Available online at www.sciencedirect.com
ScienceDirect
Geoderma 136 (2006) 774–787
GEODERMA
www.elsevier.com/locate/geoderma

Fuzzy soil mapping based on prototype category theory

Feng Qi [a,*], A-Xing Zhu [b,c], Mark Harrower [c], James E. Burt [c]

[a] Department of Political Science and Geography, University of Texas at San Antonio, 6900 NW Loop 1604, San Antonio, TX 78249, USA
[b] 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
[c] Department of Geography, University of Wisconsin–Madison, 550 North Park Street, Madison, WI 53706, USA

Received 11 January 2005; received in revised form 26 May 2006; accepted 6 June 2006
Available online 1 August 2006

Abstract

An essential component of soil mapping is classification, a process of assigning spatial soil entities to predefined categories (classes). However, by their nature soils exist as a continuum both in the spatial and attribute domains and often cannot be fitted into discrete categories without introducing errors or at least over-simplification. One approach to mitigate this problem in digital soil mapping is the combination of fuzzy logic-based class assignment with a raster GIS representation model which allows the continuous spatial variation of soils to be expressed at much greater detail than has been achieved in conventional (analog) soil survey. However, applications of fuzzy soil mapping face two significant challenges: defining the central concept of a soil category and determining the degree of membership to the central concept. Prototype category theory is presented here as a potential solution to these difficulties. Emerging from ideas of family resemblance, centrality and membership gradience, and fuzzy boundaries (fuzzy set theory), prototype category theory stresses the fact that category membership is not homogenous and that some members are better representatives of a category than others. A prototype can be viewed as a representation of the category, that 1) reflects the central tendency of the instances' properties or patterns; 2) consequently is more similar to some category members than others; and 3) is itself realizable but is not necessarily an instance. Based on this notion, we developed a prototype-based approach to acquire and represent knowledge on soil–landscape relationships and apply the knowledge in digital soil mapping under fuzzy logic. The prototype-based approach was applied in a case study to map soils in central Wisconsin, USA. Our approach created maps that were more accurate in terms of both soil series prediction and soil texture estimation than either the traditional soil survey or a case-based reasoning approach.

Keywords: Soil map; Fuzzy logic; Cognitive theory; Prototype category theory; GIS; SoLIM

1. Introduction

In traditional soil mapping it is long-standing convention to classify soils and depict soil classes as discrete polygons on an 'area-class' map (Mark and Csillag, 1989). However, soils are known to exist more or less as a continuum in both geographic space and attribute space (Burrough, 1996; Zhu, 1997a) as one soil type blends into another. Fitting this continuous spatial character of soils into discrete soil categories with full memberships over-generalizes the inherent complexity of soil variation, which in turn degrades the accuracy of soil spatial information products. In addition to the over-generalization of

* Corresponding author. Tel.: +1 210 458 5611; fax: +1 210 458 5430.
E-mail addresses: feng.qi@utsa.edu (F. Qi), azhu@wisc.edu (A.-X. Zhu), maharrower@wisc.edu (M. Harrower), jburt@geography.wisc.edu (J.E. Burt).

0016-7061/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.geoderma.2006.06.001

soil variations, manual production in traditional soil mapping also faces other limitations, primarily low speed and high cost of production (Zhu et al., 2001).

In order to overcome these limitations, many researchers started to explore the use of knowledge-based techniques and fuzzy logic concepts to improve soil mapping process and its products (see McBratney et al., 2003 for a review). Among these endeavors, the SoLIM approach (Zhu and Band, 1994; Zhu et al., 1996, 1997; Zhu, 1997a, b; Zhu et al., 2001) is a knowledge-based soil mapping system developed for mass production of soil survey. SoLIM is an automated soil inference system that combines fuzzy logic-based class assignment with a raster GIS representation model, which allows the continuous spatial variation of soils to be expressed at much greater detail so that the class transitions and within-class variations can be represented. SoLIM uses a n-dimensional similarity vector to depict soil properties at each pixel location (i, j) (Zhu, 1997a): $S_{ij}=(S_{ij}^1, S_{ij}^2,\ldots, S_{ij}^k, \ldots,S_{ij}^n)$, where S_{ij}^k represents the similarity value or fuzzy membership of the soil at location (i, j) to the prescribed soil class k, and n is the total number of these prescribed classes. The soil similarity vectors (S) are inferred under fuzzy logic based on the same concept $S=f(E)$ as in traditional soil survey, where the essential formative environment data E can be derived using GIS techniques (Zhu et al., 1996; McSweeney et al., 1994), and the relationship between soil and relevant environmental variables (the soil-landscape model) f is obtained through knowledge acquisition. f should reflect both the central concepts of soil classes and the transitions between central concepts.

Knowledge-based fuzzy digital soil mapping approaches currently face two major challenges in defining f: the representation of the central concept of a soil category and how membership to this central concept is modeled. These difficulties are largely due to two factors. First, knowledge acquisition from human experts has long been noted to be the bottleneck for the development of knowledge-based systems (Molokova, 1993; Weibel et al., 1995), especially in the case of knowledge-based soil mapping, where knowledge of the soil–landscape model largely exists as “tacit knowledge” (Hudson, 1992). Second, it is desirable that the extracted knowledge be represented in a form that is computable as well as readily communicable with soil scientists so that other soil scientists can validate and update this knowledge base. This makes it necessary for the knowledge representation to approximate the mental representation of soil scientists’ understanding of the soil classes and the knowledge acquisition to be based on the cognitive process suited soil scientists’ understanding of soil variation.

Early implementations of SoLIM, however, did not take into consideration the cognitive aspects of knowledge formulation and representation for soil classification. In previous SoLIM applications, soil scientists were required to provide either the exact forms of fuzzy membership functions (Zhu, 1999; Zhu et al., 2001) or a large set of typical cases for known soil types in the study area (Shi et al., 2004). Soil inference was then conducted through fuzzy inference or case-based reasoning. Lacking a cognitive basis, these previous approaches place unreasonable demands on soil scientists during knowledge acquisition by requiring that they reformulate their mental knowledge into the desired form. Difficulty may also arise when the acquired knowledge needs to be interpreted or updated. This paper presents an approach that addresses this issue by employing the principles of cognitive theory in both the knowledge representation and the associated knowledge acquisition process in an effort to provide guidance to practitioners of digital soil mapping using approaches similar to SoLIM.

It has been contended that better understanding of how human beings acquire, organize, and process domain knowledge eases the difficulties in acquiring knowledge from domain experts in the development of knowledge-based systems (McCracken and Cate, 1986; Ford et al., 1991; Zhu, 1999; Özesmi and Özesmi, 2004). Especially for soil classification, it has long been suggested that it would be possible and desirable to connect the design of knowledge-based classification systems with cognitive theories (McCracken and Cate, 1986). This paper presents an approach to obtaining and representing knowledge on soil–landscape relationships based on cognitive theories on human categorization, specifically, the prototype category theory (Rosch, 1973, 1978; Smith and Medin, 1981; Lakoff, 1987; Minda and Smith, 2001). We represent and acquire the knowledge of soil–landscape model from soil scientists in terms of *prototypes* and *membership gradations* from prototypes. Variations of soil properties can then be modeled through prototype-based reasoning. The next section of this paper gives a brief introduction to prototype theory and how it differs from “classical” category theory which underpins traditional approaches to soil mapping. A novel prototype-based fuzzy soil mapping method is then presented. The advantages of this new method are illustrated and evaluated through a case study.

2. Prototype theory

Developments in cognitive psychology over the last 30 years have radically changed our understanding of how humans comprehend and describe the world through the use of categories. The act of classification—the ability to distinguish that A is different from B—is possible because of categories. Whenever we see something as a kind of thing we are using categories, whether we are conscious of it or not. Categories help to (1) structure the world for us, (2) reduce detail, (3) aid memory retention and (4) facilitate the understanding of relationships.

The classical view of categories, traceable back to Aristotle, treats categories as empty containers into which we place similar items. Similarity is derived from shared or common properties and assignment is made using a mental "check-list" of these attributes (Lakoff, 1987; Hahn and Ramscar, 2001). Classical category theory further assumes:

1. Once placed in the container, all members of a category are equally good examples of that category;
2. the boundaries between categories can be sharply delineated;
3. no overlap exists between categories;
4. categories are free of individual concerns, such as experience, neurobiology, intelligence, or embodied sensory knowledge (i.e., kinesthetic intelligence).

It was not until the 1970s that the classical view of categories received intense criticism and new views started to emerge (Smith and Medin, 1981). The new theories viewed categories as a human construct created by an active mind searching for meaning rather than seeing them as pre-existing structures in the world. According to this new view, categories are conceptual structures resulted from human's perception and interaction with the environment (Hahn and Ramscar, 2001). Categorization is a process of comparing new instance to previously established (but highly malleable) mental representations of the category and classification is based on similarity of the instance to the existing representations.

Among the newly developed category theories, prototype theory enjoys the largest support. Emerging from Wittgenstein's earlier ideas on family resemblance, centrality and membership gradience (Wittgenstein, 1953) and Zadeh's fuzzy set theory (Zadeh, 1965), prototype theory (Rosch, 1973, 1978) stresses the fact that category membership is not homogenous and that some members are better representatives of a category than others, which is noted as the "prototype effects" (Lakoff, 1987). Furthermore, it is believed that people may hold more than one kind of prototype for a given category and will deploy them differently depending upon the situation. Other concepts in the prototype approach include:

1. *Internal heterogeneity.* Prototype categories allow for uneven fit. Although penguins cannot fly, they are still birds.
2. *Indeterminate boundaries.* The category "U.S. Senator" is unambiguous—one is either a senator or one is not. However, what threshold defines "rich"? Is it possible to be 5 cents short of being rich?
3. *Categories are relative.* Most qualitative categories —*hot*, *slow*, *old* for example—only have meaning in relation to their counterparts. Moreover, perception is often dependent upon experience: a big city to one person may be a village to another.
4. *Categories change.* Criteria for membership often evolve, as in the case of video games expanding our notion of "game". Classical category theory assumes that categories are static and do not change.

Characteristics of a category can be very broad and haphazard. Categories are no longer seen as passive and static "mental containers." Instead, they are constantly created, refined, and combined to create new categories, and thereby, construct new knowledge.

The prototype of a category is a composite or average of all the real instances experienced, that 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 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 its category members (Davidsson, 1992).

In the case of soil classification, the continuous soil body is categorized into soil classes (soil categories). Such categories are conceptual structures and *internal heterogeneity* is shown when soil scientists think a certain pedon is more representative of a soil class than another one, although both are classified as the same class. Although soils have been traditionally mapped as discrete polygons, it has long been recognized that soil classes have *indeterminate boundaries* in both geographic space and attribute space (Burrough, 1996). In the mapping of soil polygons in terms of soil–landscape units (Hudson, 1990), definitions have always been relative, as in a convex (versus concave) position or steep (versus not steep) slopes. Prototypes may change and criteria for determining memberships evolves through years of field work by experiencing more and more real instances of a soil category. Such a prototype, if realized as an actual instance, has the highest membership among all instances of the category under concern. Soil categorization/classification is then through the comparison of the new instance to such established prototypes.

Traditional soil inventory mapping, however, is rooted in the classical view of categories. It overlooks the prototypical characteristics of soil categories and leads to over-generalization in both the spatial and attribute domain (Zhu, 1997b). Prototype category theory, as presented here, provides an alternative to acquiring knowledge about soil categories by creating category prototypes and defining membership gradations based on the difference between a new instance and the prototype.

3. Prototype-based soil mapping

Based on prototype theory, a category can be represented by its prototype: a composite set of features that summarizes the real instances of the category. This then serves as the cognitive reference points for inference (Khalidi, 1995; Minda and Smith, 2001). New instances are identified and memberships determined by matching the features (Khalidi, 1995; Markman, 1999). Our approach to soil mapping based on prototype theory consists of two steps. The first step involves acquiring knowledge about the soil categories from soil scientists and representing the knowledge in terms of prototypes and membership gradations. The second step is prototype-based soil inference.

3.1. Knowledge representation and knowledge acquisition

Smith and Medin (1981) distinguish two kinds of features related to the representation of a category: the "core features" versus the "identification features". In the case of soil classification, the core features are those properties that are formally described in *Soil Taxonomy* (Soil Survey Staff, 1999) and reveal relations between classes, while the identification features are those that are commonly used to identify the spatial occurrences of certain soil classes. In soil survey, soils are mapped based on local soil–landscape models that describe the relationships between soil and its formative environmental factors (Hudson, 1992). In the context of predictive soil mapping (Scull et al., 2003; McBratney et al., 2003), the formative environment is usually characterized by variables such as terrain characteristics, parent materials, etc. The representation of soil classes for soil inference can thus use these environmental factors as the identification features.

In order to incorporate the prototypical aspects of soil classes, it is necessary to explicitly model the prototypes and membership gradations in the knowledge representation. Specifically, the prototypes of classes can be stored using a common frame structure (Minsky, 1975; Fillmore, 1985; Zhu, 1999), while the membership gradations can be represented with optimality curves (Zhu, 1999). An example of such frame is illustrated in Fig. 1. Each frame defines the prototype(s) of a soil class and is also linked to a set of optimality curves that describe how membership responds when the values of the identification features change. An optimality value of 1 refers to full

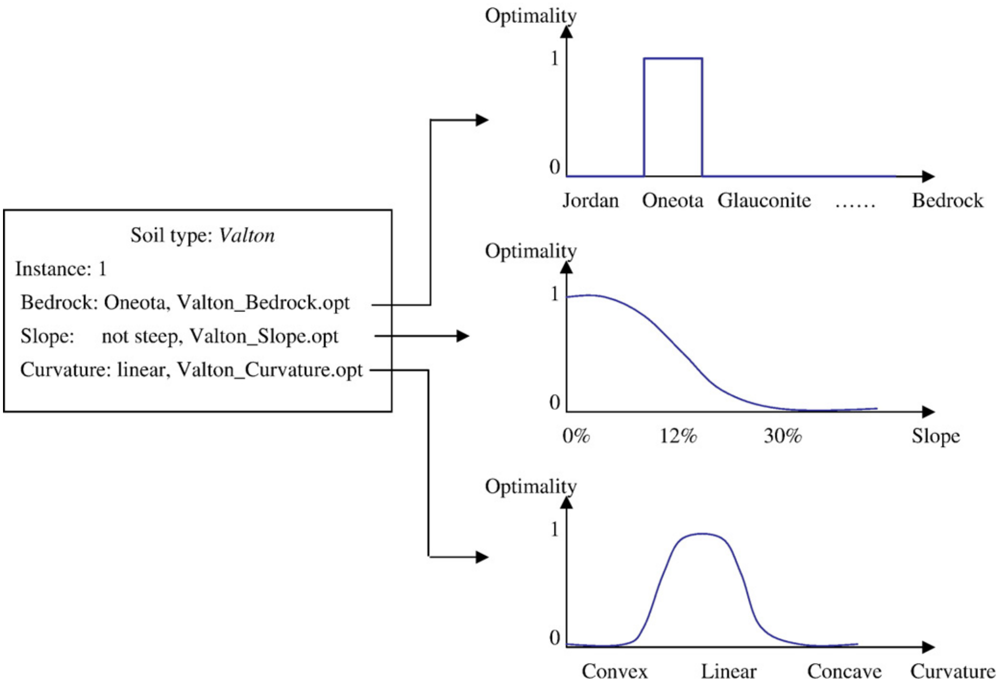

Fig. 1. Improved frame representation of the soil class *Valton*.

membership, and 0 means no membership at all. For soil class *Valton* (Fig. 1), we see that the membership drops from 1 to 0 when the bedrock changes from *Oneota* to any other type, and the membership decreases constantly when slope gets steeper.

Knowledge represented in such a way can be obtained from soil scientists through knowledge acquisition. If the prototype of a soil class is realizable as a local instance, it can be identified on a virtual landscape using a knowledge-acquisition tool such as 3dMapper (Burt and Zhu, 2004). 3dMapper is a visualization software tool that can display arbitrary 3D views of topography and allow for overlay of GIS data layers that are necessary in the identification of a typical soil landscape unit. Fig. 2 shows the identification of the prototype of soil class *Valton* on a virtual landscape using 3dMapper. In the identification of this prototype, soil scientist can load in data layers that capture soil-formative environmental factors. The values of these environmental factors for the prototype can be recorded and stored in the frame structure in Fig. 1. In many cases, however, the prototype of a soil class may not be realizable as a real instance. It is very common that certain soil classes (e.g. soil series) may be first found and defined for other areas and no typical pedons develop in the area currently under concern. In these cases, prototypes should be defined by a soil scientist in the form of descriptive knowledge in knowledge acquisition. This involves the determination of a list of relevant environmental variables (identification features) for each individual soil class and the typical environmental conditions under which the soil class is expected to occur. Once such knowledge is obtained from soil scientist through interviews, a set of frames (Fig. 1) can be constructed to store the prototypes.

Given the prototypes defined for all the possible soil classes in the mapping area, optimality curves can be modeled empirically using *a priori* heuristic curves, as known as the Semantic Import Model (SI) approach (Burrough, 1989; McBratney and Odeh, 1997). The choice for heuristic curves is based on the types of the environmental variables (identification features) (Shi et al., 2004). For a feature that is categorical (such as bedrock geology), the function is Boolean as the one shown in Fig. 1. For continuous features, the soil scientist can choose from the various widely accepted models discussed by Burrough et al. (1992) and MacMillan et al. (2000). The next section will show the use of Gaussian functions in a case study for soil mapping in Wisconsin.

3.2. Prototype-based inference

Once the expert knowledge is acquired in the form of prototype frames and membership curves, it can be used for prototype-based soil inference. For any pixel in the mapping area, the degree of class membership to any predefined soil class can be determined by their degrees of similarity to the class prototype-based on the membership curves. It seems easy to confuse prototype-based inference with case-based reasoning (Shi et al., 2004), in which conclusions are drawn based on similarities to existing classified cases. The important

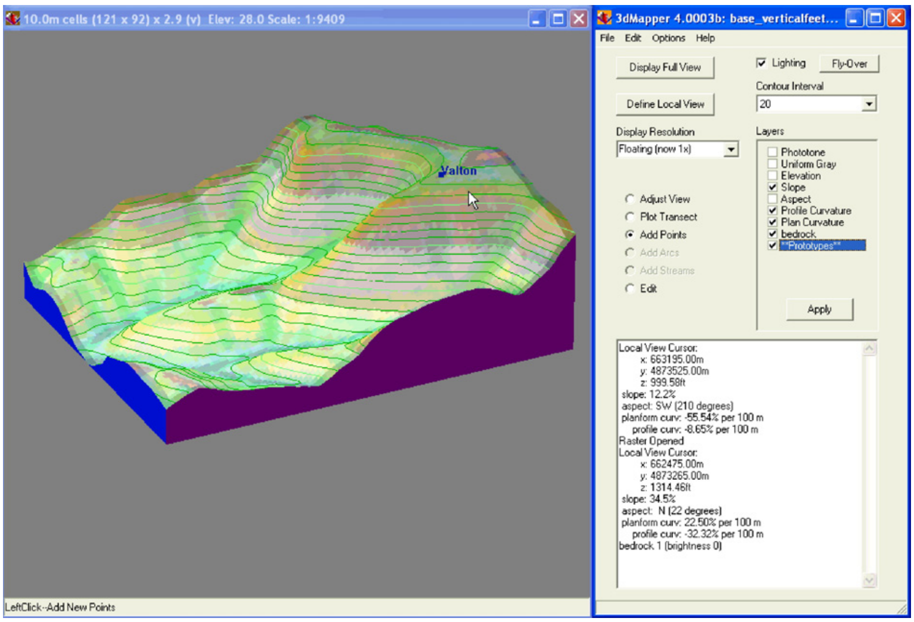

Fig. 2. Identify the prototype of soil class *Valton* on a virtual landscape using 3dMapper.

differences between these two approaches include: (1) case-based reasoning is based on the exemplar view of categories which takes specific exemplars as the category's mental representation and assumes knowledge about the category is implicitly embedded in the individual cases. In contrast, prototype theory views prototypes as the mental representation of a category where knowledge about the category (e.g. identification features) is explicitly represented (Smith and Medin, 1981); (2) case-based reasoning usually requires vast amount of cases (exemplars) while prototype-based reasoning, as will be shown in the case study, may need only one or a few prototypes (central tendency) for each category (Zeithamova, 2003); (3) case-based reasoning uses examples that actually exist as members of the categories under concern while prototypes may not be actual instances, but rather abstract representations or "textbook" examples that might not exist in a given dataset; (4) case-based reasoning uses not only similarities to the existing cases but also the typicality of the cases themselves to determine memberships of a new item, while prototype-based reasoning needs to account for only the similarity to the prototypes.

Our paper suggests that prototype-based inference is a better choice for fuzzy soil mapping due to the following considerations. First, in the case of soil mapping, abundant classified examples may not be always available in order for a case-based reasoning to achieve acceptable performance. Second, soil mapping using case-based reasoning requires the existence of local instances of the categories, but representative examples for some soil classes may not exist locally within the mapping area. Third, it is always desirable to have the knowledge explicitly represented and documented for communication and map updates. Case-based reasoning does not directly allow this because knowledge is embedded in large amount of cases. Unlike in situations where it is impossible to have an abstract summary of the category features, the knowledge of a soil–landscape model can be well represented in a holistic way using prototypes rather than being implicitly embedded in individual exemplars (which are unique to a data set and not readily transferable to new data sets). And fourth, research in cognitive psychology (Minda and Smith, 2001) has noted that exemplar theory has often favored small, poorly structured categories, whereas prototype-based models may represent better categories that are more complex and well structured. In our case study, we will show the results from both case-based reasoning and prototype-based inference to compare the two in soil mapping.

4. Case study

The prototype-based soil inference approach was implemented under the SoLIM framework to update the soil survey of Raffelson watershed in La Crosse, Wisconsin (Fig. 3). The watershed is located in the "Driftless area" of southwestern Wisconsin that has remained free of direct impact from the most recent Pleistocene era continental glaciers. The area has a typical ridge and valley terrain with relatively flat, narrow ridges. There are moderate side slopes with gradients below 20%

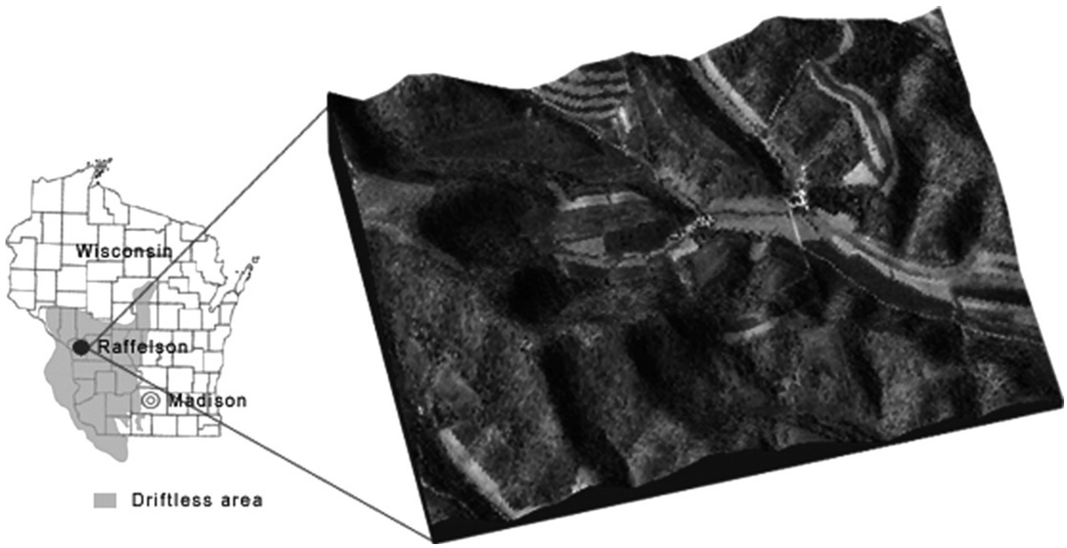

Fig. 3. Location and topography of the Raffelson Watershed, WI.

and steep slopes with high gradient values around 50%. The bedrock in this watershed is mainly of two types: (1) Prairie du Chien dolomite and (2) sandstone of Upper Cambrian age. Dolomite is only present on the high, rounded ridge areas; elsewhere it was removed by geologic erosion. Wherever the dolomite is eroded, the bedrock is sandstone, with erosion and cutting responsible for most of the differences in the landforms. The erosion of bedrock has divided what was once a fairly level plateau and has formed a relatively dissected upland with apparent relief. Most ridges and valleys in the area have been cultivated since late 19th century. Side-slopes are generally forested, though some have been cleared for pasturing.

In terms of parent material, there are as many as five distinct types within the Raffelson watershed area, which are *Oneota* (dolomite), *Glauconite* (sandstone), *Jordan* (sandstone), *Wonewoc* (sandstone) and alluvial materials. The complexity of parent material over the area has led to the development of 16 different series (classes) of soil in the watershed. In this study a senior soil scientist from the local office of U.S Department of Agriculture–Natural Resources Conservation Service (USDA–NRCS) was asked to provide expert knowledge in the form of a soil–landscape model concerning the 16 soil classes.

In acquiring the knowledge of the prototypes for the soil classes, we first interviewed the soil scientist and obtained the descriptions of the central concept for each soil class. We then identified realizations of the prototypes for four soil classes *Lamoille*, *Churchtown*, *Norden*, and *Council* on a virtual landscape because there exist the fully representative instances for these soil classes in the watershed. For the other 12 classes, we obtained the most typical values of relevant environmental variables from the soil scientist in an interview session. Finally, membership changes were modeled with Gaussian optimality curves as shown in Fig. 1. The cross-over points (Burrough, 1989) were determined empirically based on soil scientist's knowledge of the non-overlapping boundary of two adjacent central concepts. Likewise, the shapes of the curves were determined based on soil scientists' descriptions of the central concepts. For example, if the central concept of soil class *Valton* indicates it occurs on slopes that are less than 12% steep, the optimality curve is S-shaped (Fig. 1). On the other hand, if a soil occurs on slopes that are steeper than 20%, the curve has a reversed S-shape. The normal bell-shaped curve indicates the highest membership in the middle, with membership decreasing toward both tails.

In order to conduct soil inference with the acquired expert knowledge, we first created a GIS database of the environmental variables that were listed by soil scientists as the identification features for the soil prototypes. Because different soil classes may have different identification features, our GIS database is extensive to include all features identified in knowledge acquisition. The features used in this research include parent material, elevation, slope gradient, surface curvatures (profile and planform curvatures) (Zevenbergen and Thorne, 1987), and two spatial variables: topographic wetness index and percentage of colluvium from competing bedrocks. Topographic wetness index is used to combine connectivity information based on flow direction with slope dynamics to represent the hydrological topographic characteristics that influence soil formation (Moore et al., 1993; Band et al., 1993). Since colluvium from different bedrocks tends to influence soil development, we included the spatial variable that describes the percentage of colluvium from competing upslope bedrocks for footslope locations. Specifically, for a given footslope location, the relative amount of colluvium it received from a certain bedrock is approximated on the basis of the accumulated upstream drainage cells originating from the given bedrock polygon. The percentage of colluvium from multiple competing upslope bedrocks is then computed relatively.

During soil inference, our inference engine scans through all pixels within the mapping area to compute the similarity vectors. For any given pixel, the inference engine first looks up in the knowledge base the prototype and identification features defined for every soil class. The value of similarity to this class is obtained by imposing a fuzzy-AND operation (Zadeh, 1965) on the optimality values for all identification features. The use of fuzzy-AND operation follows Zhu and Band (1994) and is based on the limiting factor principle in ecology, which states that the limiting factor controls the development of soil formation, thus the environmental variable that gives the least optimality value determines the membership of the soil.

Once the similarity vectors for all locations in the mapping area are computed, a soil series map can be created through the process of defuzzification (Janikow, 1998) by assigning each location the soil class that has the highest membership value in the similarity vector. The resulting raster map is expected to be spatially more detailed than traditional soil survey based on the 'area-class' model. In addition, uncertainties associated with this classification process can be also computed from the similarity vector (Zhu, 1997b) to depict the typicalities of the classified soils and the transitions between soil classes. The continuous variation of soils can be further represented by continuous soil property maps derived

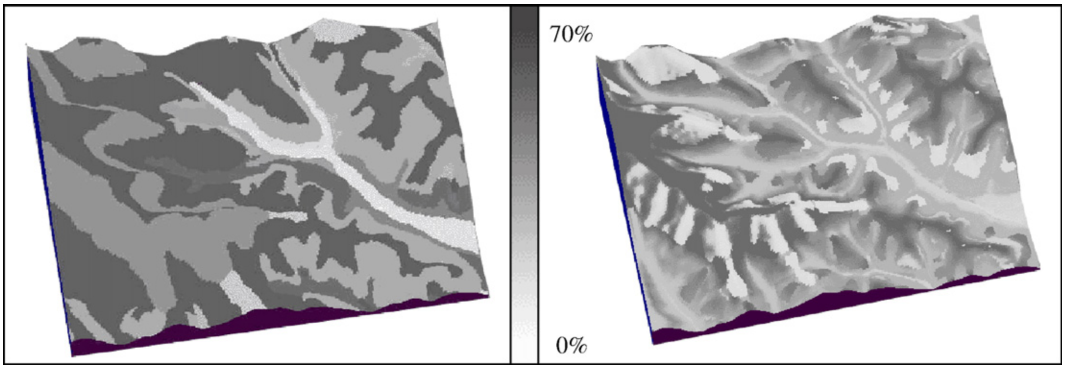

Fig. 4. Distribution of A horizon sand percentage for the Raffelson watershed: (left) from conventional soil survey map; (right) from prototype-based inference.

from the similarity vectors. A continuous soil property (e.g. A horizon depth) map can be generated with the following formula according to Zhu et al. (1997):

$$v_{ij} = \frac{\sum_{k=1}^{n} s_{ij}^{k} v^{k}}{\sum_{k=1}^{n} s_{ij}^{k}} \tag{1}$$

where v_{ij} is the property at site (i, j); v^k is the typical value (either defined in national standards or determined locally) of that property of soil class k; s_{ij}^{k} is the membership value of soil class k at (i, j); n is the total number of soil classes in the area.

Field validation was conducted in order to evaluate the prototype-based fuzzy soil mapping approach. We collected data from 99 field points in the Raffelson watershed, of which all were classified and assigned soil series names by two experienced soil scientists from USDA–NRCS local offices and 49 were given a texture analysis to determine the percentages of sand and silt in the A horizon. The surface texture property was selected because the relative sand, silt, and clay content of the surface soil is a fundamental soil property that is closely related to many other soil properties such as permeability, porosity, water holding capacity, and soil fertility (Posadas et al., 2001).

The validity of the prototype-based fuzzy mapping approach was evaluated in two aspects. First, the inference results were compared to the traditional soil survey map in terms of classification accuracy, map consistency and the level of detail both spatially and parametrically. Second, the prototype-based inference results were compared to case-based reasoning results in terms of classification accuracy based on inferred soil series maps and ability to capture continuous soil property variations based derived soil property maps.

5. Results and discussion

5.1. Comparison of prototype-based inference results with traditional soil survey map

The 99 sample points for field evaluation were chosen based on two sampling strategies: transect sampling and

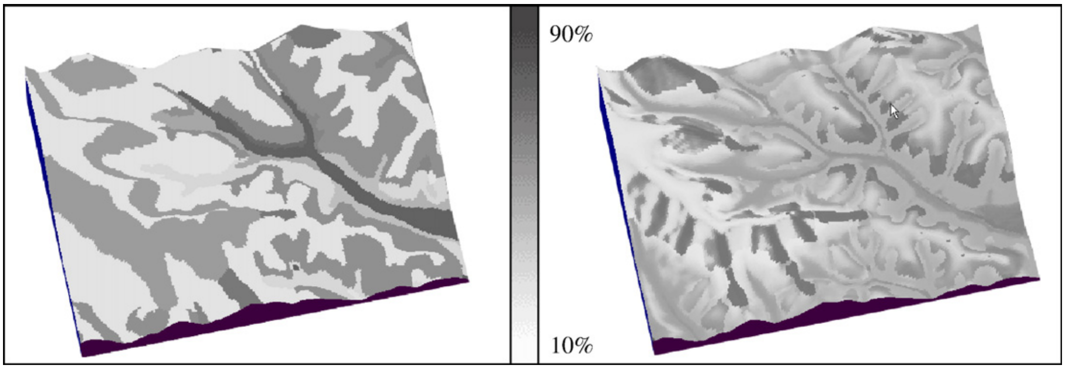

Fig. 5. Distribution of A horizon silt percentage for the Raffelson watershed: (left) from conventional soil survey map; (right) from prototype-based inference.

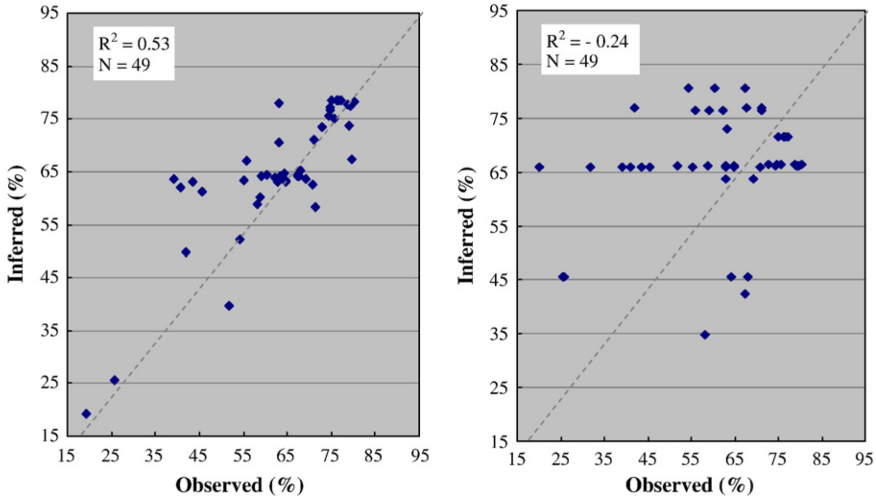

Fig. 6. Scatter plots of observed *A* horizon silt percentage values vs. the values derived from traditional soil map (left) and prototype-based inference (right) at 49 sample locations in Raffelson watershed.

point sampling. Transect sampling was employed to cover transitions between major landscape units (such as from ridge top to valley bottom and from concave draw to convex nose positions) at 53 of the 99 sites. The remainder of the 99 sites was selected by local scientists based on the topography of the watershed to cover the major landscape units (such as ridge tops, shoulder positions, valley bottoms) throughout the watershed. The soil series names given by soil scientists of these 99 sites were compared to both those mapped in conventional soil survey and those obtained from the inferred soil series map derived from prototype-based inference. Of the 99 sites, the prototype-based approach correctly inferred the soil series at 83 sites (~83.8%), while the conventional soil survey mapped 66 sites correctly (~66.7%). A thorough examination of both maps indicates that the improved performance is largely explained by three factors. First, spatial distributions of soil classes on the inferred soil map follow accurately soil scientist's knowledge about the soil–landscape relationships while conventional soil map shows evidence of misplacement of class boundaries. This is often inevitable in manual soil survey due to the difficulty in precise determination of landscape characteristics through visual perception. Second, the automated approach allows for consistent application of soil scientist's knowledge in the entire mapping area while manual soil survey may introduce inconsistency. And third, inclusions that are common in conventional soil map due to limitations of scale can be avoided with the raster based representation because soil variations can be depicted in more detail.

In addition to improved accuracy in classifying soil series, the advantage of prototype-based soil inference largely lie in its ability to capture within-class variations and transitions between class prototypes. As aforementioned, soil at each pixel location is represented by a similarity vector and the soil series map is created by defuzzifying the similarity vectors at all pixel locations. Although each pixel location is labeled as the soil class to which it has the highest similarity value, its similarity to other soil classes provides information about the soil pedon's typicality to its assigned class. In order to measure numerically the accuracy of the inferred map in capturing continuous variations of soil properties, we created continuous soil texture maps using Eq. (1). The typical soil texture values for each soil series (v^k) that appears in our study area were obtained from the sample site associated with the highest similarity value to the series according to Moore (2004). For the sake of comparison, soil texture maps based on the published soil survey map were also generated by assigning each pixel the typical texture values of the soil series as which the pixel is labeled in the map according to official soil survey records. Figs. 4 and 5 show the maps of *A* horizon texture in terms of sand and silt percentages, respectively. The apparent difference between the maps derived from the prototype-based inference and those

Table 1
Accuracy of the derived *A* horizon texture in the Raffelson watershed: the prototype-based inference result vs. the soil survey map

	Percentage of sand			Percentage of silt		
	MAE	RMSE	AC	MAE	RMSE	AC
Inference result	7.75	12.65	0.85	6.49	11.11	0.85
Soil survey map	14.87	19.83	0.58	14.51	18.08	0.46

Table 2
Comparison of soil series inferred from prototype-based approach and case-based reasoning against the field observations for the Raffelson study area

	Oneota (39 samples)		Glauconite (41 samples)		Wonewoc (8 samples)		Jordan (6 samples)		Aluvial (5 samples)	
	Correct	%	Correct	%	Correct	%	Correct	%	Correct	%
Prototype-based	37	95	36	88	5	63	3	50	2	40
Case-based	35	90	34	83	6	75	4	75	2	40

based on soil survey is the shown spatial details and continuity of soil texture variations. Inferred texture maps depict more details than those based on soil survey and tend to show continuous changes between soil types within the same bedrock breaks while the survey maps show abrupt changes along class boundaries.

Closer inspection of the texture maps also show that the prototype-based inference results provide realistic details about spatial variations of soil texture. For example, in the Raffelson watershed sandstone bedrock (*Wonewoc*, *Jordan*) tends to develop soils with coarser texture (higher sand and lower silt) than Dolomite (*Oneota*). Within a catena, however, we would expect a low percentage of sand on flat ridge tops due to the preservation of finer materials and an increase of sand on back slopes due to the erosion of finer materials (silt percentage should exhibits a reversed pattern). Furthermore, convergent areas should have a low percentage of sand and high percentage of silt due to the accumulation of fine materials and divergent areas show the opposite. This expected spatial pattern of soil texture is clearly observable from the maps derived from inference results, while the map based on soil survey lacks the same level of details. The traditional soil survey classification differentiates only two discrete soil classes occurring on ridge top and back slope positions and does not capture the continuous transitions between and within such positions. As a result, the texture maps based on soil survey fail to separate ridge and shoulder positions although they tend to have different soil textures in reality and cannot differentiate convex back slope from concave locations where differential erosion and accumulation patterns should result in distinct textures.

Particle-size analysis were conducted with samples collected at 49 field sites using the pipette method (Kilmer and Alexander, 1949) to determine *A* horizon textures. The two scatter plots in Fig. 6 compare the estimation of percentages of silt in *A* horizon from prototype-based inference with those derived from soil survey map. Fig. 6 (left) shows the tendency of the inferred values to follow the field observations, while Fig. 6 (right) shows little correlation between the observed values and those obtained from soil survey map. Three indices were also computed to evaluate the performance of prototype-based inference: mean average error (MAE), root mean square error (RMSE), and agreement coefficient (AC). The AC is defined as (Willmott, 1984):

$$AC = 1 - \frac{n \cdot \mathrm{RMSE}^2}{\mathrm{PE}}, \tag{2}$$

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

$$PE = \sum_{j=1}^{n} (|P_i - \bar{O}| + |O_i - \bar{O}|)^2, \tag{3}$$

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 indicates perfect agreement and 0 means complete disagreement between the estimated and observed values (Willmott, 1984). Table 1 lists these statistics for comparing the performance of prototype-based inference with the soil survey map in estimating soil texture values. The MAE and RMSE statistics for prototype-based inference are consistently lower than those for the soil map, and the significantly higher AC for prototype-based inference supports previous evidence that prototype-based inference is able to capture continuous variations of soil properties better than the soil survey map.

5.2. Comparison of prototype-based inference with case-based reasoning

Case-based reasoning has been implemented with the procedure outlined in Shi et al. (2004) in the Raffelson

Table 3
Accuracy of the derived *A* horizon texture in the Raffelson watershed: the prototype-based inference result vs. the case-based reasoning result

	Percentage of sand			Percentage of silt		
	MAE	RMSE	AC	MAE	RMSE	AC
Prototype-based inference	7.75	12.65	0.85	6.49	11.11	0.85
Case-based reasoning	11.07	15.92	0.76	8.93	13.23	0.75

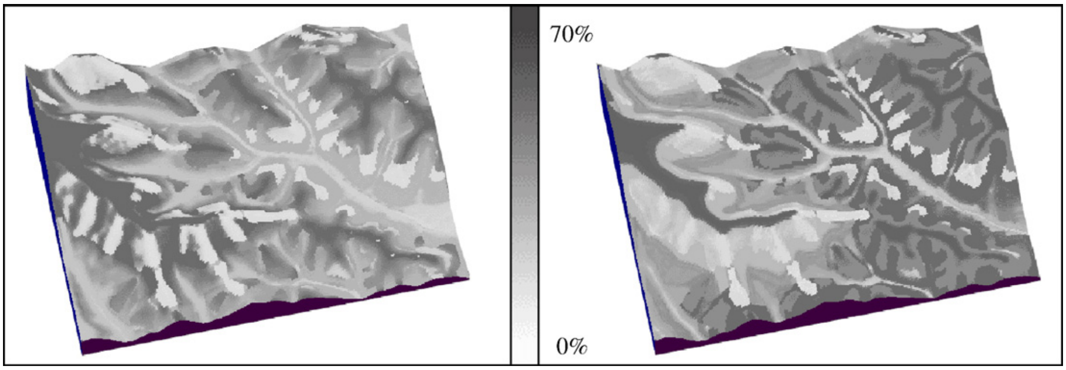

Fig. 7. *A* horizon sand percentage for the Raffelson watershed: (left) from prototype-based inference; (right) from case-based reasoning.

watershed. A total of 78 cases were used to create a case-based soil series map for the area. Of the 99 field sites, the prototype-based approach correctly inferred the soil series at 83 sites (~83.8%), while case-based reasoning inferred 81 sites correctly (~81.8%). A breakup of soil series based on parent materials provides more information about the performances of the two approaches. Table 2 lists the accuracies of the two approaches in classifying soil series developed on the five different parent materials in the watershed. The table shows that prototype-based approach performs better on the two most dominant parent materials that cover over 90% of the watershed area while case-based reasoning shows advantage in small areas (less than 5% of the total area) of the two minor parent materials.

By examining the nature of the misclassified samples, an apparent difference between the prototype-based inference and case-based reasoning were noticed. The sample points misclassified by the prototype-based inference in areas of the three minor parent materials are those whose occurrences can not be explained by soil scientist's soil–landscape model. In other words, the landscape characteristics of these sample points do not match those provided by soil scientist for the observed soil series. One possible reason is that these local soils exist as exceptions that are caused by local disturbances. A second possible reason is that soil scientist's knowledge about these soils is not well-developed since the area is small and the formulation of mental prototypes has not achieved a sufficient level. In either case, the soil class cannot be sufficiently represented in soil scientist's mind with a single prototype that summarizes the central tendency of real instances. Therefore inference based on prototypes did not perform as well as inference based on individual cases.

Prototype-based inference performed better than case-based reasoning in the two dominant parent material areas for which the prototypes have been well-developed given sufficient experience with real instances. It was noted that case-based reasoning misclassified some of the samples although the occurrences of these soils match exactly soil scientist's knowledge. This could be due to the fact that case-based reasoning requires large amount of real cases to achieve high accuracy. Although it helped in capturing exceptions to general rules, the overall accuracy sacrifices from the lack of cases at dominant landscape locations.

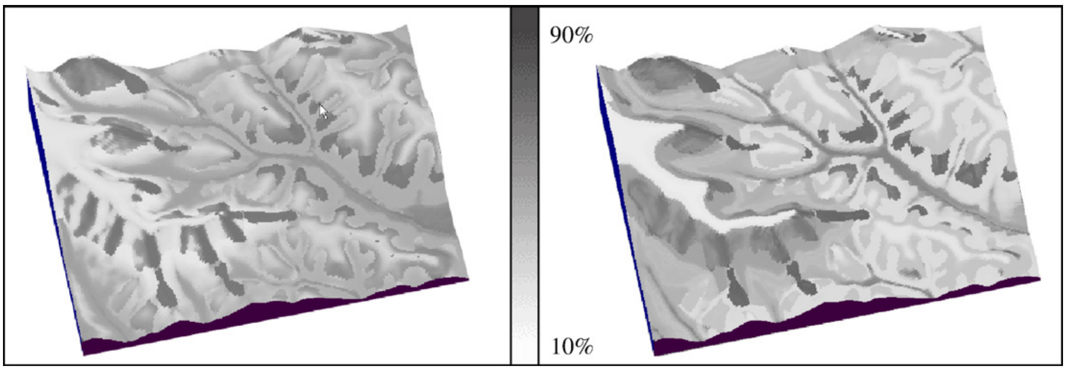

Fig. 8. (left) *A* horizon silt percentage for the Raffelson watershed: (left) from prototype-based inference; (right) from case-based reasoning.

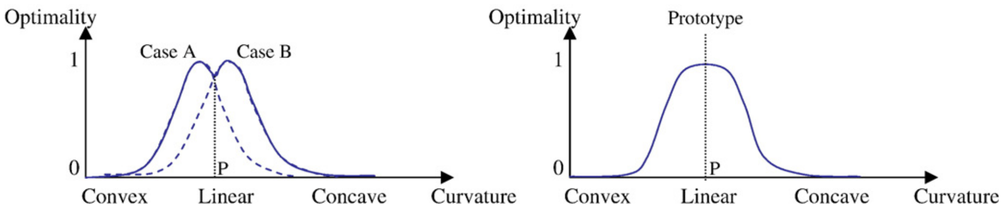

Fig. 9. Membership curves for soil *Valton* based on curvature: (left) membership curve defined by 2 cases. (right) membership curve defined by one prototype.

The performances of prototype-based inference and case-based reasoning in modeling continuous soil properties are further compared through field evaluations using *A* horizon textures at 49 sample points. Figs. 8 and 9 illustrate the *A* horizon soil texture maps derived from case-based reasoning results with Eq. (1) in comparison to those derived from prototype-based inference. And Table 3 lists the statistics computed for both approaches. It is noticeable that the property maps generated from prototype-based inference exhibits more continuous transitions than those from case-based reasoning (Figs. 7 and 8). The reason is that the prototype-based approach models membership gradations with a set of curves that give the highest membership to instances that are closest to class prototypes and make gradual decrease of memberships proportional to the deviation from prototypes. This leads to smooth transitions between class prototypes. Case-based reasoning, on the other hand, models instance memberships based on the typicalities of the cases themselves by taking into account the occurrences of cases at similar landscape locations. The lack of cases at certain landscape locations, therefore, may lead to misrepresentation of the case typicalities and results in faulty membership values. Fig. 9 shows the membership curve simulated from 3 cases used in case-based reasoning, where the lack of cases between point *A* and *B* causes the unrealistic local dip in membership at *P*. As a result, the property values derived using the similarity vectors computed from these cases may not match the observed values. As shown in Table 3, the texture values of the 49 field points obtained from prototype-based reasoning match the observed values significantly better (AC = 0.85 for sand and 0.85 for silt) than those from case-based reasoning (AC = 0.76 for sand and 0.75 for silt), although their accuracies in classifying soil series names are not significantly different.

6. Conclusions

This paper has presented a fuzzy soil mapping approach based on prototype category theory. The case study in the Raffelson watershed shows that results from the prototype-based inference are significantly more accurate than both the conventional soil map in both soil class identification and property estimation (accuracy increase of ~17% on soil series name prediction and agreement coefficient increase of ~0.3 on *A* horizon texture estimation). Compared to case-based reasoning, prototype theory shows significant improvement only in property estimation (accuracy increase of ~2% on soil series name prediction and agreement coefficient increase of ~0.1 on *A* horizon texture estimation). This case study demonstrates that the prototype-based approach is an effective method in terms of knowledge acquisition and knowledge-based fuzzy mapping. It shares the same advantages with previous knowledge-based fuzzy soil mapping methods: the automated mapping process is more efficient than manual soil survey and reduces errors introduced in manual compilation; soil scientist's knowledge of the local soil–landscape model can be consistently applied in the entire mapping area; and the fuzzy representation scheme allows for the representation of high level of details of soil information. Meanwhile, prototype category theory provides a cognitive basis for the process of knowledge acquisition, knowledge representation and fuzzy soil inference. It not only leads to more accurate classification of soil categories, but also represents the knowledge of a soil–landscape model explicitly in a holistic way that is readily reusable and transferable to new data sets. In consideration of the primary characteristic elements of human categorization, the prototype-based approach shows advantages over case-based reasoning in capturing within-class variations and transitions between soil classes. As noted by Suchan (1998, p.v–vi):

> Prototype category theory can be a prompt to imaginative thinking about the categories we craft or choose before definitions are formalized for geographic representation. A category precedes and influences data collection, analysis, synthesis, and presentation. By viewing categories differently, we may gain insights to a given category and its relations to others.

Our case study shows that the prototype-based approach applies successfully in the “driftless area” of southwestern Wisconsin, especially for soils of which the soil–landscape model is well-developed based on soil scientist's field experience. The general framework (the knowledge representation, acquisition, and soil inference schemes) should be also applicable in other parts of the world. Successful application of this approach, however, may require changes to the soil category level and identification features being used based on the local landscape characteristics and the purpose of the inventory. Applications in other landscapes are under way to further examine the prototype-based approach within the SoLIM framework.

Like other knowledge-based methods, we recognize that this approach is highly sensitive to the quality of the acquired expert knowledge. If the prototypes of soil classes are poorly defined, the performance of this approach can be less trustworthy, as shown by soils on the *Jordan* and *Wonewoc* parent material in the study watershed. In this case, proper utilization of specific knowledge in terms of cases can improve the accuracy of soil inference. The cognitive basis for inference based on specific knowledge is the exemplar category theory. Exemplar models store the specific exemplars as the category's mental representation and inference is accomplished through case-based reasoning. Studies have shown that no single theory accounts for all real-world categories: the exemplar theory is often favored for small, poorly structured categories containing low-dimensional stimuli, whereas prototype-based strategies are better for representing categories that are better structured and contain higher dimensional stimuli (Minda and Smith, 2001). For real world categories such as soil classes, the two situations may exist at the same time. As suggested by psychologist Zeithamova (2003), it should be necessary to adopt a composite model in situations where both kinds of categories are present in the categorization system. It thus is necessary in our future research to develop a more versatile system that allows for knowledge acquisition and inference based on both prototypes and specific exemplars if not all soil classes are represented as well-formulated prototypes in soil scientist's mental representation of the soil–landscape model.

Acknowledgements

The work reported in this paper is supported by the funding from the Natural Resources Conservation Service of United States Department of Agriculture, and by the Chinese Academy of Sciences through its "One-Hundred Talents" program to A-Xing Zhu, and by the University of Wisconsin–Madison. The authors also wish to thank the anonymous reviewers for their comments on the manuscript.

References

Band, L.E., Patterson, P., Nemani, R.R., Running, S.W., 1993. Forest ecosystem processes at the watershed scale: 2. Incorporating hillslope hydrology. Agricultural and Forest Meteorology 63, 93–126.

Burrough, P.A., 1989. Fuzzy mathematical methods for soil survey and land evaluation. Journal of Soil Science 40, 477–492.

Burrough, P.A., 1996. Natural objects with indeterminate boundaries. In: Burrough, P.A., Frank, A.U. (Eds.), Geographic Objects with Indeterminate Boundaries. Francis and Taylor, London.

Burrough, P.A., McMillan, R.A., Van Deursen, W., 1992. Fuzzy classification methods for determining land suitability from soil profile observations and topography. Journal of Soil Science 43, 193–210.

Burt, J.E., Zhu, A.X., 2004. 3dMapper 4.02. Terrain Analytics. LLC, Madison, WI. www.terrainanalytics.com.

Davidsson, P., 1992. Concept acquisition by autonomous agents: cognitive modeling versus the engineering approach. Lund University Cognitive Studies, vol. 12, ISSN:1101–8453. Lund University, Sweden.

Fillmore, C., 1985. Frames and semantics of understanding. Quaderni di Semantica 6, 222–253.

Ford, K.M., Petry, F.E., Adams-Webber, J.R., Chang, P.J., 1991. An approach to knowledge acquisition based on the structure of personal construct systems. IEEE Transactions on Knowledge and Data Engineering 3, 78–87.

Hahn, U., Ramscar, M., 2001. Similarity and Categorization. Oxford University Press, New York.

Hudson, B.D., 1990. Concepts of soil mapping and interpretation. Soil Survey Horizons 31, 63–73.

Hudson, B.D., 1992. The soil survey as paradigm-based science. Soil Science Society of America Journal 56, 836–841.

Janikow, C.Z., 1998. Fuzzy decision trees: issues and methods. IEEE Transactions on Systems, Man and Cybernetics. Part B. Cybernetics 28, 1–14.

Khalidi, M.A., 1995. Two concepts of concept. Mind and Language 10, 402–422.

Kilmer, V.J., Alexander, L.T., 1949. Methods of making mechanical analyses of soils. Soil Science 68, 15–24.

Lakoff, G., 1987. Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. University of Chicago Press, Chicago.

MacMillan, R.A., Pettapiece, W.W., Nolan, S.C., Goddard, T.W., 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets Syst. 113, 81–109.

Mark, D.M., Csillag, F., 1989. The nature of boundaries on 'area-class' maps. Cartographica 26, 65–78.

Markman, A.B., 1999. Knowledge Representation. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

McBratney, A.B., Odeh, I.O.A., 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma 77, 85–113.

McBratney, A.B., Mendonca Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

McCracken, R.J., Cate, R.B., 1986. Artificial intelligence, cognitive science, and measurement theory applied in soil classification. Soil Science Society of America Journal 50, 557–561.

McSweeney, K., Gessler, P.E., Slater, B.K., Petersen, G.W., Hammer, R.D., Bell, J.C., 1994. Towards a new framework for modeling the soil–landscape continuum. Factors of soil formation: a fiftieth anniversary retrospective. Soil Science Society of America Journal Special Publication 33, 127–143.

Minda, J.P., Smith, J.D., 2001. Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. Journal of Experimental Psychology. Learning, Memory, and Cognition 27, 775–799.

Minsky, M., 1975. A framework for representing knowledge. In: Winston, P.H. (Ed.), The Psychology of Computer Vision. McGraw-Hill, New York.

Molokova, O.S., 1993. Methodology for the acquisition of knowledge for expert systems: part 1. Basic concepts and definitions. Journal of Computer and Systems Sciences International 31, 1–5.

Moore, A.C., 2004. Predicting soil property variation using a soil landscape inference model. Thesis (M.S.), University of Wisconsin-Madison.

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal 57, 443–452.

Özesmi, U., Özesmi, S.L., 2004. Ecological models based on people's knowledge: a multi-step fuzzy cognitive mapping approach. Ecological Modelling 176, 43–64.

Posadas, A.N.D., Gimenez, D., Bitelli, M., Vaz, C.M.P., Flury, M., 2001. Multifractal characterization of soil particle size distributions. Soil Science Society of America Journal 65, 1361–1367.

Rosch, E.H., 1973. Natural categories. Cognitive Psychology 4, 328–350.

Rosch, E.H., 1978. Principles of categorization. In: Rosch, E.H., Lloyd, B.B. (Eds.), Hillsdale, Cognition and Categorization. Lawrence Erlbaum Associates, NJ.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Progress in Physical Geography 27, 171–197.

Shi, X., Zhu, A.X., Burt, J.E., Qi, F., Simonson, D., 2004. A case-based reasoning approach to fuzzy soil mapping. Soil Science Society of America Journal 68, 885–894.

Smith, E.E., Medin, D.L., 1981. Categories and Concepts. Harvard University Press, Cambridge, MA.

Soil Survey Staff, 1999. Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys. Natural Resources Conservation Service, U.S. Department of Agriculture, Agriculture Handbook, vol. 436. U.S. Government Printing Office, Washington D.C., pp. 10–869.

Suchan, T., 1998. Categories in Geographic Representation. Ph.D. Thesis, The Pennsylvania State University, University Park, PA.

Weibel, R., Keller, S., Reichenbacher, T., 1995. Overcoming the knowledge acquisition bottleneck in map generalization: the role of interactive systems and computational intelligence. In: Frank, A., Kuhn, W. (Eds.), Spatial Information Theory: A Theoretical Basis for GIS. Springer-Verlag, London, pp. 139–156.

Willmott, C.J., 1984. On the evaluation of model performances in physical geography. In: Gaile, G.L., Willmott, C.J. (Eds.), Spatial statistics and models. D. Reidel Publi., Dordrecht, the Netherlands, pp. 43–460.

Wittgenstein, L., 1953. Philosophical Investigations. Macmillan, New York.

Zadeh, L., 1965. Fuzzy sets. Information and Control 8, 338–353.

Zeithamova, D., 2003. Theories of categorization and category learning: why a single approach cannot be sufficient to account for the phenomenon. Studia Psychologica 45, 169–185.

Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surface Processes and Landforms 12, 47–56.

Zhu, A.X., 1997a. A similarity model for representing soil spatial information. Geoderma 77, 217–242.

Zhu, A.X., 1997b. Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. Photogrammetric Engineering and Remote Sensing 63, 1195–1202.

Zhu, A.X., 1999. A personal construct-based knowledge acquisition process for natural resource mapping. International Journal of Geographical Information Science 13, 119–141.

Zhu, A.X., Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. Canadian Journal of Remote Sensing 20, 408–418.

Zhu, A.X., Band, L.E., Dutton, B., Nimlos, T., 1996. Automated soil inference under fuzzy logic. Ecological Modeling 90, 123–145.

Zhu, A.X., Band, L.E., Vertessy, R., Dutton, B., 1997. Deriving soil property using a soil land inference model (SoLIM). Soil Science Society of America Journal 61, 523–533.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal 65, 1463–1472.