Integrating Different Types of Knowledge for Digital Soil Mapping

**X. Shi***
Department of Geography
Dartmouth College
6017 Fairchild
Hanover, NH 03755

**R. Long**
R. Dekett
J. Philippe
USDA-NRCS
481 Summer Street
Suite 202
St. Johnsbury, VT 05819

Digital soil mapping has undergone a rapid development in the past decade (McBratney et al., 2003; Grunwald, 2006; Lagacherie and McBratney, 2007). Generally there are two approaches being taken in DSM research and practice. One aims at truly automatic, objective, and quantitative mapping, taking advantage of the techniques in statistics, geostatistics, machine learning, and data mining, and generally relying heavily on dense sampling from either field or existing soil maps. McBratney et al. (2003) provides a comprehensive review of this approach. Some researchers taking this approach have also challenged the traditional class mapping paradigm and proposed to directly map soil layers and properties (e.g., McSweeney et al., 1994; Gessler et al., 1995). The other approach tries to fit within the conventional soil survey and mapping framework, including the conventional process and standard. It aims to effectively utilize the soil scientists' knowledge, while reducing the inconsistency and cost associated with the traditional manual process (Zhu et al., 2001; Shi et al., 2004). Its major digital components include the knowledge engineering techniques for knowledge acquisition, knowledge representation, and knowledge-based inference. While the two approaches are in no sense mutually exclusive (Grunwald, 2006; Walter, Lagacherie, and Follain 2007), the differences in philosophy and technical emphasis between them may lead to different plans and strategies for implementing DSM.

This paper presents an example of the knowledge-based approach. Other authors have argued that making good uses of soil scientists' tacit knowledge is efficient and economical (e.g., Hudson, 1992; Hengl and MacMillan, 2007; MacMillan et al., 2007). Here we further articulate two advantages of this approach. First, it does not rely on dense field sampling. The cost of field sampling is always a concern in small-scale soil survey and mapping conducted by a public agency like the USDA-Natural Resources Conservation Service (NRCS). Meanwhile, there is little doubt that the samples needed by an experienced soil scientist are considerably fewer than what are required by a computationally inductive method (e.g., regression, decision tree, and neural network). This might be the primary reason for that the NRCS soil survey and mapping is still largely a human—rather than computer-based practice. Second, the knowledge-based approach might be more acceptable to soil surveyors, who will be the users of DSM tools under the current work scheme of NRCS. Researches have found that users' acceptance is vital in deploying automating systems (Leake, 1996; Yeh and Shi, 1999). There are two issues under this concern. First, the user needs to trust the system and this trust usually is based on his/her understanding of the system. Second, the user often wants to be involved in the problem-solving process rather than standing outside the process. Both issues are particularly significant in a domain like soil survey and mapping that has been traditionally centered round expert knowledge and can be characterized as a combination of science and art (Hudson, 1992; McKenzie et al., 2000; Zhu et al., 2001). The two advantages of the knowledge-based approach justify its...
applications, especially in the USA and other developed countries where more soil scientists and associated soil knowledge are available (Hartermink and Nachtergaele, 2006).

Therefore, there should be demand for DSM tools that implement the knowledge-based approach. The developers of such DSM tools must consider the characteristics of soil scientists' knowledge so as to develop the tools accordingly. In this paper, we analyze soil scientists’ knowledge from the perspectives of scale and space. The scale refers to the geographical coverage of the knowledge. Some knowledge is global and covers the entire mapping area, and some is local and only covers limited regions in the mapping area. The space refers to how the knowledge is represented, as the knowledge can be represented by environmental values in parametrical space, or by locations in geographical space. Knowledge in parametrical space is likely to be global, while knowledge in geographical space is more often to be local. Table 1 lists examples of the knowledge in different categories.

The scale aspect of knowledge is particularly important in DSM, since the local knowledge can be used to identify and deal with local exceptions. Sometimes the soil scientist approves of the general pattern in the map created by a DSM tool using the global knowledge, but finds that the results at certain locations differ from his/her expectations. These local exceptions may arise from two sources. First, due to the limitation of the knowledge engineering techniques, the soil-landscape model conveyed from the soil scientist to the computer may not precisely represent the soil scientist’s actual knowledge. For example, with the rule-based knowledge acquisition process, the soil scientist may find it difficult to express the knowledge about situations that deviate from the general pattern, and the resulting soil-landscape model may therefore be too general to address some local specifics. The other source of local exceptions is problems in the data, which can be generally classified as insufficiency and incorrectness. Data insufficiency means that the data available to the computer cannot fully or correctly represent the environmental factors the soil scientist uses to characterize the soil formative environment. This may be due to the lack of data resources, e.g., a detailed geological map is not available for identifying a special parent material; or due to the limitation of the current analytical techniques, for example, the current techniques of terrain analysis cannot delineate some special terrain positions. Data incorrectness, on the other hand, may include value error, inappropriate calculation scale, and inappropriate data resolution. Specific examples of incorrectness include: errors in the Digital Elevation Model (DEM) cause incorrect values of terrain attributes; the neighborhood size used by the soil scientist in the field for measuring terrain attributes (e.g., slope) is different from that used by the computer on the DEM (Shi et al., 2007); and inappropriate resolutions smooth out important features or exaggerate insignificant details. A DSM tool should provide functions to address local exceptions resulting from all these sources.

In the space aspect, the necessity for explicitly distinguishing parametrical and geographical spaces lies in the fact that some knowledge can be more precisely, or at least more conveniently, represented in one or the other space. For example, the best way to express the knowledge that the optimal slope gradient for soil X is 8 to 20% is simply specifying these two numbers. However, when a detailed geological map is not available, drawing polygons seems to be the only way to express the knowledge that a soil is controlled by a local parent material. The knowledge engineering techniques for acquiring, representing, and using these two types of knowledge are very different. Representing knowledge in parametrical space usually does not require geographical specification. The inference using this type of knowledge is usually a pixel-by-pixel operation, which is straightforward and efficient. Representing and using knowledge in geographical space, on the other hand, requires sophisticated GIS operations.

Other authors have presented similar analyses of soil scientists’ knowledge. In Bui’s (2004) module of representation in a structured environment (MORSE) framework, the first-level default rules well match the global knowledge in this paper, and the lower-level more particular rules appear to correspond to the local knowledge. The prototype discussed by Qi et al. (2006) is essentially the global knowledge; and in parametrical space, the prototype is not substantially different from an earlier concept, instance (Zhu et al., 1996). Hengl and MacMillan (2007) and MacMillan et al. (2007) recognize that soil scientists’ tacit knowledge consists of at least two separate components: conceptual knowledge, which corresponds to the global knowledge in parametrical space, and geographic knowledge, which corresponds to the local knowledge in geographical space.

In terms of knowledge engineering techniques, RBR and CBR have been used to address the scale and space features of soil scientists’ knowledge (Skidmore et al., 1991; Cook et al., 1996; Galbraith et al., 1998; Holt and Benwell, 1999; Zhu et al., 1996; Zhu, 1999; MacMillan et al., 2000; Shi et al., 2004). Rule-based reasoning uses if-then rules. For example, the soil scientist may specify that if the slope gradient is 8 to 20%, then the soil is X. Based on this rule, RBR identifies all the locations in the mapping area whose slope gradients fall into that range and labels the soils at those locations to be X. This example illustrates that RBR is good at handling the global knowledge in parametrical space. Case-based reasoning, on the other hand, can be used to handle knowledge in geographical space. With CBR, instead of specifying the optimal environmental values for the given soil, the soil scientist pinpoints or delineates typical locations or regions for the soil. It is then the task of the computer to extract the optimal values from the environmental data, based on those locations or

<table>
<thead>
<tr>
<th>Space</th>
<th>Parametrical</th>
<th>Geographical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>Representation: Rules</td>
<td>Representation: Global cases</td>
</tr>
<tr>
<td></td>
<td>Inference: Global RBR Example: “The optimal slope gradient for soil X in this mapping area is between 8% and 20%.”</td>
<td>Inference: Global CBR Example: “All the locations in this mapping area whose environmental conditions are similar to that of this location are likely to have soil X.”</td>
</tr>
<tr>
<td>Local</td>
<td>Inference: Regional RBR Example: “In areas of geological type A, the optimal slope gradient for soil X is between 8% and 20%.”</td>
<td>Inference: Local cases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inference: Local CBR Example: “The vicinity of this location is likely to have soil X.”</td>
</tr>
</tbody>
</table>
regions. In other words, in DSM, CBR translates the knowledge from geographical space to parametrical space. Cases can be used to represent the global knowledge, that is, the soil scientist can specify that all locations in the mapping area that are similar, in terms of environmental condition, to the case’s location have soils similar to the case’s soil (Shi et al., 2004). Case-based reasoning can also be local, which restricts the influence of a case to the case’s vicinity. In this paper, the former is referred to as global CBR and the latter, local CBR.

Both RBR and CBR have advantages and limitations. Acquiring rules that represent the global knowledge in parametrical space seems straightforward to soil scientists, since in traditional mapping a soil-landscape model is usually formalized as a set of environmental values and the soil scientist would naturally try to make the values as global as possible. Additionally, compared with CBR, RBR is more computationally efficient, since the number of the rules is usually small and the calculation is usually a simple pixel-by-pixel process. However, the soil scientist may not always know precise optimal values at the beginning of a project, especially for a new mapping area. Also, rules can be too general to address some exceptional locations. On the other hand, the soil scientist may know the locations of a soil, even before he/she is able to articulate the relationship between the soil and environmental values. Directly acquiring and using this location-based knowledge may save the soil scientist’s effort in translating the knowledge from geographical space to parametrical space. Furthermore, CBR is good at handling local exceptions since it can restrict the inference within the cases’ vicinities.

Rule-based reasoning and CBR have been used separately in DSM. To fully take advantage of different types of knowledge and different inference technologies, we propose here an integrated process, in which the soil scientist first performs RBR or global CBR using the global knowledge to create draft maps, and then, if necessary, performs local CBR using the local knowledge to fine-tune the draft maps. This process seeks to achieve both high efficiency (through the use of the global knowledge) and high accuracy (through the use of local knowledge). We introduce a software package named Soil Inference Engine (SIE) that implements this integrated process. A case study in northern Vermont is presented.

MATERIALS AND METHODS

Study Area

Our study area, the La Pointe Brook watershed, lies within the town of Averill, Essex County, Vermont (Fig. 1). Its area is about 3.5 km², and the elevation ranges from 479 m at the outlet to the East Branch of the Nulhegan River, to 853 m at the summit of Sable Mountain. The watershed is entirely within the USGS Averill Lake topographic quadrangle. It was selected because its bedrock types, relief, and expected soil sequences are representative of a large area of Essex County, for which an initial soil mapping project is underway. Most of the study area is underlain by Lower Devonian-era phyllite and schist of the Gile Mountain formation. The upper elevations of Sable Mountain are underlain by granite, and are thinly covered by loamy till of Wisconsin age. Most middle and lower elevations are more deeply covered by loamy basalt till.

A few areas of very poorly drained organic materials are found on broad flats and in depressions. The vegetation consists mainly of spruce-fir forests on the mountain summit and more poorly drained lower slopes, and mixed northern-hardwood and spruce-fir forests on middle slopes. Overall the topography is characterized by a series of hills and narrow valleys interspersed with an occasional mountain. Drainage patterns are somewhat irregular and controlled by the local bedrock.

This project focuses on four major soil series which largely compose a catena formed in loamy basalt till. These soils occur in a regular pattern throughout much of northeastern Vermont on landforms that are well represented within the study area. Important physical and chemical properties of the soils, including texture, moist bulk density, organic matter content, and reaction are well understood since nearby areas have been mapped and sampled. Therefore, the knowledgebase developed in the test watershed should have applicability to the entire Essex County soil survey area.

Data

The DEM used in the case study was generated from the USGS 7.5-min topographic maps. The horizontal resolution of the DEM is 10 m. Invalid elevation anomalies, such as abrupt breaks in elevation profiles, terracing, false depressions, and nonexistent rises were identified and corrected during many quality control checks throughout the production process. An enforcing operation was employed to ensure all water flows downhill along known stream paths to avoid pooling in unwanted sinks. Four terrain attribute layers, namely slope gradient, profile curvature, planform curvature (Zevenbergen and Thorne, 1987), and wetness index (Beven and Kirkby, 1979), were derived from the DEM for formalizing the soil-landscape model in the study area. The slope gradient, profile curvature, and planform curvature layers were generated using 3dMapper (Terrain Analytics, LLC). A smoothed wetness index layer was generated using the Topocrop program (Schmidt, 2001). A 30-m neighborhood was used for all terrain attribute calculations. To use only four fairly basic terrain attributes is a decision based on numerous trials with many different possible input layers. On the one hand, the soil scientist found that the four layers worked satisfactorily for the test watershed. On the other hand, the simplicity of the environmental data facilitated the test of the proposed RBR-CBR process.
The Integrated Process

The RBR-CBR process implemented by SIE contains eight steps. Independently, but not surprisingly, MacMillan et al. (2007) employed a procedure that is almost identical to the one described here. The eight steps are as follows:

1. The soil scientist provides the global knowledge, including names of the soils he/she expects to see in the mapping area and descriptions of the environmental conditions of these soils. The environmental conditions depicted by environmental values are formalized into rules and saved into rulebases; those represented by geographical locations are formalized into cases and saved into global casebases.

2. The soil scientist or a GIS specialist prepares data layers for characterizing the environmental conditions. The data layers may cover terrain attributes, geology, vegetation, climate, and other features. These data layers are stored in a GIS database.

3. The SIE performs RBR or global CBR, using the global knowledge and the GIS database, to generate maps of the general pattern of soil distribution in the mapping area.

4. The soil scientist verifies the draft maps from Step 3. If he/she is satisfied with the maps, the mapping is done. Otherwise, he/she may go back to Step 1 to adjust the rules or global cases, or go to Step 5 to fine-tune the draft maps.

5. The soil scientist provides the local knowledge, in the form of cases, to address local exceptions. The cases are saved in local casebases.

6. The SIE performs local CBR using the local knowledge and the GIS database.

7. The soil scientist verifies the maps from Step 6. He/she can adjust the cases and run local CBR again. He/she repeats this process until he/she is satisfied with the result.

8. The soil scientist uses SIE and other GIS tools to integrate the results from Step 3 and Step 6 to generate the final maps.

The SIE provides an integrated user interface to facilitate this process. It allows easy switches between the global knowledge and the local knowledge, between RBR and CBR, and between cases in different forms, including point, line, polygon, and pixel.

Acquisition and Use of the Global Knowledge

The soil inference of SIE using the global knowledge contains three phases, represented by the three functions, \( E \), \( P \), and \( T \), in Eq. [1] (Shi et al., 2004):

\[
\hat{s}_{ij,k} = T_{k}^{\frac{m}{n}} \{ P_{c,a}^{\frac{m}{n}} (E_{c,a}(z_{ij,a}, z_{c,a})) \} \quad [1]
\]

where \( s_{ij,k} \) is the prediction value at location \((i, j)\) for soil \(k\); \( z_{ij,a} \) is the value of environmental factor \(a\) at \((i, j)\); \(z_{c,a}\) is the most optimal range given by rule/case \(c\), defining the most favoring condition of factor \(a\) for soil \(k\). In RBR, \(z_{c,a}\) is directly specified by the soil scientist, while in CBR, \(z_{c,a}\) is derived by the computer based on the case location and the environmental data layers.

Function \(E\) evaluates to what extent \(z_{ij,a}\) favors soil \(k\) and its output is herein called optimality value. If \(z_{ij,a}\) falls into the range of \(z_{c,a}\), \(E\) returns the maximum optimality value; otherwise, \(E\) uses a function to derive the optimality value based on the difference between \(z_{ij,a}\) and \(z_{c,a}\). The SIE provides four generic functions for \(E\), including nominal, ordinal, cyclic, and continuous. The nominal function outputs zero once \(z_{ij,a}\) is outside the range of \(z_{c,a}\). It can be used for mapping under crisp logic or for handling categorical data, for example, geological type and vegetation type. The ordinal function is for descriptions like “well drained,” “moderately well drained,” and “poorly drained.” The SIE handles such data using a stair algorithm, which drops the optimality value discretely as \(z_{ij,a}\) departs from \(z_{c,a}\). The cyclic function is for handling a factor like slope aspect, whose maximum value goes back to the minimum value. The continuous function uses Gaussian curves to continuously adjust the optimality value (Shi, Zhu, and Wang, 2005):

\[
\hat{s}_{ij,k,a} = e^{(\frac{(z_{ij,a} - v_{1})}{w_{1}})} & \text{if } z_{ij,a} < v_{1} \\
1 & \text{if } z_{ij,a} < v_{1} \\
\hat{s}_{ij,k,a} = e^{(\frac{(z_{ij,a} - v_{2})}{w_{2}})} & \text{if } z_{ij,a} > v_{2} \quad [2]
\]

where \(s_{ij,k,a}\) is the optimality value for environmental factor \(a\) at location \((i, j)\) for soil type \(k\); and \(e\) is the base of natural logarithms (2.71828...). The meanings of the parameters in this Eq. [2] are given in Table 2. The basic idea of Eq. [2] is to provide two sets of parameters so that the soil scientist can define an asymmetric curve to address situations \(z_{ij,a} < z_{c,a}\) and \(z_{ij,a} > z_{c,a}\) differently.

Function \(P\) in Eq. [1] integrates the optimality values from individual environmental factors to produce an overall predicted value for soil \(k\). Under the crisp logic, \(P\) outputs “yes” or “no,” while in fuzzy soil mapping (Burrough, 1989; Burrough 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; MacMillan et al., 2000), \(P\) gives a fuzzy membership. The minimum operator has been used for \(P\) (Zhu et al., 1996; Zhu, 1997; Shi et al., 2004). This operator uses the minimum among all the optimality values from individual factors as the overall predicted value. It is the simplest option for \(P\) and has the limiting-factor principle in ecology as the theoretical basis. However, it does not allow the soil scientist to express the knowledge that some environmental factors are more important than others. To address this problem,

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>(z_{ij,a})</td>
<td>The value of environmental factor (a) at location ((i, j)), read from the GIS database.</td>
</tr>
<tr>
<td>(v_{1}) and (v_{2})</td>
<td>The two user-specified central values that define the lower and upper limits of the most optimal range of environmental factor (a) for soil (k). In other words, if (z_{ij,a}) falls between (v_{1}) and (v_{2}), in terms of a location ((i, j)) has the maximum optimality value for soil (k). Graphically, (v_{1}) and (v_{2}) determine the width of the flat top of the function curve.</td>
</tr>
<tr>
<td>(w_{1}) and (c_{1})</td>
<td>The user can adjust the shape of the left half of the curve by specifying that if (z_{ij,a}) is smaller than (v_{1}) and the difference between (z_{ij,a}) and (v_{1}), the output optimality value is to be (c_{1}). To simplify the adjusting operation, SIE uses a fixed value, 0.5, for (c_{1}).</td>
</tr>
<tr>
<td>(w_{2}) and (c_{2})</td>
<td>Similar to (w_{1}) and (c_{1}), but for the situation that (z_{ij,a}) is greater than (v_{2}).</td>
</tr>
<tr>
<td>(r_{1}) and (r_{2})</td>
<td>They control the behavior of the optimality value as the values for a deviate from the most optimal range. Graphically, the higher these two values, the flatter the tops and the steeper the sides of the curves.</td>
</tr>
</tbody>
</table>
besides the minimum operator SIE provides a weighted-average operator and implements the Analytical Hierarchy Process (AHP, Saaty, 1980; Saaty and Vargas, 2001) to reduce the inconsistency in the weighting process. The SIE also provides a multiplication operator that uses the product of all the optimality values from individual factors as the overall predicted value.

It is possible for a soil to occur under different environmental conditions. For example, a soil may occur on both gentle and steep slopes, but on a steep slope the soil only occurs in a concave area. The SIE allows the soil scientist to use different sets of rules (called instances by Zhu et al., 1996) or different cases to describe those different conditions. Function $T$ in Eq. [1] integrates the predicted values from different instances or cases to obtain the final predicted value. For the global knowledge, SIE only implements the maximum operator for $T$, which uses the maximum value from individual instances or cases as the final value for a location.

The soil scientist working on this project identified four major soil series in the study area that largely compose a catena formed in loamy basal till: Dixfield (Coarse-loamy, isotic, frigid Aquic Haplorthods; moderately well drained), Colonel (Loamy, isotic, frigid, shallow Aquic Haplorthods; somewhat poorly drained), Cabot loamy, mixed, active, nonacid, frigid, shallow Histic Humaquepts; very poorly drained). Figure 2 shows an example of how the soil scientist used SIE to express his global knowledge in parametrical space about Cabot soil in the study area. Cabot occurs under two series in the study area that largely compose a catena formed in loamy basal till: Dixfield (Coarse-loamy, isotic, frigid Aquic Haplorthods; moderately well drained), Colonel (Loamy, isotic, frigid, shallow Aquic Haplorthods; somewhat poorly drained), Cabot loamy, mixed, active, nonacid, frigid, shallow Histic Humaquepts; very poorly drained). Figure 2 shows an example of how the soil scientist used SIE to express his global knowledge in parametrical space about Cabot soil in the study area. Cabot occurs under two different environmental conditions: The majority of Cabot occurs on gentle slopes (<8%), while some also occurs on poorly drained steeper slopes. The soil scientist created two instances (Major Cabot and Steeper Cabot) to represent these two conditions. He loaded data for three environmental factors, slope, planform curvature, and smoothed wetness index to characterize Cabot's formative environment, but for Major Cabot, he actually used only slope and wetness (Planform curvature is checked off in the factor list). Figure 2 shows the rule (i.e., the fuzzy membership function, represented by the green curve) about the currently selected environmental factor, slope gradient, for the currently selected instance, Major Cabot. Also referring to Table 2, for Major Cabot the soil scientist picked a continuous function that favors gentle slope and specified that when slope is greater than 8% ($v_2$), the optimality value decreases as slope increases. The rate of decrease is defined by the setting that when the slope equals 20% ($v_2 = 8 + 12 = 20$), the optimality drops to 0.5. He defined a rule for wetness index in a similar way (not shown here). For Major Cabot, he chose to use Multiplication to integrate the optimality values from slope and wetness index, because he believed that the effects of the two factors are interrelated.

**Acquisition and Use of the Local Knowledge**

The local CBR process is represented as follows:

$$s_{i,j,c} = T_k \left( \prod_{c=1}^{n} \left\{ P_r \{ E_{c,a} (z_{i,j,a} , z_{j,a}) , d_{ij,c} \} \} \right\} \right)$$  [3]

Compared with Eq. [1], Eq. [3] has an extra function, $D$, for adjusting the optimality value using $d_{ij,c}$, the geographical distance between location $(i,j)$ and case $c$. The assumption underlying $D$ is that the closer two locations are, the more likely they are to have similar soils, which is justified by the observations of spatial autocorrelation in soils (e.g., Burrough and McDonnell, 1998). The SIE uses a Gaussian function to implement $D$:

$$D_{c,a} (E_{c,a} (\cdot) , d_{ij,c}) = E_{c,a} (\cdot) e^{d_{ij}^2 / d_{ij}^2} \ln(\varepsilon)$$  [4]

In Eq. [4], $d_{ij,c}$ is a search distance specified by the soil scientist for case $c$. The search distance eventually defines the influence region of the case. Parameter $\varepsilon$ is a very small value (e.g., 0.0001). When $(i,j)$ is within the influence region, the output from the $\varepsilon$ part of Eq. [4] varies between 0 (when $d_{ij,c} = d_{ij}$) and 1 (when $d_{ij,c} = 0$), thus achieving a continuous adjustment to the optimality value generated by the $E$ function. The soil scientist can also specify the value for $\varepsilon$, which adjusts the shape of the Gaussian curve.

In SIE, $d_{ij,c}$ is calculated along the terrain surface represented by the DEM (Shi, Zhu, and Wang, 2005). This surface distance helps limit the influence region of a case within the case’s own slope facet, which is a desired feature in soil inference.

The $T$ function is particularly important in local CBR, as a location may fall into the influence regions of multiple cases and the relationships among these cases may be complicated. The SIE provides a group of algorithms for $T$, some using a single case to determine the final value and the others integrating information from multiple cases.

Local CBR can serve two purposes. The first is to map those local soils controlled by certain environmental factors (e.g., parent material) for which no data layers are available for defining rules or running global CBR. To express the knowledge of these soils, the soil scientist has to pinpoint locations or delineate areas. Local CBR then uses these cases to map the soils. For our study area, even before the soil scientist ran RBR for the four major soil series, he knew that there were local exceptions caused by local parent materials. Specifically, well drained Hogback and Tunbridge soils were known to occur in places where the depth to the underlying bedrock is shallow or moderately deep. These soils are formed in friable till, unlike the dense basal till found in most of the study area. Since detailed surface materials and geological data layers are not available, the soil scientist created polygon cases to spatially delineate those parent materials and run local CBR using these cases. Figure 3 shows the local CBR result for soil Hogback.
The second use of local CBR is to fine-tune the soil maps created using the global knowledge. In our case study, the soil scientist indeed found small areas where the fuzzy membership values from RBR did not well match his expectations. He then used positive cases to increase the membership values in certain areas and negative cases to decrease the values in some other areas. A positive case specifies that the more similar a given location to the case’s location, the higher the fuzzy membership at the given location. The positive CBR result was integrated with the RBR result through a maximum operation, that is, the maximum value from the two results was to be used as the final value. A negative case defines a disfavoring condition for a soil, that is, the more similar to this condition, the lower the membership. The soil scientist found that a common use of negative cases was to decrease fuzzy membership values for those major soils in places where local soils occur. The negative CBR result was integrated with the RBR result through a minimum operation. Figures 4 and 5 illustrate the uses of positive and negative cases.

Validation

We distinguish two kinds of validations, one for assessing the capability of the DSM methodology and tool, and one for evaluating the actual quality of the map product. We also identify three sources of error: the soil scientist’s knowledge, the input environmental data, and the DSM methodology and its associated software tool. This differentiation is necessary, since the quality of the final soil map from DSM is not only determined by the DSM methodology and tool, but also highly relies on the quality of the knowledge and the quality of data; Simply evaluating the actual quality of the map cannot tell the sources of the error. Since the purpose of this study is to test a specific DSM methodology and the software tool, but not to evaluate the soil scientist’s knowledge or the environmental data, we chose not to base our validation on field samples, but largely on the soil scientist's judgment. A primary concern in the validation is to control for two (knowledge and data) of the three factors and test the other (DSM methodology and tool).

Our validation contains four components. The first is the soil scientist’s subjective evaluation of the SIE process, software tools, and products. The soil scientist in this project was asked to judge, based on his best knowledge and after making a reasonable effort to work with SIE, whether the process and the software tools would work in practice and whether the maps he created using SIE are of satisfactory quality. We considered this subjective evaluation primarily important, because of two reasons: 1. The SIE is designed to be used by soil scientists in daily practice, and whether this goal is achieved can only be judged by the soil scientists who have tried it; 2. since it is well known that soil mapping is a fairly subjective practice, asking the same soil scientist who created the SIE products to judge their qualities seemed to be the best way to control for the knowledge factor in the validation.

The other three components of the validation are relatively objective and are based on two benchmark products: a traditional soil map created by the soil scientist who would test SIE, and the fuzzy membership values assigned by the soil scientist to a set of validation points. Before the soil scientist started to use SIE, he was asked to (i) create a traditional map for the test area in a manual way; and (ii) predict fuzzy memberships for the soils at 100 locations (validation points) selected...
through a stratified sampling. He was asked to use all his knowledge, all the available data layers, and all the available GIS visualization tools to create these two products. For the 100 validation points, a software tool was developed to help the soil scientist specify how similar the soil at a location is to a certain soil series, and translate the natural-language specification into fuzzy membership values. Figure 6 shows an example that the soil scientist specified that the soil at validation point 3 is mostly Cabot, somewhat Colonel, and hardly Dixfield; the tool would save 0.7, 0.4, and 0.1 as the fuzzy memberships for the three series, respectively. These fuzzy membership values are assumed to fully and correctly represent the soil scientist’s knowledge of the soil at that location.

The second component of the validation is a visual comparison of the map created by SIE (referred to as the SIE map herein) and the map created through the manual process (referred to as the manual map herein). This visual inspection is the quickest way for the soil scientist, as well as other people, to compare and contrast the two maps. The third component is a spatial overlay of the two maps, based on which the mismatches between them can be precisely presented and a mismatch matrix can be calculated. The fourth component of the validation is a statistical comparison of the fuzzy membership values assigned by the soil scientist to the 100 validation points and their counterparts from SIE.

**Creation of a SSURGO Map**

The immediate output from SIE is a series of raster fuzzy membership maps. To create a map that meets the Soil Survey Geographic Database (SSURGO) standard of NRCS (http://soildatamart.nrcs.usda.gov, accessed 30 Apr. 2007), the following procedure was applied. First, the fuzzy membership maps of different soil series were integrated through a “hardening” process (Zhu, 1997), in which the soil series with the highest fuzzy membership value at a given location was designated as the dominant soil series at that location. The hardened raster map was then overlaid with a slope layer to create soil map units on different slope phases. Based on the National Cooperative Soil Survey standards regarding map unit purity (Soil Survey Division Staff, 1993), and also considering cartographic clarity, we selected 12,000 m² as the minimum size for a map unit polygon in our study area. Using SIE’s sliver-removing tool, patches smaller than 12,000 m² were merged into their neighboring patches. Finally, the raster map was vectorized, the zigzag boundaries caused by square pixels were smoothed, and map unit symbols (MUSYM) were assigned to the resulting polygons.

**RESULTS AND DISCUSSION**

Although it took time for the soil scientist to formalize rules and cases, to explore options for the E, P, and T functions, and to test parameter values, he considered that workload and time expense were reasonable when compared with the traditional manual process. The SIE allowed him to easily fine-tune the representation of his
knowledge and conduct many trials conveniently and rapidly. He believes that the work efficiency can be further improved as he becomes more adept at SIE. More importantly, he was generally satisfied with the output from SIE. He pointed out that the RBR was able to correctly map the catena pattern of the four major soil series, and that the local CBR was effective in handling local exceptions. He found that when a detailed surface material data layer is unavailable, creating local cases was the most effective, if not the only, way to represent his knowledge of those soils controlled by local parent materials (e.g., Hogback and Tunbridge). He also used local CBR to depict an area of Peacham that was “missed” by RBR and had been predicted to be Cabot.

Figure 7 shows the manual map and the SIE map side by side for comparison. Table 3 is a legend of the map units in the two maps. Visual inspections found that the two maps show fairly similar catena patterns of the four major soil series. The differences between the two maps pertain largely to the Colonel series, an intermediate soil in the catena between poorly drained Cabot and moderately well-drained Dixfield. Specifically, in some places where one map has unit 73 that represents a Colonel-Cabot complex, the other map has unit 23 that is dominated by Cabot, or unit 160 that is dominated by Dixfield. The soil scientist attributed these mismatches to the complexity of unit 73 (Colonel-Cabot) and the quality of the environmental data, rather than to the knowledge acquisition and inference processes. He explained that on the one hand, Colonel occurs over transitional landscapes and it is hard to formalize explicit rules for the environmental condition of Colonel; on the other hand, the 10-m DEM used in this project was not precise enough for him to geographically identify Colonel locations for creating cases. He acknowledged that the boundaries between Unit 23 (Cabot-dominant) and Unit 73 (Colonel-Cabot) and between 73 and 160 (Dixfield-dominant) in both maps contain considerable uncertainties, and thus he would not call the mismatches mistakes for either map.

Another apparent difference between the two maps is that the polygons in the SIE map are more fragmented. In the area 8C-7D in Fig. 7, the SIE map shows three strips of 73D that do not exist in the manual map. These units, identified mainly by the planform curvature data, represent the Colonel-Cabot complex occurring in relatively steep drainage ways. Referring to this example, the soil scientist praised DSM’s capability of and consistency in mapping small landform segments, as these small segments would be considered minor components, or inclusions, in manual mapping, and different soil scientists may treat them differently.

The mismatch matrix (Table 4) quantifies the comparison of the two maps. For example, the matrix shows that 43% of the area of unit 73 in the manual map was mapped by SIE as unit 23; and 53% of the area of unit 23 in the SIE map

---

**Table 3. Legend of map units in the study area.**

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Map unit name</th>
</tr>
</thead>
<tbody>
<tr>
<td>92E</td>
<td>Hogback-Rawsonville complex, 35 to 60% slopes, very rocky.</td>
</tr>
<tr>
<td>92D</td>
<td>Hogback-Rawsonville complex, 15 to 35% slopes, very rocky</td>
</tr>
<tr>
<td>92C</td>
<td>Hogback-Rawsonville complex, 8 to 15% slopes, very rocky</td>
</tr>
<tr>
<td>105F</td>
<td>Lyman-Rock outcrop complex, 60 to 90% slopes, very stony</td>
</tr>
<tr>
<td>160E</td>
<td>Dixfield sandy loam, 35 to 60% very stony</td>
</tr>
<tr>
<td>160D</td>
<td>Dixfield sandy loam, 15 to 35% very stony</td>
</tr>
<tr>
<td>160C</td>
<td>Dixfield sandy loam, 8 to 15% very stony</td>
</tr>
<tr>
<td>160B</td>
<td>Dixfield sandy loam, 3 to 8% very stony</td>
</tr>
<tr>
<td>73D</td>
<td>Colonel-Cabot complex, 15 to 35% very stony</td>
</tr>
<tr>
<td>73C</td>
<td>Colonel-Cabot complex, 8 to 15% very stony</td>
</tr>
<tr>
<td>73B</td>
<td>Colonel-Cabot complex, 3 to 8% slopes, very stony</td>
</tr>
<tr>
<td>23C</td>
<td>Cabot silt loam, 3 to 8% slopes, very stony</td>
</tr>
<tr>
<td>23B</td>
<td>Cabot silt loam, 0 to 3% slopes, very stony</td>
</tr>
<tr>
<td>24A</td>
<td>Peacham muck, 0 to 3% slopes, very stony</td>
</tr>
<tr>
<td>363E</td>
<td>Tunbridge-Dixfield complex, 35 to 60% slopes, very stony</td>
</tr>
<tr>
<td>363D</td>
<td>Tunbridge-Dixfield complex, 15 to 35% slopes, very stony</td>
</tr>
<tr>
<td>363C</td>
<td>Tunbridge-Dixfield complex, 8 to 15% slopes, very stony</td>
</tr>
<tr>
<td>363B</td>
<td>Tunbridge-Dixfield complex, 3 to 8% slopes, very stony</td>
</tr>
</tbody>
</table>

---

**Table 4. Mismatch matrix for the manual map vs. the SIE map (Area unit is m²).**

<table>
<thead>
<tr>
<th>The Manual Map</th>
<th>Percentage, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>23</td>
</tr>
<tr>
<td>23</td>
<td>432628</td>
</tr>
<tr>
<td>73</td>
<td>175796</td>
</tr>
<tr>
<td>160</td>
<td>36900</td>
</tr>
<tr>
<td>Total</td>
<td>645324</td>
</tr>
</tbody>
</table>

---

†The percentage is calculated as the area of each category divided by the total area in the same row or column.
was mapped by the manual map as unit 73. Note that the mismatch matrix does not include the two local soils, Hogback and Tunbridge, because their delineations in both maps were directly controlled by the soil scientist and their consistencies between the two maps are apparent.

The relationships between the fuzzy membership values assigned by the soil scientist to the 100 validation points and their counterparts from SIE are plotted (Fig. 8). We only plotted for Cabot, Colonel, and Dixfield, since the area of Peacham is small and only a few points have non-zero membership values for this soil. The correlation coefficients ($R$) are as follows: Cabot = 0.80; Colonel = 0.42; Dixfield = 0.62; and Peacham = 0.66. After reviewing the membership values in both sets, the soil scientist confirmed that the low correlation in Colonel was still largely due to the high uncertainty in the model for this soil. Again, the correlation was not evaluated for the two local soils (Hogback and Tunbridge), since only a few validation points received membership values for these soils.

Generally, the mismatches between the results directly from the soil scientist and from SIE can be attributed to the three factors we have been considering in this study: knowledge, data, and the DSM process and tool. To control for the knowledge, we had the same soil scientist create the manual map, assign fuzzy membership values, and create the SIE map. However, if the knowledge itself contains considerable uncertainty, it is not unexpected to see inconsistency in the results. We also tried to control for the data factor by making the same environmental data available to the soil scientist when he generated the three different knowledge representations (the manual map, the fuzzy membership values, and the knowledgebase for SIE). However, we found that two data issues may still lead to mismatches between the different representations: 1. Some terrain attributes commonly used in DSM, such as various types of curvatures and wetness index, are not conventionally used in manual mapping. As a result, the soil scientist may not be familiar enough with these attributes to assign proper values for them when creating the knowledgebase. 2. The ways that the soil scientist and the computer use data are quite different. The human capability of incorporating contextual information may mitigate data quality problems, while the computer’s precision in handling local values guarantees consistency, but not necessarily well represents what the soil scientist has in mind.

The data issues are also related to another source of mismatches, the DSM process and tool. Through this project, we realized that the shift from vector to raster mapping is more profound than its technical appearance. Particularly, in vector mapping the map is seen as a “salad bowl,” that is, each location in the map unit has one and only one type of soil, even if a map unit is a “complex” one that contains more than one soil. In raster mapping, when a pixel is assigned fuzzy memberships for multiple soil types, it is thought to be a “melting pot,” that is, the soil at the pixel is similar to those soil types in terms of properties, but not a spatial collage. The SSURGO standard was established based on the “salad bowl” model, and therefore in a DSM project like ours, there has to be a conversion from “melting pot” to “salad bowl,” during which problems may occur. For example, the operations that group pixels into “complex” units to meet the SSURGO standard can be subjective and ad hoc.

**CONCLUSION**

This paper presents an analysis of soil scientists’ knowledge from the scale and space perspectives. The analysis sets the basis
for a formal framework for constructing structured knowledge acquisition and inference processes, and leads to a DSM methodology specifically designed to effectively capture and apply the knowledge of local soil experts. We found that this analysis of knowledge is particularly necessary for DSM. When drawing the manual map, the soil scientist never felt that he needed to first analyze his knowledge into different categories. However, when he tried to articulate his knowledge and fit it into prescribed forms and formats required for computerized representation and inference, he found the framework presented in this paper helpful. Furthermore, the structured process helped the soil scientist identify “holes” or uncertainties in his knowledge. From the tool developer’s perspective, this analysis provided guidelines to incorporating different knowledge engineering technologies for fully acquiring and utilizing soil scientists’ knowledge. Although the quantitative comparison between the manual map and the DSM map showed numerous mismatches, the soil scientist was convinced that the proposed DSM process and the associated software tool are effective. He was generally satisfied with the resulting maps, in terms of both quality and cost.

We actually consider the mismatches between the results directly from the soil scientist and that from SIE to be more meaningful than their coincidence, as the mismatches reveal that the two approaches were indeed different and may produce apparently different products. As the soil scientist evaluated, the mismatches do not necessarily indicate which map is “better,” but provide hints on how DSM will impact the traditional soil survey and mapping paradigm, including its standards, processes, and applications.

While we find the outcome of this pilot project encouraging, we are well aware of its limitations: 1. The test watershed is fairly small with only a few major soil series and relatively consistent soil–landscape relationships. 2. We have largely relied on one soil scientist’s judgment in the validation, which is subject to bias, especially when the validation involves considerable subjectivity. When dealing with a vastly larger area containing significantly heterogeneous environments, the process of building and applying a knowledgebase can be much more complicated.

ACKNOWLEDGEMENTS
Support from USDA-Natural Resources Conservation Service (NRCS), under Agreement No. 53-7482-5-466 is gratefully acknowledged.

REFERENCES


