A Case-based Reasoning Approach to Fuzzy Soil Mapping

Xun Shi,* A-Xing Zhu, James E. Burt, Feng Qi, and Duane Simonson

ABSTRACT

Some problems in traditional soil mapping—high cost, high subjectivity, poor documentation, and low accuracy and precision—have motivated the development of a knowledge-based fuzzy soil mapping system, named SoLIM (Soil Land Inference Model). The rule-based method of the current SoLIM has its limitations. It requires explicit knowledge of the details of soil–environment relationships and it assumes that the environmental variables are independent from each other. This paper presents a case-based reasoning (CBR) approach as an alternative to the rule-based method. Case-based reasoning uses knowledge in the form of specific cases to solve a new problem, and the solution is based on the similarities between the new problem and the available cases. With the CBR method, soil scientists express their knowledge by providing locations (cases) indicating the association between a soil and a landscape or environmental configuration. In this way, the soil scientists avoid the difficulties associated with depicting the details of a soil–environment relationship and assuming the independence of environmental variables. The CBR inference engine computes the similarity between the environmental configuration at a given location and that associated with each case representing a soil type, and then uses these similarity values to approximate the similarity of the local soil at the given location to the given soil type. A case study in southwestern Wisconsin demonstrates that CBR can be an easy and effective way for soil scientists to express their knowledge. For the study area, the result from the CBR inference engine is more accurate than that from the traditional soil mapping process. Case-based reasoning can be a good solution for a knowledge-based fuzzy soil mapping system.

Soil mapping is basically an inference process based on Jenny's model (Jenny, 1941, 1980). In routine soil survey and mapping, this model can be represented as

$$S = f(E) \quad [1]$$

where S denotes soil, E denotes environmental variables, and f denotes the soil–environment relationship (soil–landscape model). According to this model, if the environmental conditions at a given location and the soil–environment relationship are known, then it is possible to infer the conditions of soil at that given location. With today's spatial information technologies, including geographic information systems (GIS), remote sensing, and the Global Positioning System (GPS), it is possible to characterize the environmental conditions in details. Defining the soil–environment relationship, however, has always been a challenge in soil mapping (Hole and Campbell, 1985; Hudson, 1992; McKenzie et al., 2000). To date defining the soil–environment relationship for soil mapping purpose is still largely a mental process (Hudson, 1992; McKenzie et al., 2000). A well-trained and experienced soil scientist is capable of properly grasping the soil–environment relationships in a certain area and using these relationships to infer the spatial distribution of soils over the area.

Associated with this mental process is the manual process in creating soil maps: with the built soil–landscape model for a mapping area in mind, a soil scientist manually delineates soil polygons on orthophotos under stereoscopes. Several problems are associated with this manual process. The first is the high cost (on money, labor, and time). Zhu et al. (2001) indicated that with the current rate of soil survey updating, updating all of the soil surveys in the USA requires 220 yr. The second problem is the high subjectivity. Researchers have noticed that different soil scientists may map the same area in significantly different ways (Bie and Beckett, 1973; Burrough et al., 1997; McBratney and Odeh, 1997; MacMillan et al., 2000), and this is at least partially due to the inconsistency in the manual mapping process. Another problem is that the knowledge is hard to preserve in this field and training a qualified soil scientist is expensive. This is because the manual mapping is largely a personal operation that lacks a scheme to guarantee good documentation of the knowledge. Still another problem is with the polygon-based model. This model assumes that the soils are the same everywhere within a polygon and are to be of the type assigned to this polygon, and they change abruptly at the polygon boundary. Apparently, in most situations this assumption is not valid, as soils often change continuously over both geographical and property spaces (e.g., Burrough et al., 1997; McBratney and Odeh, 1997; Zhu, 1997a). The manual mapping does not allow this continuous variability of soils to be precisely represented, even if the soil scientists do know the continuous nature of soil variation.

These problems have motivated the development of knowledge-based systems and the application of fuzzy logic in this field. Knowledge-based systems aim at making a good utilization of domain experts' knowledge, meanwhile trying to avoid the problems associated with a manual process, such as inconsistency, tediousness, and loss of knowledge due to personnel change. Researchers have used knowledge-based systems to classify soil samples (Galbraith and Bryant, 1998; Galbraith et al., 1998; Holt and Benwell, 1999), predict soil properties (Cook et al., 1996), and map soil-landscape units

X. Shi, Dep. of Geography, Dartmouth College, 6017 Fairchild, Hanover, NH 03755; A-X. Zhu, State Key Lab of Resources and Environmental Information Systems, Inst. of Geographical Sciences and Natural Resources Res., Chinese Academy of Sciences, Building 917, Datun Road, An Wai, Beijing 100101, China; A.-X. Zhu, J.E. Burt, and F. Qi, Dep. of Geography, University of Wisconsin-Madison, 550 North Park Street, Madison, WI 53706; D. Simonson, NRCS-USDA, 1850 Bohmann Drive, Suite C, Richland Center, WI 53581. Received 25 Apr. 2002. *Corresponding author (xun.shi@dartmouth.edu).

Published in Soil Sci. Soc. Am. J. 68:885–894 (2004).

677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: 3D, three-dimensional; CBR, case-based reasoning; GIS, geographic information systems; SoLIM, soil land inference model.

(Skidmore et al., 1991). Most of these authors recognized the necessity of employing certain techniques to represent soil scientists' knowledge of the continuity in the soil distribution. Some of them (Galbraith et al., 1998; Holt and Benwell, 1999) explicitly pointed out the usefulness of fuzzy logic. On the other hand, the applicability and advantages of fuzzy logic in soil survey and mapping have been systematically studied and well justified (Burrough, 1989; Burrough et al., 1992; Burrough et al., 1997; Mays et al., 1997; McBratney and De Gruijter, 1992; McBratney and Odeh, 1997; De Bruin and Stein, 1998; Zhu and Band, 1994; Zhu et al., 1996).

Among the pioneer experiments, the SoLIM (Zhu and Band, 1994; Zhu et al., 1996, 1997 2001; Zhu, 1997a, 1997b, 1999) may be the first knowledge-based fuzzy mapping system that can be used in routine soil mapping practice. There are, however, two interrelated limitations with the current SoLIM. First, it requires explicit knowledge from the soil scientist—the soil scientist needs to create fuzzy rules to explicitly depict in detail how a soil varies in accordance with an environmental variable. Second, it needs the variable independence assumption—when creating fuzzy rules for a specified variable, the soil scientist has to assume that all the environmental variables are independent from each other, because he/she can work only on one variable at a time. In practice, a soil scientist may often be unable to give details of the relationship between a soil and an individual environmental variable. This may be because the soil scientist has not formulated the explicit rules for the soil and the environmental variable in the mapping area, or because the soil scientist knows there are significant interactions among environmental variables, but has no way to depict this complexity when working on a single variable. Although the authors of SoLIM understand the importance of the interactions among environmental variables (Zhu and Band, 1994; Zhu et al., 1996), they consider it not feasible to have the soil scientist simultaneously handle multiple environmental variables using the rule-based method of the current SoLIM (Zhu et al., 1996).

This paper presents the use of a CBR method as an alternative to the rule-based method used by the current SoLIM. Case-based reasoning refers to a concept and the corresponding technology in the knowledge-based system discipline. It uses the knowledge represented in specific cases to solve a new problem (Aamodt and Plaza, 1994; Kolodner, 1993; Leake, 1996; Watson, 1997). A case in CBR contains two basic parts: the description of the problem and the solution of the problem (Kolodner, 1993). The description part is for evaluating the similarity between the case and a new problem. If the case and the new problem are similar enough, then the solution part of the case is used to solve the new problem. The possibility of using CBR to solve spatial problems and the advantages of CBR-GIS hybrid systems in certain application domains have been discussed by Yeh and Shi (1999), Shi and Yeh (1999), and Holt and Benwell (1999).

The applicability of CBR in soil mapping can be justified through examining the two assumptions of CBR. The first assumption is that cases are capable of representing domain experts' knowledge. Hudson (1992) finds that a large part of soil scientists' knowledge can be subsumed to *tacit knowledge*, which is learned from practical work, especially from field experiences. The tacit knowledge of soil scientists is often the most important knowledge in soil mapping, yet is the most difficult knowledge to learn by a new soil scientist and by the computer, because it is usually hard to articulate and generalize, due to the fact that the soil-forming process can be highly complicated and has not been fully understood. As a result, a large part of the knowledge of the soil–environment relationship is empirical and exists in the form of cases. It might be difficult to generalize these cases to form explicit and general rules. However, according to the studies in the knowledge-based system field (e.g., Schank, 1982; Kolodner, 1993) CBR can be very effective in capturing and representing knowledge existed in the form of specific cases.

The second assumption of CBR is that a new problem can be solved by referring to similar cases. The concept of landscape unit in traditional soil survey and mapping provides a basis for using the similarity-based method to conduct soil inference. Hudson (1992) listed several basic characteristics of landscape unit, of which two are most relevant to applying CBR to this field: "Generally, the more similar two units are, the more similar their associated soils tend to be; conversely, dissimilar units tend to have dissimilar soils"; and "Same or similar units can occur again and again in space." These two principles provide the basis for inferring soils by referring to soils (cases) with similar environmental conditions.

In this research, a complete methodology of using CBR to conduct knowledge-based fuzzy soil mapping is developed. The methodology contains two major parts: the case-based knowledge acquisition process and the case-based soil inference process. The main objective of this research is to study effectiveness of this CBR method in capturing knowledge on soil–environmental relationships and in mapping spatial distribution of soils under the SoLIM framework.

MATERIALS AND METHODS

Study Site

The study area of this research is the Pleasant Valley, a watershed in southwestern Wisconsin (Fig. 1 and 2). The area of this study area is about 5 km^2. It is located in the eastern portion of the Driftless Area, which was not directly overridden by continental ice sheets during the Quaternary. The major bedrock in the Pleasant Valley is Jordan Sandstone capped by Prairie du Chein Dolostone. The topography is primarily narrow, alluvial valleys, steep slopes, and broad ridges (Irvin et al., 1997). Most ridges and valleys have been under cultivation since the latter part of the 19th century. Side-slopes are generally forested, though some have been cleared for pasturing. The soils in this area have formed from multiple layers of aeolian loess of recent origin (Pleistocene era) deposited over ancient bedrock residuum. The soils can be considered relatively young, because most soil forming processes of surface layers have taken place in the last ten to twelve thousand years. Major soil forming processes include

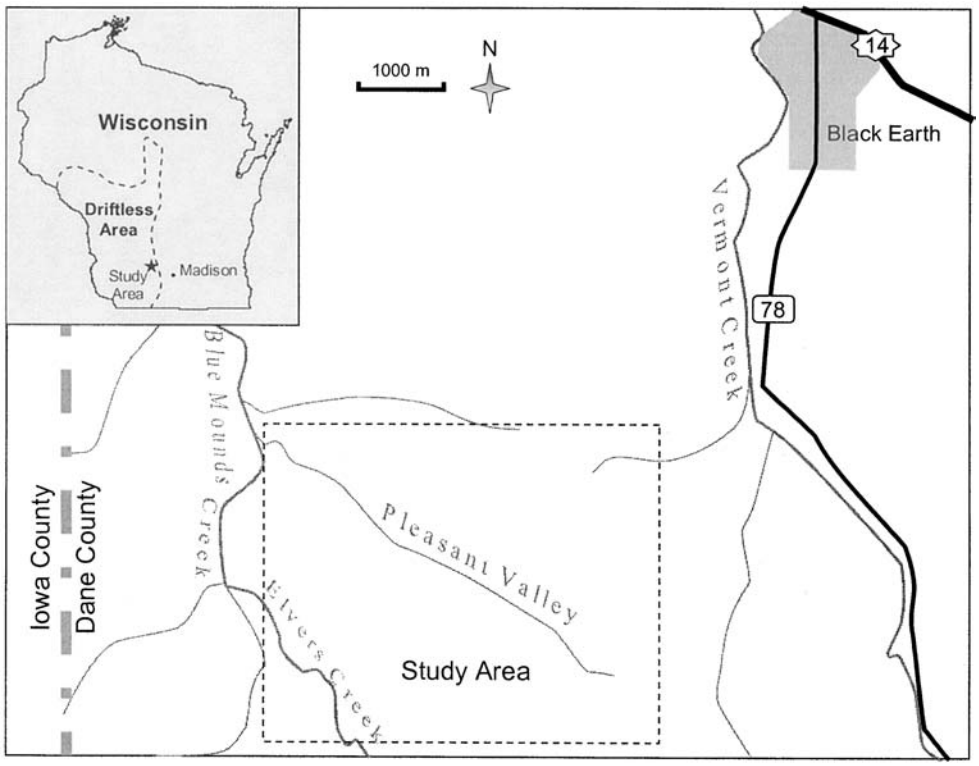


Fig. 1. Location of the study area.

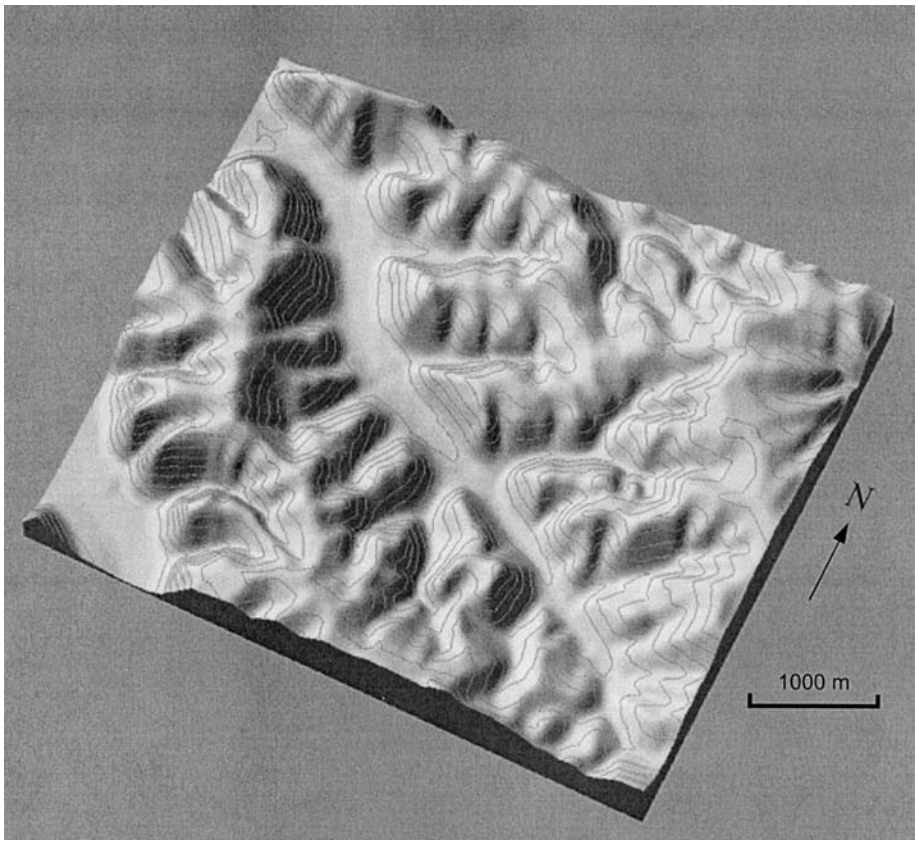


Fig. 2. Topography of the Pleasant Valley study area, Wisconsin.

eluviation, erosion, and mass-wasting (the downslope movement of rock, regolith, and soil, under the influence of gravity, see Gore, 1998). Ridges typically have a relatively thin mantle of loess with substantial residuum; the soils on the side slopes tend to be relatively thin; and valleys have thick alluvial and colluvial deposits (Knox et al., 1990; Slater and McSweeney, 1992; Clayton and Attig, 1997).

The Soil Scientist and Soil Classification Unit

A senior soil scientist from the local office of National Resources Conservation Service (NRCS) was asked to provide knowledge for this case study. This soil scientist has extensive experience of soil survey and mapping in Wisconsin.

In this research, soil series is used as the taxonomic unit for differentiating soils. It is believed that other units or differentiating criteria (e.g., argillic horizon color) can also be used in this CBR methodology without much difficulty. The soil series is chosen, because it is the classification unit used in routine soil survey and mapping projects at county level. Choosing soil series has several advantages: first, the soil scientist working for this research (as well as many other soil surveyors working on routine soil survey and mapping projects) is more familiar with soil series than other classification units; second, the methodology developed in this research can be more applicable and acceptable in routine soil mapping practices; and third, the result from this research will be more comparable with the existing soil maps.

Case-Based Reasoning Approach to Acquisition and Representation of Knowledge for Soil Mapping

With CBR, the acquisition and representation of knowledge mean creating cases. In this research, the soil scientist creates cases through a knowledge-acquisition tool called 3dMapper (Burt and Zhu, 2002). The 3dMapper is a software tool that creates three dimensional (3D) representations of topography using digital elevation model (DEM), and allows the user to drape other data layers, such as air photos, geological types, and terrain attributes (e.g., slope gradient, aspect, curvatures, etc.), over the topography, thus brings 3D views of landscapes to the user. In addition, and more importantly, the 3dMapper allows the user to do heads-up digitization on these 3D views; that is, the user can draw points, lines, and polygons over the landscape on the 3D views (Fig. 3). The main purpose of the 3dMapper is to provide a simulation of the field environment to a soil scientist, which might help him/her recall his/her tacit knowledge. Meanwhile, the heads-up digitization function of the 3dMapper provides an easy way for the soil scientist to express this knowledge.

A soil scientist can use the 3DMapper to create *tacit points* (cases). Each tacit point represents a case that contains the information from three spaces: geographical space, attribute space, and solution space. In geographical space, a tacit point corresponds to a location on the earth's surface, which can be located by its geographical coordinates. In attribute space, it corresponds to a combination of values of certain environmental variables. In solution space, it corresponds to a certain soil or a grade of similarity to the given soil. In this research, the soil scientist is asked to give only the most typical cases for the soils found in the mapping area. In other words, each tacit point should represent only one soil, thus reducing the subjectivity in case generation.

There is no predefined standard on the number of tacit points for a mapping area. Theoretically, each tacit point should represent a unique association between an environmental configuration and a soil. A complete casebase (i.e., collection of tacit points) for a mapping area should exhaust this kind of association in that area. Technically, in a case-based inference each tacit point is used individually and there is no statistical significance to achieve. In practice, how many tacit points are needed for an area will be determined by the soil scientist based on his/her understanding of the soil–environment relationships in that area.

Case-Based Reasoning Approach to Soil Inference

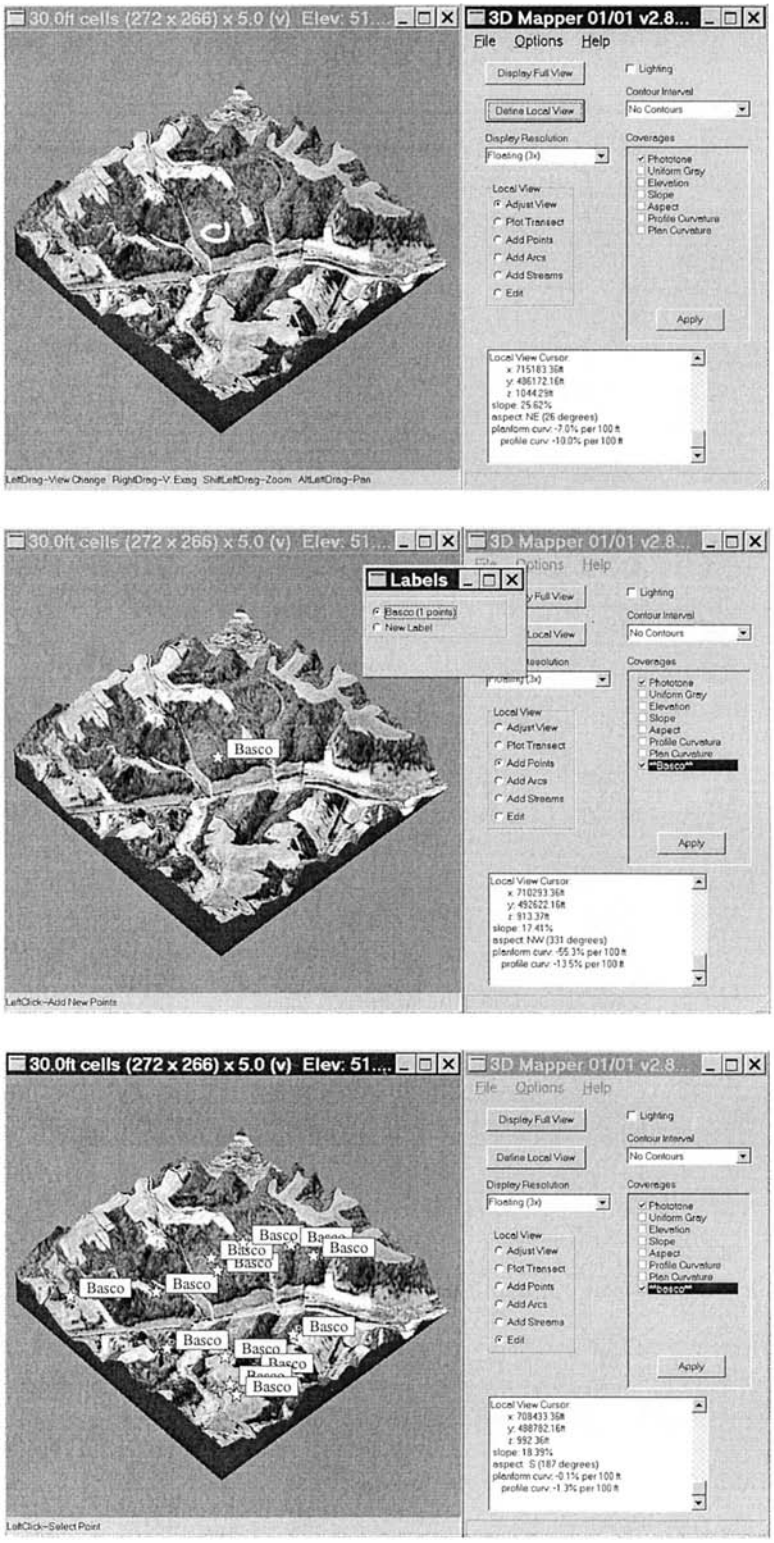
The goal of soil inference under fuzzy logic is to derive, for every location in the mapping area, the fuzzy membership values of all the soils found in the area. With the CBR method, these fuzzy membership values will be computed based on the similarity between the environmental configuration of the given location and that of each tacit point. The technical details of computing the fuzzy membership value for a certain soil at a specific location can be represented with a generic equation:

$$s_{ij}^{k} = \overset{n}{\underset{t=1}{T_{ij}^{k}}} \{ \overset{m}{\underset{v=1}{P_{ij}^{t}}} [E_{ij}^{v,t} (e_{ij}^{v}, e^{v,t})] \} \qquad [2]$$

where s_{ij}^{k} is the fuzzy membership value at location (i, j) for soil k; m is the number of environmental variables taken into account, and n is the number of tacit points for soil k; e_{ij}^{v} is the value of the vth environmental variable at location (i, j), and $e^{v,t}$ is the value of the vth variable at the tth tacit point for soil k; E is the function for evaluating the similarity on the vth variable, and this function can be specific for variable v, tacit point t, and location (i, j); P is the function for evaluating the similarity at the case level (based on all the environmental variables, that is, the configuration of environmental conditions), and can be specific for tacit point t and location (i, j); and T is the function for deriving the final fuzzy membership value based on the similarities between site (i, j) and all the tacit points for soil k, and can be specific for soil k and location (i, j).

There can be various choices for functions T, P, and E in Eq. [2]. In this research, the maximum operator is used for function T, which is the simplest possible form for T under the nearest neighbor principle. Among the similarity values from all the tacit points for soil k, the maximum operator selects the highest one as the fuzzy membership value for soil k at the given location. For function P, the minimum operator is used. This follows Zhu and Band (1994) and is based on the limiting factor principle in ecology. The limiting factor principle assumes that the limiting factor controls the development of soil formation, thus no additional information about the relative importance of each factor at a local point is needed. While the limiting factor method is probably the easiest and simplest choice for function P, more research, nevertheless, is needed to find out the most reasonable way to integrate the influences of different environmental variables on soil formation. The choice for function E should be based on the data type of the environmental variable. For a variable whose values are categorical, Boolean operators can be used. For the variables whose values are continuous, the soil scientist can choose from the models discussed by Burrough et al. (1992) and MacMillan et al. (2000).

The environmental variables used in this research for soil inference include parent material (from geological data), elevation, slope gradient, surface curvatures (profile and planform curvatures) (Zevenbergen and Thorne, 1987), and wetness index (Beven and Kirkby, 1979). The selection of these variables was based on the knowledge of the local soil scientist and all variables are treated equally. Although attempts have been made to assign realistic weights to different variables, the soil scientist working for this project found it difficult to quantify the importance of each variable in this area.



a. "I know on that spur the soil series is Basco." – An NRCS soil scientist.

b. The soil scientist pinpoints a labeled point (tacit point or case) on that spur.

c. In this way, the soil scientist creates tacit points (cases) for soil series Basco.

Fig. 3. Procedure of creating tacit points (cases) using 3dMapper.

With the tacit points and the environmental data, the CBR inference engine produces a fuzzy membership map for each soil series found in the study area. The soil scientist can examine these fuzzy membership maps to see if they match what he/she expected for the area. If problems are found, the soil scientist goes back to adjust the tacit points, including moving or removing existing tacit points, or adding new tacit points, and run the inference engine again. This process is repeated until the soil scientist is satisfied with the result.

Validating Methods

In this research, data of 91 field points in the study area are used to validate the final soil maps. These field data have not been used to adjust tacit points. Of these 91 points, all

were assigned soil series names by a group of soil scientists from NRCS local offices; 59 were given complete profile descriptions; and 44 were given a texture analysis (the percentages of sand and silt in *A* horizon).

One method to check the capability of the CBR approach in capturing the major pattern of soil distribution is to compare the fuzzy memberships derived by the inference engine for the sample points and those given by the soil scientists. Since the fuzzy membership values given by soil scientists can be very subjective, in this research the soil scientists were asked to simply name the soil at each sample point as what they would do in a conventional soil mapping process. These soil series names are referred to as *observed names* herein. To get the soil series names for the sample points from the CBR result, a "hardening" method is used (Zhu, 1997a), that is, the soil series with the highest fuzzy membership value at a sample point is used as the soil series for that sample point. These names are referred to as *inferred names* herein. Meanwhile, we also compared the observed names and the names given by the published soil survey map at the sample points (referred to as *mapped names* herein). We understand that soil maps display map units, which are spatial units but not soil classification units. Fortunately, in our study area all the map units are single-type units, that is, the soil in one unit belongs to only one soil series. Thus we are able to read soil series from the soil map. One problem in the comparison is that some of the soil series names used in the soil map are no longer in use due to the dated nature of the soil map. As a result, we were only able to use the sites whose soil series names are still in use to do the comparison. There are a total of 57 of these sites.

Testing the capability of the CBR approach in representing the continuity of soils needs soil properties whose values are continuous. In this study, the depth to C horizon and the texture of A horizon are used for this purpose. Profile depth can be an indication of the degree of soil formation and development. However, the depths of C and horizons below can be highly variable and the descriptions of these horizons can be highly subjective. Therefore, we choose to use only the depth to the top of C horizon. The typical value of the depth to C horizon for each soil series that appears in our study area are taken from the Official Soil Descriptions (OSD) (Soil Survey Division, NRCS, USDA, available online at http://ortho.ftw.nrcs.usda.gov/cgi-bin/osd/osdname.cgi, verified 19 Jan. 2004). The data in the National Soil Information System (NASIS) (Information Technology Center, NRCS, USDA) are not used because the information in NASIS is incomplete for the study area at the time of writing. A map of depth to C horizon based on the CBR results for the study area is derived using the formula below (Zhu and Band, 1994):

$$D_{ij} = \frac{\sum_{k=1}^{n} s_{ij}^{k} d^{k}}{\sum_{k=1}^{n} s_{ij}^{k}} \quad [3]$$

where D_{ij} is the depth value at site (i, j), s_{ij}^{k} is the fuzzy membership value of soil series k at site (i, j), d^{k} is the typical depth value of soil series k, and n is the total number of soil series prescribed in the soil-landscape model used by the inference engine. Meanwhile, a depth map based on the published soil survey map is generated by assigning each pixel the typical depth value of the soil series as which the pixel is labeled in the soil survey map. The two maps are used to compare the spatial patterns of depth to C horizon derived from different sources. Field observations of depth to C horizon are used to compare the accuracies of the above two maps. Respective maps of percentage of sand and silt are created in the same way as described above. Laboratory results of percentages of sand and silt at the 44 field sites are used to examine the accuracies of these maps.

RESULTS AND DISCUSSION

Of the 57 sites used for conducting the comparison on soil series name, the inferred names match the observed names at 49 sites (about 86%), and the published soil survey map matches the observed names at 26 sites (46%). Among the 57 sites, there are 32 sites for which the inference engine and the soil survey map give different soil series names. Among these 32 sites, the inferred names match the observed names at 25 sites (78%) while the mapped names match the observed names at only two sites (6%).

Figure 4 shows the maps of the depth to C horizon created based on the CBR result and the soil survey

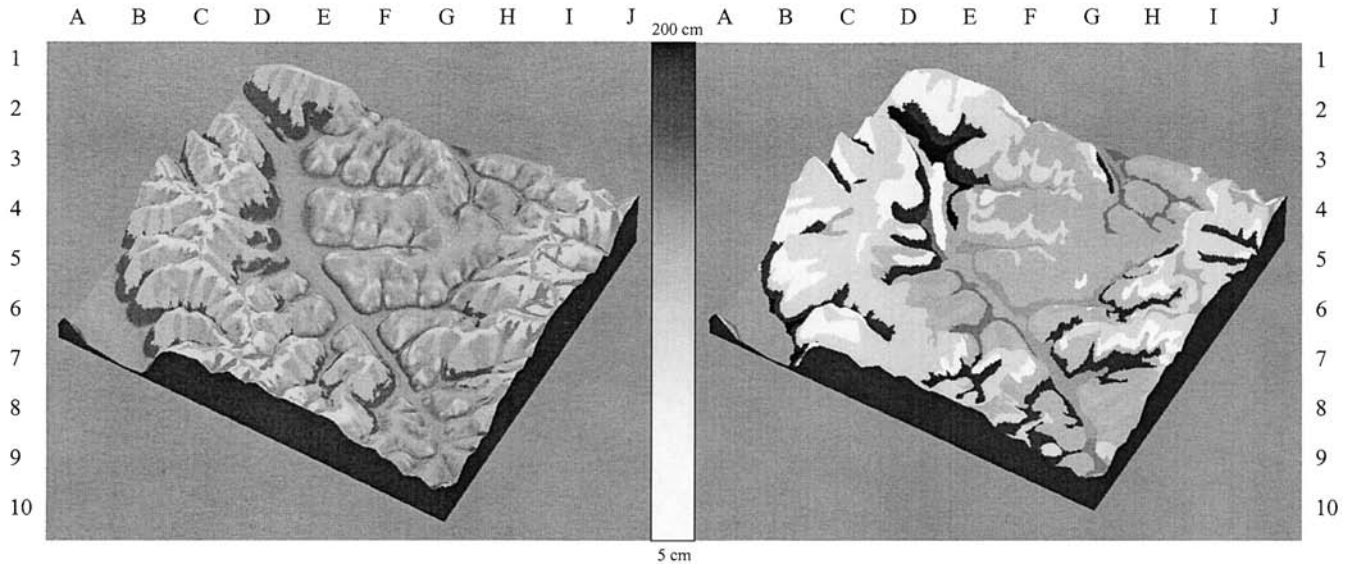


Fig. 4. (left) Map of depth to C horizon of the Pleasant Valley study area, based on the CBR result. (right) Map of depth to C horizon of the Pleasant Valley study area, based on the soil survey map.

map, respectively. While the major spatial patterns displayed by the two maps are similar, the difference is also apparent. The most obvious difference is that the map based on the CBR result (Fig. 4 [left]) shows the spatial variation in a more continuous way. Also in this map, the change of depth to C horizon from deeper on the relative flat ridge area to shallower on the steep back slope to deeper again on the foot slope and toe slope follows the topography better. This is particularly clear in Area F6. Over this area while the depth based on the CBR result follows topography well, in the map based on the survey map, relatively high values on the flat ridge expand and cover the whole back slope. In the area from C2 to D7, the change of depth from the narrow ridge to the shoulder slope is clear on the map based on the CBR result, but the map based on the survey map provides very little information on this. Another expected pattern, that the depth value is higher in a convergent area due to material accumulation and lower in a divergent area due to erosion, is also well represented in the map based on the CBR result, but is almost not recognized in the map based on the survey map (e.g., in Areas F4 and F6).

Depths to C horizon at the 59 field sites were read from the two maps and are compared with the observed depth values. Scatter plots are created to illustrate how well the inferred values and the mapped values match the observed values (Fig. 5). Figure 5 (left), which is the scatter plot for the inferred values, clearly shows the tendency of the inferred values to follow the observed values. In Fig. 5 (right), which is the scatter plot for the mapped values, the tendency is very weak. The mean absolute error (MAE) and the root mean square error (RMSE) of the inferred values against the observed values are 22.7 and 32.3, respectively, while their counterparts of the mapped values are 27.6 and 42.7, respectively.

Figures 6 and 7 are maps of *A* horizon texture. The maps based on the CBR result again have advantages in providing detailed information about spatial variation and representing realistic spatial patterns of the soils in the study area. In the areas from F4 to F6 and E6, an expected pattern is that the sand percentage is relatively low on the flat ridge due to the preservation of finer materials, and relatively high on the back slope due to the erosion of finer materials, while the silt percentage has a reverse pattern. The maps based on the CBR result clearly show these patterns, but the map based on the survey map again mixes the ridge area and the middle slope area. There also should be difference between the texture patterns of convergent areas and those of divergent areas: In a convergent area, due to the accumulation of fine materials, the sand percentage should be relatively low and the silt percentage should be relatively high; In a divergent area, the patterns should be reversed. The maps based on the CBR result again perform better than the maps based on the survey result in representing this pattern (e.g., the area from E3 to F3). Statistics are calculated for the 44 sample points whose actual texture values are available (Table 1).

When tracing the sources of the errors in the CBR results, besides the subjectivity of the soil scientists, we took into account two important factors. First, at this time the inference engine can only compute the similar-

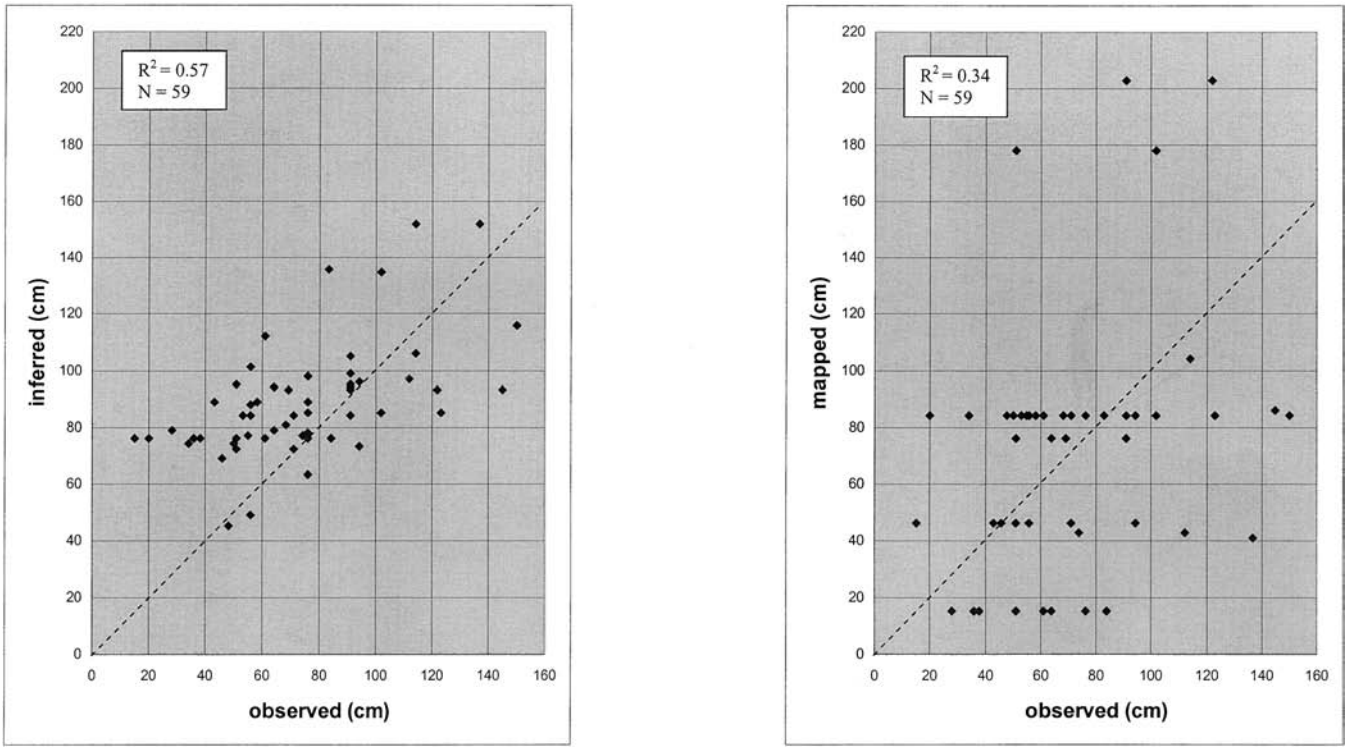


Fig. 5. (a) Scatter plot of observed depths to C horizon vs. the depth to C horizon derived from the case-based reasoning result at 59 sample locations in the Pleasant Valley study area. (b) Scatter plot of observed depth to C horizon vs. the depth to C horizon derived from the soil survey map at 59 sample locations in the Pleasant Valley study area.

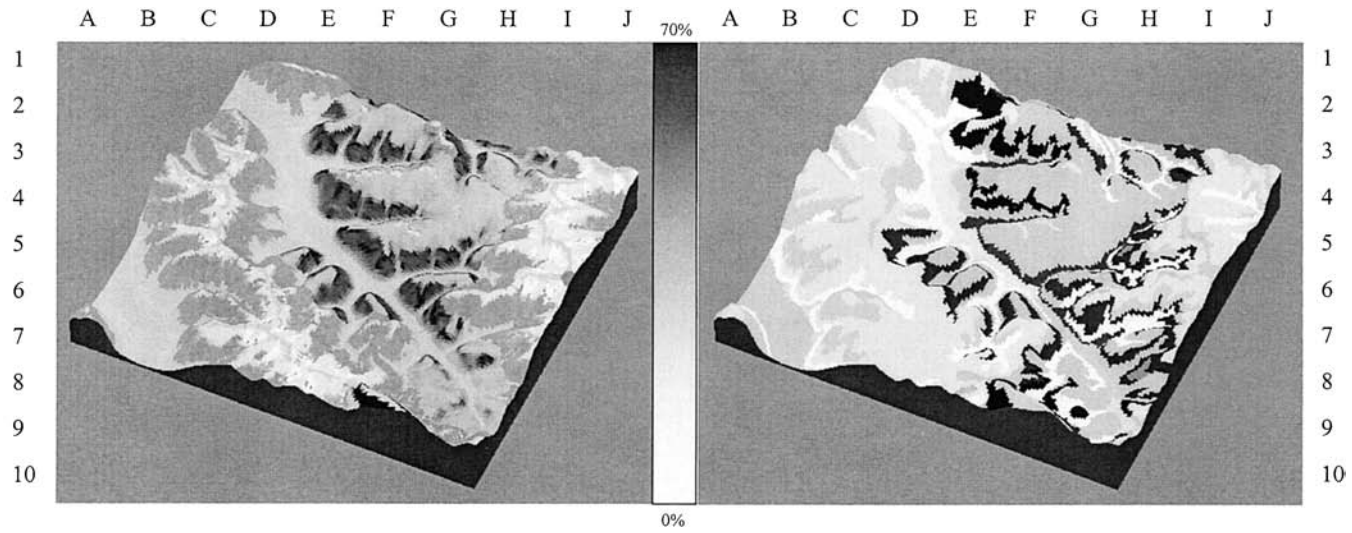


Fig. 6. (left) Sand percentage in *A* horizon derived from the case-based reasoning result. (right) Sand percentage in *A* horizon derived from the soil survey map.

ity between a given location and a tacit point in the attribute space. No similarity in the geographic space, that is, spatial similarity (Holt and Benwell, 1999), has been considered. The lack of spatial consideration may lead to incomplete characterization of the tacit points (cases), which consequently, may cause error in the inference. Second, some of soil scientists' knowledge has not been well utilized in the current inference process. Particularly, some critical information used by soil scientists in modeling the soil–environment relationship, such as the information about catena, the information about slope positions, and the information about some special terrain features, is still not available under current spatial analysis techniques. These problems indicate potential research directions.

CONCLUSIONS

In the case study in the Pleasant Valley area, the comparisons on the soil series names, on the depth to C horizon, and on the texture of *A* horizon consistently show that the inference result from the CBR process is more accurate than the published soil survey map. This case study demonstrates that CBR can be an effective approach to knowledge acquisition, knowledge representation, and soil inference for soil mapping under fuzzy logic. Apparently, the CBR method inherits some advantages from previous computerized knowledge-based fuzzy mapping methods: A computerized mapping process has a much higher efficiency than that of a manual process; The computerized approach can maintain a high consistency during the whole mapping process; Soil scientists' knowledge can be stored and accumulated in a computerized knowledgebase; the fuzzy representation scheme can represent and present accurate and precise information. Meanwhile, this research reveals some unique advantages of the CBR method. When a friendly and appropriate interface (like 3dMapper) is provided, the CBR approach allows soil

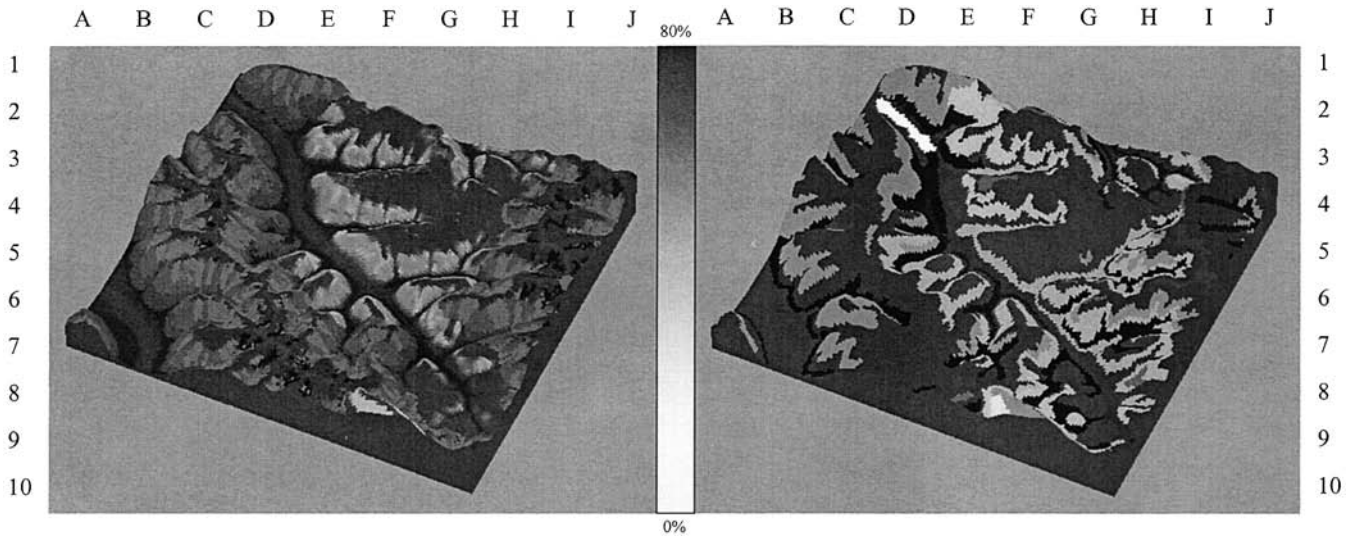


Fig. 7. (left) Silt percentage in *A* horizon derived from the case-based reasoning result. (right) Silt percentage in *A* horizon derived from the soil survey map.

Table 1. Accuracy of the derived texture of *A* horizon in the Pleasant Valley study area: The case-based reasoning (CBR) result vs. the soil survey map. Mean average error (MAE) and root mean square error (RMSE) are calculated using derived values (based on the CBR result and the soil survey map, respectively) as estimates and using the lab analysis result as true values.†

	Percentage of sand		Percentage of silt	
	MAE	RMSE	MAE	RMSE
Inference result	14.6	20.3	18.6	23.6
Soil survey map	17.3	24.9	20.2	25.9

† Number of samples: 44.

scientists easily to express their tacit knowledge by pinpointing specific locations in geographic space, without having to make efforts to generalize the knowledge into rules in attribute space. Pinpointing tacit points can be easier than providing general rules, because the tacit knowledge, which is learned by investigating soils at specific locations during fieldwork, is likely to be in the form of cases in a soil scientist's mind; also, pinpointing tacit points allows the soil scientist to avoid depicting the details of the soil–environmental relationship in attribute space and assuming the "variable independence." This feature, combining with the similarity-based inference strategy, allows soil scientists easily to adjust the representation of their knowledge and control the inference results. The soil scientist can easily find out which tacit point is controlling which part of the area and can easily adjust the tacit point to modify the inference result. Finally, the tacit points are technically independent from each other in the inference process, thus adding and removing any tacit points would not impact the other tacit points and the inference setting (i.e., no retraining is needed). This makes knowledge accumulation and update easy.

Although this research focuses on soil scientists' specific knowledge (i.e., knowledge of the association between a certain soil and a specific landscape), it has been found that the knowledge that can be provided by soil scientists is not always of one single type. The knowledge can exist in both specific (as cases) and general (as rules) forms. Proper utilization of general knowledge can improve the efficiency of knowledge acquisition and soil inference. Therefore, to fully extract and utilize soil scientists' knowledge and to improve the performance of the soil inference, research is needed to develop a more versatile system that contains knowledge acquisition tools and inference engine capable of accommodating and taking advantages of various types of knowledge.

ACKNOWLEDGMENTS

Support from the Walter and Constance Burke Award, Dartmouth College, is gratefully acknowledged. Support from the Graduate School, University of Wisconsin-Madison and from Natural Resources Conservation Service (NRCS), USDA under Agreement No. 69-5F48-9-00186 is gratefully acknowledged. Support from the one hundred talents Program of Chinese Academy of Sciences is greatly appreciated.

REFERENCES

Aamodt, A., and E. Plaza. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications 7:39–52.

Beven, K.J., and M.J. Kirkby. 1979. A physically-based, variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24:43–69.

Bie, W.W., and P.H.T. Beckett. 1973. Comparison of four independent soil surveys by air photo interpretation, Paphos area (Cyprus). Photogrammetria 29:189–202.

Burt, J.E., and A.X. Zhu. 2002. 3dMapper 2.11. Dep. of Geography, Univ. of Wisconsin-Madison, Madison.

Burrough, P.A. 1989. Fuzzy mathematical methods for soil survey and land evaluation. J. Soil Sci. 40:477–492.

Burrough, P.A., R.A. MacMillan, and W. van Deursen. 1992. Fuzzy classification methods for determining land suitability from soil profile observations and topography. J. Soil Sci. 43:193–210.

Burrough, P.A., P.F.M. Van Gaans, and R. Hootsmans. 1997. Continuous classification in soil survey: Spatial correlation, confusion and boundaries. Geoderma 77:115–135.

Clayton, L., and J.W. Attig. 1997. Pleistocene geology of Dane County, Wisconsin. Wisconsin Geological and Natural History Surv. Bull. 95. Wisconsin Geological and Natural History Survey, Madison, WI.

Cook, S.E., R.J. Corner, G. Grealish, P.E. Gessler, and C.J. Chartres. 1996. A rule-based system to map soil properties. Soil Sci. Soc. Am. J. 60:1893–1900.

De Bruin, S., and A. Stein. 1998. Soil-landscape modelling using fuzzy c-means clustering of attribute data derived from a digital elevation model (DEM). Geoderma 83:17–33.

Galbraith, J.M., and R.B. Bryant. 1998. A functional analysis of soil taxonomy in relation to expert system techniques. Soil Sci. 163: 739–747.

Galbraith, J.M., R.B. Bryant, and R.J. Ahrens. 1998. An expert system for soil taxonomy. Soil Sci. 163:748–758.

Gore, P.J.W. 1998. Mass Wasting. Available at http://www.gpc.peachnet.edu/~pgore/geology/geo101/masswasting.html (accessed 15 Oct. 2002; verified 19 Jan. 2004).

Hole, F.D., and J.B. Campbell. 1985. Soil landscape analysis. Rowman & Allanheld, Totowa, NJ.

Holt, A., and G.L. Benwell. 1999. Applying case-based reasoning techniques in GIS. Int. J. Geographical Information Sci. 13:9–25.

Hudson, B.D. 1992. The soil survey as paradigm-based science. Soil Sci. Soc. Am. J. 56:836–841.

Information Technology Center, NRCS, and USDA. 2001. National Soil Information System (NASIS) [Online]. Available at http://nasis.nrcs.usda.gov (accessed 5 July 2001; verified 19 Jan. 2004). USDA-NRCS, Washington, DC.

Irvin, B.J., S.J. Ventura, and B.K. Slater. 1997. Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. Geoderma 77:137–154.

Jenny, H. 1941. Factors of soil formation: A system of quantitative pedology. McGraw-Hill, New York.

Jenny, H. 1980. The soil resource: Origin and behavior. Springer-Verlag, New York.

Knox, J.C., D.S. Leigh, and T.A. Frolking. 1990. Roundtree formation (new). p. 64–67. *In* L. Clayton and J.W. Attig (ed.) Geology of Sauk County, Wisconsin. Wisconsin Geol. and Natural History Surv., Madison, WI.

Kolodner, J. 1993. Case-based reasoning. Morgan Kaufmann Publ., San Mateo, CA.

Leake, D.B. 1996. CBR in context: The present and future. p. 3–30. *In* D.B. Leake (ed.) Case-based reasoning: Experiences, lessons, and future directions. MIT Press, Cambridge, MA.

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

Mays, M.D., I. Bogardi, and A. Bardossy. 1997. Fuzzy logic and risk-based soil interpolations. Geoderma 77:299–315.

McBratney, A.B., and J.J. De Gruijter. 1992. A continuum approach to soil classification by modified fuzzy *k*-means with extragrades. J. Soil Sci. 43:159–175.

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

McKenzie, N.J., P.E. Gessler, P.J. Ryan, and D.A. O'Connell. 2000. The role of terrain analysis in soil mapping. p. 245–265. *In* J.P. Wilson and J.C. Gallant (ed.) Terrain analysis: Principles and applications. John Wiley & Son, New York.

Schank, R. 1982. Dynamic memory: A theory of reminding and learning in computers and people. Cambridge University Press, Cambridge.

Shi, X., and A.G.O. Yeh. 1999. The integration of case-based systems and GIS in development control. Environ. Plan. B: Plan. Design 26:345–364.

Skidmore, A.K., P.J. Ryan, W. Dawes, D. Short, and E. O'loughlin. 1991. Use of an expert system to map forest soils from a geographical information system. Int. J. Geographical Information Syst. 5: 431–445.

Slater, B.K., and K. McSweeney. 1992. Modeling soil horizon stratigraphy, Pleasant Valley, Wisconsin. *In* E. Nater (ed.), Soils-geomorphology: Pre-conference tour guidebook for the SSSA Annual Meeting, Mineapolis, Minnesota, 1–6 Nov. 1992. ASA, Madison, WI.

Soil Survey Division, NRCS, and USDA. 2001. Official Soil Series Descriptions [Online]. Available at http://ortho.ftw.nrcs.usda.gov/osd/(accessed 12 May 2001; verified 19 Jan. 2004). USDA-NRCS, Washington, DC.

Watson, I. 1997. Applying case-based reasoning: Techniques for enterprise systems. Morgan Kaufman Publ., San Mateo, CA.

Yeh, A.G.O., and X. Shi. 1999. Applying case-based reasoning to urban planning: A new planning support system tool. Environ. Plan. B: Plan. Design 26:101–116.

Zevenbergen, L.W., and C.R. Thorne. 1987. Quantitative analysis of land surface topography. Earth Surf. Processes 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. Photogramm. Eng. Remote Sens. 63:1195–1202.

Zhu, A.X. 1999. A personal constructed-based knowledge acquisition process for natural resource mapping using GIS. Int. J. Geographic Information Syst. 13:119–141.

Zhu, A.X., and L.E. Band. 1994. A knowledge-based approach to data integration for soil mapping. Can. J. Remote Sens. 20:408–418.

Zhu, A.X., L.E. Band, B. Dutton, and T.J. Nimlos. 1996. Automated soil inference under fuzzy logic. Ecol. Modell. 90:123–145.

Zhu, A.X., L.E. Band, R. Vertessy, and B. Dutton. 1997. Derivation of soil properties using a Soil Land Inference Model (SoLIM). Soil Sci. Soc. Am. J. 61:523–533.

Zhu, A.X., B. Hudson, J. Burt, K. Lubich, and D. Simonson. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. Am. J. 65:1463–1472.

SSSAJ

Soil Science Society of America Journal

Vol. 68 No. 3 May–June 2004

SSSJD4 68(3):713–1130 (2004)