is designed to mimic Charles Darwin's model of evolution by simulating the natural processes of reproduction, mutation, competition and selection. These processes account for the perpetual variation, survival and improvement of all natural

populations by survival of the fittest members of the population (Fogel, 1996). Each variable that describes a process (in this case, each variable that contributes to triggering a landslide) corresponds to a gene. A string of genes or a chromosome represents a raster cell in our data set and contains information that can be used to predict the likelihood that the cell is prone to landslides.

Genetic programming is used to implement genetic algorithm concepts. Genetic programming starts with a series of simple classification algorithms. In effect, these algorithms may indicate that if predictive variable X has a particular value, assign the cell a landslide hazard rating of Y , or if the difference between predictive variables A and B is greater than C , assign the cell a landslide hazard rating of Z . Once an initial set of algorithms is created, the programming process proceeds iteratively, with each iteration referred to as a generation. In a given generation, each classification algorithm is applied to all of the chromosomes in the data set, and the “fitness” of the algorithm (i.e., its ability to correctly predict hazard ratings from the independent variables) is evaluated. At the end of the generation, the “fitter” classification algorithms are more likely to be copied into the set of algorithms that will be used in the next generation (this process is referred to as reproduction). By biasing reproduction in favor of the “fitter” algorithms, algorithms that do not do a good job of predicting hazard ratings from the independent variables (i.e., the less fit algorithms) tend to fall out of the set of algorithms.

In addition to reproductions, algorithms can combine with one another to form hybrid algorithms (this process is called crossover) and mutate into entirely new algorithms (mutation). Given these processes, the set of algorithms continually changes from generation to generation, and over multiple generations, only algorithms that do a good job predicting hazard ratings survive.

After some number of generations (or some other termination criteria is reached), the genetic programming process is stopped and the current set of classification algorithms is returned. These algorithms can then be applied to any set of independent variable values to produce predicted landslide hazard ratings.

PROGRESS TO DATE AND FUTURE PLANS

At the present time, data is still being collected and preprocessed for use in this study. It is expected that the final data set will be developed by January 2001. The data will then be divided into model building and model validation subsets. The model building subset will be used to construct a genetic programming-based classification system that predicts landslide hazard ratings from the set of available independent variables. In addition, the model building subset will be used to build a second landslide hazard rating prediction system based on standard regression and/or discriminant analysis procedures. Both classification systems will then be applied to the validation subset, and the resulting predicted landslide hazard ratings will be compared to the known ratings for this area. The absolute accuracies (i.e., how well each predictive model did in predicting known landslide hazard ratings) and the relative accuracies (i.e., how well the two prediction systems performed relative to each other) of the two predictive models will be noted. These accuracy assessments will be used to evaluate the success of the combined GIS/genetic programming approach.

LITERATURE CITED

- Dikau, R., D. Brunsten, L. Schrott, and M. Ibsen, (Eds.). 1996. *Landslide Recognition: Identification, Movement and Causes*, John Wiley & Sons, New York, NY.
- Dunne, Thomas, and L. Leopold, 1978. *Water in environmental planning* W.H. Freeman, San Francisco, CA.
- Fogel, David B., 1995. *Evolutionary Computing: Toward a New Philosophy of Machine Intelligence*, IEEE Press, New York, NY.
- Jan, J., 1997. *Classification of Remotely Sensed Data Using Adaptive Machine Learning Methods*. Diss., Colorado State University.
- Kalvoda, Jan, and Charles Rosenfeld, (Eds.) 1998. *Geomorphological Hazards in High Mountain Areas*. Dordrecht, Kluwer Academic, Boston, Ma, 1998.
- Larsen, Matt, and Angelo Torres-Sanchez, 1998. The Frequency and Distribution of Recent Landslides in three Montane Tropical Regions of Puerto Rico. *Geomorphology* 24(1998) 309-331.
- Mantovani F., R. Soeters, and Cees. J. Van Westen, 1996. Remote Sensing Techniques for Landslide Studies and Hazard Zonation in Europe. *Geomorphology*. 15 (3-4): 213-225
- Mejia-Navarro, Mario, 1993. *Geological hazard and risk evaluation using GIS : methodology and model applied to Medellin, Columbia*. Thesis, Colorado State University.
- Ritter, D.F., R.C. Kochel, and J.R. Miller, 1995. *Process Geomorphology*, Wm. C. Brown Publishers, Dubuque, IA.
- Tsukamoto, Yoshinori and Hirohiko Minematsu, 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*. Publ. No. 167, 1987.
- USAID. 2000. Venezuela - Floods Fact Sheet #11 (FY 2000). 4/10/2000
www.info.usaid.gov/hum_response/ofda/venezfl_fs11_fy00.html
- Van Westen, Cees J., N. Rengers, M.T.J. Terlien, and R. Soeters, 1997. Prediction of the Occurrence of Slope Instability Phenomena through GIS-based Hazard Zonation. *Geologische Rundschau*. 86:404-414.