The Representation of Geoscience Information for Data Integration

Chang-Jo F. Chung[1] and Andrea G. Fabbri[2]

[1]Geological Survey of Canada, 601 Booth St., Ottawa, Ontario K1A 0E8, Canada, and [2]International Institute for Aerospace Surveys and Earth Sciences (ITC), 350 Boulevard 1945, P.O. Box 6, 7500 AA Enschede, The Netherlands

In mineral exploration, resource assessment, or natural hazard assessment, many layers of geoscience maps such as lithology, structure, geophysics, geochemistry, hydrology, slope stability, mineral deposits, and preprocessed remotely sensed data can be used as evidence to delineate potential areas for further investigation. Today's PC-based data base management systems, statistical packages, spreadsheets, image processing systems, and geographical information systems provide almost unlimited capabilities of manipulating data. Generally such manipulations make a strategic separation of spatial and nonspatial attributes, which are conveniently linked in relational data bases. The first step in integration procedures usually consists of studying the individual characteristics of map features and interrelationships, and then representing them in numerical form (statistics) for finding the areas of high potential (or impact).

Data representation is a transformation of our experience of the real world into a computational domain. As such, it must comply with models and rules to provide us with useful information. Quantitative representation of spatially distributed map patterns or phenomena plays a pivotal role in integration because it also determines the types of combination rules applied to them.

Three representation methods—probability measures, Dempster-Shafer belief functions, and membership functions in fuzzy sets—and their corresponding estimation procedures are presented here with analyses of the implications and of the assumptions that are required in each approach to thematic mapping. Difficulties associated with the construction of probability measures, belief functions, and membership functions are also discussed; alternative procedures to overcome these difficulties are proposed. These proposed techniques are illustrated by using a simple, artificially constructed data set.

Key words:
Dempster-Shafer belief function
Representation
Geographic information system (GIS)
Data integration
Spatially distributed map pattern
Favorability function
Fuzzy set
Certainty factor
Geopotential map
Prediction
Natural resources
Natural hazard
Probability

Introduction

In the earth sciences, spatially distributed data are represented as maps to convey how geographic location, shape, size, distribution, topology, and association of features and processes interrelate to make a comprehensible landscape. Whereas some maps mainly reflect the characteristics of the earth's surface, for example, geomorphology and land use, other maps contain information about deeper layers of the earth's crust, for example, bedrock geology, Bouguer gravity, and aeromagnetic anomaly.

Maps also represent observations and interpretations of nature—including human interference with nature—and are constructed by specialists who associate different levels of confidence, accuracy, and extensibility to observations and to interpretations (Brodaric, 1992). Constructing thematic maps involves a delicate decision process whereby selected features are combined and their associated attributes—including accuracy and confidence—must also be combined. Such attributes can have both a spatial nature, for example, characteristics related to spatial frequency or distribution, and a nonspatial nature, for example, characteristics related to the properties that identify the mapping units as homogeneous distinct objects (McMaster, 1991).

Furthermore, maps may represent different types of observations: (1) continuous measurements, for example, gravitational field; (2) discrete samples, for example, geochemical stream sediment data; (3) distinct geometric objects (polygons, segments, and points), for example, individual landslides, lineaments, mineral occurrences; and (4) complete polygonization of map space, for example, bedrock geology. Frequently, maps consist of many layers of different types of observations overprinted on the same planar surface, where spatial and inherent process relationships are represented for human perception. One layer represents the main theme, whereas the other layers provide the secondary or accessory information.

In some cases, it is desirable to construct special-purpose or thematic maps, in which various aspects of complex associations of mappable features are grouped into classes of importance for human use, such as the assessment of nonrenewable resources, natural hazards, and environmental impacts. The first task is to define the purpose of the integration; the second task is to represent the map information in a uniform way according to the defined purpose. We propose a general background for systematic representation of geoscience spatial information for computational decision processes in support of data integration.

The field of geographic information systems arose from the need to computerize the process of map data capture, management, and analysis for theme representation and modeling. The approach proposed here is based on the definition of favorability functions, on three different interpretations of such functions in terms of certainty, belief, and possibility, and using several methods of estimation. An application on an artificially constructed example is described, and a processing strategy is presented with a discussion of the combination rules required in predictive integration.

As will become evident in the remainder of this article, the approach we propose is equally applicable in exploration and in environmental studies. The term "geopotential maps" was proposed by Luttig (1987) to cover under one roof the thematic nature of maps representing a potential either for natural resource hazard or for environmental impact (van Westen, 1992).

Capturing and Modeling Spatial Data

Capturing spatially distributed data is generally done by computer by assigning a geographic location to one or more points belonging to each feature and by associating to each feature codes for its identification and description. A synthetic form of computer representation is termed the "vector model." In this model, a minimum number of points is used to locate point data, segments, or polygonal features. The elementary objects or features are identified by labels, which are also used in relational tables to associate several attributes to them. Complex hierarchical data structures are generated to express the topological relationships of points, segments, and polygons. The vector model is very efficient for representing objects with low spatial variability, that is, map units with constant value over large areas.

Many types of data, however, are characterized by high spatial variability, such as geophysical or geochemical anomalies, remotely sensed images, or digital elevation. For the latter data types, a satisfactory representation is the "raster model." In this model, a rectangular matrix of numbers is used in which each element, termed a picture element or pixel, is in one-to-one correspondence with a small area on the ground or in the original picture material. The value of the pixel may represent the geophysical intensity field, its brightness, or its membership in a class or in a map unit. Also, point, segment, and polygonal objects can be represented in this model. Absolute location of the pixels is obtained by supplying the coordinates of one pixel, say, at the top left of a rectangular matrix, the pixel size (or better, the size of the area corresponding to the pixel), and the orientation of the matrix or raster. This convention avoids the repetition of such coordinates for each pixel. As with the vector model, topological relationships and other nonspatial attributes can be stored in relational tables where the pixel values are also stored.

In spatial data integration studies, some map data layers may be in raster model format (or variable rasters), whereas other layers may be in vector model format. To compare data sets from different data model formats and to integrate them into a map showing target areas for further study, they should all be conveniently represented in a single data model format. Each of the two basic data models have computational advantages and disadvantages (Aronoff, 1989); however, the raster model seems to be the most practical representation for statistical interpretation and analysis, and is therefore the representation we will use.

Another interesting aspect of the computer representation of map data is the assignment of feature identification to polygons (two dimensions), segments (one dimension), and points (zero dimensions). In the vector model, a point is located by coordinate values in a very fine virtual grid; therefore, a virtual zero dimension is implied. In the raster model, a point is assigned to a pixel containing it and is, therefore, equated to the pixel itself—its location within the pixel is not necessarily specified. In processing spatially distributed data, map unit features or samples must be quantized, that is, the range of their characteristic values, including location coordinates, must be divided into intervals; all the values within an interval must be represented by a single quantization level.

The main consequence of quantization is that in the assessment of a mineral potential theme, we will identify the pixels (or the subpolygons) that are likely to "contain" point-like undiscovered mineral occurrences, whereas in the assessment of landslide hazard, we will identify the points (pixels) that are likely to "be contained" within a future landslide (set of pixels or a polygon). This also depends on the relationship of pixel size to deposit size. In some cases, to "be contained" would apply to mineral deposits. Such a duality of theme representation is simply due to the fact that the target of some themes can be considered as having zero dimensions, whereas in other themes the target can have one or more dimensions.

Consider a study region A and m layers, $L_1,...,L_m$, of spatial data in A. The whole area is divided into small cells, and the data are recorded at each pixel for each layer. Obviously, the size of the area corresponding to the pixel plays an important role and also determines what types of representation and measurement are suitable for describing the data and their location. This becomes particularly crucial if the size of the pixel is very large. For continuous data the average minimum, maximum, and/or range (maximum-minimum) values within each pixel may be used. For polygonal data, binary (presence/absence) representations for each polygon type (for example, color) at each pixel within it can be the effective description.

Most map data not only represent real observations but also abstractions or interpretations from a limited number of observations. Thus even if a map shows the coverage of the entire area for a particular map unit, real observations of the unit are usually made at a few locations, for example, along flight lines for airborne geophysical surveys, at one or two sample points for each station in lake or stream sediment surveys, or along traverses by geologists in reconnaissance mapping. For these reasons, maps can differ in quality and in the level of uncertainty assigned to each occurrence of a map unit. Varnes (1974) describes engineering geological maps in which different degrees of uncertainty are assigned to different map units and to the contacts between such units. Similarly, more recent and detailed geological maps separate the outcrop pattern from the interpretations, thus making it easier to integrate map data with remotely sensed data (airborne or spaceborne). In addition, the description and homogeneity of mapping units is frequently represented in lengthy legends, which, in some cases, lead to a synthetic and structured description of the map symbology and characterization. Often, the understanding of such symbology is an essential key to quantitative representation and statistical modeling. We can conclude, therefore, that the computer representation of mappable data requires various levels of abstraction, redefinition, and absolute positioning.

Representation of Favorability Functions

Consider m layers of map data each containing one set of map units (that is, a single theme) in a study area A where an exploration target for a specific mineral deposit type D is sought. Consider a pixel p in A and a proposition:

$$T_p: \text{"}p \text{ contains a deposit of type D."} \qquad (1)$$

The layers can be grouped into two classes, one of continuous and the other of noncontinuous measurements. For each kth layer L_k containing continuous measurements, the quantized value at p for L_k is a finite interval $[\min_k, \max_k]$, where $\min_k$ and $\max_k$ are known. For each layer L_k containing noncontinuous measurements, such as lithology, without loss of generality, we may assume that the quantized value at p for L_k takes one integer value among $\{1, 2, \ldots, n_k\}$, where n_k is the maximum number of the map units in L_k. One of these n_k units may represent the areas not covered by observed values in the kth layer (that is, it may represent an "others" class). Then, the m layers of map data at every p in **A** are represented, in a quantized form, by

$$\{(v_k(p), k = 1,...,m), p \in \mathbf{A}\}, \tag{2}$$

where $v_k(p)$ is the quantized value for the kth layer $\mathbf{L}_k$ at p. Here the m quantized values $(v_k(p), k = 1, \ldots, m)$ in equation 2 are regarded as m pieces of evidence of the proposition that p contains a deposit.

However, one has to be very careful in interpreting the value of $v_k(p)$ because very often it only represents a unit in that layer at p; for example, on a geological map, it represents the presence of a specific rock type. Therefore, $v_k(p) < v_k(q)$ does not mean anything except that p belongs to the $v_k(p)$ unit and q belongs to the $v_k(q)$ unit.

We may regard the quantization, v_k, as a function of **A** into a finite interval for the kth layer:

$$v_k: \begin{cases} \mathbf{A} \to [\min_k, \max_k], & \text{if } \mathbf{L}_k \text{ is continuous pattern,} \\ \mathbf{A} \to \{1, 2, \ldots, n_k\}, & \text{if } \mathbf{L}_k \text{ is noncontinuous pattern.} \end{cases} \tag{3}$$

For our exploration target, mineral deposit of type **D** in **A**, the value $v_k(p)$ alone at p may have very little meaning [that is, $v_k(p)$ means only that p belongs to the $v_k(p)$th map unit in the kth layer], although $v_k(p)$ could have a significant meaning with respect to the other values within the kth layer, $\mathbf{L}_k$. In addition, when we compare $v_k(p)$ with the value $v_l(p)$ of the lth layer $\mathbf{L}_l$ at p, it is again difficult to comprehend the significance of the comparison with respect to a mineral exploration target of deposit type **D** in **A**, because $v_k(p)$ may represent the map unit of the rock type, whereas $v_l(p)$ may symbolize aeromagnetic intensity observation at p, for example. Only a physical model might justify assigning importance to such a comparison (Green and Craig, 1984).

As a first step in quantitative modeling for data integration in exploration and environmental studies, we propose to define a relative favorability index function, r_k, for each kth layer $\mathbf{L}_k$. We first define for each layer:

$$r_k: \begin{cases} [\min_k, \max_k] \to [\alpha, \beta], & \text{if } \mathbf{L}_k \text{ is continuous pattern,} \\ \{1, 2, \ldots, n_k\} \to [\alpha, \beta], & \text{if } \mathbf{L}_k \text{ is noncontinuous pattern.} \end{cases} \tag{4}$$

where α and β are two known constants, and $r_k(\delta)$ ($\in [\alpha, \beta]$) represents a measurement related to the "sureness"—probability, certainty, belief, plausibility, possibility, or compatability—that the proposition (p contains a deposit of type **D**) is true given the evidence of δ at p in the kth layer $\mathbf{L}_k$. We define the compound function f_k of r_k and v_k at each p in **A**: $f_k(p) = r_k(v_k(p))$, for all $p \in \mathbf{A}$, that is,

$$f_k: \begin{cases} \mathbf{A} \xrightarrow{v_k} [\min_k, \max_k] \xrightarrow{r_k} [\alpha, \beta], & \text{if } \mathbf{L}_k \text{ is continuous pattern,} \\ \mathbf{A} \to \{1, 2, \ldots, n_k\} \to [\alpha, \beta], & \text{if } \mathbf{L}_k \text{ is noncontinuous pattern.} \end{cases} \tag{5}$$

where f_k is termed a favorability function for the kth layer $\mathbf{L}_k$. A $f_k(p)$ value near α indicates that the sureness that the proposition is true is very low, that is, given the evidence $v_k(p)$ at p, the support that p belongs to mineralized areas with deposits of type **D** is very low. However, a $f_k(p)$ value near β means that the evidence $v_k(p)$ strongly supports the proposition being true.

Many different interpretations, such as probability, certainty, belief, plausibility, or possibility, of the measurement related to the favorability function can be given; the values α and β depend on the interpretation. We will discuss this in the next section.

For a given proposition, suppose that we have defined m favorability functions $f_k(p)$ $(k = 1,...,m)$, one for each layer. Contrary to the quantized value, $v_k(p)$ for the kth layer $\mathbf{L}_k$ at p, $f_k(p)$ then contains a significant meaning with respect to the proposition. For example, $f_k(p) < f_k(q)$ implies that within the kth layer $\mathbf{L}_k$, the support for the proposition from the evidence $v_k(q)$ at the pixel q is stronger than the evidence $v_k(p)$ at the pixel p. Another example is that $f_k(p) < f_l(p)$ implies that, at the pixel p, the support for the proposition from the evidence $v_l(p)$ of the lth layer $\mathbf{L}_l$ is stronger than the evidence $v_k(p)$ of the kth layer $\mathbf{L}_k$. In other words, if we wish to study the contributions of the evidence of the kth and the lth layers to the proposition of **D** at a pixel p, we may examine $f_k(p)$ and $f_l(p)$ instead of $v_k(p)$ and $v_l(p)$. We will elaborate on this in the next sections.

At each pixel p we have

$$(f_1(p), f_2(p), \ldots, f_m(p)), \tag{6}$$

indicating how each of the m pieces of evidence $(v_k(p), k = 1,...,m)$, in equation 2 supports the sureness that the proposition is true at p. We will discuss how to estimate these m functions in the section on estimation of favorability functions.

Therefore, the m layers of map data (or individual themes) are now represented by

$$\{(f_k(p), k = 1,...,m), p \in \mathbf{A}\}, \tag{7}$$

instead of equation 2.

This proposed representation (eq. 7) is more useful more relevant in the exploration of deposits of type D in A than the original representation in equation 2. Similar considerations can be easily made in the analysis of a set of maps for landslide hazard assessment.

Interpretation of Favorability Functions

The favorability function defined in equation 5 represents a measurement related to the sureness—probability, certainty, belief, plausibility, or possibility—that the proposition is true given the evidence at a pixel.

Probabilistic Interpretation

Given a proposition, the favorability function $f_k(p)$ is interpreted as the conditional probability, denoted by $\mathrm{Prob}_k\{\mathrm{T}_p \text{ is true} \mid v_k(p)\}$ that p contains at least one deposit given the evidence $v_k(p)$ at p in L_k; that is, for $k = 1,2,\ldots,m$,

$$f_k(p) = \mathrm{Prob}_k\{\mathrm{T}_p \text{ is true} \mid \text{evidence } v_k(p) \text{ at } p\}. \tag{8}$$

Another possible interpretation of $f_k(p)$ as a probabilistic representation is that, for $k = 1,2,\ldots,m$,

$$f_k(p) = \mathrm{Prob}_k\{\text{evidence } v_k(p) \text{ is observed at } p \mid \mathrm{T}_p \text{ is true}\}. \tag{9}$$

Here we define $f_k(p)$ as a conditional probability that the evidence $v_k(p)$ is observed at p, given that the proposition is true. This representation is somewhat related to the inversion theory in geophysics. When we compare these two interpretations in equations 8 and 9, the conditional probability in equation 8 is a much more natural choice than that in equation 9 for the favorability function $f_k(p)$. Hence, we propose from now on $f_k(p) = \mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\}$ in equation 8 as our probabilistic interpretation of the favorability function, and that α and β in equation 5 are 0 and 1, respectively.

However, using Bayes' theorem, we also establish the relationship between these two probability interpretations in equations 8 and 9:

$$\begin{aligned}&\mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\}\ \mathrm{Prob}_k\{v_k(p)\}\\ &\quad= \mathrm{Prob}_k\{v_k(p) \mid \mathrm{T}_p\}\ \mathrm{Prob}_k\{\mathrm{T}_p\},\end{aligned} \tag{10}$$

where T_p denotes the proposition that p contains a deposit of type D. $\mathrm{Prob}_k\{\mathrm{T}_p\}$ is the prior probability that a pixel p contains a deposit before we have any evidence (layers), and $\mathrm{Prob}_k\{v_k(p)\}$ is the probability that p has evidence $v_k(p)$. Both $\mathrm{Prob}_k\{\mathrm{T}_p\}$ and $\mathrm{Prob}_k\{v_k(p)\}$ can be relatively easily estimated from the data and do not play any crucial roles. Particularly, $\mathrm{Prob}_k\{\mathrm{T}_p\}$ should be a constant for all k's and all pixels, because it relates to neither any specific layers nor any specific pixels.

Which one of the two conditional probabilities in equations 8 or 9 we choose is obviously crucial to the final results, although we can convert one interpretation to the other easily using equation 10.

Certainty Factors (CF) Interpretation

In a rule-based system (or expert system), uncertainty is usually associated with the rule from evidence of a proposition. A certainty factor (CF) at p for the kth layer, denoted by $\mathrm{CF}_k(p)$ is defined as the change in certainty that the proposition (a pixel p contains deposits of type D) is true, from without the evidence $v_k(p)$ at p to given the evidence $v_k(p)$ at p in the kth layer. Certainty factors (CF's) originally proposed by Shortliffe and Buchanan (1975), for the medical expert system MYCIN, range between -1 and $+1$. Positive numbers for CF correspond to an increase in certainty in a proposition after the evidence $v_k(p)$ is observed, whereas negative numbers correspond to a decrease in certainty. In this interpretation, α and β in equation 5 are -1 and $+1$, respectively, and for $p \in \mathrm{A}$ and the kth layer L_k,

$$f_k(p) = \mathrm{CF}_k(p). \tag{11}$$

However, CF is defined as a function of probability. The original CF proposed by Shortliffe and Buchanan (1975) was defined as

$$\mathrm{CF}_{\mathrm{org}}(p) = \begin{cases} \dfrac{\mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\} - \mathrm{Prob}_k\{\mathrm{T}_p\}}{1 - \mathrm{Prob}_k\{\mathrm{T}_p\}}, \\ \quad \text{if } \mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\} > \mathrm{Prob}_k\{\mathrm{T}_p\}, \\ \dfrac{\mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\} - \mathrm{Prob}_k\{\mathrm{T}_p\}}{\mathrm{Prob}_k\{\mathrm{T}_p\}}, \\ \quad \text{if } \mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\} < \mathrm{Prob}_k\{\mathrm{T}_p\}. \end{cases} \tag{12}$$

From this definition, reasonable strategies for data integration could not be established for $\mathrm{CF}_{\mathrm{org}}(p)$ when uncertainties are propagated through several pieces of evidence. Heckerman (1986) discussed the desiderata of CF for propagating uncertainties through a complex inference network, and proposed the following modified certainty factor in terms of probabilities:

$$\mathrm{CF}_k(p) = \begin{cases} \dfrac{\mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\} - \mathrm{Prob}_k\{\mathrm{T}_p\}}{\mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\}(1 - \mathrm{Prob}_k\{\mathrm{T}_p\})}, \\ \quad \text{if } \mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\} > \mathrm{Prob}_k\{\mathrm{T}_p\}, \\ \dfrac{\mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\} - \mathrm{Prob}_k\{\mathrm{T}_p\}}{\mathrm{Prob}_k\{\mathrm{T}_p\}(1 - \mathrm{Prob}_k\{\mathrm{T}_p\})}, \\ \quad \text{if } \mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\} < \mathrm{Prob}_k\{\mathrm{T}_p\}. \end{cases} \tag{13}$$

By following the probability rules, CF defined in equation 13 satisfies the desiderata to propagate uncertainties through a complex inference network. An excellent discussion on this subject was provided by Heckerman (1986). According to this definition, CF is equal to zero if the conditional probability is equal to the prior probability; the absolute value of CF increases if the conditional probability is far from the prior probability. Thus, CF is a measure of certainty with respect to the prior probability only. From now on when referring to CF, we will use the definition in equation 13.

Dempster-Shafer Belief Function Interpretation

In this approach (Shafer, 1976), two favorability functions for each layer, denoted by $\mathrm{Bel}_k\{\mathbf{T}_p \mid v_k(p)\}$ and $\mathrm{Pls}_k\{\mathbf{T}_p \mid v_k(p)\}$, are defined to represent the minimum and maximum degree of belief that the evidence $v_k(p)$ of $\mathbf{L}_k$ supports the proposition that a pixel p contains deposits. These functions are called a belief function and a plausible function; they range from 0 to 1. Whereas the belief function represents the minimum belief that $v_k(p)$ of $\mathbf{L}_k$ supports the proposition that the pixel p contains deposits, the plausible function represents the maximum belief. The difference $\mathrm{Pls}_k\{\mathbf{T}_p \mid v_k(p)\} - \mathrm{Bel}_k\{\mathbf{T}_p \mid v_k(p)\}$ represents ignorance of one's belief that p contains deposits given evidence $v_k(p)$ in $\mathbf{L}_k$ (Shafer, 1976).

Another interpretation of these two functions is associated with the uncertainty of the conditional probability that the proposition is true given the evidence $v_k(p)$ as discussed in equation 8. Instead of specifying one probability measure, $\mathrm{Prob}_k\{\mathbf{T}_p \mid v_k(p)\}$, we present an interval $[\mathrm{Bel}_k\{\mathbf{T}_p \mid v_k(p)\}, \mathrm{Pls}_k\{\mathbf{T}_p \mid v_k(p)\}]$. The distance, $\mathrm{Pls}_k\{\mathbf{T}_p \mid v_k(p)\} - \mathrm{Bel}_k\{\mathbf{T}_p \mid v_k(p)\}$, of the interval represents the uncertainty associated with $\mathrm{Prob}_k\{\mathbf{T}_p \mid v_k(p)\}$.

We have the following relationship between these two functions. Let $\mathrm{Bel}_k\{\bar{T}_p \mid v_k(p)\}$ represent the degree of belief that the evidence $v_k(p)$ in $\mathbf{L}_k$ supports the proposition that the pixel p does *not* contain any deposit. Then the plausible function can be expressed (Shafer, 1976) as

$$\mathrm{Pls}_k\{\mathbf{T}_p \mid v_k(p)\} = 1 - \mathrm{Bel}_k\{\bar{T}_p \mid v_k(p)\}. \tag{14}$$

Comparing this approach with the probabilistic interpretation in the previous section, we would expect the following relationship (Wally, 1987):

$$\mathrm{Bel}_k\{\mathbf{T}_p \mid v_k(p)\} < \mathrm{Prob}_k\{\mathbf{T}_p \mid v_k(p)\} < \mathrm{Pls}_k\{\mathbf{T}_p \mid v_k(p)\}. \tag{15}$$

If $\mathrm{Pls}_k\{\mathbf{T}_p \mid v_k(p)\} - \mathrm{Bel}_k\{\mathbf{T}_p \mid v_k(p)\} = 0$, and hence

$$\mathrm{Bel}_k\{\mathbf{T}_p \mid v_k(p)\} = 1 - \mathrm{Bel}_k\{\bar{T}_p \mid v_k(p)\}, \tag{16}$$

then Bel_k in equation 16 is termed the Bayesian belief function, and we would expect all three mappings, $\mathrm{Pls}_k\{\mathbf{T}_p \mid v_k(p)\}$, $\mathrm{Prob}_k\{\mathbf{T}_p \mid v_k(p)\}$, and $\mathrm{Bel}_k\{\mathbf{T}_p \mid v_k(p)\}$, to be identical.

Fuzzy Logic Interpretation

Consider a fuzzy set S consisting of all the pixels ($p \in \mathbf{A}$) where the proposition is "likely" to be true (Zadeh, 1965). The fuzzy set S is defined by a membership function μ_S:

$$\mu_S\colon \mathbf{A} \to [0, 1], \tag{17}$$

which represents grade of membership or degree of compatibility. We may also interpret the membership function as a possibility function (Zadeh, 1978) of p, where the proposition is likely to be true (Zadeh, 1965) and is denoted by a set of ordered pairs,

$$S = \{(p, \mu_S(p)), p \in \mathbf{A}\}. \tag{18}$$

The main difference between randomness (probability) and fuzziness is as follows: Randomness has to do with uncertainty concerning membership or nonmembership of an object in a nonfuzzy (ordinary) set that has precisely defined criteria of membership (any object is either in the set or not in the set), and the probability is the measurement of the uncertainty. Fuzziness, on the other hand, has to do with sets in which objects may have grades of membership intermediate between full membership and nonmembership (a fuzzy set does not have precisely defined criteria of membership; it has only fuzzily defined criteria), and the membership function is the measurement of the fuzziness (Bellman and Zadeh, 1970).

The problem of estimating μ_k from the prior knowledge of the set of m pairs, $(p_1, \mu_S(p_1))$, $(p_2, \mu_s(p_2))$, $\ldots$, $(p_m, \mu_S(p_m))$, is called abstraction (Bellman and others, 1966).

At each kth layer $\mathbf{L}_k$, we define a fuzzy set S_k by a membership function μ_k based on the evidence in $\mathbf{L}_k$:

$$\mu_k\colon \mathbf{A} \to [0, 1]. \tag{19}$$

That is, at each layer, we define a fuzzy set S_k:

$$S_k = \{(p, \mu_k(p)), p \in \mathbf{A}\}. \tag{20}$$

In this interpretation, the favorability function is $f_k(p) = \mu_k(p)$ for each kth layer $\mathbf{L}_k$; it represents the grade of membership of the pixel p in which the proposition is likely to be true. If the membership function $\mu_k(p)$ is near 1, it implies that p is likely to contain a deposit.

Estimation of Favorability Functions

In a study area, the construction and estimation of favorability functions for map layers are complex but cru-

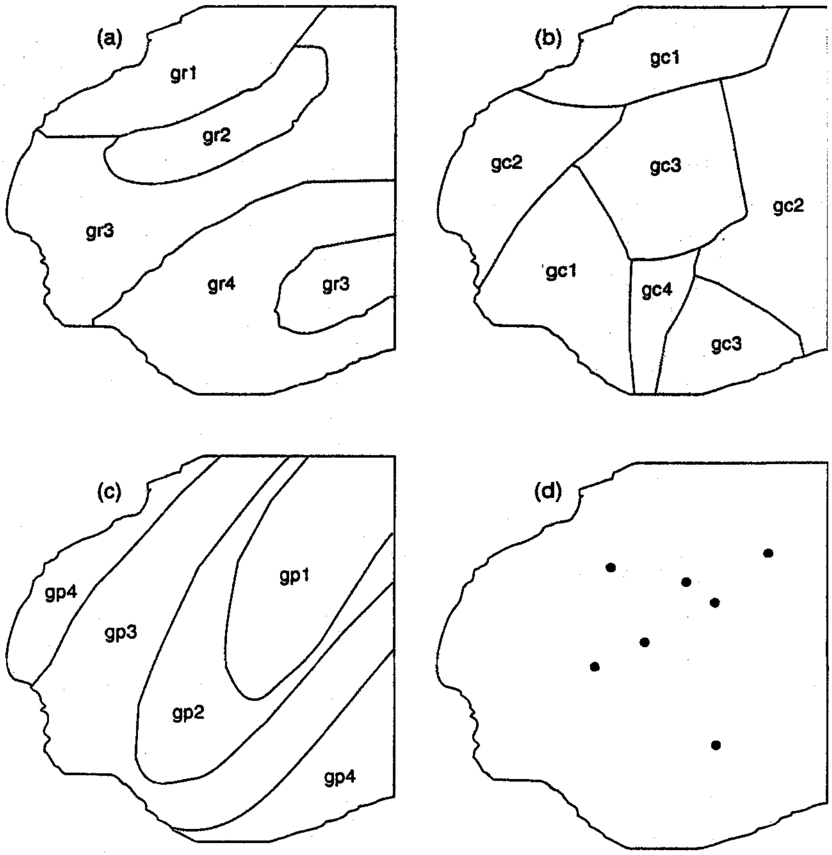

Figure 1. Plots of three hypothetical simple thematic maps: *a*, a geological map with four units, labeled gr1–gr4; *b*, a geochemical map with four units labeled gc1–gc4; and *c*, a geophysical interval map with four units, labeled gp1–gp4. *d*, Seven known mineral deposits. See tables 1 and 2 for the statistics on the maps and on their overlays to the known mineral deposits.

cially important tasks in data integration. To obtain favorability functions, the first step is defining a proposition T_p, such as "A pixel contains at least one mineral occurrence of a given type," or "A pixel will be affected by the environment."

Favorability functions depend on the corresponding propositions. The more precisely the proposition is defined, the easier it is to construct a favorability function. For instance, it would be much easier to construct a favorability function for a proposition stating that "A pixel will be affected by active landslides of the type flowslide" than from one simply stating that "A pixel will be affected by a landslide." On the other hand, the usefulness of a study is limited by the scope of the proposition. Hence, it is important to balance the ease of constructing favorability functions and the simplicity of formulating propositions.

A favorability function must be constructed for each layer in a study area. For a given layer and a given proposition, there is no unique procedure for constructing the favorability function. In this article we consider three separate situations: (1) few or no recorded occurrences of the proposition in environments similar to the one in the study area; (2) known occurrences within the study area; and (3) known occurrences in similar environments outside the study area.

To illustrate the estimation procedures of the favorability functions in this section, we construct an artificial study area that contains these four layers: (i) lithological map; (ii) geochemical anomaly map based on a stream sediment geochemical survey and a drainage basin map; (iii) geophysical anomaly map; and (iv) location map of the known volcanogenic massive sulphide deposits. They are shown in figure 1; statistics related to the spatial data are shown in table 1. To simplify the discussion in this example, we view the deposits as points in two-dimensional space.

No Known Occurrences in Study Area

In the absence of known occurrences in a similar environment that could be used as a control area, the construction of favorability functions depends on the evidence in the study area and on the opinion of experts; therefore, it is largely subjective. In the absence of a strict rule, we discuss only one possible scenario in the following example, shown in figure 1. Knowledge acquisition techniques described by McGraw and Harbison-Briggs (1989) and Butler and Carter (1986) should play an important role, although such techniques are not used here.

For each map layer, the first step is to order the map units with respect to their importance in the proposition. Let us assume that the proposition T_p is

Table 1. Pixel counts for the map units in figure 1, the three maps of a hypothetical study area for mineral exploration.

a. Geology			*b.* Geochemistry			*c.* Geophysics		
Type	Name	No. of pixels	Type	Name	No. of pixels	Type	Name	No. of pixels
gr1	Intrusive	7,777	gc1	Low	15,304	gp1	High	10,733
gr2	Mixed volcanics	5,236	gc2	No data	18,654	gp2	Medium	9,579
gr3	Acidic volcanics	21,455	gc3	High	13,219	gp3	Low	18,828
gr4	Sedimentary	15,255	gc4	Very low	2,546	gp4	Very low	10,583

"A pixel p contains at least one volcanogenic massive sulphide (VMS) deposit." (21)

Suppose that the lithological map layer in figure 1*a* represents the bedrock geology map units as follows: (1) intrusive rocks, (2) mixed volcanic rocks, (3) acidic volcanic rocks, and (4) sedimentary rocks. For the proposition, a possible order of the map units is mixed volcanics (2), acidic volcanics (3), intrusives (1), and sediments (4). Under this order, the favorability function for the layer should satisfy the expression

$$\beta \geq f(2) \geq f(3) \geq f(1) \geq f(4) \geq \alpha. \qquad (22)$$

The next step is to determine which map units are unrelated to the proposition and the relative importance of space-time contacts between any two map units.

We assign weights as follows: $f(4) = \alpha$, because the sedimentary rocks are not genetically related to the mineralization. Suppose that our experts say that the ratio $f(2)/f(3) = 3$, approximately, that is, that mixed volcanics are three times more important than acidic volcanics. Suppose that our experts say that $f(3)/f(1) = 10$, that is, acidic volcanics are one order of magnitude more important than intrusives from the point of view of VMS mineralization. Then we can set the following relations:

$$\begin{aligned} &f(4) = \alpha, \\ &f(2)/f(3) = 3, \qquad (23) \\ &f(3)/f(1) = 10. \end{aligned}$$

The expressions in relations 22 and 23 imply that

$$\begin{aligned} &f(2) = 30t < \beta, \\ &f(3) = 10t, \\ &f(1) = t, \qquad (24) \\ &f(4) = \alpha, \end{aligned}$$

where t is a constant term assigned to $f(1)$, and the weights in equations 24 reflect the experts' view about the relative properties of deposits per unit area of the given rock units.

The final step in the construction of a favorability function for this layer is to determine α and β. In this example, the value of t in equations 24 depends on how much support the layer can provide to the proposition of the presence of volcanogenic massive sulphide deposits. It may also depend on the size of the area corresponding to the pixel and on our interpretation of the favorability function (or of the model underlying it).

Probability Function. As discussed previously, we may assume that the favorability function is regarded as the conditional probability that the proposition is true given that the pixel or point belongs to a map unit in the map layer. Returning to our example, we can write

$$\begin{aligned} f(i) &= \text{Prob}\{\text{the pixel contains at least one VMS} \\ &\qquad \text{deposit} \mid \text{a pixel is assigned to unit } k\} \qquad (25) \\ &= \text{Prob}\{\text{T}_p \mid V_k(p)\}. \end{aligned}$$

First we assign $\alpha = 0$ and $\beta = 1$. Although we may assume that all the values assigned to the $f(i)$'s in equations 24 hold here, we must be careful in selecting the value for t to reflect equation 25. For example, if five VMS deposits known in unit 3 are believed by the experts to be all that exist, then, considering the area occupied by acidic volcanics (21,455 pixels as shown in table 1*a*) and using equation (A1) discussed in the Appendix, t should satisfy

$$f(3) = 10t = 1 - (1 - (1/21{,}455))^5 = .000233, \qquad (26)$$

where the area covered by acidic volcanics is 2,145,500 m^2, the area occupied by a pixel (that is, the pixel size) is 100 m^2, and $f(3)$ illustrates the conditional probability that a pixel (100 m^2) contains at least one of the five deposits given that the pixel is covered by acidic volcanics. Equation 26 implies that $t = .000023$.

However, if the experts also believe that unit 1, intrusive rock, contains one VMS deposit, then we have

$$f(1) = t = 1 - (1 - (1/7777))^1 = .000129, \qquad (27)$$

where the area occupied by intrusive rocks is 777,700 m^2. Equation 26 implies that $t = .000129$.

Obviously, there is a conflict in determining t by equation 26 and by equation 27. In this situation, one or two of the following three assumptions must be modified to estimate t consistently:

(i) $f(3) = 10t$ and $f(1) = t$,

(ii) five deposits exist in unit 3,

(iii) one deposit exists in unit 1.

Suppose that we accept (i) and (iii). We have $f(3) = .00129$. Then the estimated number of deposits in unit 3, by using (A3) in the Appendix, is

$$n = \text{integer part of } \{.5 + \log(0.99871)/\log(21{,}454/21{,}455)\} = 27, \quad (28)$$

which is quite different than (ii) in which a total of only five deposits is expected. Hence, we may construct the favorability function in terms of probability by accepting (i) and (ii), or (i) and (ii)′, where for (ii)′, the experts state "Ten deposits exist in unit 3." If we accept (i) and (ii)′, then

$$t = (1 - (21{,}454/21{,}455)^{10})/10 = .000047, \quad (29)$$

and all $f(i)$'s are assigned according to equations 24. Hence, in this example,

$$\begin{aligned} f(1) &= \text{Prob}\{T_p \mid p \text{ is in intrusive rock}\} \\ &= t = .000047, \\ f(2) &= \text{Prob}\{T_p \mid p \text{ is in mixed volcanic rock}\} \\ &= 30t = .00141, \\ f(3) &= \text{Prob}\{T_p \mid p \text{ is in acidic volcanic rock}\} \\ &= 10t = .00047, \\ f(4) &= \text{Prob}\{T_p \mid p \text{ is in sedimentary rock}\} = 0. \end{aligned} \quad (30)$$

By following this type of iterative procedure, we may guide the experts to bring out and logically modify or adapt their opinions to determine the favorability functions.

Certainty Factor. Certainty factors can be estimated two ways. The first procedure is direct estimation, where α and β in equation 5 are -1 and $+1$, positive numbers for CF correspond to an increase in certainty in a proposition, and negative numbers correspond to a decrease in certainty. We do not discuss this procedure here, because similar lines of thought as in the section on no known occurrences are applied in this procedure. The other procedure is through equation 13 by first estimating the probability as discussed in the previous section on probability function.

In this latter procedure using equation 13, we need to estimate the prior probability Prob{p contains a deposit}, which represents the probability that a pixel contains a deposit prior to obtaining any evidence, in addition to the conditional probabilities discussed in the previous section. Obviously,

$$\text{Prob}_k\{p \text{ contains at least one deposit}\} = \text{Prob}\{T_p\}, \text{ for all } k, \quad (31)$$

and this prior probability depends on the expected total number of deposits in the study area regardless of any evidence.

Following the discussion in the previous section, from equations 30, in our example we obtain the estimated number of deposits in each rock type by using equation A3 in the Appendix:

$$\left.\begin{aligned} n(1) &= \#\text{ of deposits in intrusive rock} = 1, \\ n(2) &= \#\text{ of deposits in mixed volcanic rock} = 8, \\ n(3) &= \#\text{ of deposits in acidic volcanic rock} = 10, \\ n(4) &= \#\text{ of deposits in sedimentary rock} = 0. \end{aligned}\right\} \quad (32)$$

The expected total number of deposits is approximately 19. Using equation A1 in the Appendix, we estimate

$$\text{Prob}\{T_p\} = 1 - (1 - (1/49{,}723))^{19} = .000382. \quad (33)$$

Comparing equations 33 and 30, we see that the intrusive and sedimentary rocks provide negative evidence for the proposition, whereas the mixed and acidic volcanic rocks provide positive evidence. Using equations 30, 33, and 13, we have

$$\begin{aligned} CF(1) &= -.87, \\ CF(2) &= .74, \\ CF(3) &= .19, \\ CF(4) &= -1.00. \end{aligned} \quad (34)$$

Belief Function and Plausible Function. Instead of specifying one probability measure, $\text{Prob}_k\{T_p \mid v_k(p)\}$, as we did in the section on probability functions, we construct an interval $[\text{Bel}_k\{T_p \mid v_k(p)\}, \text{Pls}_k\{T_p \mid v_k(p)\}]$. One of the difficulties in constructing the interval is that $\mathrm{Bel}_k\{\bar{T}_p \mid v_k(p)\}$ $(=1 - \mathrm{Pls}_k\{\mathrm{T}_p \mid v_k(p)\}) + \mathrm{Bel}_k\{\mathrm{T}_p \mid v_k(p)\} \leq 1$, contrary to $\mathrm{Prob}_k\{\bar{T}_p \mid v_k(p)\} + \mathrm{Prob}_k\{\mathrm{T}_p \mid v_k(p)\} = 1$. To overcome this difficulty, we emphasize the following assumptions: (i) the interval is related to uncertainty on the conditional probability that the proposition is true given the evidence $v_k(p)$; (ii) the distance, $\mathrm{Pls}_k\{\mathrm{T}_p \mid v_k(p)\} - \mathrm{Bel}_k\{\mathrm{T}_p \mid v_k(p)\}$ is a measure of the uncertainty of the conditional probability; and (iii) a function of the certainty factor (CF) defined in equation 13 may be used as a measure of the uncertainty.

The certainty factor CF [assumption (iii) above] is a measure of uncertainty with respect to the prior probability and *not* of uncertainty with respect to the conditional probability that we are discussing here. In the absence of an uncertainty measure on the conditional probability, we use assumption (iii) instead. This undesirable assumption can produce unreliable and uninterpretable results. Whenever possible, we should try to estimate the uncertainty on the conditional probability, following the similar argument previously discussed in the section on no known occurrences.

We have already obtained the conditional probabilities for our example in equations 30 and CF's in equations 34. Having only the evidence of the rock types with respect to the proposition that p contains a deposit, and following the arguments that we have used in previous sections, we conclude that the uncertainties associated with the conditional probabilities are very high. That is, the probability values in equations 30, except perhaps for $\mathrm{Prob}\{\mathrm{T}_p \mid p \text{ is in sedimentary rock}\} = 0$, are unstable.

We first state the belief function equal to the probability function in this example as shown in equation 30. We also state the distance of the interval equal to $(1 - \mid \mathbf{CF} \mid)$. We obtain

$$\begin{aligned}
&[\mathrm{Bel}\{\mathrm{T}_p \mid p \text{ is in intrusive}\}, \mathrm{Pls}\{\mathrm{T}_p \mid p \text{ is in intrusive}\}] \\
&\qquad = [0.000047, 0.130047], \\
&[\mathrm{Bel}\{\mathrm{T}_p \mid p \text{ is in mixed volcanics}\}, \\
&\quad \mathrm{Pls}\{\mathrm{T}_p \mid p \text{ is in mixed volcanics}\}] \\
&\qquad = [0.00141, 0.26141], \\
&[\mathrm{Bel}\{\mathrm{T}_p \mid p \text{ is in acidic volcanics}\}, \qquad (35) \\
&\quad \mathrm{Pls}\{\mathrm{T}_p \mid p \text{ is in acidic volcanics}\}] \\
&\qquad = [0.00047, 0.81047], \\
&[\mathrm{Bel}\{\mathrm{T}_p \mid p \text{ is in sedimentary}\}, \\
&\quad \mathrm{Pls}\{\mathrm{T}_p \mid p \text{ is in sedimentary}\}] \\
&\qquad = [0, 0].
\end{aligned}$$

Fuzzy Logic Membership Function. Considering the example again, we now define the fuzzy set S_1 for the geological map shown in Figure 1*a*, which consists of all the pixels likely to contain VMS deposits:

$$S_1 = \{(p, \mu_1(p)), p \in \mathbf{A}\}, \qquad (36)$$

where $\mu_1\colon \mathbf{A} \rightarrow [0, 1]$ is the membership (or possibility) function based on the geological map.

For a pixel in unit i, assuming that the possibility that a pixel belongs to S is identical to the probability that the proposition is true given that the pixel is in unit i; the membership function is the same as the probabilities (eqs. 30) discussed previously. Logic more than calculus could be our guideline in defining possibility! Here we can repeat the arguments previously made (see Zadeh, 1968).

The assumption that the possibility function is identical to the probability function must be carefully examined. An excellent comparison between these two functions, including the possibility/probability consistency principle, was made by Zadeh (1978). In fact, Zadeh (1978) has suggested that this principle provides a basis for computing the possibility function of the conditional probability, and that the possibility function can be used as the uncertainty of the conditional probability needed to construct belief and plausible functions.

Known Occurrences in the Study Area

Suppose that there are N instances in which the proposition is true within the study area. Among N, we have n known events (or observations where the proposition is true), and $N - n$ unknown events (or unobserved situations in which the proposition is true). Combining knowledge from experts and n known events, we wish to identify the areas where the $N - n$ unknown events are likely to occur.

Traditionally, n known events and m layers within the study area are studied using multivariate statistical analysis (for example, Agterberg and others, 1972; Chung, 1978, 1983), such as regression analysis, discriminant analysis, and canonical correlation analysis. The established multivariate relationships between the known events and m layers are used to delineate areas where other unknown events are likely to occur. The advantages of the multivariate approach are as follows:

a1. The relationships between the known events and the all-combined layers (by overlaying and merging the information) are studied at the same time. Thus, the relationships among the layers are also taken into account when the relationships between the known events and the all-combined layers are established using multivariate analysis.

a2. The procedures are well established, and many existing computer software programs, such as SPSS (SPSS Inc., 1990), STATGRAPHICS (STSC Inc., 1991), SYSTAT (SYSTAT Inc., 1990), and SIMSAG (Chung, 1983), are available.

a3. The multivariate approach requires relatively fewer assumptions than those for the approach that we are discussing here.

However, the disadvantages of the multivariate approach are the following:

d1. We assume that $n = N$; that is, to establish the multivariate relationships, we assume that $N - n$ unknown events will not occur.

d2. The number of known events n must be reasonably large.

If n is not large, multivariate relationships established may not reflect the true relationships. This is particularly so if the number of layers is also large (for example, we have 10 or more layers).

Nevertheless, the first step must be the application of the multivariate analysis techniques to establish the statistical relationships between the known events and the combined layers. When the results are not satisfactory (that is, associations are not statistically significant), the following procedure is recommended.

For each layer, the table containing statistics related to the layer is computed. In our example, statistics corresponding to figures 1*a*, 1*b*, and 1*c* are shown in tables 1*a*, 1*b*, and 1*c*, respectively. In addition, based on the seven known events shown in figure 1*d*, the values in the last column of tables 2*a*, 2*b*, and 2*c* represent the estimators of the conditional probabilities that the proposition (assuming a pixel size of 100 m²) is true given that the pixel is contained in the unit specified in the corresponding first column, using (A1) in the Appendix.

Using the statistics (for example, the conditional probabilities in the last column of table 2), we construct the favorability function for the geological map of the study area with the following order:

$$\beta > f_1(3) \geq f_1(2) \geq f_1(1) \geq f_1(4) \geq \alpha, \tag{37}$$

which contradicts the order, based on the experts' conceptual VMS deposit model, of equation 22.

The presence of acidic volcanic rocks appears more significant than that of mixed volcanic rocks. Assuming that figure 1*a* is correct, we may interpret it in two ways: (i) the genetic model that implies equations 23 by the experts is wrong; or (ii) many more unknown deposits are hidden in mixed volcanic rocks in the area.

If we were to accept (ii), then the value .000191 corresponding to $f_1(2)$ in table 2*a* should be changed to a value larger than .000233 corresponding to $f_1(3)$ in table 2*a*. As discussed in equations 24, we also did expect that $f_1(3)$ is about 10 times larger than $f_1(1)$, the conditional probability related to $f_1(3)$ is only 1.8 times larger than that of $f_1(1)$ as shown in the last column of table 2a. We infer that these differences are due to undiscovered VMS deposits in the study area.

Using these arguments, we may modify equations 24 to

$$\begin{aligned} f_1(2) &= 10t < \beta, \\ f_1(3) &= 5t, \\ f_1(1) &= t, \\ f_1(4) &= \alpha. \end{aligned} \tag{38}$$

We will use these relationships to construct the favorability function for the geological map of the study area.

Although the values in equations 38 now satisfy equations 23, they do not satisfy the conditions in equations 24. We can now modify the values in equations 38 again or accept the values to construct the favorability function. Obviously equations 38 will provide the necessary guidelines to construct the favorability function.

Suppose that we perform a statistical analysis of the stream geochemical data from the study area. We divide the study area into the mosaic of drainage basins, as shown in figure 1*b*, and then classify the basins into five classes (high, medium, low, very low, and no data) for the potential of VMS deposits, according to the results of the statistical analysis. In figure 1*b*, we do not have a basin classified as medium.

The relationships among the conditional probabilities in the last column of table 2*b* are reasonable; we obtain the following favorability function for the geochemical map:

$$\begin{aligned} f_2(3) &= 3r < \beta, \\ f_2(2) &= r, \\ f_2(1) &= 0.5r, \\ f_2(4) &= \alpha. \end{aligned} \tag{39}$$

Similarly, we obtain the following relationships among the favorability functions for the geophysical anomaly map, which has been divided into four classes (high, me-

Table 2. Pixel counts and mineral deposit counts (shown in figure 1*d*) for the map units in figure 1.

Type	Name	No. of pixels	No. of mineral deposits in fig. 1*d*	Probability (A1)
		a. Geology		
gr1	Intrusive	7,777	1	.000129
gr2	Mixed volcanics	5,236	1	.000191
gr3	Acidic volcanics	21,455	5	.000233
gr4	Sedimentary	15,255	0	.0
		b. Geochemistry		
gc1	Low	15,304	1	.000065
gc2	No data	18,654	2	.000107
gc3	High	13,219	4	.000303
gc4	Very low	2,546	0	.0
		c. Geophysics		
gp1	High	10,733	3	.000279
gp2	Medium	9,579	2	.000208
gp3	Low	18,828	2	.000106
gp4	Very low	10,583	0	.0

dium, low, and very low) for potential VMS deposits as shown in figure 1*c*:

$$\begin{aligned} f_3(1) &= 4k < \beta, \\ f_3(2) &= 2k, \\ f_3(3) &= k, \\ f_3(4) &= \alpha. \end{aligned} \tag{40}$$

The last, but not the least important, problem is how to assign the values for t in equations 38, r in equations 39, and k in equations 40. These values should be assigned according to the importance of the map layers with respect to the proposition on the VMS deposits.

Probability Function. By accepting equations 38, which are based on the conditional probabilities computed from the seven discovered VMS deposits in the study area, as well as a genetic VMS mineral deposit model, we must now determine a value for t to construct the favorability function for the geological map. In this example, t is directly related to the number of deposits expected in intrusive rocks in the study area. Suppose that the discovered deposit is the only VMS deposit existing in intrusive rocks (that is, we assume that no more undiscovered VMS deposits exist in intrusive rocks in the study area). Then from table 1*a*, we have

$$t = 1 - (7{,}776/7{,}777) = .000128, \tag{41}$$

and all $f(i)$'s are assigned according to equations 38. Hence, in this example,

$$\begin{aligned} f_1(1) &= \text{Prob}\{\text{T}_p \mid p \text{ is in intrusive rock}\} \\ &= t = .000128, \\ f_1(2) &= \text{Prob}\{\text{T}_p \mid p \text{ is in mixed volcanic rock}\} \\ &= 10t = .001280, \\ f_1(3) &= \text{Prob}\{\text{T}_p \mid p \text{ is in acidic volcanic rock}\} \\ &= 5t = .000640, \\ f_1(4) &= \text{Prob}\{\text{T}_p \mid p \text{ is in sedimentary rock}\} = 0. \end{aligned} \tag{42}$$

We can also estimate the number of deposits from equations 42 by using equation A3:

$$\begin{aligned} &\#\text{ in intrusive rock} = 1, \\ &\#\text{ in mixed volcanic rock} = 7, \\ &\#\text{ in acidic volcanic rock} = 14, \\ &\#\text{ in sedimentary rock} = 0. \end{aligned} \tag{43}$$

Under these conditions, we expect six more deposits in mixed volcanic rock and nine more deposits in acidic volcanic rock.

To find the favorability function for the geochemical map in figure 1b, we simply assume that the probability, Prob{T_p | p is in "no data"} = r, that a pixel in a "no-data" area (class 2 in figure 1*b*) contains a VMS deposit is the same as the prior probability, Prob{T_p}.

To obtain Prob{T_p}, we use the estimated total number (22) of VMS deposits in the study area, including the seven known deposits obtained in equations 43. Hence, using (A1) again, we obtain Prob$\{T_p\}$ = .000442. Now using equations 39, we have

$$\begin{aligned}
f_2(3) &= \text{Prob}\{T_p \mid p \text{ is in "high"}\}\\
&= 3r = .001326,\\
f_2(2) &= \text{Prob}\{T_p \mid p \text{ is in "no data"}\}\\
&= r = .000442,\\
f_2(1) &= \text{Prob}\{T_p \mid p \text{ is in "low"}\}\\
&= 0.5r = .000221,\\
f_2(4) &= \text{Prob}\{T_p \mid p \text{ is in "very low"}\} = 0.
\end{aligned} \tag{44}$$

Using (A3) and assuming equations 44 are "correct," we can again obtain the number of deposits in each unit:

$$\begin{aligned}
&\#\text{ in "high" geochemical anomaly} = 18,\\
&\#\text{ in "no data"} = 9,\\
&\#\text{ in "low" geochemical anomaly} = 4,\\
&\#\text{ in "very low" geochemical anomaly} = 0.
\end{aligned} \tag{45}$$

Now the total number of deposits in the study area A is 31 from equations 44, whereas the total number of deposits is 22 from equations 42 based on the geological map. If we assume that two layers, geology and geochemistry, provide equally weighted evidences, then the two numbers, 22 and 31, are contradictory. To avoid this contradiction, we must modify either equations 38 or equations 39. Suppose that equations 38 are "correct," and we wish to modify equations 39 as follows:

$$\begin{aligned}
f_2'(3) &= 2r < \beta,\\
f_2'(2) &= r,\\
f_2'(1) &= 0.1r,\\
f_2'(4) &= \alpha.
\end{aligned} \tag{46}$$

Then by modifying equations 44 according to equations 46, we obtain

$$\begin{aligned}
f_2(3) &= \text{Prob}\{T_p \mid p \text{ is in "high"}\}\\
&= 2r = .000884,\\
f_2(2) &= \text{Prob}\{T_p \mid p \text{ is in "no data"}\}\\
&= r = .000442,\\
f_2(1) &= \text{Prob}\{T_p \mid p \text{ is in "low"}\}\\
&= 0.1r = .000044,\\
f_2(4) &= \text{Prob}\{T_p \mid p \text{ is in "very low"}\} = 0.
\end{aligned} \tag{47}$$

Using (A3) and assuming equations 47, we calculate the modified number of deposits in each unit:

$$\begin{aligned}
&\#\text{ in "high" geochemical anomaly} = 12,\\
&\#\text{ in "no data"} = 9,\\
&\#\text{ in "low" geochemical anomaly} = 1,\\
&\#\text{ in "very low" geochemical anomaly} = 0.
\end{aligned} \tag{48}$$

Now the total number of deposits in the study area A is 22, as we had in equations 43.

For the favorability function for the geophysical map shown in figure 1*c*, we assume that Prob$\{T_p \mid p$ is in "medium" anomaly$\} = 2k =$ Prob$\{T_p\}$ = .000442, and then we apply equations 40 and obtain

$$\begin{aligned}
f_3(1) &= \text{Prob}\{T_p \mid p \text{ is in "high"}\}\\
&= 4k = .000884,\\
f_3(2) &= \text{Prob}\{T_p \mid p \text{ is in "medium"}\}\\
&= 2k = .000442,\\
f_3(3) &= \text{Prob}\{T_p \mid p \text{ is in "low"}\}\\
&= k = .000221,\\
f_3(4) &= \text{Prob}\{T_p \mid p \text{ is in "very low"}\} = 0.
\end{aligned} \tag{49}$$

Using equation A3 and assuming equations 49, we calculate the number of deposits in each unit from figure 1*c*:

$$\begin{aligned}
&\#\text{ in "high" geophysical anomaly} = 10,\\
&\#\text{ in "medium" geophysical anomaly} = 5,\\
&\#\text{ in "low" geophysical anomaly} = 5,\\
&\#\text{ in "very low" geophysical anomaly} = 0.
\end{aligned} \tag{50}$$

The total number of deposits is 20, which is compatible with the previous total number 22, so we accept equations 49 as a favorability function of the geophysical layer.

Certainty Factor. Using equation 13 and Prob$\{T_p\}$ = .000442, we obtain the CF's for the geological map by equations 42, the CF's for the geochemical map by equations 47, and the CF's for the geophysical map by equations 49. For the geological map,

$$\begin{aligned} &\mathbf{CF}_1(1) = -0.710,\\ &\mathbf{CF}_1(2) = 0.655,\\ &\mathbf{CF}_1(3) = 0.310,\\ &\mathbf{CF}_1(4) = -1. \end{aligned} \tag{51}$$

For the geochemical map,

$$\begin{aligned} &\mathbf{CF}_2(1) = 0.50,\\ &\mathbf{CF}_2(2) = 0,\\ &\mathbf{CF}_2(3) = -0.900,\\ &\mathbf{CF}_2(4) = -1. \end{aligned} \tag{52}$$

For the geophysical map,

$$\begin{aligned} &\mathbf{CF}_3(1) = 0.50,\\ &\mathbf{CF}_3(2) = 0,\\ &\mathbf{CF}_3(3) = -0.50,\\ &\mathbf{CF}_3(4) = -1. \end{aligned} \tag{53}$$

Belief Function and Plausible Function. Instead of specifying one probability measure, $\text{Prob}_k\{\mathbf{T}_p \mid v_k(p)\}$, as we did previously, we construct an interval $[\text{Bel}_k\{\mathbf{T}_p \mid v_k(p)\}, \text{Pls}_k\{\mathbf{T}_p \mid v_k(p)\}]$. We again emphasize these three points: (i) the interval is related to uncertainty on the conditional probability that the proposition is true given the evidence $v_k(p)$; (ii) the distance $\text{Pls}_k\{\mathbf{T}_p \mid v_k(p)\} - \text{Bel}_k\{\mathbf{T}_p \mid v_k(p)\}$ is a measure of the uncertainty of the conditional probability; and (iii) the certainty factor (CF) is defined in equation 13 and we may use a function of the CF as a measure of the uncertainty in the previous section. The CF is a measurement of uncertainty with respect to the prior probability and *not* of uncertainty with respect to the conditional probability. But in the absence of an uncertainty measurement on the conditional probability, we use (iii) instead.

Returning to the example for the geological layer, we have already obtained the conditional probabilities in equations 42 and CF's in equations 51. We first define the belief function as equal to the probability function in this example, as shown in equations 42. We also define the distance of the interval as equal to $(1 - \mid \mathbf{CF} \mid)$. We obtain

$$\begin{aligned} &[\text{Bel}\{\mathbf{T}_p \mid p \text{ is in intrusive}\}, \text{Pls}\{\mathbf{T}_p \mid p \text{ is in intrusive}\}]\\ &\quad = [0.000128, 0.290128],\\ &[\text{Bel}\{\mathbf{T}_p \mid p \text{ is in mixed volcanics}\},\\ &\quad \text{Pls}\{\mathbf{T}_p \mid p \text{ is in mixed volcanics}\}]\\ &\qquad = [0.00128, 0.34628],\\ &[\text{Bel}\{\mathbf{T}_p \mid p \text{ is in acidic volcanics}\},\\ &\quad \text{Pls}\{\mathbf{T}_p \mid p \text{ is in acidic volcanics}\}]\\ &\qquad = [0.00064, 0.69064],\\ &[\text{Bel}\{\mathbf{T}_p \mid p \text{ is in sedimentary}\},\\ &\quad \text{Pls}\{\mathbf{T}_p \mid p \text{ is in sedimentary}\}]\\ &\qquad = [0, 0]. \end{aligned} \tag{54}$$

Fuzzy Logic Membership Function. All the discussions of fuzzy logic in the section on no known occurrences apply here. Using similar arguments, we assume that the membership functions are identical to the corresponding probability measures discussed in equations 42, 47, and 49.

Known Occurrences Outside Study Area (Control Area)

Suppose that there are no or very few known occurrences for the proposition in the study area, but we have known occurrences in other areas where geological environments are similar to the study area. Usually such areas have been well explored with respect to the proposition, and the known occurrences within the areas are well recorded. The areas are called control areas for the proposition.

Although it is highly desirable to have one control area that contains all the layers considered in the study area and the occurrences for the given proposition, in practice, we rarely have such control areas. When all the information is contained in one control area, traditionally, a multivariate statistical analysis (Agterberg and others, 1972; Chung and Agterberg, 1980; Chung, 1983), such as regression analysis, discriminant analysis, or canonical correlation analysis, is applied to establish multivariate relationships between the known events and the combination of all input layers. Then the established multivariate relationships are applied to the study area to identify the areas where unknown events are likely to occur. The three advantages of the multivariate approach discussed previously are also valid here.

In practice, we take one control area for each layer to estimate the favorability function. For the example discussed previously, we show three separate control areas in figure 2, one for each layer for the three maps shown in figure 1, to construct the favorability functions. We assume that the mineral deposits shown in figure 1*d* are unknown.

Table 3 presents the statistics for the control areas of

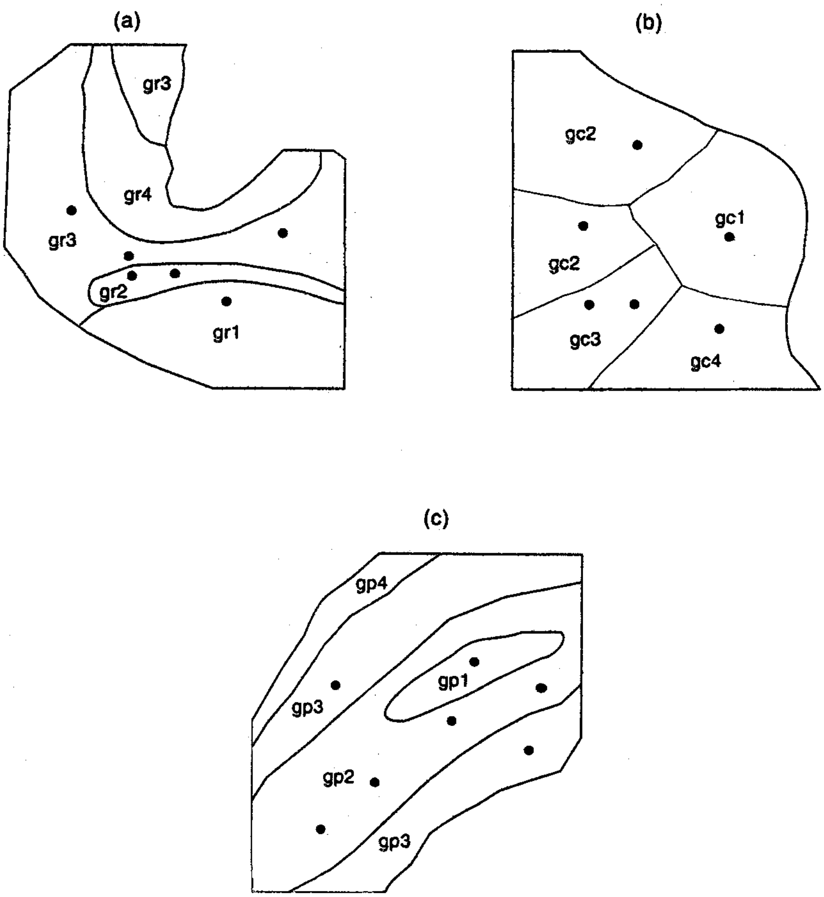

Figure 2. Plots of three different hypothetical control areas of geology, geochemistry, and geophysics from terranes of similar characteristics to those represented in figure 1, but where mineral deposits were discovered. Map unit labels correspond to the labels in figures 1*a*, 1*b*, and 1*c*. The mineral deposits are indicated by the points in these control areas. See table 3 for the statistics.

figure 2. The values in the last column of tables 3*a*, 3*b*, and 3*c* represent the estimators of the conditional probabilities that the proposition (assuming the size of a pixel is 100 m^2) is true given that the pixel is contained in the unit specified in the corresponding first column using equation A1 in the Appendix.

If we use the statistics (for example, the conditional probabilities in the last column of table 3*a*) to construct

Table 3. Pixel and mineral deposit counts for the map units in the three hypothetical control areas shown in figures 2*a*, 2*b*, and 2*c*.

Type	Name	No. of pixels	No. of mineral deposits in fig 1*d*	Probability (A1)
		a. Geology		
gr1	Intrusive	3,646	1	.000274
gr2	Mixed volcanics	932	2	.002145
gr3	Acidic volcanics	6,512	3	.000461
gr4	Sedimentary	2,896	0	.0
		b. Geochemistry		
gc1	Low	3,722	1	.000269
gc2	No data	5,706	2	.000350
gc3	High	2,284	2	.000875
gc4	Very low	2,690	1	.000372
		c. Geophysics		
gp1	High	1,116	1	.000896
gp2	Medium	6,570	4	.000609
gp3	Low	6,364	2	.000314
gp4	Very low	1,210	1	.000826

the favorability function for the geological map of the study area, then the favorability function should have the following order:

$$\beta > f_1(2) \geq f_1(3) \geq f_1(1) \geq f_1(4) \geq \alpha, \tag{55}$$

which is the same as in expression 22, based on a conceptual VMS deposit model.

From table 3*a*, we obtain

$$\begin{aligned} f_1(2) &= 7.83t < \beta, \\ f_1(3) &= 1.68t, \\ f_1(1) &= t, \\ f_1(4) &= \alpha. \end{aligned} \tag{56}$$

Except for $f_1(3) = 1.68\ t$, equations 56 are similar to equations 38. If we were to accept equations 56 as the basis for constructing a favorability function for the geology layer shown in figure 1*a*, we would have somewhat different functions from those discussed in the section on known occurrences in the study area. Using similar arguments, we may wish to modify equations 56 as follows:

$$\begin{aligned} f_1(2) &= 8t < \beta, \\ f_1(3) &= 3t, \\ f_1(1) &= t, \\ f_1(4) &= \alpha. \end{aligned} \tag{57}$$

As we discussed previously, equations 57 will provide the necessary guidelines to construct the favorability function for the geological map of the study area. The probabilities in table 3*a* are much higher than those in table 2*a* because we do not expect any unknown deposits in figure 2*a* by the definition of the control area, whereas we do expect to discover many more deposits in the study area in figure 1.

Looking at the conditional probabilities shown in the last column of table 3*b*, which correspond to the control area in figure 2*b*, we see that the relationships among the conditional probabilities are not acceptable. For example, the "very low" geochemical anomaly area has much higher probability than the "low" area. This may suggest that the geochemical anomalies considered in the control area are not significant for the mineralization. We have also assigned $f_2(4) = \alpha$ in equations 39, implying that if the areas are classified as "very low," we do not expect to discover any deposits. From the control area in figure 2*b*, we must allow the possibility that there might be deposits, therefore, we modify equations 57 and obtain the following favorability function for the geochemical map:

$$\begin{aligned} f_2(3) &= 2r < \beta, \\ f_2(2) &= r, \\ f_2(1) &= r, \\ f_2(4) &= 0.5r > \alpha. \end{aligned} \tag{58}$$

Similarly, we obtain the following relationships among the favorability functions for the geophysical anomaly map from table 3*c* for figure 2*c*:

$$\begin{aligned} f_3(1) &= 3k < \beta, \\ f_3(2) &= 2k, \\ f_3(3) &= k, \\ f_3(4) &= \alpha. \end{aligned} \tag{59}$$

Again, it is critical to assign the values for *t* in equations 57, *r* in equations 58, and *k* in equations 59. These values should be assigned according to the importance of the map layers with respect to the proposition on the VMS deposits.

By using arguments similar to those used in the section on known occurrences in the study area, as well as the guidelines in equations 57, 58, and 59, we can construct the favorability functions for probability measures, certainty factors, belief and plausible functions, and membership functions.

Discussion

1. In this article we provide a fundamental concept of establishing a favorability function for each layer as a first step to integrate several layers of spatial geoscience data. The proposed approach establishes a base for developing integration rules in a geographical information system (GIS), based on logical, statistical, probabilistic, evidence theoretical, and fuzzy logic operations. This approach is particularly useful for predicting potential areas for mineral exploration and geological hazards. The prediction of point-like objects, such as mineral deposits, as a function of their multilayered context, is just one example of a much greater variety of "geopotential" maps that can be generated (Luttig, 1987). This representation can be used for hazard assessment of soil erosion, land use,

and landslide risk, as well as environmental impact assessment.

2. If the different layers in a multiple data set represent expert knowledge in the best available terms (that is, enhanced images, extracted knowledge or features, degrees of belief, spatial characteristics of topological relationships, direct or indirect symptoms of hazardous or beneficial events), then it is feasible to consider the geometric probabilities associated with the different map patterns.
3. The artificially constructed example maps may seem simple at first. However, the problem of representation is a complex one, particularly in view of real-life situations with many layers and with some ill-defined relationships between layers, within layers, and between layers and the known occurrences.
4. As mentioned in the discussion of the estimation procedures of the favorability functions, there is no unique method possible for constructing favorability functions. A particular difficulty is how to combine the expert opinions and the known data. Knowledge acquisition techniques (see Butler and Carter, 1986; and McGraw and Harbison-Briggs, 1989) have an important role in constructing favorability functions.
5. In the fuzzy logic approach, the problem of estimating μ_k from the known data set of m pairs, $(p_1, \mu_S(p_1))$, $(p_2, \mu_S(p_2))$, . . . , $(p_m, \mu_S(p_m))$, is called abstraction (Bellman and others, 1966). Although we have not used this method in this article, we should have followed this abstraction to estimate the membership function when we have the control areas.
6. We had many difficulties estimating the belief and the plausible functions in the Dempster-Shafer approach. We may be able to overcome these difficulties by following the possibility/probability consistency principle (Zadeh, 1978), which provides a basis for the computation of the possibility function of the conditional probability; the possibility function can be used as the uncertainty of the conditional probability. This uncertainty is of course needed to construct the belief and the plausible functions.
7. The combination (or integration) rules are determined depending on the interpretation of the favorability functions (Chung and Moon, 1991; Moon and others, 1991). For example, if the favorability functions are constructed as the CF's (certainty factors), then there exists a rule to combine several types of evidence for the proposition.
8. Real applications are bound to include a mixture of geological and conceptual modeling, and expert knowledge and statistical analysis of the known occurrences.

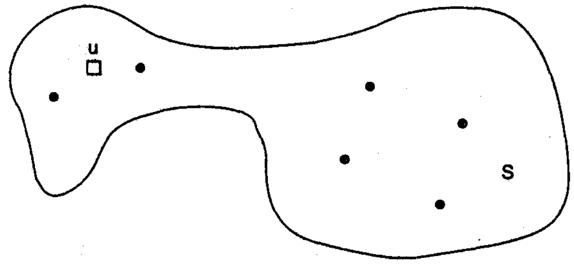

Figure 3.

Appendix

In figure 3, S denotes the whole area and u denotes a subarea in S. Let σ be the size of S, and let υ be the size of u. Suppose that we have n random points in S. Let p be the probability that u contains at least one of the n points. Then it can be shown that

$$p = \text{Prob}\{\text{at least one of the } n \text{ points is in } u\} \\ = 1 - (1 - \alpha)^n, \qquad \text{(A1)}$$

where $\alpha = \text{minimum}(\upsilon/\sigma, 1)$.

However, if $\upsilon \ll \sigma$ (for example, $50n\upsilon < \sigma$), then p can be approximated by the expected number of the n points in u:

$$p \approx n\upsilon/\sigma = E(\# \text{ of the } n \text{ points in } u). \qquad \text{(A2)}$$

The approximation (A2) is shown here in addition to the true probability in equation A1 because (i) relation A2 is mistakenly but commonly used as the true probability, and (ii) under the restricted condition ($50n\upsilon < \sigma$), relation A2 approximates the probability well.

Inversely, for a given probability φ that a subarea u of size υ in S contains one of the λ points in S needed to achieve the probability φ,

$$\lambda = \text{integer part of}\left\{0.5 + \frac{\log(1 - \varphi)}{\log(1 - \alpha)}\right\}, \qquad \text{(A3)}$$

where $\alpha = \text{minimum } (\upsilon/\sigma, 1)$.

References

Agterberg, F.P., Chung, C.F., Fabbri, A.G., Kelly, A.M., and Springer, J.S., 1972, Geomathematical evaluation of copper and zinc potential of the Abitibi area, Ontario and Quebec: Geological Society of Canada Paper 71-41, 55 p.

Aronoff, S., 1989, Geographic information systems—A management perspective: Ottawa, Canada, WDL Publications, 294 p.

Bellman, R.E., Kalaba, R., and Zadeh, L.A., 1966, Abstraction and pattern classification: Journal of Mathematical Analysis and Applications, v. 13, p. 1–7.

Bellman, R.E., and Zadeh, L.A., 1970, Decision-making in a fuzzy environment: Management Science, v. 17, no. 4, p. 141–164.

Brodaric, B., 1992, Map compilation with CAD for geological field mapping, *in* Proceedings of Computer and Mineral Exploration Symposium on "Mapping to Mining": Toronto, March 1992, p. 16–25.

Butler, K., and Carter, J., 1986, The use of psychometric tools for knowledge acquisition: A case study, *in* Gale, W., ed., Artificial intelligence and statistics: Reading, Mass., Addison Wesley, p. 295–319.

Chung, C.F., 1978, Computer program for the logistic model to estimate the probability of occurrence of discrete events: Geological Society of Canada Paper 78–11, 23 p.

———, 1983, SIMSAG—Integrated computer system for use in evaluation of mineral and energy resources: Mathematical Geology, v. 15, no. 1, p. 47–58.

Chung, C.F., and Agterberg, F.P., 1980, Regression models for estimating mineral resources from geological map data: Mathematical Geology, v. 12, no. 5, p. 473–488.

Chung, C.F., and Moon, W.M., 1991, Combination rules of spatial geoscience data for mineral exploration: Geoinformatics, v. 2, p. 159–169.

Green, A.A., and Craig, M., 1984, Integrated analysis of image data for mineral exploration, *in* Proceedings of the International Symposium on Remote Sensing of the Environment, 3rd Thematic Conference on Remote Sensing and Exploration Geology, Colorado Springs, April 16–19, 1984, p. 131–137.

Heckerman, D., 1986, Probabilistic interpretations for MYCIN's certainty factors, *in* Kanal, L.N., and Lemmer, J.F., eds., Uncertainty in artificial intelligence: New York, Elsevier, p. 167–196.

Katz, S., 1991, Emulating the PROSPECTOR expert system with a raster GIS: Computers and Geosciences, v. 17, no. 7, p. 1033–1050.

Luttig, G.G., 1987, Conclusions: Geology versus mineral, groundwater and soil resources' management—Approach to the public—Education and training questions—Types and acceptance of geopotential maps, *in* Arndt, P., and Luttig, G.W., eds., Mineral resources' extraction, environmental protection and land-use planning in the industrial and developing countries: Stuttgart, E. Schweizerbart'sche Verlagsbuchhandlung (Nagele u. Obermiller), p. 319–331.

McGraw, K.L., and Harbison-Briggs, K., 1989, Knowledge acquisition—Principles and guidelines: Englewood Cliffs, N.J., Prentice-Hall, 376 p.

McMaster, R.B., 1991, Conceptual frameworks for geographic knowledge, *in* Butterfield, B.P., and McMaster, R.B., eds., Map generalization—Making rules for knowledge representation: Harlow, Essex, England, Longman Scientific & Technical, p. 21–39.

Moon, W.M., Chung, C.F., and An, P., 1991, Representation and integration of geological, geophysical and remote sensing data: Geoinformatics, v. 2, p. 177–182.

Shafer, G., 1976, A mathematical theory of evidence: Princeton, N.J., Princeton University Press, 297 p.

Shortliffe, E.H., and Buchanan, B.G., 1975, A model of inexact reasoning in medicine: Mathematical Biosciences, v. 23, p. 351–379.

Varnes, D.J., 1974, The logic of geological maps, with reference to their interpretation and use for engineering purposes: U.S. Geological Survey Professional Paper 837, 48 p.

Wally, P., 1987, Belief function representations of statistical evidence: Annals of Statistics, v. 15, p. 1439–1465.

van Westen, C.J., 1992, Medium scale landslide hazard analysis using a PC-based GIS. A case study from Chinchina, Colombia, *in* van Westen, C.J., van Duren, I., Kruse, H., and Terlien, eds., UNESCO-ITC project on mountain hazard mapping in the Andean environment, using geographical information systems. Part I—Theoretical introduction, training package: Enschede, The Netherlands, ITC, p. 100–115.

Zadeh, L.A., 1965, Fuzzy sets: IEEE Information and Control, v. 8, p. 338–353.

———1968, Probability measures of Fuzzy events: Journal of Mathematical Analysis and Applications, v. 10, p. 421–427.

———1978, Fuzzy sets as a basis for a theory of possibility: Fuzzy Sets and Systems, v. 1, p. 3–28.

Received September 7, 1992; revised December 1, 1992; accepted January 21, 1993.

Early reviews by D.A. Singer of the U.S. Geological Survey and Kiyoji Shiono of Osaka City University helped us not only to express our ideas better but also to correct several mistakes. Geological Survey of Canada contribution number 32692.