

Knowledge Discovery from Area-Class Resource Maps: Capturing Prototype Effects*

Feng Qi, A-Xing Zhu, Tao Pei,
Chengzhi Qin, and James E. Burt

ABSTRACT: This paper presents a knowledge discovery approach to extracting knowledge from area-class resource maps. Prototype theory forms the basis of the approach which consists of two major components: (1) a scheme for organizing knowledge used in categorizing geographic entities which allows for the modeling of indeterminate boundaries and non-uniform memberships within categories; and (2) a data mining method using the Expectation Maximization (EM) algorithm for extracting such knowledge from area-class maps. A case study on knowledge discovery from a soil map demonstrates the details of the approach. The study shows that knowledge for classifying geographic entities with indeterminate boundaries is embedded in area-class maps and can be extracted through data mining; and that continuous spatial variation of geographic entities can be better modeled if the knowledge discovery process retains knowledge of within-class variations as well as transitions between classes.

Introduction

The development of geographic knowledge discovery arose out of the need to intelligently and automatically transform geographic *data* into *information* and synthesize geographic *knowledge* (Yuan et al. 2001). Research in this area has received continuous attention during the past decade. Moreover, studies have extended the scope of knowledge discovery from remote sensing data sources to traditional maps (Malerba et al. 2002; Qi and Zhu 2003).

Feng Qi, Department of Geology and Meteorology, Kean University, 1000 Morris Ave., Union, New Jersey 07083. Phone: 908 737 3668; Fax: 908 737 3699. E-mail: <fqqi@kean.edu>. **A-Xing Zhu**, State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Building 917, Datun Road, An Wai, Beijing 100101, China and Department of Geography, University of Wisconsin-Madison, 550 N. Park St., Madison, Wisconsin 53706. Phone: 608 262 0272; Fax: 608 265 3991. E-mail: <azhu@wisc.edu>. **Tao Pei**, State Key Laboratory of Resources and Environmental Information System Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Building 917, Datun Road, An Wai, Beijing 100101, China. E-mail: <peit@1reis.ac.cn>. **Chengzhi Qin**, State Key Laboratory of Resources and Environmental Information System Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Building 917, Datun Road, An Wai, Beijing 100101, China. E-mail: <qincz@1reis.ac.cn>. **James E Burt**, Department of Geography, University of Wisconsin-Madison, 550 N. Park St., Madison, Wisconsin 53706. Phone: 608 263 4460; Fax: 608 265 3991. E-mail: <jeburt@wisc.edu>.

The map is a powerful medium for presenting spatial information and geographical relationships. Much of our understanding of the relationships among spatial phenomena is embedded in maps. In natural resource management practice, for example, decades of survey practice have generated and amassed large volumes of inventory maps. In particular, maps of natural resources such as soils, wildlife habitats, and potential environmental hazards are usually created through knowledge-driven environmental modeling (Skidmore 2002). That is, the distributions of these resources are inferred from other easily observable environmental conditions using expert knowledge of the relationships between the resources and these relevant environmental conditions (Zhu et al. 2001). For example, wildlife habitat maps can be modeled using species-environmental relationships where the environmental factors include climate, vegetation, landscape characteristics, and human factors. Wildfire assessment maps can be modeled using three main environmental factors: fuel (biomass type, moisture levels, etc.), weather, and topography (Skidmore 2002). Based on these models, the natural resources are classified and represented as area-class maps (Mark and Csillag 1989). These maps thus contain rich *knowledge* on such relationships.

Previous studies (Moran and Bui 2002; Qi and Zhu 2003; Qi 2004) have demonstrated the use of inductive learning methods to extract these relationships from area-class resource maps. The extracted knowledge is valuable in at least two ways. First, it facilitates traditional, resource inventory map updates. Inventory maps need to

be updated when relevant environmental conditions change or better reference maps exist. The resource–environment relationships used to create the original map often exist as domain experts' tacit, undocumented knowledge. When the domain expert is not available during the update (which is often the case with soil mapping where the update cycle is usually longer than the career span of a soil expert), a new local model needs to be developed from scratch. This would involve tremendous amounts of fieldwork. Therefore, if the knowledge implicitly embedded in the original map could be retrieved and represented in a quantitative format, it would facilitate the update process. Second, if the extracted knowledge could model continuous spatial variations of the natural resource, it might lead to maps that are more accurate or to improved representations of the spatial distribution of such resources through automated knowledge-based inference.

The area–class presentation of the resource maps, however, limited previous studies (Moran and Bui 2002; Qi and Zhu 2003; Qi 2004) to extract knowledge that can only model natural resources in a traditional way, such as generalized and discrete classes (categories). That is, a particular location is associated with one and only one class with full membership. As a result, spatial entities are delineated as polygons with crisp boundaries and no overlaps, within which areas share a membership of unity to each of the prescribed categories. In reality, though, most geographic entities have indeterminate boundaries in both attribute space and geographic space (Burrough 1996) and exist more or less as continuums. Fitting such continuums into discrete categories with uniform memberships limits the portrayal of geographic entities in terms of both spatial details and attribute accuracy (Burrough 1996), thus failing to meet the needs of many environmental management applications.

In the area of soil mapping, scientists have begun developing new approaches to map continuous spatial variations of soil using fuzzy logic (Zhu et al. 1996; De Gruijter et al. 1997; Zhu et al. 2001; Shi et al. 2004; Qi et al. 2006). De Gruijter and colleagues mapped soil properties using the fuzzy k-means clustering method with large amounts of field points (De Gruijter et al. 1997). Other studies (Zhu et al. 1996; Zhu et al. 2001; Shi et al. 2004; Qi et al. 2006) followed the traditional knowledge-driven soil modeling approach, but they adopted a number of fuzzy inference methods to infer soil descriptions based on soil–environment relationships acquired directly from local soil experts.

For example, Zhu and colleagues (1996, p. 2001) worked with local soils experts to acquire knowledge on membership gradations within soil classes through interactive graphic interfaces. In a later study, Shi and colleagues (2004) used case-based reasoning where soil experts provided the typical cases of soil classes. Qi and colleagues (2006) later addressed some of the limitations of the previous methods and designed a new knowledge acquisition methods based on prototype theory.

The common ground of these methods is that they all acquire knowledge directly from local experts, either through on-site interview techniques or through expert-controlled computer input with graphic interfaces. Knowledge acquisition directly from domain experts, however, has always been considered as a bottleneck for the development of knowledge-based approaches and systems (Molokova 1993). In an area where no experienced soil expert is available, this type of knowledge acquisition would be difficult. An alternative way to overcoming this difficulty is to acquire knowledge that has already been expressed by experts but exists in other carrier forms. This paper expands on previous work in knowledge discovery from resource maps (Moran and Bui 2002; Qi and Zhu 2003; Qi 2004) by exploring these forms and focusing on the extraction of membership gradations within classes on them. We present an approach to extracting the knowledge for classifying natural resources in a way that explicitly reflects the indeterminate nature of class boundaries and membership gradations within classes. The belief is that such knowledge is embedded in the area–class map presentation and can be extracted by taking into consideration how humans cognitively use geographic knowledge in categorizing the continuous geographic entities and creating the maps.

Our approach consists of two major components: 1) a scheme for organizing knowledge used in categorizing geographic entities, which allows for the modeling of indeterminate boundaries and non-uniform memberships within categories; and (2) a data mining method for extracting such knowledge from map data. The scheme for organizing knowledge is based on prototype theory (Rosch 1973; Lakoff 1987), a well endorsed cognitive theory of human categorization. The data-mining method is based on an empirical measure of the typicality of category members that were first introduced by psychologists Rosch and Mervis (1975) within the context of prototype theory.

Household Item	Listed Features				Family Resemblance Measure
	F ₁ (5) ^a	F ₂ (4)	F ₃ (3)	F ₄ (2)	
Chair	F ₁ (5) ^a	F ₂ (4)	F ₃ (3)	F ₄ (2)	14
Sofa	F ₁ (5)	F ₂ (4)	F ₃ (3)	F ₅ (2)	14
Cushion	F ₁ (5)	F ₂ (4)	F ₆ (1)	F ₇ (2)	12
Rug	F ₁ (5)	F ₃ (3)	F ₇ (2)	F ₈ (1)	11
Vase	F ₁ (5)	F ₅ (2)	F ₉ (1)	F ₁₀ (1)	9
Telephone	F ₂ (4)	F ₄ (2)	F ₁₁ (1)	F ₁₂ (1)	8

a. Numbers in parentheses indicate how often each feature occurs in set of instances.

Table 1. Experimental Results of the Family Resemblance analysis. [Source: Rosch and Mervis (1975).]

Prototype Theory

One of the central concerns of cognitive psychology is the process of categorization (and subsequent classification) by which humans organize and represent their knowledge of the world (Rosch 1978; Tversky and Hemenway 1984). Classical category theory views the world as being comprised of natural partitions, and the purpose of categorization is to allocate objects into the appropriate partitions (Hahn and Ramscar 2001). Under this view, all instances of a category share common features that are singly necessary and jointly sufficient for defining the category (Smith and Medin 1981). Thus, an instance that possesses all of the defining features belongs to the category with full membership, while an instance that lacks any of the defining features must be excluded from the category completely.

This classical view of categories received intense criticism in the 1970s as new theories started to emerge (Smith and Medin 1981). Of the newly developed theories of categorization, prototype theory (Rosch 1973; 1978) is the most widely endorsed. Prototype theory emphasizes the fact that category membership is not homogenous and that some members are better representatives of a category than others—observations that are noted as the “prototype effects” (Lakoff 1987). To quantify prototype effects, Rosch and Mervis (1975) experimented with natural concepts, developing a measure of “family resemblance” as the determinant of an item’s typicality. In their experiments, subjects were asked to list features of various subsets of a superordinate concept, such as *furniture*, for which the subsets varied in typicality (*chair* is typical, *lamp* atypical). Table 1 illustrates their analysis on the distribution of listed features. Each feature listed for a subset is weighted by the total number of subsets that it is listed for; then, for each subset, the weights of all of its features

are summed, yielding a family resemblance measure. In short, an item has a greater degree of family resemblance if it contains features shared by many other members of that same concept group. Rosch and Mervis’s study shows that these family resemblance measures are very highly correlated with typicality ratings of the subsets determined by other means, and it allows for the quantification of prototype effects.

In terms of knowledge representation, prototype theory assumes that categories are mentally represented by prototypes that summarize a person’s previous exemplar experiences (Minda and Smith 2001). The categorization of possible members is based on how similar they are to the prototype (Hampton 1995). The prototype of a category is a composite or average of all the real instances experienced which 1) reflects some measure of central tendency of the instances’ properties; 2) consequently, is more similar to some category members than others; and 3) is itself realizable as, but may not necessarily be, an instance (Smith and Medin 1981). Compared to the classical view of categories, the prototype is a summary representation of a category in terms of features that may be only probable to the members of the category.

Implications for Geographic Knowledge Discovery

Due to the complexity and continuity of most geographic features and phenomena, geographic classification is a complicated abstraction process. In an ideal situation, an indefinite number of categories should be used to approximate one-to-one correspondences for the modeling of spatial variations in full detail. However, we typically categorize in such a way that few concepts capture rich situations, and the categories are, as a result, the most informative (MacEachren 1995). In this case, we often discretize the continuous geographic features to a limited number of

classes (categories). This leads to an irresolvable indefinite correspondence between the concepts and the represented world. Because this discretization results in inherent gradation of memberships to categories, it generates prototype effects of the geographic categories (Lakoff 1987).

We traditionally partition our geographic environment into various kinds of categories that exhibit such prototype effects. For example, in the case of landform recognition, the natural landscape is commonly seen as hills and valleys, or, on a larger scale, summits, shoulders, back slopes, foot slopes, and drainage ways. Prototype effects are reflected in that a particular foot slope location may be perceived to be a better example of a foot slope than some other locations. Moreover, in the case of soil classification, the continuous soil body is categorized into soil classes. Prototype effects are shown when soil scientists claim that a certain pedon is more representative of one soil class than another, although both are classified in the same way.

Our traditional products of the classification practice, however, are rooted in the classical view of categories. The area–class representation of soil classes, for example, does not explicitly model the prototypical characteristics of the soil categories that exist in soil scientists' mental representations and is an over-generalization of soil variations in both the spatial and attribute domains (Zhu 1997). Previous studies on knowledge discovery from such maps overlooked the explicit prototypical characteristics and extracted knowledge that fails to capture the indeterminate boundaries of the soil classes. To overcome this deficiency, we developed a knowledge discovery approach that explicitly considers the characteristics of geographic categories delineated by prototype theory.

Previous work by geographers has examined the connections of prototype theory and geographic categorization from several perspectives. For example, aspects of map reading and human knowledge construction have been studied (Lloyd 1994; Lloyd and Carbone 1995). Building GIS data models that reflect prototype effects (Usery 1993) and modeling spatial relations between geographic entities (Mark and Egenhofer 1994) have also been topics of interest, as has comparing spatial categories in the context of GIS interoperability (Rodríguez and Egenhofer 2004; Feng and Flewelling 2003; Ahlqvist 2005). Few, however, have been reported on the application of prototype theory in geographic knowledge discovery. This research brings a start in this direction by introducing an approach to representing knowledge for classifying geographic

entities with indeterminate boundaries and extracting such knowledge from an area–class map.

For knowledge representation schemes to be capable of retaining the prototype effects of geographic classes, they should take into consideration the way that human experts organize their knowledge mentally. For individual domain experts, the development of their knowledge about the categories of their specialty is a process of constructing the categories' prototypes based on accumulated instance experiences (MacEachren 1995; Minda and Smith 2001). This gives insight to knowledge representation in terms of prototypes. Classification is then performed through the comparison of new instances to such established prototypes, and membership of an instance is determined based on its similarity to the prototype.

Development of a knowledge discovery algorithm capable of deriving prototype effects from classified examples is the next step. As implied by prototype theory, prototypes are defined by modal feature bundles; and the "family resemblance" measure developed by Rosch and Mervis (1975) is a direct estimate of the typicality of members. While Rosch and Mervis (1975) experimented with only binary features (absence or presence of an attribute), the notion of a family resemblance score can be easily adapted for multi-valued and numerical features. The value of a feature that leads to the highest partial family resemblance is the modal value among all category members. In other words, when we obtain the frequency distribution of the values of a feature for all instances of the same category, the prototype of this category resides at the point with the highest frequency, and the shape of the frequency distribution curve indicates how membership changes when an instance departs from the prototype.

Geographic Knowledge Discovery Based on Prototype Theory

Knowledge Representation

According to Markman (1999), different representation models are best for different cognitive tasks. For example, long-term memory uses semantic nets, while categorization tasks use featural models. In a featural model, a category is represented by a composite set of features (properties). Based on prototype theory, such features should summarize the real instances

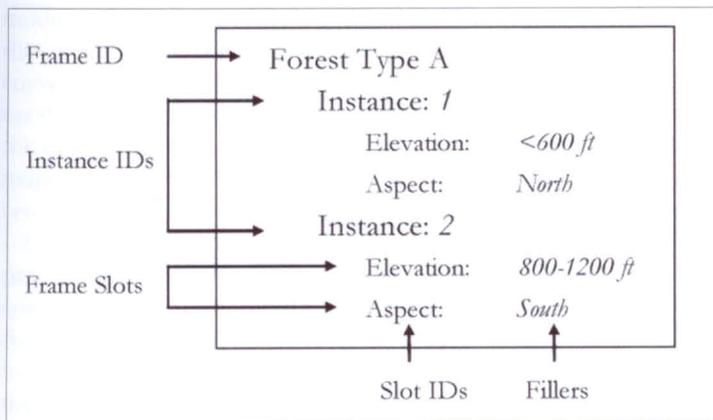


Figure 1. Frame representation of a forest class.

of the category which serve as the cognitive reference points for inference (Minda and Smith 2001).

A widely adopted knowledge representation of categories using the featural model organizes the set of features of a category into frames (Minsky 1975; Fillmore 1985). Figure 1 shows an example of using a frame to represent the distribution of a forest resource category with its environmental features. In such a representation, a frame is composed of a number of instances that define the typical featural configurations of the category. Each instance consists of a list of slots to describe the values of the features. The slots and fillers together then define the forest category in terms of its geographical location. To incorporate the prototypical properties of geographic classes, the most straightforward way is to explicitly model the prototypes and membership gradations in a frame representation. This means storing the prototypes of classes using a common frame structure, while the membership gradations are represented with optimality functions (Zhu 1999).

Figure 2 shows an example of such an extended frame representation for soil classes. The slots and fillers of the frame define the prototype(s) of a class; each slot also has a link to an optimality function that describes how membership responds when the value of the feature changes. Specifically, if the value of a feature corresponds to an optimality value of 1, the possession of such a feature will most probably lead to the highest membership in the class under review. On the other hand, an optimality of 0 means that the corresponding feature value does not favor the class at all. In other words, the optimality value directly models the partial family resemblance (normalized to [0, 1]) of an individual feature. For soil class *Valton* (Figure 1), we see that the optimality value drops

immediately from 1 to 0 when the bedrock changes from *Oneota* to any other type, meaning that an instance that is found on any bedrock type other than *Oneota* will lead to a 0 membership of the instance to *Valton*. Class membership to *Valton* also depends on the slope and curvature features: steeper slopes limit the development of *Valton*, and linear curvatures provide an optimistic condition for the soil.

In our extended frame representation, inter-frame links represent spatial relationships between geographic categories. Figure 2 illustrates the representation of four soil classes. It shows explicit spatial relationships among the four soil classes: directly downslope from *Valton* one may find *Lamoille*; further below there are *Elbaville* and *Dorvorton*. Knowledge represented in such a way can be obtained through traditional knowledge engineering. Domain experts can define prototypes either in the form of descriptive knowledge (Type I knowledge in Zhu 1999) or in the form of typical cases (Shi et al. 2004). These experts can also define optimality functions (Type II knowledge in Zhu 1999) or model them by using *a priori* heuristic curves (Qi et al. 2006). This paper presents an alternative approach to obtain knowledge on prototypes and membership gradations of geographic classes through knowledge discovery from classified examples. Such an approach is especially attractive when an experienced domain expert is not available to provide the knowledge.

Classified examples exist in two formats: field samples and existing inventory maps. For the former, data mining algorithms can be directly applied to derive class characteristics. While in most situations the cost associated with massive field sampling prohibits the collection of a large number of field examples, existing inventory maps are an acceptable source of classified examples: each location enclosed within a polygon is an example of the class indicated by the polygon label, although the memberships of the examples labeled as the same class may vary.

This paper illustrates the knowledge discovery approach with existing inventory maps, using soil maps as examples. Such maps are created by soil experts through predictive soil mapping: the distribution of soil classes is inferred from the distribution of relevant environmental variables. The basic idea of knowledge discovery from these maps is a reverse engineering of the mapping process: by associating the map with relevant environmental

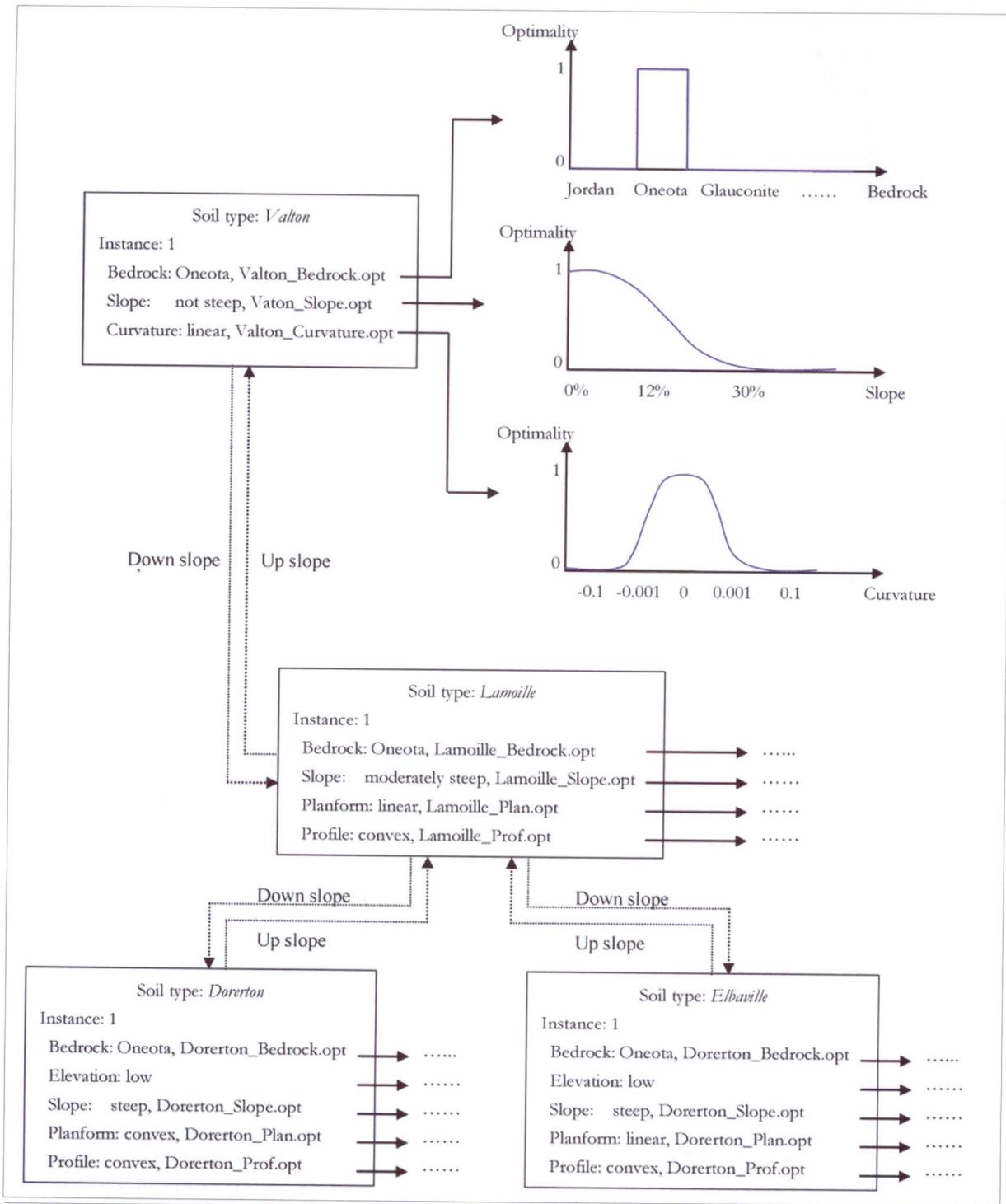


Figure 2. Frame representation of the knowledge about the distribution of four soil classes.

data layers, the knowledge for classifying the soil classes will be revealed.

Data Mining

The family resemblance measure developed by Rosch and Mervis (1975) quantifies an instance's

typicality based on a sum of features that an instance possesses, then the result is weighted by how many other category members also possess those features. An item is a typical member of a concept if it contains features shared by many other members of the same concept. Applying this to continuously valued features, a set of

modal values of the defining features will define the class prototype. Frequency distribution curves of these features then model the membership gradations. The data-mining algorithm for extracting knowledge from natural resource maps is directly based on this notion of family resemblance.

In deriving the prototypes and optimality functions of the mapped classes, histograms are first constructed to obtain the frequency distribution of the feature values of all existing examples belonging to each individual class. The modal values define the prototype(s), and the optimality functions are obtained by fitting a curve to each of the data histograms—under the assumption that the examples are representative of the population of the classes.

A smooth curve can be obtained by fitting a histogram in a number of ways. For example, a spline curve can be fitted to pass through the data points. One concern with using a spline is that its shape completely depends on the shape of the data histogram; thus, it is sensitive to the number of intervals used in plotting a histogram. A second way is to fit *a priori* heuristic curves using the data that generate the histogram. For example, in the fields of natural resource mapping and land use modeling, Gaussian curves are often employed as fuzzy membership functions (Burrough et al. 1992). A Gaussian curve can be easily fitted using methods such as maximum likelihood. However, it is often the case that the histogram may not exactly follow the curve because the data do not follow a normal distribution. Figure 3 shows a histogram plotted using the data of a soil class. The histogram of soil class *Dorerton* based on feature profile curvature, for example, is clearly unsymmetrical, and the Gaussian curve fitted with maximum likelihood fails to capture the shape well (Figure 3a).

To obtain curves that better approximate the shapes of data histograms, kernel functions can be used. Figure 3b shows a curve fitted to the underlying histogram using Gaussian kernel functions for individual data points. A kernel function, however, does not give a holistic representation of the fitted curve with a single formula, thus clutters the extracted knowledge with stored kernel points. To obtain a holistic but realistic representation of the fitted curve, each histogram can be fitted using the composite of two Gaussian curves. The combination of two curves shows a better approximation of the unsymmetrical shape of a histogram than a single curve fitted using the maximum likelihood method (Figure 3c). Furthermore, the case study in the next section will also show that

the decomposition of the curve into two separate curves enables bimodal situations to be accounted for and improves the accuracy of the knowledge discovery.

When a single Gaussian curve is expressed with the following formula:

$$D \sim N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

where D follows the Gaussian (Normal) distribution, with μ being the mean and σ the variance. Our combined model is:

$$D \sim p \cdot N(\mu_1, \sigma_1^2) + (1-p) \cdot N(\mu_2, \sigma_2^2) \quad (2)$$

where p is the proportion of the two distributions when combining the two. The two curves are fitted using the Expectation Maximization (EM) method (Dempster et al. 1977). As an iterative optimization method, the EM algorithm is known to best deal with problems of estimating parameters in a distribution function when information on the observed data is incomplete. In the case of fitting two Gaussian distributions to describe the feature values of a group of examples, the incomplete information is the Gaussian distribution to which an individual example belongs. Set $\delta_i \in [0, 1]$ for each data point, where $\delta_i = 1$ if the i th data belongs to $N(\mu_1, \sigma_1^2)$ and $\delta_i = 0$ if it belongs to the other distribution. Thus, each data point has a feature value and an unknown δ_i . The EM algorithm is then used to estimate the parameters of $p, \mu_1, \sigma_1, \mu_2, \sigma_2$ with the unknown δ_i .

With each iteration, we compute an expectation of the unknown δ_i based on the current setting of parameters with the following formula (the expectation step):

$$E(\delta_i^{(t+1)}) = \frac{p^{(t)} N(\mu_1^{(t)}, (\sigma_1^2)^{(t)})}{p^{(t)} N(\mu_1^{(t)}, (\sigma_1^2)^{(t)}) + (1-p) N(\mu_2^{(t)}, (\sigma_2^2)^{(t)})} \quad (3)$$

where t refers to the number of iterations so far. During the maximization step we estimate the parameters using maximum likelihood, given the current value of δ_i , thus:

$$\mu_1^{(t)} = \frac{\sum_{i=1}^N x_i \delta_i^{(t)}}{\sum_{i=1}^N \delta_i^{(t)}} \quad (4)$$

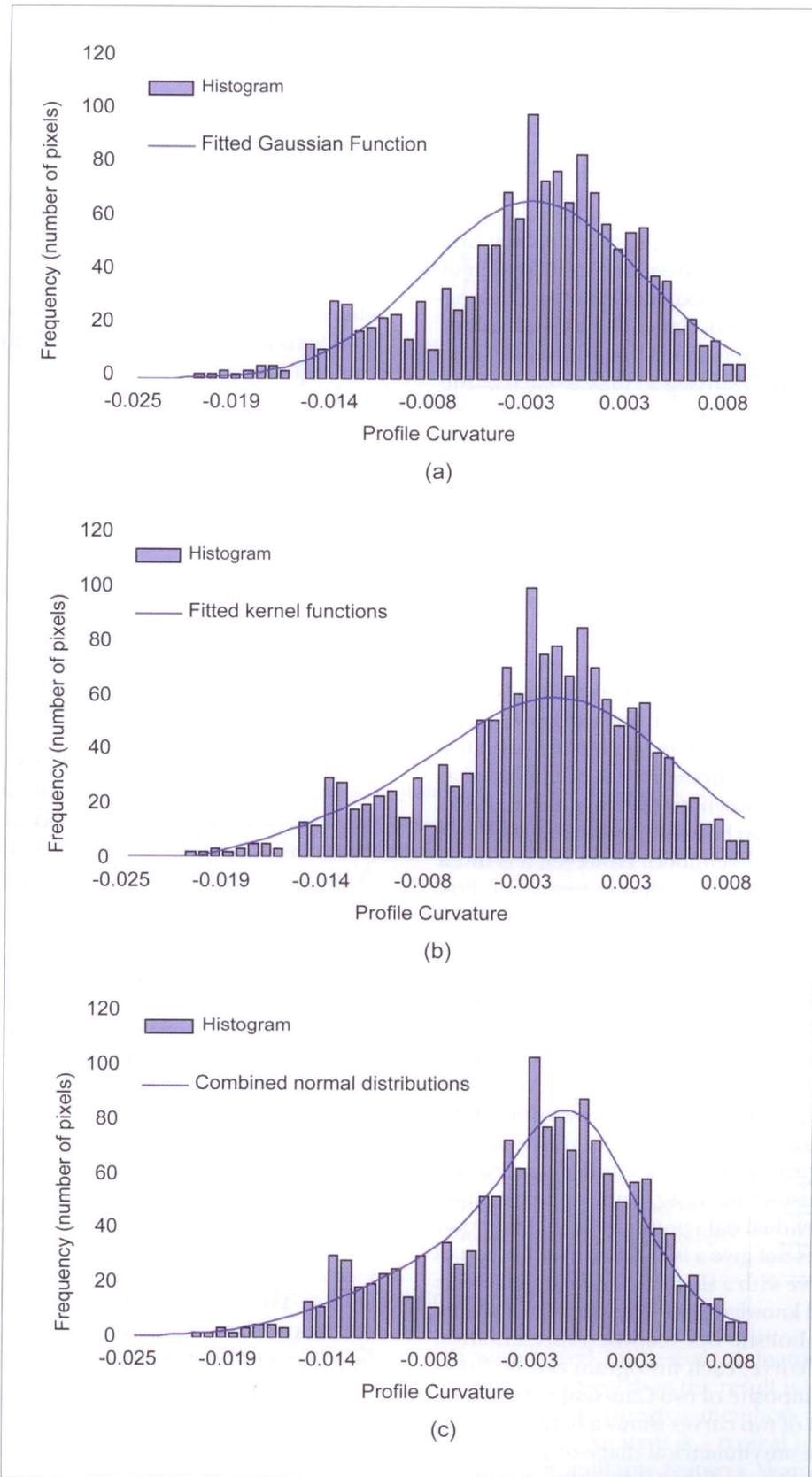


Figure 3. Fitting a histogram to an optimality function using: a) Gaussian curve; b) kernel functions; and c) combination of two Gaussian curves.

Case Study

To better illustrate the knowledge discovery method, we present here a case study on knowledge discovery from a soil map. In many countries of the world, the spatial distribution of soils is routinely collected, presented, and archived during soil surveys. The products are soil maps of different scales. Previous research indicates that valuable knowledge is embedded in archived soil maps and that such knowledge can be revealed

through knowledge discovery (Moran and Bui 2002; Qi and Zhu 2003). More specifically, the maps may yield knowledge about the environmental conditions under which each of the mapped soil class developed. Environmental conditions are typically described using such variables as bedrock, elevation, and

slope, and these variables are the very features that soil experts used to identify soil classes in soil mapping.

Our study area is the Raffelson watershed in southwestern Wisconsin. A recent soil survey produced a soil series map that shows 16 different soil series (classes) in the area. Figure 4 shows the soil series map of the watershed. We used this soil series map to extract the knowledge required to classify soils at the watershed. A uniquely constructed GIS database captured the local soil-formative environment. The database contains five commonly used soil-formative environmental variables at

$$\mu_2^{(t)} = \frac{\sum_{i=1}^N x_i (1 - \delta_i^{(t)})}{\sum_{i=1}^N (1 - \delta_i^{(t)})} \quad (5)$$

$$(\delta_1^2)^{(t)} = \frac{\sum_{i=1}^N [\delta_i^{(t)} (x_i - \frac{\sum_{i=1}^N \delta_i^{(t)} x_i}{\sum_{i=1}^N \delta_i^{(t)}})]^2}{\sum_{i=1}^N \delta_i^{(t)}} \quad (6)$$

$$(\delta_2^2)^{(t)} = \frac{\sum_{i=1}^N [(1 - \delta_i^{(t)}) (x_i - \frac{\sum_{i=1}^N (1 - \delta_i^{(t)}) x_i}{\sum_{i=1}^N (1 - \delta_i^{(t)})})]^2}{\sum_{i=1}^N (1 - \delta_i^{(t)})} \quad (7)$$

$$p^{(t)} = \frac{\sum_{i=1}^N \delta_i^{(t)}}{N} \quad (8)$$

where N is the number of data, and $(\cdot)^{(t)}$ represents the t th step in the iterative process. The calculation ceases when values of the parameters in the M -step stabilize. Once the parameters are estimated using the EM algorithm, the final curve is then represented in the form of Equation (2) and stored in the knowledge base.

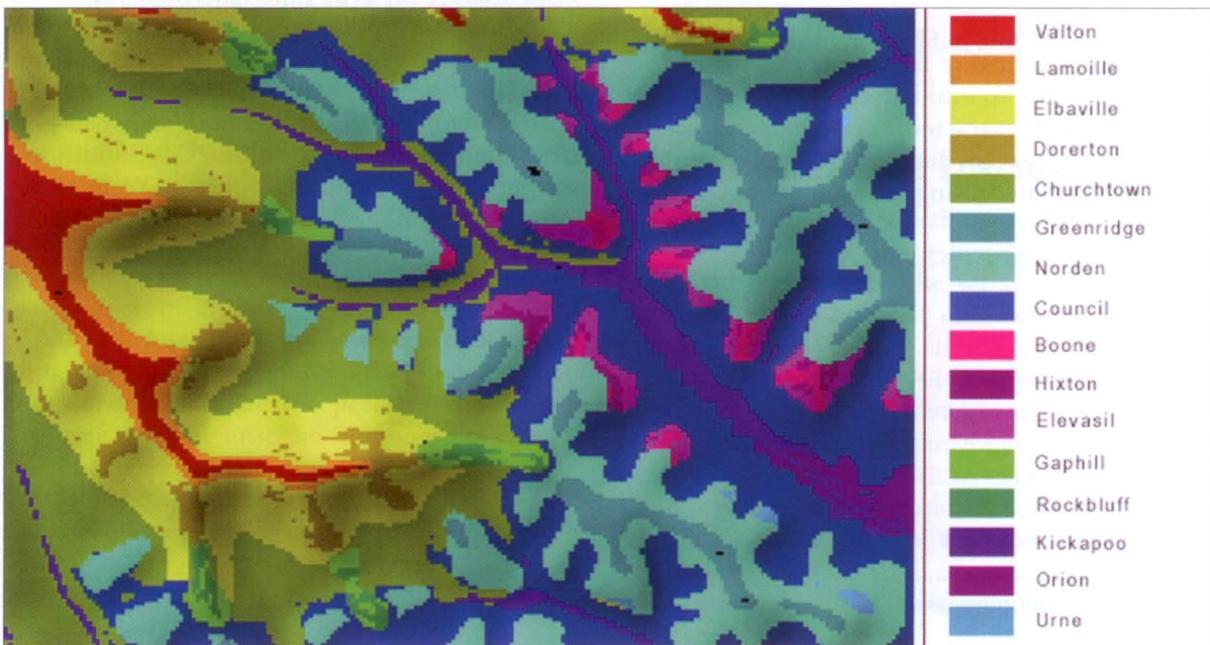


Figure 4. Soil series map of the Raffelson watershed.

the watershed scale and five spatial variables. The five primitive variables are elevation, slope gradient, plan-form curvature, profile curvature, and geology. Three of the spatial variables capture the spatial relations of the soil-formative environmental factors: distance to streams, topographic wetness index, and percentage of colluvium from competing bedrocks. The remaining two spatial variables capture topological and directional relations between soil classes—the upslope and downslope neighbors of each mapped location.

Pixels in the study area were grouped according to the different soil series, as assigned on the soil map. We then constructed histograms for the values of all pixels in a particular class for every individual feature. The frequency histogram can be either unimodal or bimodal. Data preprocessing detected bimodal cases and separated the double prototypes through visualization. Human experts examined the histograms and identified the soil-feature pairs with significant bimodal histograms. When a bimodal case was detected, the two modes were regarded as two prototypes, and the two Gaussian functions derived with the EM method were viewed independently, as were their associated optimality functions. Figure 5 shows an example of a bimodal case.

Evaluation

Extracted knowledge, especially knowledge in the form of optimality functions, is difficult to validate directly. To evaluate this knowledge discovery method, extracted knowledge was used for soil inference so that the inference results could be evaluated using field samples. Qi and

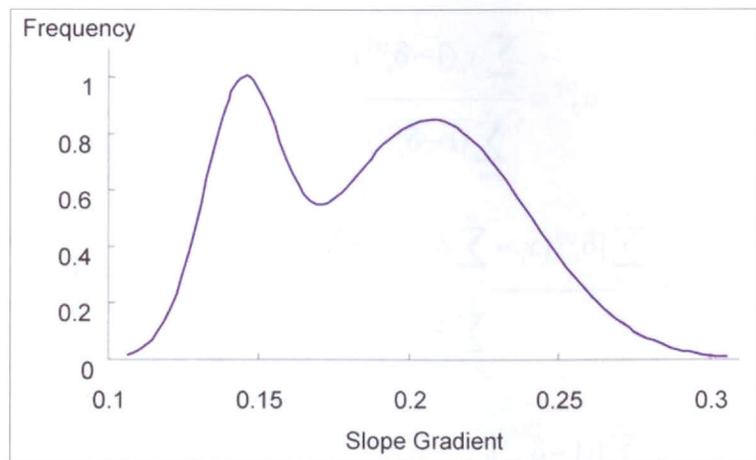


Figure 5. Histogram Curve for Soil Series Lamoille based on Slope Gradient.

colleagues (2006) used prototype-based inference to infer spatial distributions of soils and their properties from knowledge provided by soil experts. With prototype-based inference, the degrees of similarity to the class prototypes for any pixel in the mapping area determine their respective degrees of class membership to the soil series. Every pixel is then associated with a set of membership values to all 16 soil series. Soil classification in the traditional sense can be easily conducted by assigning each location the soil class that has the highest membership value among all classes. Figure 6 illustrates the classification process.

The inferred membership values were evaluated in an indirect way as no experienced soil expert was available to rate the samples using the assessment method described by Gopal et al. (2001). In our study, the continuous variation of soils was represented by continuous soil property maps derived from these membership values. Continuous soil property maps of surface soil texture in terms

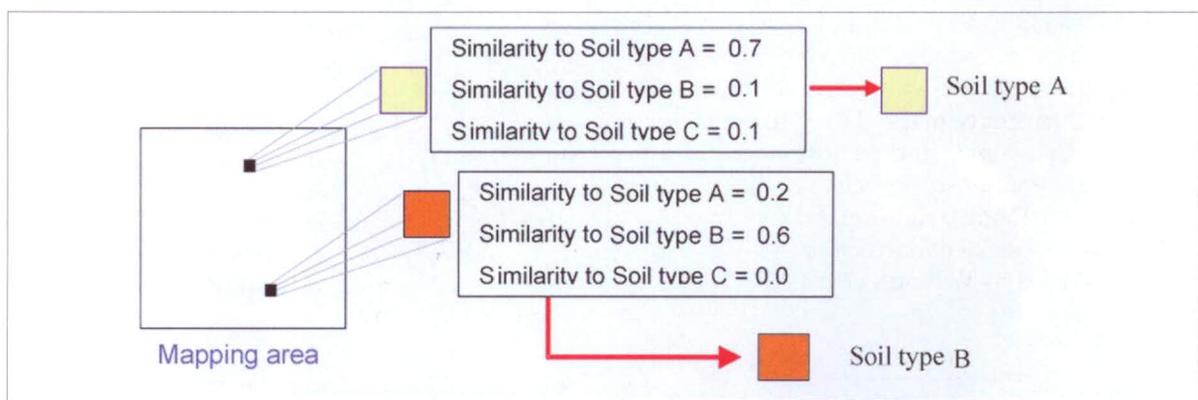


Figure 6. Classifying soil pixels based on membership values.

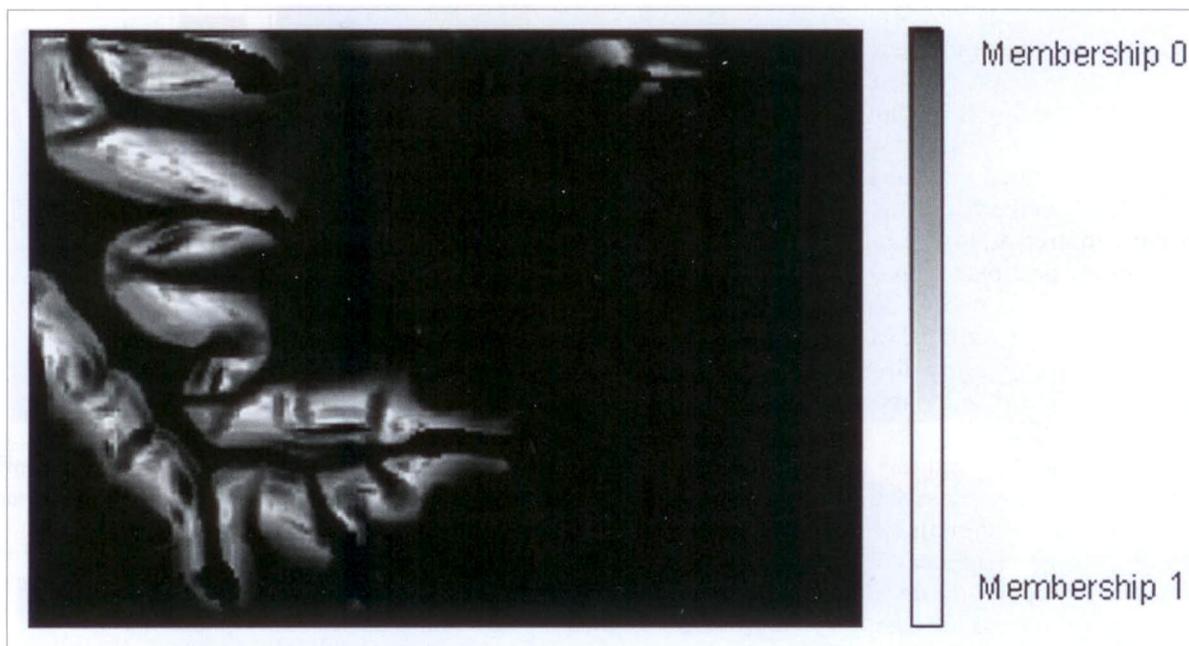


Figure 7. Membership map of soil *Elbaville*.

of percentage of sand and silt in the *A* horizon were generated in this case study following the approach used in Zhu et al. (1997) and Qi et al. (2006). For the sake of comparison, soil texture maps were also derived from the original map by assigning each pixel the typical texture values of the labeled soil series, based on official soil survey records.

To evaluate the inference results, we collected data from 99 field points in the Raffelson watershed; of these, all were classified and assigned soil series names by the USDA-NRCS local office and 49 were given a texture analysis to determine the percentages of sand and silt in the *A* horizon. The inference results were then evaluated in two aspects. First, the soil series map was compared to the original map (from which the knowledge was extracted) to assess classification accuracy. Second, the property maps derived from the original soil map and from knowledge discovery were compared to examine the estimation accuracies of both sets of property maps. Three indices were computed to evaluate the performances of all property maps against the field samples: mean absolute error (MAE), root mean squared error (RMSE), and agreement coefficient (AC). The AC index is defined by Willmott (1984) as:

$$AC = 1 - \frac{n \cdot RMSE^2}{PE} \quad (9)$$

where n is the number of observations and PE the potential error variance defined as:

$$PE = \sum_{j=1}^n (|P_i - \bar{O}| + |O_i - \bar{O}|)^2 \quad (10)$$

given that \bar{O} is the observed mean, and P_i and O_i are the estimated and observed value, respectively. AC values vary between 0 and 1, with 1 indicating perfect agreement and 0 means complete disagreement between the estimated and observed values (Willmott 1984).

Results and Discussion

A set of optimality functions was fitted for the 16 soil series in the watershed using the method detailed above; then the set was used for soil inference. The inferred membership values to a soil series *Elbaville* of all pixels in the area are illustrated in Figure 7. The lighter pixels are those with higher membership values than those of the darker ones. White zones are usually the typical positions at which to expect a particular soil series; typicality gradually fades out to black zones where memberships to the soil series are zero.

A traditional soil series map was created by assigning each location the soil class that has the highest membership value among all classes (Figure 8).

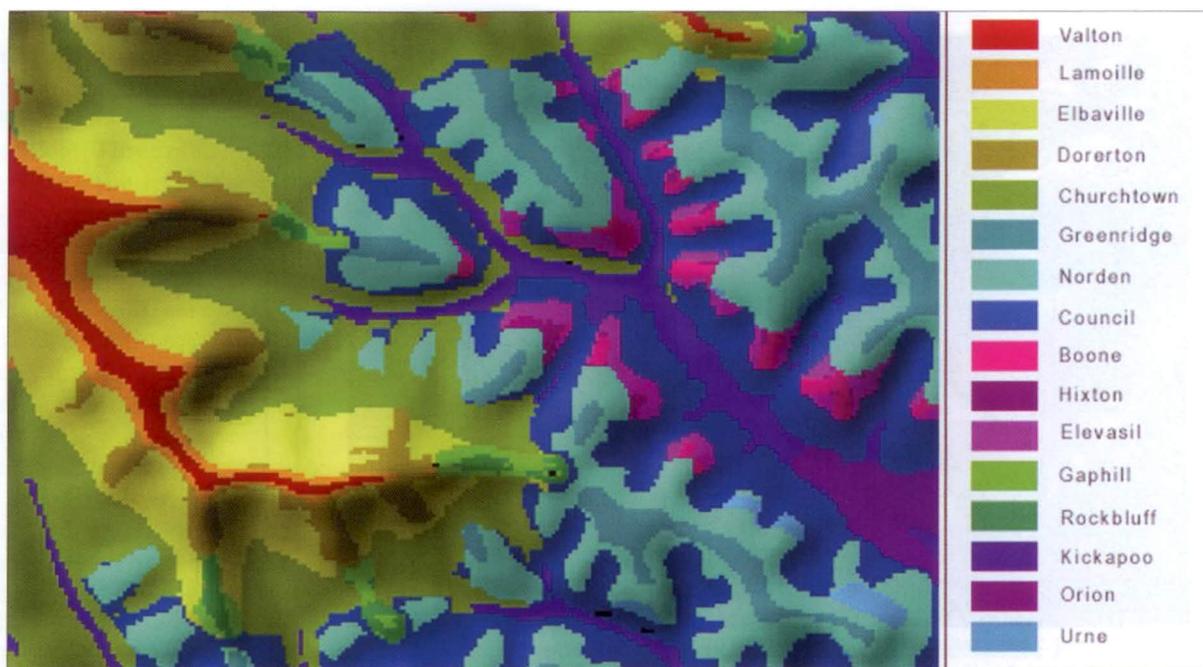


Figure 8. Classified soil series map.

Comparing Figure 8 to Figure 4, we note that the two maps are largely identical. Most soil series appear to be on similar terrain positions on both maps, with a few showing slightly different spatial extents (soil series *Orion*, for example). It was found that the original map correctly classified the soil series at 83 out of the 99 sites, while the inferred soil series map named 80 sites correctly. The inferred map mis-classified only three sites that were correctly mapped by the original map. This indicates that the extracted knowledge is able to capture the spatial distribution of soil series over the mapped area to a considerable degree of accuracy.

To validate the extracted knowledge in its ability to capture within-class variations and transitions between class prototypes, horizon texture properties were derived using the membership values. Figures 9 and 10 show the resultant soil texture maps in juxtaposition with the corresponding texture maps, based on the original soil map. The maps derived from the inference results and those based on the original soil map show comparable spatial details. One noticeable difference between the two sets of maps is the continuity of soil texture variations. The inferred texture maps tend to illustrate more continuous changes of the texture values than those based on the original map. With the inferred maps, abrupt changes of texture only occur when parent material changes.

Table 2 lists the computed statistics for field evaluation of the two sets of property maps. The

MAE and RMSE statistics for the inference results are lower than those for the original soil map. This, together with the higher AC for the inference result, implies better performance by the inferred property maps in terms of estimating continuous soil properties. This could be attributed to their ability to capture the transitions between soil prototypes, especially when the inferred maps' accuracy in predicting soil series names is even lower than that of the original map.

Conclusions

The outcome of this case study shows that the knowledge needed for classifying natural resources (such as soil) with indeterminate boundaries is embedded in existing area-class maps and can be extracted using an appropriate data mining method. The extracted knowledge, when used to infer soil series types, was able to achieve a level of accuracy comparable to that of the original map. Moreover, the extracted knowledge (particularly the prototype effect element of the knowledge) can be used to infer class memberships to different soil series. The soil property maps generated with these membership values are more continuous and more accurate than property maps derived from the original map, even though the accuracy of the inferred soil series types is slightly less than that

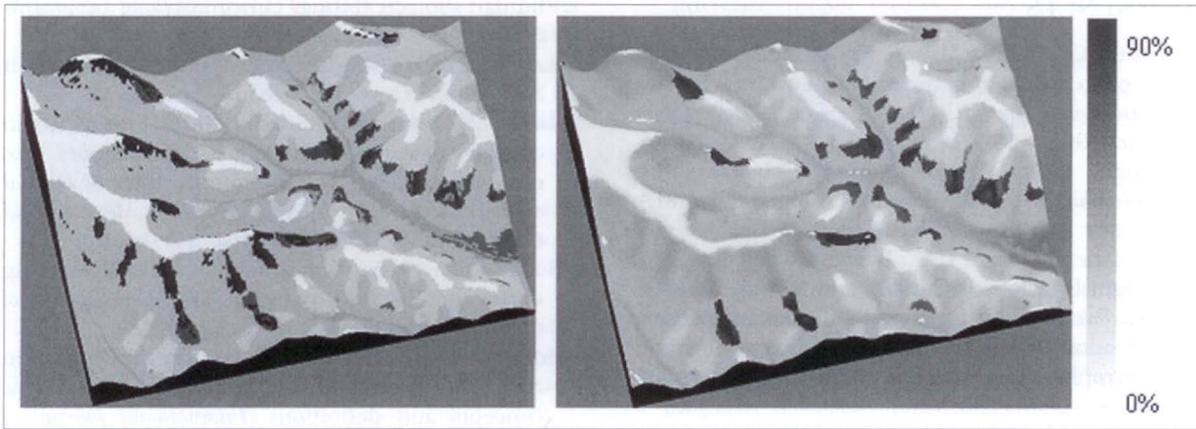


Figure 9. Left: A horizon sand percentage derived from the original raster map using field property values. Right: A horizon sand percentage inferred from data mining results.

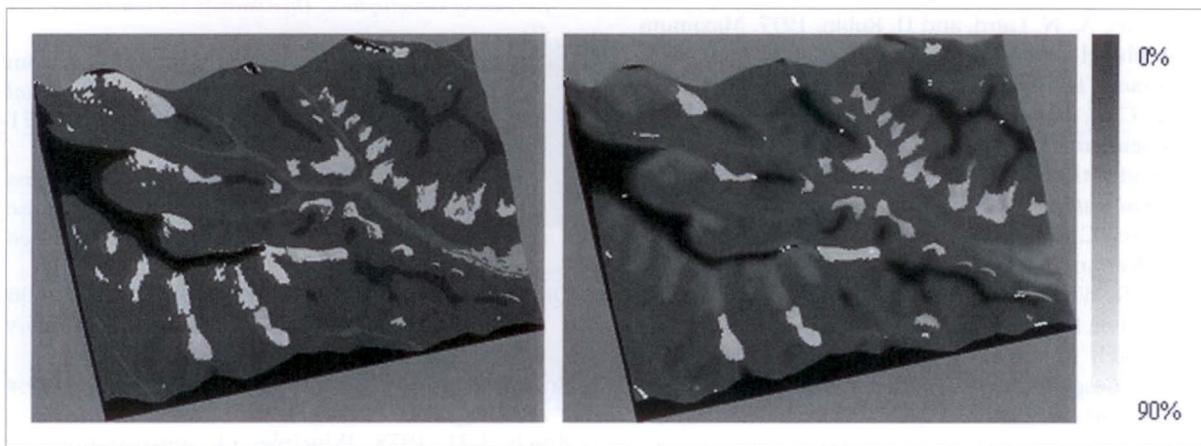


Figure 10. Left: A horizon silt percentage derived from the original raster map using field property values. Right: A horizon silt percentage inferred from data mining results.

	Percentage of sand			Percentage of silt		
	MAE	RMSE	AC	MAE	RMSE	AC
Inference Result	9.69	14.47	0.81	8.17	12.14	0.82
Original soil map	10.66	16.63	0.67	9.51	14.31	0.67

Table 2. Accuracy of the derived A horizon texture in the Raffelson watershed: the inference result vs. the original map.

of the source map. These results indicate that knowledge of prototype effects is important for classifying geographic entities with indeterminate boundaries.

ACKNOWLEDGMENTS

The research reported in this paper was a result of an international collaboration between the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Department of Geology and Meteorology, Kean University, and Department of Geography,

University of Wisconsin-Madison. The support through the following programs are greatly appreciated: “National Basic Research Program of China” (2007CB407207, Ministry of Science and Technology of China), International Partnership Project “Human Activities and Ecosystem Changes” (CXTD-Z2005-1, Chinese Academy of Sciences), the “Hundred Talents” Program of Chinese Academy of Sciences and Natural Resources Conservation Services, United States Department of Agriculture. We also like to thank the two anonymous reviewers and the editor for their constructive comments that have greatly improved the quality of this paper.

REFERENCES

- Ahlqvist, O. 2005. Using uncertain conceptual spaces to translate between land cover categories. *International Journal of Geographical Information Science* 19(7): 831-57.
- Burrough, P.A. 1996. Natural objects with indeterminate boundaries. In: P. A. Burrough and A. U. Frank (eds), *Geographic Objects with Indeterminate Boundaries*. London, U.K.: Francis and Taylor.
- Burrough, P.A., R.A. McMillan, and W. Van Deursen. 1992. Fuzzy classification methods for determining land suitability from soil profile observations and topography. *European Journal of Soil Science* 43: 193-210.
- De Gruijter, J.J., D.J.J. Walvoort, and P.F.M. van Gams. 1997. Continuous soil maps—a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. *Geoderma* 77: 169-95.
- Dempster, A., N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 39: 1-38.
- Feng, C.C., and D.M. Flewelling. 2003. Assessment of semantic similarity between land use/land cover classification systems. *Computers, Environment and Urban Systems* 28: 229-46.
- Fillmore, C.J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* 6(2): 222-254.
- Gopal, S., W. Liu, and C. Woodcock. 2001. Visualization based on fuzzy ARTMAP neural network for mining remotely sensed data. In: H. J. Miller and J. Han (eds), *Geographic Data Mining and Knowledge Discovery*. New York, New York: Taylor & Francis.
- Hahn, U., and M. Ramscar. 2001. *Similarity and Categorization*. New York, New York: Oxford University Press.
- Hampton, J.A. 1995. Testing the prototype theory of concepts. *Journal of Memory and Language* 34: 686-708.
- Lakoff, G. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, Illinois: University of Chicago Press.
- Lloyd, R. 1994. Learning spatial prototypes. *Annals of the Association of American Geographers. Association of American Geographers* 84: 418-40.
- Lloyd, R., and G. Carbone. 1995. Comparing human and neural network learning of climate categories. *The Professional Geographer* 47(3): 237-50.
- MacEachren, A.M. 1995. *How maps work*, New York, New York: Guilford Press.
- Malerba, D., A.L. Esposito, and F.A. Lisi. 2001. Machine learning for information extraction from topographic maps. In: H. J. Miller and J. Han (eds), *Geographic Data Mining and Knowledge Discovery*. New York: Taylor & Francis. pp. 291-314.
- Mark, D.M., and F. Csillag. 1989. The nature of boundaries on 'area-class' maps. *Cartographica*. 26: 65-78.
- Mark, D.M., and M.J. Egenhofer. 1994. Modeling spatial relations between lines and regions: Combining formal mathematical models and human subjects testing. *Cartography and Geographic Information Science*. 21(4): 195-212.
- Markman, A.B. 1999. *Knowledge representation*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Minda, J.P., and J.D. Smith. 2001. Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 27(3): 775-99.
- Minsky, M. 1975. A framework for representing knowledge. In: P. H. Winston (ed), *The Psychology of Computer Vision*. New York: McGraw-Hill.
- Molokova, O.S. 1993. Methodology for the acquisition of knowledge for expert systems: part 1. Basic concepts and definitions. *International Journal of Computer and Systems Sciences*. 31: 1-5.
- Moran, C.J., and E.N. Bui. 2002. Spatial data mining for enhanced soil map modeling. *International Journal of Geographical Information Science* 16(6): 533-49.
- Qi, F., and A.X. Zhu. 2003. Knowledge discovery from soil maps using inductive learning. *International Journal of Geographical Information Science* 17(8): 771-95.
- Qi, F. 2004. Knowledge discovery from area-class resource maps: Data preprocessing for noise reduction. *Transactions in Geographic Information Science* 8(3): 297-308.
- Qi, F., A.X. Zhu, M. Harrower, and J.E. Burt. 2006. Fuzzy soil mapping based on prototype category theory. *Geoderma* 136: 774-87.
- Rosch, E.H. 1973. Natural Categories. *Cognitive Psychology* 4(3): 328-50.
- Rosch, E.H. 1978. Principles of categorization. In: E. H. Rosch and B. B. Lloyd (eds), *Cognition and Categorization*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Rosch, E.H., and C. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7: 573-605.
- Shi, X., A.X. Zhu, J.E. Burt, D. Simonson, and F. Qi. 2004. A case-based reasoning approach to fuzzy soil mapping. *Soil Science Society of America Journal* 68: 885-94.
- Skidmore, A. 2002. *Environmental Modelling with GIS and Remote Sensing*. London, U.K.: Taylor & Rancis.
- Smith, E.E., and D.L. Medin. 1981. *Categories and concepts*. Cambridge, Massachusetts: Harvard University Press.
- Tversky, B., and L. Hemenway. 1984. Objects, parts, and categories. *Journal of Experimental Psychology. General* 113: 169-93.
- Usery, E.L. 1993. Category theory and the structure of features in GIS. *Cartography and Geographic Information Science* 20(1): 5-12.
- Willmott, C.J. 1984. On the evaluation of model performances in physical geography. In: G. L. Gaile and C. J. Willmott (eds), *Spatial statistics and models*. Dordrecht, Boston, and Hingham, MA: D. Reidel Publishing Company. pp. 43-460.

Yuan, M., B. Battenfield, M. Gahegan, and H. Miller. 2001. Geospatial data mining and knowledge discovery. *A USGS White Paper on Emergent Research Themes*.

Zhu, A.X., L.E. Band, B. Dutton, and T.J. Nimlos. 1996. Automated soilinference under fuzzy logic. *Ecological Modelling* 90(2): 123-45.

Zhu, A.X. 1997. Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. *Photogrammetric Engineering and Remote Sensing* 63(10): 1195-202.

Zhu, A.X. 1999. A personal construct-based knowledge acquisition process for natural resource mapping. *International Journal of Geographical Information Science* 13: 119-41.

Zhu, A.X., L.E. Band, R. Vertessy, and B. Dutton. 1997. Derivation of soil properties using a soil land

inference model (SoLIM). *Soil Science Society of America Journal* 61(2): 523-33.

Zhu, A.X., B. Hudson, J. Burt, K. Lubich, and D. Simonson. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal* 65: 1463-72.

* Note: "Feature" here and throughout this paper refers to a particular attribute or property of a category, following the use of this term in the category theory literature. It is thus different from the traditional notion of "feature" in common geographic information system (GIS) literature, where this term often refers to a geographic entity.



GIScience position at Penn State

OPEN RANK
(FULL, ASSOCIATE, OR ASSISTANT PROFESSOR) GEOGRAPHY (GISCIENCE)
THE PENNSYLVANIA STATE UNIVERSITY

Tenure track faculty position at open rank (Full, Associate, or Assistant Professor) specializing in geographic information science (GIScience). We are interested in candidates who will strengthen the Department of Geography's research and teaching program and help build strong connections to other relevant science communities. GIScience research in the Department is coordinated through the GeoVISTA Center (www.geovista.psu.edu), an interdisciplinary GIScience unit based in Geography involving five associated departmental faculty and nine faculty from departments across the University Park campus. Candidates with research expertise in any area of GIScience will be considered. Excellence in teaching, research, and service is expected, as is an established record of extramural funding. A willingness to participate in the online Masters of Geographic Information Systems (MGIS) graduate degree program is advantageous (<http://gis.e-education.psu.edu>). Applicants should submit, in digital form: 1) a letter describing how they would contribute to the Department's teaching and research program; 2) a complete curriculum vitae; 3) a maximum of five reprints; and 4) the names and addresses (including e-mail and fax) of three to five referees. Review of applications will begin November 5, 2008, but applications will be accepted until the position is filled. Applications from women and under-represented groups are encouraged. Ph.D. must be in-hand at time of application.

Send application to:
Dr. Alan M. MacEachren
Chair, Search Committee, Department of Geography
The Pennsylvania State University
302 Walker Building, University Park, PA 16802
Phone: (814) 865-7491 (main office: 814-865-3433). Fax: (814) 863-7943
E-mail: maceachren@psu.edu

Penn State University is committed to affirmative action, equal opportunity, and the diversity of its workforce