Probabilistic Prediction Models for Landslide Hazard Mapping

Chang-Jo F. Chung and Andrea G. Fabbri

Abstract

A joint conditional probability model is proposed to represent a measure of a future landslide hazard, and five estimation procedures for the model are presented. The distribution of past landslides was divided into two groups with respect to a fixed time. A training set consisting of the earlier landslides and the geographical information system-based multi-layer spatial data in the study area was used to construct the prediction maps. The predictions were then cross-validated by comparing them with the remaining later landslides. When the database falls short of providing sufficient support for the prediction, the model allows the introduction of the expert's knowledge to modify the observed frequencies of the landslides with respect to the spatial data. The additional information should improve the prediction results. A case study from the Rio Chincina region in Colombia was used to illustrate the methodologies.

Introduction

Using spatial data sets based on geographical information systems (GIS) quantitative prediction models have been proposed for landslide hazard mapping (Wang and Unwin, 1992; Carrara *et al.*, 1992; Chung and Fabbri, 1993; van Westen, 1993; Jibson *et al.*, 1998). We propose a unified probabilistic framework for predictive modeling using GIS-based multi-layer spatial data. In the probability models for the prediction of landslide hazard, the hazard at each point or pixel is considered as the joint conditional probability that the pixel will be affected by a future landslide given (conditional to) the information from the spatial input data at the pixel. We present five estimation procedures for the models and also offer a new strategy for visualizing, interpreting, and validating the results of predictions.

The five procedures are (1) direct estimation of the joint conditional probability for every pixel based on the past landslides; (2) estimation of the bivariate conditional probabilities for the thematic classes in each layer using the past landslides and then, based on them, computation of the joint conditional probability at each pixel by the Bayesian formula under the conditional independence assumption; (3) estimation as in (2) of the bivariate conditional probabilities for the thematic classes in each layer but under the assumption that the joint conditional probability for every pixel is a linear function of the bivariate conditional probabilities (the linear function is estimated using regression analysis); (4) estimation identical to (2) except that the estimated bivariate conditional probabilities using the past landslides are modified using expert's knowledge before being used to compute the joint conditional probability; and (5) the combination of (3) and (4), again assuming that the joint conditional probability for every pixel is a linear function of the modified bivariate conditional probabilities (here, too, the linear function is estimated using regression analysis).

Bayesian formulas for geologic prediction models were used by Spigelhalter (1986) and Agterberg *et al.* (1990). Chung and Fabbri (1993) have adapted the formulas for geologic hazard zonation as a part of "favorability function" approaches, and the method has been applied to landslide prediction by Chung and Leclerc (1994), Leclerc (1994), Luzi (1995), and Luzi and Fabbri (1995). Multivariate regression analysis for landslide hazard was proposed by Carrara (1983), Carrara *et al.* (1992), and more recently by Chung *et al.* (1995).

Although some layers of spatial data represent continuous measurements, such as slope angles and distances, as discussed by Chung *et al.* (1995), a map layer containing continuous measurements is usually converted into a number of classes, i.e., "thematic classification," for producing a new map representing geologic hazard. In general, we may assume that each layer represents a classification map containing a number of thematic classes. A case study from a region in central Colombia, which is affected by rapid debris avalanches, is used to compare these five procedures.

Study Area and Test Data Set in the Rio Chincina Area in Central Colombia

The catchment of the Rio Chincina, located on the western slope of the central Andean mountain range (Cordillera Central) in Colombia, near the Nevado del Ruiz volcano, was used as a test for various landslide hazard zonation techniques. Van Westen (1993) made an extensive study of the region and constructed the database of the study area. Since then it was made available as an "ideal" case-study data set for many kinds of exercises and experiments on landslide hazard zoning by van Westen *et al.* (1993), with the name of GISSIZ: training package of Geographic Information Systems on Slope Instability Zonation. It is with that data set that Chung *et al.* (1995) applied a variety of methods of multivariate regression and reviewed some of those settings as the basis of the analysis. This study broadens the approach to a comparison of other methods in which data-driven approaches and knowledge-driven approaches are considered in isolation and in combination to identify the most successful strategies for hazard prediction.

The input data for landslide hazard zonation consist of several layers of map information. Each layer may be the result of map updating by experts, of field verification, and of interpretation of aerial photographs. The prepared maps for the analysis

C-J F. Chung is with the Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario K1A 0E8, Canada. (chung@gsc.nrcan.gc.ca).

A. G. Fabbri is with the International Institute of Aerospace Survey and Earth Sciences (ITC), Hengelosestraat 99, P.O. Box 6, 7500 Enschede, The Netherlands, (e-mail: fabbri@itc.nl).

Photogrammetric Engineering & Remote Sensing
Vol. 65, No. 12, December 1999, pp. 1389–1399.

0099-1112/99/6512–1389$3.00/0

usually describe surficial and bedrock geology, including shear-strength measurements for geologic units (Jibson *et al.*, 1998), soil type, slope, land use, geomorphology, mass movements, distance from active faults, and other features which are relevant to slope instability. The preparation and the selection of input layers for the analysis are obviously a crucial and important component of building prediction models for landslide hazard, but these are not the subject of this paper. In addition, the identification of types and dates of landslide phenomena is critical to the application of predictive techniques.

For the Rio Chincina study area, van Westen (1993 and personal communications) has suggested that the seven data layers-(1) bedrock lithological map, (2) geomorphologic map, (3) slope map, (4) land-use map, (5) three maps containing distance from the nearest valley head, (6) road, and (7) fault-are "causal factors" and are significantly related to landslide hazards among the many layers described in van Westen (1993). The corresponding classes in each layer are shown in the first column of Table 1.

In the spatial database, it was assumed that the time of the study was the year 1960 and that all the spatial data available in 1960 were compiled, including the distribution of the scarps of the landslides shown in blue in Plate 1, which had occurred prior to that year. The occurrences play a pivotal role in constructing prediction models by establishing probabilistic relationships be-tween the pre-1960 landslides and the remainder of the input data set. The predictions based on those relationships were then evaluated by comparing the estimated hazard classes with the distribution of the scarps of the landslides that had occurred after 1960, i.e., during the period 1961 to 1988 shown in red in Plate 1. We have also used these seven layers to develop other predictive models for landslide hazard in Chung *et al.* (1995) and Fabbri and Chung (1996).

Probability Model

Let **A** denote the whole study area. Suppose that we have m layers of spatial map data containing "causal" factors which are known to correlate with the occurrences of future landslides in **A**. Consider a pixel p in **A** with m pixel values, $v_1(p) = c_1 \cdots$, $v_m(p) = c_m$, one for each layer. The prediction problem can be represented by the following task: aggregate the m pixel values at pixel p in **A** as a function describing the support for the condition that p is likely to be affected by a future landslide.

To construct a probability model for landslide hazard, consider the following proposition:

$$F_p\text{: "}p\text{ will be affected by a future landslide of a given type D."} \quad (1)$$

We propose that the hazard at each pixel p be expressed as the following joint conditional probability:

$$\textbf{\textit{Prob}}\{F_p | v_1(p), v_2(p), \cdots, v_m(p)\} \quad (2)$$

that p will be affected by future landslides given the m pixel values, $(v_1(p) = c_1, \cdots, v_m(p) = c_m)$.

At pixel p, the pixel value $v_1(p)$ of the first layer is c_1 which is one of the n_1 classes (map units), $\{1, 2, \cdots, n_1\}$. Consider a set of all pixels whose value in the first layer is c_1. The set is the thematic class in the first layer whose pixel value is c_1. The set is denoted by $\mathbf{A}_{1c_1}$ and it is one of the non-overlapping n_1 subareas $\{\mathbf{A}_{11}, \mathbf{A}_{12}, \cdots, \mathbf{A}_{1n_1}\}$ in the first layer. Similarly, we have $\mathbf{A}_{2c_2}$ for the second layer. Finally, we have m thematic classes $\mathbf{A}_{1c_1}, \cdots, \mathbf{A}_{mc_m}$, one for each layer, which correspond to the m input pixel values, $v_1(p)\ (=c_1), \cdots, v_m(p)\ (=c_m)$ at p. The pixel p is one of the common pixels contained in all m thematic classes $\mathbf{A}_{1c_1}, \cdots, \mathbf{A}_{mc_m}$.

TABLE 1. FREQUENCY RATIOS OF PRE-1960 OCCURRENCES OF RAPID DEBRIS AVALANCHES IN EACH CLASS AS AN ESTIMATOR OF THE BINARY CONDITIONAL PROBABILITY FUNCTION (COLUMN 2) AND A MODIFICATION (COLUMN 3) OF THE CONDITIONAL PROBABILITY FUNCTION BY EXPERT'S KNOWLEDGE FOR EACH CLASS USED IN THE ANALYSIS (COLUMN 1).

Lithological Units	Pre-1960 Data	Expert's Knowledge
unmapped area	0.020	0.010
alluvial sediments	0.015	0.020
gneissic intrusives	0.004	0.004
flow materials, alluvial, ashes	0.010	0.010
lake deposits	0.017	0.020
weathered debris flow	0.018	0.020
gabbro and diorite	0.001	0.001
mix of pyroclastic, debris flow	0.015	0.015
metasedimentary	0.020	0.020
andesitic intrusives	0.000	0.000
schists	0.001	0.000
tertiary sediments	0.016	0.016
volcanic	0.017	0.017
lahar deposits	0.051	0.050
pyroclastic flow deposits	0.002	0.002
Geomorphological Units	**Pre-1960 Data**	**Expert's Knowledge**
unmapped area	0.020	0.010
Western hills	0.011	0.010
Romeral fault zone	0.015	0.020
terrace	0.010	0.005
Land-Use Units	**Pre-1960 Data**	**Expert's Knowledge**
traditional farming	0.005	0.006
technified farming	0.017	0.012
modern intermediate farming	0.000	0.000
other crops	0.014	0.010
construction	0.012	0.010
bare	0.011	0.010
grass	0.008	0.008
forest	0.006	0.008
Slope (degree)	**Pre-1960 Data**	**Expert's Knowledge**
0–10	0.005	0.000
10–20	0.010	0.010
20–30	0.020	0.020
30–40	0.025	0.030
40–50	0.023	0.040
50–60	0.089	0.050
60–70	0.005	0.030
70–80	0.002	0.020
80–90	0.000	0.010
Valley Head Distance	**Pre-1960 Data**	**Expert's Knowledge**
>50m	0.011	0.010
25–50m	0.023	0.025
0–25m	0.036	0.050
Road Distance	**Pre-1960 Data**	**Expert's Knowledge**
>50m	0.013	0.010
25–50m	0.012	0.015
0–25m	0.012	0.020
Fault Distance	**Pre-1960 Data**	**Expert's Knowledge**
>100m	0.012	0.010
75–100m	0.017	0.014
50–75m	0.013	0.016
25–50m	0.014	0.018
0–25m	0.014	0.020

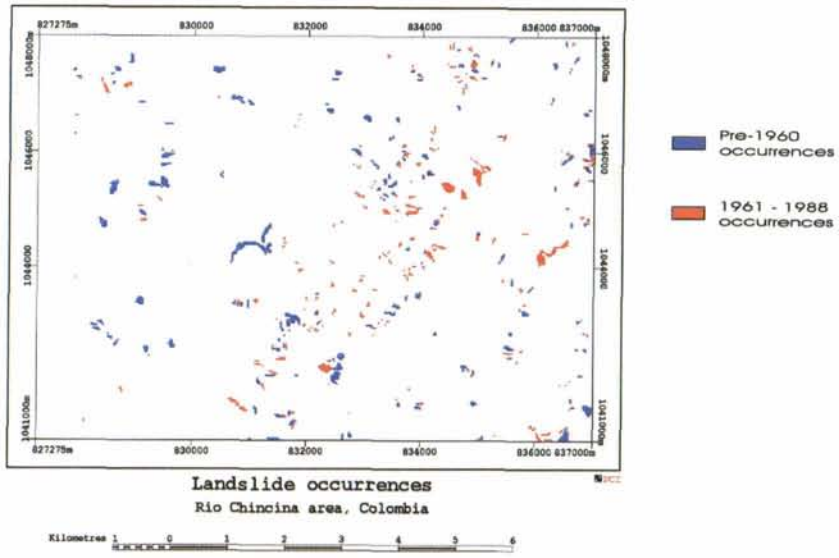

Plate 1. Distribution of the scarps of rapid debris avalanches in the Rio Chincina study area, Colombia. The landslides that occurred prior to 1960 are shown in blue and those that occurred during 1961–1988 are shown in red. The blue, pre-1960, occurrences have been used to obtain the bivariate conditional probability functions and subsequently to obtain all predictions, including those illustrated in Plates 2a, 2b, and 2c. They were also used to obtain the success rates in Table 2 and the corresponding Plate 3a. The landslides in red, which occurred during 1961–1988, have been used to perform the cross-validation test to obtain the prediction rates in Table 3 and the corresponding Plate 3b.

Suppose that **F** denotes the unknown areas which will be affected by future landslides yet to occur within **A**. The joint conditional probability at p is simply given by

$$\begin{aligned} &\textbf{\textit{Prob}}\{F_p|c_1, c_2, \cdots, c_m\} \\ &= \textbf{\textit{Prob}}\left\{p \in \mathbf{F} | p \in \bigcap_{k=1}^{m} \mathbf{A}_{kc_k}\right\} \\ &= \text{size of} \bigcap_{k=1}^{m} (\mathbf{F} \cap \mathbf{A}_{kc_k}) / \text{size of} \bigcap_{k=1}^{m} \mathbf{A}_{kc_k} \end{aligned} \tag{3}$$

where $\mathbf{F} \cap \mathbf{A}_{kc_k}$ represents the unknown area to be affected by future landslides within $\mathbf{A}_{kc_k}$ and "size of **B**" represents the size of the surface area covered by any subarea **B** in **A**. We present here the five procedures to estimate $\textbf{\textit{Prob}}\{F_p|c_1, c_2, \cdots, c_m\}$.

To estimate the joint conditional probability, let us first introduce the counterpart of $\textbf{\textit{Prob}}\{F_p|c_1, \cdots, c_m\}$ for the past landslides. Let

$$S_p\text{: "}p\text{ has been affected by a past landslide of a given type D."} \tag{4}$$

Knowing that the m pixel values at p are $(c_1, c_2, \cdots, c_m)$, the joint conditional probability that p has been affected by a past landslide conditional to that p has the m pixel values $(c_1, \cdots, c_m)$ is simply given by

$$\begin{aligned} &\textbf{\textit{Prob}}\{S_p|c_1, c_2, \cdots, c_m\} \\ &= \textbf{\textit{Prob}}\left\{p \in \mathbf{S} | p \in \bigcap_{k=1}^{m} \mathbf{A}_{kc_k}\right\} \\ &= \text{size of} \bigcap_{k=1}^{m} (\mathbf{S} \cap \mathbf{A}_{kc_k}) / \text{size of} \bigcap_{k=1}^{m} \mathbf{A}_{kc_k} \end{aligned} \tag{5}$$

where **S** represents the areas affected by the past landslides within **A**. In the following procedures, we will make extensive use of $\textbf{\textit{Prob}}\{S_p|c_1, \cdots, c_m\}$ to estimate $\textbf{\textit{Prob}}\{F_p|c_1, \cdots, c_m\}$.

Direct Estimation

The simplest estimate for the joint conditional probability in Equation 2 is obtained by using $\textbf{\textit{Prob}}\{S_p|c_1, c_2, \cdots, c_m\}$ directly. The first estimator is

$$\textbf{\textit{Prob}}_1\{F_p|c_1, c_2, \cdots, c_m\} = \textbf{\textit{Prob}}\{S_p|c_1, c_2, \cdots, c_m\} \tag{6}$$

Although the estimator is simple to compute and does not require any mathematical assumption, it fails badly as a predictor of the occurrences of future landslides. It is the best description of the past landslides, however, in terms of the spatial input data, as we will see in the case study. The estimator should not be computed as a predictor but it should be used as a benchmark for the performance of the spatial input data as "causal factors" of the future landslides.

Bayesian Estimation Under the Conditional Independence

Using the Bayesian rules, the joint conditional probability in Equation 2 can be shown as

$$\begin{aligned} &\textbf{\textit{Prob}}\{F_p|v_1(p), V_2(p), \cdots, v_m(p)\} \\ &= \frac{\textbf{\textit{Prob}}\{F_p\}\ \textbf{\textit{Prob}}\{v_1(p), v_2(p), \cdots, v_m(p)|F_p\}}{\textbf{\textit{Prob}}\{v_1(p), v_2(p), \cdots, v_m(p)\}} \end{aligned} \tag{7}$$

When we assume that $v_1(p), v_2(p), \cdots, v_m(p)$ are conditionally independent given the condition F_p (p will be affected by a future landslide), we have

$$\begin{aligned}&\boldsymbol{Prob}\{v_1(p), v_2(p), \cdots, v_m(p)|F_p\}\\&\quad = \boldsymbol{Prob}\{v_1(p)|F_p\}\boldsymbol{Prob}\{v_2(p)|F_p\}\cdots\boldsymbol{Prob}\{v_m(p)|F_p\}\end{aligned} \quad (8)$$

Hence, under the above conditional independence assumption, the joint conditional probability in Equation 7 becomes

$$\begin{aligned}&\boldsymbol{Prob}\{F_p|v_1(p), v_2(p), \cdots, v_m(p)\}\\&= \frac{\boldsymbol{Prob}\{F_p\}\,\boldsymbol{Prob}\{v_1(p)|F_p\}\,\boldsymbol{Prob}\{v_2(p)|F_p\}\cdots\boldsymbol{Prob}\{v_m(p)|F_p\}}{\boldsymbol{Prob}\{v_1(p), v_2(p), \cdots, v_m(p)\}}\\&= \frac{\boldsymbol{Prob}\{v_1(p)\}\cdots\boldsymbol{Prob}\{v_m(p)\}}{\boldsymbol{Prob}\{v_1(p), \cdots, v_m(p)\}}\,\boldsymbol{Prob}\{F_p\}\\&\quad \frac{\boldsymbol{Prob}\{F_p|v_1(p)\}}{\boldsymbol{Prob}\{F_p\}}\cdots\frac{\boldsymbol{Prob}\{F_p|v_m(p)\}}{\boldsymbol{Prob}\{F_p\}}\end{aligned} \quad (9)$$

Under the conditional independence assumption in Equation 8, the joint conditional probability in Equation 2 can be expressed in terms of three components as shown in Equation 9. The first component, the ratio of $\boldsymbol{Prob}\{v_1(p)\}\cdots\boldsymbol{Prob}\{v_m(p)\}$ and $\boldsymbol{Prob}\{v_1(p), \cdots, v_m(p)\}$, consists of the probabilities related to the input spatial data. The second component, the prior probability $\boldsymbol{Prob}\{F_p\}$, is the probability that a pixel p will be affected by a future landslide prior to having any evidence. The third component consists of m factors, and each factor, the ratio of bivariate conditional probability $\boldsymbol{Prob}\{F_p|v_k(p)\}$ and the prior probability $\boldsymbol{Prob}\{F_p\}$, indicates a contribution of each pixel value to future landslide hazard. We will examine each component in detail in Appendix A.

The first component is easily obtained by computing

$$\begin{aligned}&\boldsymbol{Prob}\{v_k(p) = c_k\} = \boldsymbol{Prob}\{p \in \mathbf{A}_{kc_k}\} = \text{size of } \mathbf{A}_{kc_k}/\text{size of } \mathbf{A};\\&\boldsymbol{Prob}\{v_1(p) = c_1, \cdots, v_m(p) = c_m\}\\&\quad = \boldsymbol{Prob}\left\{p \in \bigcap_{k=1}^{m}\mathbf{A}_{kc_k}\right\} = \text{size of } \bigcap_{k=1}^{m}\mathbf{A}_{kc_k}/\text{size of } \mathbf{A}.\end{aligned} \quad (10)$$

However, not having $\mathbf{F}$, the areas to be affected by future landslides, we cannot compute the two probabilities, $\boldsymbol{Prob}\{F_p\}$ and $\boldsymbol{Prob}\{F_p|c_k\}$, in Equation 9, which are

$$\begin{aligned}\boldsymbol{Prob}\{F_p\} &= \boldsymbol{Prob}\{p \in \mathbf{F}\} = \text{size of } \mathbf{F}/\text{size of } \mathbf{A},\\\boldsymbol{Prob}\{F_p|c_k\} &= \boldsymbol{Prob}\{p \in \mathbf{F}|p \in \mathbf{A}_{kc_k}\}\\&= \text{size of } \mathbf{F} \cap \mathbf{A}_{kc_k}/\text{size of } \mathbf{A}_{kc_k}\end{aligned} \quad (11)$$

We may substitute these unknown probabilities by their counterparts, $\boldsymbol{Prob}\{S_p\}$ and $\boldsymbol{Prob}\{S_p|c_k\}$, for the past landslides: i.e.,

$$\begin{aligned}\boldsymbol{Prob}\{S_p\} &= \boldsymbol{Prob}\{p \in \mathbf{S}\} = \text{size of } \mathbf{S}/\text{size of } \mathbf{A},\\\boldsymbol{Prob}\{S_p|c_k\} &= \boldsymbol{Prob}\{p \in \mathbf{S}|p \in \mathbf{A}_{kc_k}\}\\&= \text{size of } \mathbf{S} \cap \mathbf{A}_{kc_k}/\text{size of } \mathbf{A}_{kc_k}.\end{aligned} \quad (12)$$

For the Colombian study area, using the pre-1960 occurrences data and the input spatial data, $\boldsymbol{Prob}\{S_p|c_k\}$ was obtained by computing size of $\mathbf{S} \cap \mathbf{A}_{kc_k}$/size of $\mathbf{A}_{kc_k}$ which is shown in the first column "*Pre-1960 Data*" of Table 1. From Equation 9, by substituting the probabilities in Equation 10 and replacing $\boldsymbol{Prob}\{F_p\}$ and $\boldsymbol{Prob}\{F_p|c_k\}$ by $\boldsymbol{Prob}\{S_p\}$ and $\boldsymbol{Prob}\{S_p|c_k\}$ in Equation 12, we have the second estimate for the joint conditional probability at each pixel: i.e.,

$$\begin{aligned}\boldsymbol{Prob}_2\{F_p|c_1, \cdots, c_m\} &= \frac{\boldsymbol{Prob}\{c_1\}\cdots\boldsymbol{Prob}\{c_m\}}{\boldsymbol{Prob}\{c_1, \cdots, c_m\}}\\&\quad \boldsymbol{Prob}\{S_p\}\,\frac{\boldsymbol{Prob}\{S_p|c_1\}}{\boldsymbol{Prob}\{S_p\}}\cdots\frac{\boldsymbol{Prob}\{S_p|c_m\}}{\boldsymbol{Prob}\{S_p\}}\\&= \frac{\mathbf{s}}{\text{size of}\left(\bigcap_{k=1}^{m}\mathbf{A}_{kc_k}\right)}\,\frac{\mathbf{s}_{1c_1}}{\mathbf{s}}\cdots\frac{\mathbf{s}_{mc_m}}{\mathbf{s}}.\end{aligned} \quad (13)$$

where $\mathbf{s}$ denotes the size of $\mathbf{S}$ and $\mathbf{s}_{kc_k}$ denotes the size of $\mathbf{S} \cap \mathbf{A}_{kc_k}$.

Although the estimator in Equation 13 is one of most (if not the most) popular techniques for integrating spatial data (Agterberg *et al.*, 1990; Aspinall, 1992; Bonham-Carter, 1994) for predicting occurrences, it may not produce "good" predictions, as we will see in the case studies. It is noted that $\boldsymbol{Prob}_2\{F_p|c_1, \cdots, c_m\}$ is the joint conditional probability $\boldsymbol{Prob}\{S_p|c_1, \cdots, c_m\}$ under the conditional independence assumption of $v_1(p), v_2(p), \cdots, v_m(p)$ given the condition S_p (p has been affected by a past landslide), the counterpart of Equation 9 for the past landslides. Advantages of this estimator are that it is simple to compute and it depends only on the bivariate conditional probabilities of the occurrences of the past landslides given the pixel values at each layer separately.

Regression Model Based on Bivariate Conditional Probabilities

A general multivariate linear regression model for the conditional joint probability in Equation 2 for a pixel p can be postulated by

$$\begin{aligned}\boldsymbol{Prob}\{F_p|v_1(p), \cdots, v_m(p)\} &= \beta_0 + \beta_1 v_1(p) + \beta_2 v_2(p)\\&\quad + \cdots + \beta_m v_m(p) + \varepsilon_p\end{aligned} \quad (14)$$

where $(\beta_0, \beta_1, \cdots, \beta_m)$ are unknown parameters to be estimated and ε_p is an error associated with the linear approximation of the joint conditional probability. The model in Equation 14 may be valid only if $v_1(p), \cdots, v_m(p)$ represent continuous measurements, however, and not the thematic classification data used here.

To overcome this difficulty, we can proceed in several different ways. One approach is to transform all thematic classification data into a series of binary variables as proposed by Chung and Fabbri (1993). Here we propose another approach where the bivariate conditional probabilities, $\boldsymbol{Prob}\{F_p|v_k(p)\}$s are used instead of the $v_k(p)$s in Equation 14. The linear model in Equation 14 is modified to

$$\begin{aligned}\boldsymbol{Prob}\{F_p|v_1(p), \cdots, v_m(p)\} &= \beta_0 + \beta_1\,\boldsymbol{Prob}\{F_p|v_1(p)\} + \cdots\\&\quad + \beta_m\,\boldsymbol{Prob}\{F_p|v_m(p)\} + \varepsilon_p\end{aligned} \quad (15)$$

using the m bivariate conditional probabilities, $\boldsymbol{Prob}\{F_p|v_1(p)\}$, $\cdots$, $\boldsymbol{Prob}\{F_p|v_m(p)\}$.

Not knowing the occurrences of future landslides, it is impossible to estimate $(\beta_0, \cdots, \beta_m)$ in Equation 15 directly. To make estimation possible, let us consider the counterpart of Equation 15 for the past landslides: i.e.,

$$\begin{aligned}\boldsymbol{Prob}\{S_p|v_1(p, \cdots, v_m(p)\} &= \alpha_0 + \alpha_1\,\boldsymbol{Prob}\{S_p|v_1(p)\}\\&\quad + \cdots + \alpha_m\,\boldsymbol{Prob}\{S_p|v_m(p)\} + \varepsilon_p\end{aligned} \quad (16)$$

where S_p is defined in Equation 4, the associated probabilities

are defined in Equation 12, and $(\alpha_0, \alpha_1, \cdots, \alpha_m)$ are the unknown parameters.

We use the least-squares method (Draper and Smith, 1981) to estimate $(\alpha_0, \alpha_1, \cdots, \alpha_m)$ in Equation 16, and we may choose any subareas in the whole study area as the training area. For each pixel p, the third estimator of the joint conditional probability, $\boldsymbol{Prob}\{F_p|v_1(p), \cdots, v_m(p)\}$, is obtained by

$$\boldsymbol{Prob}_3\{F_p|v_1(p), \cdots, v_m(p)\} = \hat{\alpha}_0 + \hat{\alpha}_1 \boldsymbol{Prob}\{S_p|v_1(p)\} + \cdots + \hat{\alpha}_m \boldsymbol{Prob}\{S_p|v_m(p)\} \quad (17)$$

where $(\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_m)$ are the least-squares estimators of $(\alpha_0, \alpha_1, \cdots, \alpha_m)$ in Equation 16. When the number of pixels in the training area is very large (e.g., a wide area with pixels representing small areas on the ground), the weighted least-squares estimators discussed in Appendix B simplify the computational procedure. It is based on the unique condition subareas and produces results identical to those of the least-squares estimators.

Modified Bayesian Estimation under the Conditional Independence

For the Colombian study area, using the pre-1960 landslide occurrences data and the input spatial data, $\boldsymbol{Prob}\{S_p|c_k\}$ was obtained and is shown in column "*Pre-1960 Data*" of Table 1. We are proposing that the resulting bivariate conditional probability functions should be reviewed and modified by experts if these probabilities, based on the past landslides, are used as estimators of $\boldsymbol{Prob}\{F_p|v_k(p)\}$ for future landslides. The modified bivariate conditional probability functions by an expert are shown in column "*Expert's Knowledge*" of Table 1. We repeat the Bayesian procedure previously studied under conditional independence using the modified bivariate conditional probability functions instead of the bivariate conditional probability functions based on the past landslides.

Instead of the estimation in Equation 12, where $\boldsymbol{Prob}\{F_p\}$ and $\boldsymbol{Prob}\{F_p|v_k(p)\}$ in Equation 9 were estimated by $\boldsymbol{Prob}\{S_p\}$ and $\boldsymbol{Prob}\{S_p|v_k(p)\}$ in Equation 12, here $\boldsymbol{Prob}\{F_p\}$ and $\boldsymbol{Prob}\{F_p|v_k(p)\}$ are estimated by $\boldsymbol{Prob}_e\{F_p\}$ and $\boldsymbol{Prob}_e\{F_p|v_k(p)\}$ which are obtained by expert's knowledge. Hence, we obtain the fourth estimate for the joint conditional probability at each pixel by

$$\boldsymbol{Prob}_4\{F_p|c_1, \cdots, c_m\} = \frac{a}{\text{size of}\left(\bigcap_{k=1}^{m} \mathbf{A}_{kc_k}\right)} \frac{a_{1c_1}}{a} \cdots \frac{a_{mc_m}}{a} p_e e_1 \cdots e_m \quad (18)$$

where a denotes the size of $\mathbf{A}$, a_{kc_k} denotes the size of $\mathbf{A}_{kc_k}$, and

$$p_e = \boldsymbol{Prob}_e\{F_p\} \text{ and } e_k = \boldsymbol{Prob}_e\{F_p|c_k\} \text{ for all } k, \quad (19)$$

are obtained from the expert's knowledge. The term "expert's knowledge" is used here to indicate modifications of the frequencies of occurrence of map units of mass movements over those units, which represent more closely the mental models of experts. For instance, knowledge of the landscape may lead the expert to reinterpret the map of slope when it is felt that it does not represent satisfactorily the topographic contour lines. Some angles may be less frequent than expected due to computational limitations. For instance, slope angles and the frequency of mass movements over different angles may be biased for steeper slopes. This can be corrected by modifying the histogram of the number of landslide occurrences for the steeper slopes.

Furthermore, after some initial predictions, an expert may want to reconsider the legend of litho-stratigraphic units into more litho-technical units, which may better represent the geomorphologic setting. Such modifications may improve the prediction results once a prediction validation strategy is set up. Although much work is needed in this area of interaction expert/prediction, it is clear that the model in Equation 18 is the first step toward incorporating expert's knowledge into prediction models. The introduction of such knowledge is particularly significant or even necessary whenever a database insufficiently represents the observed geomorphologic setting of the mass movements.

Modified Regression-Combination of Input Data and Expert's Knowledge

From the model in Equation 15, we have used $\boldsymbol{Prob}\{S_p|v_k(p)\}$ and $\boldsymbol{Prob}\{S_p|v_1(p), \cdots, v_m(p)\}$ in Equation 12 instead of $\boldsymbol{Prob}\{F_p|v_k(p)\}$ and $\boldsymbol{Prob}\{F_p|v_1(p), \cdots, v_m(p)\}$ to construct the model in Equation 16. In the following modified model, we have also used $\boldsymbol{Prob}\{S_p|v_1(p), \cdots, v_m(p)\}$ instead of $\boldsymbol{Prob}\{F_p|v_1(p), \cdots, v_m(p)\}$ as before, but $\boldsymbol{Prob}\{F_p|v_k(p)\}$ was replaced by $\boldsymbol{Prob}_e\{F_p|v_k(p)\}$ in Equation 19 as we have done in the modified Bayesian estimation procedure. Hence, a new regression model is given by

$$\boldsymbol{Prob}\{S_p|v_1(p), \cdots, v_m(p)\} = \beta_0 + \beta_1 \boldsymbol{Prob}_e\{F_p|v_1(p)\} + \cdots + \beta_m \boldsymbol{Prob}_e\{F_p|v_m(p)\} + \varepsilon_p \quad (20)$$

To estimate $(\beta_0, \beta_1, \cdots, \beta_m)$ in Equation 20, we again use the least-squares method or the weighted least-squares method discussed in Appendix B, and we may choose any subareas in the whole study area as the training area. The fifth estimator of the joint conditional probability, $\mathbf{Prob}\{F_p|c_1, \cdots, c_m\}$ for each pixel p is obtained by

$$\mathbf{Prob}_5\{F_p|c_1, \cdots, c_m\} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{Prob}_e\{F_p|v_1(p)\} + \cdots + \hat{\beta}_m \mathbf{Prob}_e\{F_p|v_m(p)\} \quad (21)$$

where $(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_m)$ are the least-squares estimators of $(\beta_0, \beta_1, \cdots, \beta_m)$ in Equation 20.

Case Study in Rio Chincina Area in Colombia

Assumption and Visualization

For the Colombian study area, described earlier, we have assumed that the year of study was 1960. Hence, the set of pixels affected by the pre-1960 landslide occurrences, shown in blue in Plate 1, is regarded as the distribution of the occurrences of the past landslides. These pixels are used to compute the conditional probabilities, related to S_p in Equations 5 and 12. The red pixels in Plate 1, indicating the distribution of the later landslides, which occurred between 1961 and 1988, are considered as indicating the presence of "future" landslides. They are used to evaluate the prediction patterns only. The five prediction patterns obtained are based on the estimated joint conditional probabilities computed from Equations 6, 13, 17, 18, and 21, respectively.

For the visualization of a prediction pattern, two methods can be applied to the pattern; one using the estimated probabilities of the pixels directly and the other using the relative ranks of the estimators. To interpret the estimated probabilities directly, the assumptions such as $\boldsymbol{Prob}\{F_p|c_k\} = \boldsymbol{Prob}\{S_p|c_k\}$ required for Equations 13 and 17 or the linear models for Equations 17 or 21 must be "absolutely true." If those assumptions are only "approximately true," then the use of the relative ranks is a more promising way of interpreting the results, because the same assumptions are applied to all the pixels. No significance is directly attached to gradient of the estimators. What matters is the ranking sequence of the estimators. In addition, when two different prediction models are compared,

the ranks of the pixels, which are independent of the models, provide more neutral statistics of the predictions than from the estimated values directly.

To obtain the relative ranks for each prediction pattern, the estimated probabilities of all pixels in the study area were sorted in descending order. Then the ordered pixel values were divided into 11 classes (colored red to blue) as follows. The pixels with the highest 5 percent estimated probability values were classified as the "0 to 5 percent" class, shown as "*purple-red*" in the illustrations, occupying 5 percent of the study area. The pixels with the next highest 5 percent values were represented in "*red*," occupy an additional 5 percent of the study area, and were classified as the "5 to 10 percent" class. We repeated the classification eight more times, for classes 5 percent apart, and the resulting ten classes are shown in the ten corresponding colors: *red-purple, red, orange, yellow, light green, green, dark green, light blue, blue*, and *dark blue*. Finally, the "*purple-blue*" color was assigned to the remaining 50 percent of the area. Only three of the five patterns are shown here as Plates 2a, 2b, and 2c. They correspond to the "direct," "regression," and "modified Bayesian" procedures, respectively.

Success and Prediction Rates

For each prediction pattern, we first compared, in terms of proportions of corresponding pixels, the 11 classes obtained with the occurrences of the pre-1960 landslides shown as blue in Plate 1. There are 5515 blue pixels that indicate the areas affected by the pre-1960 landslides. We counted the number of those blue pixels present in each class. The cumulative distribution function of the 5515 pixels with respect to the eleven classes is shown in the columns of Table 2. To refer to each column, we will use the term, "*success rates*" for these 11 classes with respect to the pre-1960 occurrences. The success rates are also illustrated as line-graphs in Plate 3a.

We applied the above procedure to the landslides which occurred later, during the period 1961 through 1988, by comparing the 11 classes obtained with the distribution of the 1961 to 1988 "future" landslides shown in red in Plate 1. There are 4589 red pixels in the illustration, indicating the areas affected

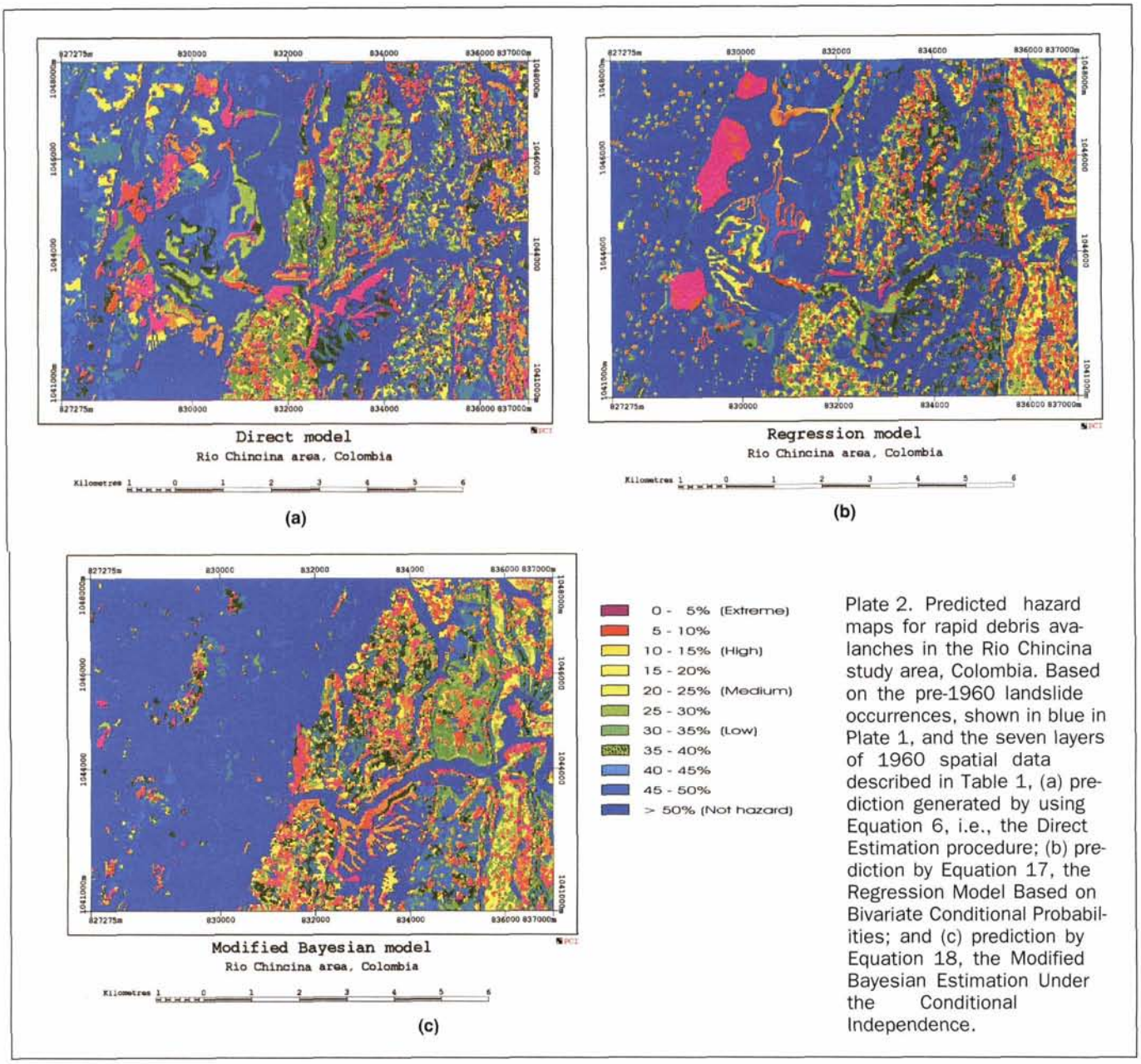

Plate 2. Predicted hazard maps for rapid debris avalanches in the Rio Chincina study area, Colombia. Based on the pre-1960 landslide occurrences, shown in blue in Plate 1, and the seven layers of 1960 spatial data described in Table 1, (a) prediction generated by using Equation 6, i.e., the Direct Estimation procedure; (b) prediction by Equation 17, the Regression Model Based on Bivariate Conditional Probabilities; and (c) prediction by Equation 18, the Modified Bayesian Estimation Under the Conditional Independence.

TABLE 2. SUCCESS RATES OF THE CLASSES (5 PERCENT APART) FOR THE FIVE ESTIMATION PROCEDURES OF THE CONDITIONAL PROBABILITIES ASSOCIATED WITH RAPID DEBRIS AVALANCHES IN THE RIO CHINCINA STUDY AREA, COLOMBIA. THE CORRESPONDING EXPRESSIONS AND FIGURES FOR THE PROCEDURES ARE ALSO INDICATED.

	Column #	I	II	III	IV	V
Row #	Classes	Direct Estimation (Eq. 6) Plate 2a	Bayesian under C.I. (Eq. 13)	Regression Model (Eq. 17) Plate 2b	Modified Bayesian (Eq. 18) Plate 2c	Modified Regression (Eq. 21)
1	0–5%	0.531	0.12186	0.23638	0.12455	0.26308
2	0–10%	0.6819	0.22545	0.36792	0.22921	0.33781
3	0–15%	0.77276	0.32993	0.43943	0.32366	0.42903
4	0–20%	0.83746	0.40036	0.48943	0.39839	0.53459
5	0–25%	0.88817	0.44982	0.58566	0.44337	0.58656
6	0–30%	0.91774	0.48602	0.64426	0.49068	0.63656
7	0–35%	0.94319	0.5509	0.72706	0.57993	0.7052
8	0–40%	0.96326	0.59677	0.78154	0.66004	0.77599
9	0–45%	0.97491	0.61631	0.79785	0.67509	0.79642
10	0–50%	0.98871	0.66272	0.83297	0.74194	0.8319

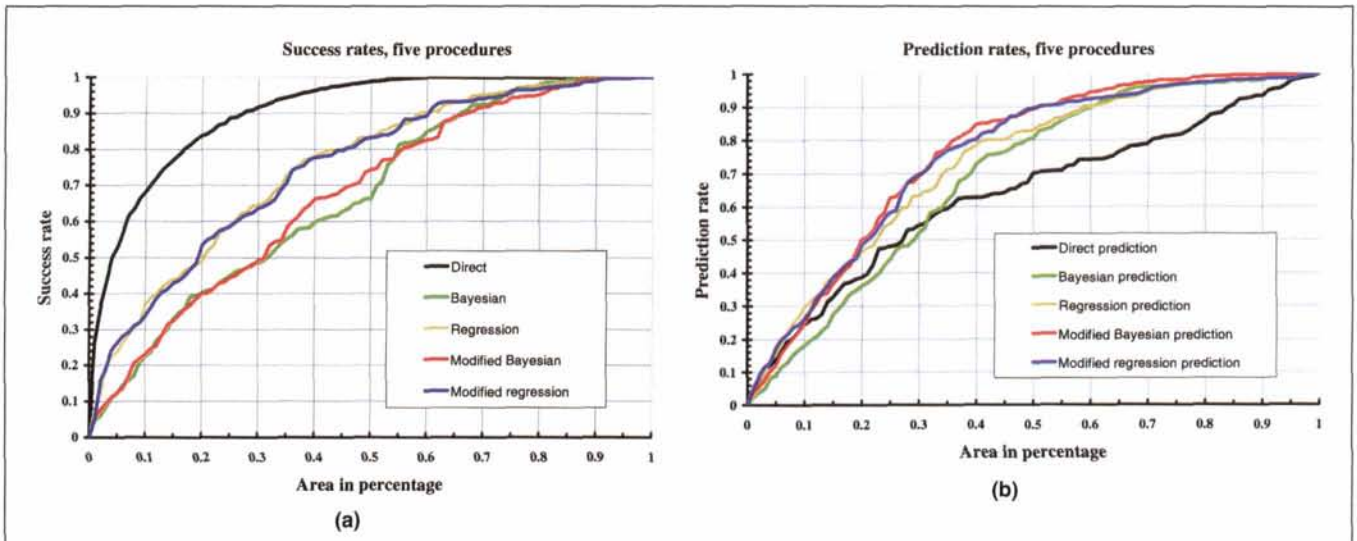

Plate 3. Success rates of predicting the hazard of rapid debris avalanches in the Rio Chincina study area, Colombia. (a) Success rates for pre-1960 landslide occurrences (training data) based on the bivariate conditional probability functions and the five estimation models using both the pre-1960 data and expert's knowledge; the corresponding values are shown in Table 2. (b) Prediction rates for the 1961–1988 landslide occurrences based on the bivariate conditional probability functions and the five estimation models using both the pre-1960 data and the expert's knowledge; the corresponding values are shown in Table 3.

by the later landslides. The cumulative distribution function of those 4589 pixels with respect to the 11 classes is shown in the columns of Table 3. In contrast to the success rates in Table 2, to refer to the columns in Table 3, we will use the term, "*prediction rates*" of the 11 classes with respect to the later landslides. The prediction rates are illustrated as line-graphs in Plate 3b.

On the one hand, the success rates in Table 2 illustrate how well the estimators perform with respect to the pre-1960 landslides used to construct the estimators. The prediction rates in Table 3, on the other hand, are used as measurements of how well the probability model in Equation 2 and of its estimators predict the distribution of future landslides. It is implicit that the prediction rates, shown in Table 3, and the corresponding graphs, shown in Plate 3b, are the only significant statistics of the model and the estimator procedure for the prediction of the distribution of future landslides.

First Prediction: Direct Estimation (Equation 6)

Plate 2a contains the prediction pattern of the 11 classes obtained by the "direct" procedure based on Equation 6. Column I in Table 2 shows the success rates of the pattern in Plate 2a with respect to the pre-1960 landslide occurrences. Column I in Table 3, shows the prediction rates of the pattern in Plate 2a with respect to the 1961 to 1988 landslide occurrences. These two columns are represented as black solid lines, in Plates 3a and 3b, respectively.

With respect to the pre-1960 "past" occurrences, obviously the success rates of the "direct estimation" using Equation 6, in Plate 3a, show the best performance among the five estimation procedures considered. The prediction rates with respect to the "future" 1961 to 1988 occurrences of "direct estimation," however, are the worst among the five procedures (Plate 3b). This fact indicates that the direct procedure should not be used to estimate the probability model of future landslides. As a pre-

TABLE 3. PREDICTION RATES OF THE CLASSES (5 PERCENT APART) FOR THE FIVE ESTIMATION PROCEDURES OF THE CONDITIONAL PROBABILITIES ASSOCIATED WITH RAPID DEBRIS AVALANCHES IN THE RIO CHINCINA STUDY AREA, COLOMBIA. THE CORRESPONDING EXPRESSIONS AND FIGURES FOR THE PROCEDURES ARE ALSO INDICATED.

	Column #	I	II	III	IV	V
Row #	Classes	Direct Estimation (Eq. 6) Plate 2a	Bayesian under C.I. (Eq. 13)	Regression Model (Eq. 17) Plate 2b	Modified Bayesian (Eq. 18) Plate 2c	Modified Regression (Eq. 21)
1	0–5%	0.1385	0.0895	0.15244	0.11672	0.17465
2	0–10%	0.24477	0.1814	0.29573	0.24456	0.26328
3	0–15%	0.32709	0.27722	0.38654	0.36476	0.38828
4	0–20%	0.38698	0.36106	0.46298	0.49956	0.48563
5	0–25%	0.4804	0.4412	0.5466	0.62456	0.58145
6	0–30%	0.5453	0.52134	0.63132	0.68902	0.69686
7	0–35%	0.59408	0.61999	0.71167	0.77809	0.77091
8	0–40%	0.62827	0.72801	0.78375	0.84495	0.80183
9	0–45%	0.63959	0.76089	0.80031	0.85736	0.84647
10	0–50%	0.70209	0.80618	0.82992	0.89373	0.89721

diction tool, Plate 2a produced a disappointing result. The relatively poor performance for future landslides may be caused by one of the following three situations: (1) the 1961 to 1988 landslides may not be directly related to the pre-1960 landslides, (2) preventive measures were possibly put in place sometime after the year 1960, and/or (3) the direct estimation for the joint conditional probability in Equation 6 is not a good procedure. The last situation was considered the most likely, because the types of the pre-1960 and the post-1960 landslides are identical (van Westen, personal communication) and it is unlikely that preventive measures had been put in place in 1960 considering that we had selected the year 1960 arbitrarily.

Second and fourth predictions: Bayesian Equation 13 and modified Bayesian Equation 18

The results from the second prediction pattern based on Equation 13 are shown in Column II of Tables 2 and 3, and appear as green lines in Plates 3a and 3b, respectively. As a prediction measure, the 0 to 15 percent class in Table 3 (Column II/Row 3) contains 27.72 percent of the 1961 to 1988 occurrences and it occupies 15 percent of the study area. The value of 27.72 percent is better than the 15 percent expected, but the prediction rates of 27.72 percent is worse than the 32.71 percent in Table 3 (Column I/Row 3) from the results of the direct estimation procedure. Although the prediction rates for the last four classes (0 to 35 percent, 0 to 40 percent, 0 to 45 percent, and 0 to 50 percent) from Equation 13 are little better than those from Equation 6, clearly, the direct procedure produced somewhat better results than the Bayesian procedure under the conditional independence assumption. For the overall performances, Equation 6 is simpler; it requires fewer assumptions, and it leads to better results than Equation 13.

Modification of the bivariate conditional probability functions by the expert's knowledge is particularly important and necessary when the database appears to under-represent the natural setting of the mass movements. In this study, the expert's knowledge was simulated by analyzing the frequency distribution of all pre-1960 input data, shown in Table 1, and then fitting a more regular or smoothed distribution of frequencies. The process is similar to fitting a simple model to a noisy distribution. While more geological criteria could be used, this simulation tests the influence of elementary changes to weight assignment. The modified values are shown in the second column of Table 1. Although the improvement of the prediction is significant, modifications by expert's knowledge on the bivariate conditional probabilities are subjective and can be arbitrary. Obviously, much work is required regarding how to incorporate subjective expert's knowledge into the prediction models. In addition, the conditional independence assumption is still imposed by the model.

Plate 2c shows the fourth prediction patterns based on the estimated conditional probabilities by Equation 18. By comparing the 11 classes in Plate 2c with respect to the pre-1960 landslides and the 1961 to 1988 landslides, we have also constructed Column IV in Table 2 and Column IV in Table 3. These two columns are illustrated as red lines in Plates 3a and 3b, respectively.

Let us compare the new results obtained from Equation 18 to the ones obtained from Equation 13 using the original bivariate conditional probability functions. While the two success rates (green and red lines) in Plate 3a for the pre-1960 data are similar, the two predictions in Plate 3b are very different for the 1961 to 1988 occurrences from the modified bivariate conditional probability functions. Obviously, out of the five lines in Plate 3b, the prediction rates using the modified Bayesian conditional probability function provide one of the two best performances. Again, let us compare these new results from Equation 18 to the results obtained from Equation 13 and consider the two prediction rates for the 0 to 20 percent class in Table 3 (Columns II and IV/Row 4). We have 36.1 percent versus 50.0 percent. As a prediction measure, Equation 18 provides better prediction results than those from Equation 13. Hence, we can conclude that the expert's knowledge provided a significantly better prediction.

Third and Fifth Predictions: Regression Equation 17 and modified regression Equation 21

The results of the linear regression model in Equation 17 are shown in Plate 2b. For this experiment, one pixel from every 4 by 4 window was systematically selected into a training data set consisting of 1473 pixels. Among the 1473 pixels, 355 pixels have been affected by the pre-1960 landslides and the remaining 1118 pixels have not been affected by the pre-1960 landslides. In summary, we apply regression analysis to the data set consisting of 1473 pixels, and obtain a set of regression estimators ($\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_m$) in Equation 17. By comparing the 11 classes in Plate 2b with respect to the pre-1960 landslides and the later 1961 to 1988 landslides, we have constructed Column III in Table 2 containing the success rates and Column III in Table 3 containing the prediction rates. These two columns are shown as brown lines in Plates 3a and 3b, respectively. As a prediction measure, for instance, the 0 to 20 percent in Table 3 class (Column III/Row 4) contains 46.3 percent of the 1961 to 1988 occurrences, while it occupies 20 percent of the whole area. The prediction rate of 46.3 percent is much better than the rate of 38.7 percent in Table 3 (Column I/Row 4), from the

results of the direct estimation procedure, and better than the rate of 36.11 percent in Table 3 (Column II/Row 4), from the results of the Bayesian estimation procedure.

To look at the overall prediction performance, let us compare three lines in Plate 3b, black from the direct procedure in Equation 6, green from the Bayesian procedure in Equation 13, and brown from the regression procedure in Equation 17. Of the three lines, the brown regression line produces the best overall performance.

By comparing the 11 classes from Equation 21 with respect to the pre-1960 landslides and to the 1961 to 1988 landslides, we have constructed Column V in Table 2 and Column V in Table 3. These two columns are illustrated as brown lines in Plates 3a and 3b, respectively. Let us compare these new results obtained from Equation 21 to the results obtained from Equation 17 using the original bivariate conditional probability functions. While the two success rates (brown and blue lines) in Plate 3a for the pre-1960 data are similar, the two predictions in Plate 3b for the 1961 to 1988 occurrences are slightly different. Obviously, as we have observed earlier, from the five lines in Plate 3b, the prediction rates using the modified regression provide one of the two best performances among the five considered in this paper. The two prediction rates behave in a similar fashion. Hence, we can conclude that the expert's knowledge provided a marginally better prediction model than the original regression in Equation 13.

Considerations on the Merit of the Predictions

The successes and prediction rates in Tables 2 and 3 have been expressed in terms of the distribution of landslide pixel proportions corresponding to the different hazard classes. The dynamic character of mass movements, however, shows that, while the activity originates at higher elevations and steeper slopes, the landslides move in the direction of lower elevations and shallower slopes. For this reason, the distribution of pixels representing the scarp of a landslide generally covers more than one hazard class. This can be easily observed in a three-dimensional display of the predictions with the image of the later landslides draped over the digital elevation image. In our application, the rates in the Tables might seem not to indicate a strong predictive power of the estimation procedures. A simple count of the cumulative proportion of individual landslides per class, however, provides another way of evaluating the rates. For instance, the values in Table 4 are the landslide counts of success and prediction rates for the modified Bayesian probability and the regression estimators. These values can be used to better interpret the corresponding values in Tables 2 and 3. For instance, the first two classes in Table 4, 0 to 5 percent and 0 to 10 percent, show prediction values 35.4 percent (99 out of 280 "future" landslides to occur) and 52.5 percent (147 out of 280) for the modified Bayesian probability model, and 38.2 percent and 55.0 percent for the regression model, respectively. Those values are more than twice as high as the corresponding ones in Table 3. Also, in Table 4 we can observe very small differences between success and prediction rates.

In addition, it must be remarked that: (1) this study does not represent a simplified simulation but it is based on a real data set which of necessity is a partial representation of a more complex situation in nature (and given the database, it is, in fact, surprising that the models presented here produce such high prediction rates); and (2) the same data set is used to compare the results of several different estimation procedures.

Concluding Remarks

We have used a spatial database for landslide hazard zoning in Colombia to compare and validate five different predictive methods based on probability models. The results of such a comparison allow one to consider general application strategies for geographical information systems.

- A spatial database for predictive modeling (i.e., with all the landslide characteristics, including topographic, geotechnical, geological, infrastructural, and temporal settings) must be built so that each information layer clearly contributes to the characterization of the typical setting of one event to be predicted. It must be recognized that, no matter how good the information available may be, the database will always contain incomplete information. In addition, with regard to the predictive methods considered here, it is irrelevant whether the data domain is in raster or vector form: the computations can be performed on the attribute tables in either domain.
- To analyze and compare the results of predictions, it is critical to partition the database in time and/or in space. Failing to do this, the models will remain poorly known and untested, even if we consider the database to be a satisfactory representation.
- It seems that, when the database provides "reasonable" support, multivariate regression generates better results than Bayesian probability methods and it also avoids the assumption of conditional independence of the input layers.
- When the database falls short of providing "reasonable" support for a prediction, the introduction of the expert's knowledge, to modify the observed frequencies of the input data relationships, appears to improve the results of the predictions. This can be demonstrated by comparative analysis and sensitivity analysis.

TABLE 4. SUCCESS AND PREDICTION RATES OF THE CLASSES (5 PERCENT APART) FOR TWO ESTIMATION PROCEDURES IN TERMS OF THE CORRESPONDING PROPORTIONS OF THE NUMBER OF LANDSLIDES ASSOCIATED WITH RAPID DEBRIS AVALANCHES IN THE RIO CHINCINA STUDY AREA, COLOMBIA.

	Number of Landslides Intersected for the Regression Method in Eq. 17. Ratio in Bracket. Plate 2b		Number of Landslides Intersected for the Modified Bayesian Probability in Eq. 18. Ratio in Bracket. Plate 2c.	
Classes	"Success Rate" Out of 177 Landslides Which Occurred Prior to 1960	"Prediction Rate" Out of 280 Landslides Which Occurred Between 1961 and 1988	"Success Rate" Out of 177 Landslides Which Occurred Prior to 1960	"Prediction Rate" Out of 280 Landslides Which Occurred Betweeen 1961 and 1988
0–5%	76 (0.429)	107 (0.382)	68 (0.384)	99 (0.354)
0–10%	107 (0.605)	154 (0.550)	93 (0.525)	147 (0.525)
0–15%	119 (0.672)	184 (0.657)	104 (0.588)	174 (0.621)
0–20%	127 (0.718)	196 (0.700)	117 (0.661)	197 (0.704)
0–25%	138 (0.780)	214 (0.764)	125 (0.706)	210 (0.750)
0–30%	143 (0.808)	221 (0.789)	130 (0.734)	215 (0.768)
0–35%	148 (0.836)	237 (0.846)	134 (0.757)	233 (0.832)
0–40%	152 (0.859)	247 (0.882)	139 (0.785)	244 (0.871)
0–45%	153 (0.864)	248 (0.886)	142 (0.802)	246 (0.879)
0–50%	153 (0.864)	255 (0.911)	155 (0.876)	255 (0.911)

This situation will require extensive experimentation with similar data sets and also with more methods and their modifications using the same data set. In all cases, the computational strategy requires data in the form of hypotheses which can be tested.

- Once the preliminary statistical analysis of the database has been performed, the statistical results, showing the frequency distribution of the occurrences of the past landslides with respect to the supporting pieces of evidence, should be reviewed by experts. This is particularly important and necessary when the database appears to be under-representing the natural setting of the mass movements.
- For the least-squares estimation of the regression coefficients in Equations 17 and 21, we have also studied several different training data sets ranging from about 1,000 pixels to the whole study area (43,7019 pixels) using the weighted least-squares estimation in Appendix B. The prediction rates with respect to the size of the training data set appear to be robust and the study on the effect will be a subject to a future contribution.

This research provides a unified framework to predictive modeling with GIS using probability concepts. The authors of this contribution are currently studying the results from subsets of the spatial database and of inverting the control data in time, i.e., using the later landslides to predict the location of the older ones. The latter process is expected to lead to poorer results. In addition, the application of prediction models based on fuzzy set techniques and Dempster-Shafer's evidential theory has been proposed by Chung and Fabbri (1993). We are currently evaluating these models.

Acknowledgments

Two referees provided useful comments, which helped us to express our ideas better and clearer. We wish to thank Dr. C. J. van Westen, International Institute of Aerospace Survey and Earth Science (ITC), The Netherlands, who provided the spatial data and the expert's knowledge in Table 1. We also wish to acknowledge a NATO Collaborative Research Grant awarded to the authors in November 1997. The study was also partly funded from a research grant provided to the Spatial Data Analysis Laboratory of the Geological Survey of Canada by PCI Inc., Richmond Hill, Canada.

References

Agterberg, F.P., G.F. Bonham-Carter, and D.F. Wright, 1990. Statistical pattern integration for mineral exploration, *Computer Applications in Resource Estimation, Prediction and Assessment of Metals and Petroleum* (G. Gaal and D.F. Merriam, editors), Pergamon Press, New York, pp. 1–21.

Aronoff, S., 1989. *Geographic Information Systems: A Management Perspective*, WDL Publications, Ottawa, Canada, 294 p.

Aspinall, R.J., 1992. An inductive modeling procedure based on Bayes' theorem for analysis of pattern in spatial data, *International Journal of Geographic Information Systems*, 6(2):105–121.

Bonham-Carter, G.F., 1994. *Geographic Information Systems for Geoscientists. Modeling with GIS*, Pergamon, New York, 398 p.

Carrara, A., 1983. Multivariate models for landslide hazard evaluation, *Mathematical Geology*, 15(3):403–427.

Carrara, A., M. Cardinali, and F. Guzzeti, 1992. Uncertainty in assessing landslide hazard and risk, *ITC Journal*, 1992-2:172–183.

Chung, C.F., and A.G. Fabbri, 1993. The representation of geoscience information for data integration, *Nonrenewable Resources*, 2(2):122–139.

Chung, C.F., A.G. Fabbri, and C.J. van Westen, 1995. Multivariate regression analysis for landslide hazard zonation, *Geographical Information Systems in Assessing Natural Hazards* (A. Carrara and F. Guzzetti, editors), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 107–133.

Chung, C.F., and Y. Leclerc, 1994. A quantitative technique for zoning landslide hazard, *Papers and Extended Abstracts for Technical Programs of IAMG'94, 1994 International Association for Mathematical Geology Annual Conference*, Mont Tremblant, Quebec, Canada, 3–5 October, pp. 87–93.

Draper, N.R., and H. Smith, 1981. *Applied Regression Analysis, Second Edition*, Wiley, N.Y., 709 p.

Fabbri, A.G., and C.F. Chung, 1996. Predictive spatial data analysis in the geosciences, *Spatial Analytical Perspectives on GIS in the Environmental and Socio-Economic Sciences* (M. Fisher, H.J. Scholten, and D. Unwin, editors), GISDATA Series No. 3, Taylor & Francis, London, pp. 147–159..

Jibson, R.W., E.L. Harp, and J.A. Michael, 1998. *A Method for Producing Digital Probabilistic Seismic Landslide Hazard Maps: An Example from the Los Angeles, California, Area*, U.S. Geological Survey Open-File Report 98-113, 17p., 2 plates.

Leclerc, Y., 1994. *The Design of FM: Data Integrating Package for Zoning Natural Hazards in the Developing Countries*, unpublished M.E. Des. Thesis, Environmental Science, Faculty of Environmental Design, University of Calgary, Canada, 127 p.

Leclerc, Y., and C.F. Chung, 1993. *FavMod for Integration of Spatial Geoscience Information*, Geological Survey of Canada, Calgary, Open File 2577 (software), 87 p.

Luzi, L., 1995. *GIS for Slope Stability Zonation in the Fabriano Area, Central Italy*, unpublished M.Sc. Thesis, ITC, Enschede, The Netherlands, 261 p.

Luzi, L., and A.G., Fabbri, 1995. Application of favourability modeling to zoning of landslide hazard in the Fabriano area, central Italy, *Proc. Joint European Conference and Exhibition on Geographic Information, JEC-GIS'95: "From Research Application Through Cooperation,"* Den Haag, The Netherlands, 26–31 March, 1:398–403.

Spiegelhalter, D.J., 1986. A statistical view of uncertainty in expert systems, *Artificial Intelligence and Statistics*, (W.A. Gale, editor), Addison-Wesley Publ. Co., Reading, Massachusetts, pp. 17–55.

van Westen, C.J., 1993. *Application of Geographic Information Systems to Landslide Hazard Zonation*, Ph.D. Thesis, Technical University of Delft, International Institute for Aerospace Survey and Earth Sciences, Enschede, The Netherlands, ITC Publication 15, Vol. 1, 245 p.

van Westen, C.J., H.M.G. van Duren, I. Kruse, and M.T.J. Terlien, 1993. *GISSIZ: Training Package for Geographic Information Systems in Slope Instability Zonation*, ITC, Publication 15, ITC, Enschede, The Netherlands, V. 1 - Theory, 245 p., v. 2, Exercises, 359 p., with 10 diskettes.

Wang, S.Q., and D.J. Unwin, 1992. Modeling landslide distribution on loss soils in China: An investigation, *International Journal of Geographical Information Systems*, 6(5):391–405.

(Received 10 June 1998; accepted 07 October 1998; revised 07 January 1999)

Appendix A. Three Components

Using the Bayesian rules under the condition that $v_1(p)$, $v_2(p)$, $\cdots$, $v_m(p)$ are conditionally independent given the condition that F_p (p will be affected by a future landslide), the joint conditional probability for each pixel p in Equation 2, as shown in the last term of Equation 9, is

$$\begin{aligned}&\boldsymbol{Prob}\{F_p|v_1(p), v_2(p), \cdots, v_m(p)\}\\&= \frac{\boldsymbol{Prob}\{v_1(p)\}\cdots\boldsymbol{Prob}\{v_m(p)\}}{\boldsymbol{Prob}\{v_1(p), \cdots, v_m(p)\}}\,\boldsymbol{Prob}\{F_p\}\,\frac{\boldsymbol{Prob}\{F_p|v_1(p)\}}{\boldsymbol{Prob}\{F_p\}}\\&\quad\cdots\frac{\boldsymbol{Prob}\{F_p|v_m(p)\}}{\boldsymbol{Prob}\{F_p\}}\end{aligned}\qquad(A.1)$$

The first component, as shown in Equation 10, the ratio of $\boldsymbol{Prob}\{v_1(p)\}\cdots\boldsymbol{Prob}\{v_m(p)\}$ and $\boldsymbol{Prob}\{v_1(p), \cdots, v_m(p)\}$, consists of the probabilities related to the spatial input, and is easily obtained. To clearly understand it, however, we must temporarily suppose that the m thematic classes, $\mathbf{A}_{1c_1}, \cdots, \mathbf{A}_{mc_m}$ for p are statistically independent. Then

$$\textbf{Prob}\{v_1(p), \cdots, v_m(p)\} = \textbf{Prob}\{v_1(p)\} \cdots \textbf{Prob}\{v_m(p)\},$$

and, hence, the first component, the ratio, is equal to 1. A meaning for the independence is that the m patterns, $\mathbf{A}_{1c_1}, \cdots, \mathbf{A}_{mc_m}$, one for each layer, are not "statistically related to" each other. Hence, they are random patterns, although all patterns have the pixel p in common. Such an assumption of independence is certainly unrealistic, however, and should not be made in practice. The independence implies the conditional assumption made in Equation 8, where the m patterns are independent provided that the pixel p will be affected by a future landslide. Hence, this is much stronger and stricter than the conditional independence assumption.

Completely opposite to the independence assumption is the situation if we assume that the m patterns, $\mathbf{A}_{1c_1}, \cdots, \mathbf{A}_{mc_m}$, are completely correlated, i.e., $\mathbf{A}_{1c_1}, \cdots, \mathbf{A}_{mc_m}$, are identical. Then,

$$\textbf{Prob}\{v_1(p), \cdots, v_m(p)\} = \textbf{Prob}\{v_k(p)\} \text{ for any } k,$$

the ratio in the first component is simply $\textbf{Prob}\{v_k(p)\}^{m-1}$ which becomes very small, nearly zero, whenever the pattern $\mathbf{A}_{kc_k}$ is reasonably smaller than the whole study area. In this situation, the conditional probability in Equation A.1 becomes nearly zero.

The second component is the prior probability $\textbf{Prob}\{F_p\}$:

$$\textbf{Prob}\{F_p\} = \textbf{Prob}\{p \in \mathbf{F}\} = \text{size of } \mathbf{F}/\text{size of } \mathbf{A}, \quad \text{(A.2)}$$

which is the probability that a pixel p is contained in a future landslide prior to that for which we have any evidence. It does not depend either on the location of the pixel p or on the pixel values at p. It is an identical value for all pixels and, hence, it is not critical when we compare the relative significance of pixels with respect to the landslide prediction. It can only be determined by expert's knowledge where the areas to be affected by future landslides in the study will be hypothesized.

We will examine the m factors in the third component now. It is the component showing how each of the m evidences is related to the prediction model, i.e., the joint conditional probability in Equation A.1. If the areas to be affected by future landslides are known (impossible in practice), then the bivariate conditional probabilities in the component are obtained by

$$\begin{aligned}\textbf{Prob}\{F_p|v_k(p)\} &= \textbf{Prob}\{p \in \mathbf{F}|p \in \mathbf{A}_{kc_k}\} \\ &= \text{size of } \mathbf{F} \cap \mathbf{A}_{kc_k}/\text{size of } \mathbf{A}_{kc_k}. \quad \text{(A.3)}\end{aligned}$$

In the k^{th} factor of the component, the ratio of $\textbf{Prob}\{F_p|v_k(p)\}$ and $\textbf{Prob}\{F_p\}$, let us assume that $\textbf{Prob}\{F_p|v_k(p)\}$ is greater than $\textbf{Prob}\{F_p\}$. This would indicate that the pattern $\mathbf{A}_{kc_k}$ (the evidence in the k^{th} layer) corresponding to $v_k(p)$ $(=c_k)$ can be considered as a positive effect toward having future landslides. Hence, the ratio, the k^{th} factor would be greater than 1. Otherwise, the bivariate conditional probability has to be smaller than the prior probability. If the pattern $\mathbf{A}_{kc_k}$ does not have any effect on future landslides, then the two probabilities should be identical, i.e., $\textbf{Prob}\{F_p|v_k(p)\} = \textbf{Prob}\{F_p\}$. These arguments for the k^{th} factor can be applied to each of the m factors in the third component.

Appendix B. Weighted least-squares method for data from unique condition subareas.

The least-squares estimator $(\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_m)$ in Equation 17 is obtained by

$$\begin{pmatrix} \hat{\alpha}_0 \\ \vdots \\ \hat{\alpha}_m \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{nm} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

$$x_{ik} = \textbf{Prob}\{S_{p_i}|v_k(p_i)\}, \quad y_i = \textbf{Prob}\{S_{p_i}|v_1(p_i), \cdots, v_m(p_i)\}, \quad \text{(B.1)}$$

p_i represents the ith pixel, n is the number of pixels in the training area, m is the number of data layers, and the size of the matrix $\mathbf{X}$ is $n \times (m + 1)$.

In the training area, when m layers are overlaid, the area becomes divided into h non-overlapping subareas and the pixels in each subarea have m identical pixel values. Such subareas can be termed "unique-condition subareas" and, in most GIS applications, their number is much smaller than that of the pixels. The whole Colombian study area consisted of 437,019 pixels, but after overlaying the seven available data layers, only 4,728 unique-condition subareas could be identified.

Suppose that we have h unique condition subareas in the training area and n_j pixels in the j th unique condition subarea $(n_1 + n_2 + \cdots + n_h = n)$. The same regression coefficients, $(\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_m)$ in Equation B.1 can be obtained by using a much smaller $(h \times (m + 1))$ matrix $\mathbf{V}$ rather than the large $(n \times (m + 1))$ matrix $\mathbf{V}$: i.e.,

$$\begin{pmatrix} \hat{\alpha}_0 \\ \vdots \\ \hat{\alpha}_m \end{pmatrix} = (\mathbf{V}'\mathbf{N}\mathbf{V})^{-1}(\mathbf{V}'\mathbf{N}\mathbf{W}) \quad \text{(B.2)}$$

where

$$\mathbf{V} = \begin{pmatrix} 1 & v_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & v_{h1} & \cdots & v_{hm} \end{pmatrix}, \quad \mathbf{N} = \begin{pmatrix} n_1 & & 0 \\ & \ddots & \\ 0 & & n_h \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} w_1 \\ \vdots \\ w_h \end{pmatrix},$$

$$v_{jk} = \textbf{Prob}\{S_{p_j}|v_k(p_j)\}$$

$$w_j = \textbf{Prob}\{S_{p_j}|v_1(p_j), \cdots, v_m(p_j)\},$$

p_j represents a pixel in the j th unique-condition subarea, h is the number of unique-condition subareas in the training area, m is the number of data layers, and $\mathbf{N}$ is a diagonal matrix with diagonal values $(n_1, \cdots, n_h)$.