INT. J. GEOGRAPHICAL INFORMATION SCIENCE, 2003
VOL. 17, NO. 8, 771–795

Research Article

Knowledge discovery from soil maps using inductive learning

FENG QI[1] and A-XING ZHU[1,2]

[1]Department of Geography, University of Wisconsin-Madison, 550 North Park Street, Madison, WI 53706, USA; e-mail: fqi@wisc.edu

[2]State Key Laboratory of Resources and Environmental Information System Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Building 917, Datun Road, An Wai, Beijing 100101, China; e-mail: axing@geography.wisc.edu

(*Received 10 July 2002; accepted 16 May 2003*)

Abstract. This paper develops a knowledge discovery procedure for extracting knowledge of soil-landscape models from a soil map. It has broad relevance to knowledge discovery from other natural resource maps. The procedure consists of four major steps: data preparation, data preprocessing, pattern extraction, and knowledge consolidation. In order to recover true expert knowledge from the error-prone soil maps, our study pays specific attention to the reduction of representation noise in soil maps. The data preprocessing step has exhibited an important role in obtaining greater accuracy. A specific method for sampling pixels based on modes of environmental histograms has proven to be effective in terms of reducing noise and constructing representative sample sets. Three inductive learning algorithms, the See5 decision tree algorithm, Naïve Bayes, and artificial neural network, are investigated for a comparison concerning learning accuracy and result comprehensibility. See5 proves to be an accurate method and produces the most comprehensible results, which are consistent with the rules (expert knowledge) used in producing the soil map. The incorporation of spatial information into the knowledge discovery process is found not only to improve the accuracy of the extracted knowledge, but also to add to the explicitness and extensiveness of the extracted soil-landscape model.

1. Introduction

It is well established that the map is a powerful medium for presenting spatial information and geographical relationships. Much of our understanding of the relationships among spatial phenomena is inexplicitly embedded in maps. It is often desirable to have these understandings explicitly stated for complex map interpretation as well as for future map updates. With developments in both geographic information processing techniques and geographic data warehousing, it is possible to extract explicitly the knowledge embedded in maps. Malerba *et al.* (2002) used machine learning tools to extract information from topographic maps. Compared to the general purpose topographic maps, thematic maps concern specific geographic features and contain specialized domain knowledge. For example, natural

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2003 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810310001596049

resource maps are usually created by experts through a modeling process, and thus convey knowledge about the particular models.

Soil maps are one example of these resource maps. In soil survey, soils are mapped based on the concept that soil is the result of the interaction of its formative environment: $S=f(E)$, as referred to as the soil factor equation by Dokuchaeiv (Glinka 1927) and Hilgard (Jenny 1961). Hudson (1992) generalized this soil factor equation to a soil-landscape paradigm, which is now the guiding paradigm for soil surveys in the USA. When creating soil maps in soil survey, soil experts through great effort work out the relationships between soil and its landscape conditions and draw soil polygons based on the perceived distribution of landscape units (Hudson 1992). The spatial configuration of the resulting soil polygons thus implies the relationships between soil and the environmental conditions over the landscape. Information on how the soil types are related to each other, and why certain soil is mapped at certain landscape locations, are the implicit knowledge embedded in the soil map. This implicit knowledge is considered to be the soil-landscape model (Hudson 1990, 1992).

The knowledge of the soil-landscape model embedded in soil maps is valuable in at least two ways. First, it has the potential to facilitate traditional soil survey updates. The conventional soil survey is a manual and time-consuming process. It is very unlikely that the soil scientist(s) who initially mapped the soils over an area would be the person(s) to conduct the soil survey update for the area, since the update cycle is often longer than the career span of a soil scientist. On the other hand, the soil-landscape model used to create the soil map often exists as the soil experts' tacit, undocumented knowledge. Therefore, when local soil experts retire or move out of an area, they take the knowledge with them. To remap the area during soil survey updates, new soil experts would have to develop their own model from scratch. This would involve a tremendous amount of fieldwork. However, if the knowledge of the experienced soil experts could be retrieved and presented in a proper form, the new soil scientists could then build upon it. This would greatly facilitate the update of soil survey. Second, the knowledge of local soil-landscape relationships, once extracted and properly formulated, could be used for automated soil mapping, modelling, and classification. Moran and Bui (2002) used machine learning generated rules of soil-landscape relationships to mimic the mental process used by soil surveyors and managed to reproduce the original soil map to a considerable extent. Zhu and Band (1996) developed a soil-land inference model (SoLIM) to combine the expert knowledge on soil-landscape relationships with geographical information system (GIS) and artificial intelligence (AI) techniques under fuzzy logic to map soils. Like other knowledge-based systems, SoLIM builds upon expert knowledge for automated inferences, and is only suitable for areas where there are experienced local soil experts from whom the needed knowledge on soil-landscape relationships can be obtained (Zhu 1996). For regions where there is no experienced human expert available to provide the knowledge, a possible alternative is to extract knowledge from other data sources. The soil map produced in previous surveys is one such potential source. The U. S. Department of Agriculture (USDA) has been maintaining a soil survey geographic (SSURGO) database that contains soil surveys of much of the nation's land. The availability and accessibility of digital soil maps make it possible to extract useful information, in this case, knowledge on soil-landscape relationships, from the maps.

In this paper, we present a knowledge discovery procedure that uses inductive learning to extract the knowledge embedded in natural resource maps. We use a soil map as an example to illustrate this procedure. Soil maps created through manual soil surveys, like other natural resource maps, are prone to two kinds of errors: inclusions and misplacements of boundaries. Therefore, the map product may not represent the soil expert's true knowledge, but may contain noise. The procedure presented here pays particular attention to this unavoidability of errors. To reduce the impact of errors (noise) in the maps and recover the expert knowledge, as part of the whole procedure we designed a sampling strategy for data preprocessing. This paper examines knowledge discovery from three perspectives: (1) the impact of data preprocessing, particularly from the perspective of noise reduction; (2) the effectiveness of three basic types of inductive learning algorithms: a decision tree learning algorithm, the neural network backpropagation algorithm, and the Naive Bayes algorithm; and (3) the effect of the incorporation of spatial information on the knowledge discovery process. The rest of this paper is organized as follows: we begin with a brief introduction to the conventional soil survey process to examine how local soil scientists' knowledge is encoded into the soil maps; this provides the basis for our knowledge discovery methodology. The knowledge discovery procedure is then presented. The process is illustrated through a case study. Last, we discuss our results and draw some conclusions.

2. Basis for extracting knowledge from soil maps

The conventional soil survey is based on Jenny's classic model:

$$S=f(d, o, r, p, t, \ldots),$$

where S is the soil, d is the climatic factor, o is the biotic factor, r represents the topographic factor, p refers to the parent material, and t is the time factor. The ellipses after t represent unspecified factors that might be important locally or even regionally. Basically, the equation says that soil is the result of the interactions of its formative environmental conditions, and the factors of soil formation interact in a distinctive manner within the so-called soil-landscape units so that soils are considered to be predictable and mappable objects based on landscape units (Hudson 1990, 1992). The soil survey and mapping process is actually an inference process based on observable landscape conditions and expert knowledge of the local soil-landscape model. Work by Deka *et al.* (1995), McLeod *et al.* (1995), Wright (1996), Bruin *et al.* (1999), and others demonstrates the continuing success of the soil-landscape model concept for various terrain types. Therefore, the soil-landscape paradigm is widely adopted in soil survey practice to produce soil maps. During soil survey, soil experts first conduct intensive fieldwork to study the local soils toward the construction of a soil-landscape model. They then characterize the landscape conditions by studying aerial photos using stereoscopes, and examining the geological information. Soil surveyors then delineate the spatial extents of different soils or combinations of different soils based on their understanding of the soil-landscape model and the observed landscape conditions.

During the soil mapping process, the soil-landscape relationships are elaborately worked out and implicitly applied to the soil polygon delineation. The spatial positions of the soil polygons thus imply the relationships between different soil types and their underlying environmental conditions. When soil experts draw the polygon boundaries, they implicitly integrate multiple environmental data layers:

the geology layer, the topographic layers and the land use layer observed through stereoscoping. The basic idea of extracting knowledge from these polygon-based soil maps is to reverse this mapping process. In other words, the relationships between soil type and landscape characteristics can be revealed through a knowledge discovery approach by analysing soil maps together with the landscape characteristics captured using GIS.

3. The knowledge discovery procedure

Knowledge discovery or data mining is 'the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data' (Fayyad 1996). The knowledge discovery procedure employed in this study is a modified version of the general steps presented by Fayyad (1996), who states that a complete knowledge discovery process includes 'data warehousing, target data selection, cleaning, preprocessing, transformation and reduction, and finally consolidation and use of the extracted knowledge'. Our knowledge discovery process consists of four major steps: data preparation, data cleaning and preprocessing, pattern extraction, and finally, knowledge examination and interpretation. The implementation flow is illustrated in figure 1.

3.1. *Data preparation: data selection and compilation of a GIS database*

To discover knowledge of soil-landscape relationships embedded within existing soil maps, the first task is to choose the relevant variables that effectively describe the soil formative environment. While environmental conditions include various factors, only some of them influence soil-formation. According to literature in pedology (Hudson 1992, McSweeney *et al.* 1994), soil-formative environmental factors include climate, parent material, geomorphology, biology, and human interference. While at the watershed scale the practical environmental variables used in a soil-landscape model are usually bedrock geology, topographical characteristics, and vegetation conditions.

It is important to point out that the specific list of environmental variables to be used is area specific, depending on the pedogenesis in the local area. But among the most commonly used variables in the construction of soil-landscape models are the bedrock geology and the basic topographic variables: elevation, slope gradient, slope aspect, and surface curvature (planform and profile). Given no other information concerning the local soil pedogenesis, these variables should be the starting list of environmental variables used for knowledge discovery. Most likely, further information on local soil pedogenesis can be obtained from soil survey reports. Furthermore, since one of our purposes in extracting knowledge from existing soil maps is to help inexperienced local soil experts to build their own soil-landscape model, these soil experts may want to add some potential variables to examine. Auxiliary variables other than the basic ones are thus added in most cases. In §4, we will add two kinds of additional variables in a case study. One kind serves to describe complex topographic characteristics by taking into account the spatial relations of basic topographic variables. The other kind considers topological attributes of the soil polygons to describe spatial relations between different soil series. It will be shown that the inclusion of variables that describe spatial relations will not only improve the accuracy of the extracted soil-landscape model, but also lead to a more comprehensible knowledge representation.

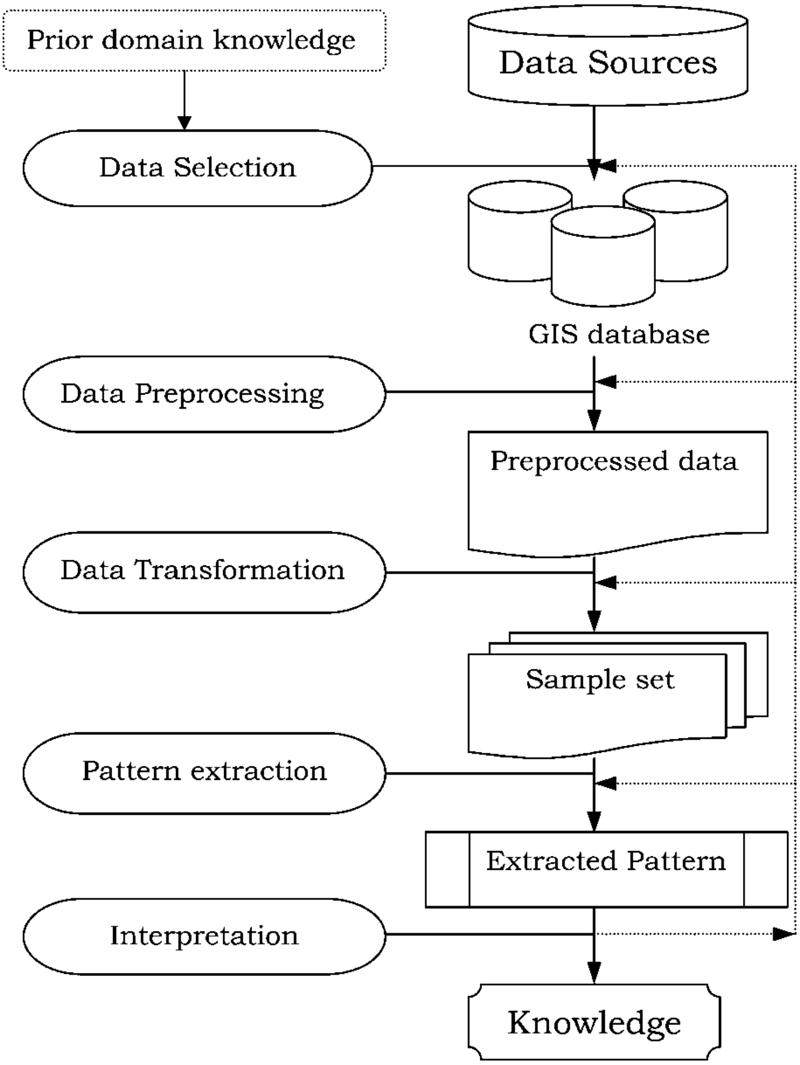


Figure 1. The knowledge discovery process.

Upon the determination of soil-formative environment variables in the mapped area, the target dataset for knowledge discovery is a GIS database consisting of the soil map and the data layers of the identified environmental variables. The elevation, slope, aspect, and curvature data layers can all be derived directly from a digital elevation model (DEM). Geology is often very important in soil formation. However, detailed description of spatial variation of geology is very difficult to obtain, particularly for surficial geology. Information on geology often exists in the form of polygon maps, as is the soil map generated from traditional soil survey. These polygon-based maps suffer from the same kind of drawbacks (subjectivity and scale limitation) as the conventional soil maps do (Zhu 1996). Due to its importance in soil formation, geological information is often used to indicate different areas of major pedogensis. Vegetation information can be obtained from vegetation maps or from remote sensing images, such as tree canopy coverage (Nemani *et al.* 1993) and leaf area index (Nemani *et al.* 1993, Fassncht *et al.* 1997). However, the usefulness of vegetation information very much depends on the ability of these data layers in relating to soil formation or soil conditions in the given area.

It is worth noting that the variables included in our GIS database are not necessarily the same variables used by the soil experts when they created the soil map. Actually, it is impossible to obtain the same exact base maps that they used. We may miss some of the layers they used because we don't have their knowledge of the actual soil-landscape model yet. On the other hand, it is usually the case that we may add new layers that explicitly describe certain aspects of the landscape characteristics. Furthermore, the data layers we use in our knowledge discovery may be of greater accuracy or resolution than the original data soil experts used, due to the constant improvement of data capture and data representation in GIS. Although the use of different variables may lead to the extraction of a soil-landscape model different in form from the original tacit knowledge of soil experts, the better accuracy and explicit representation of landscape characteristics may lead to a more extensive model. In section 6, we will see from a case study how the inclusion of auxiliary data layers improves the extracted soil-landscape model to make it more comprehensive.

3.2. *Data cleaning and preprocessing*

Tasks in this step usually include removing noise or outliers and eliminating invariant or redundant representations of the data. As aforementioned, a conventional soil map is produced through a manual mapping process that is not only time-consuming, but also error-prone and inconsistent. Most soil mappers base their soil unit delineation on visual interpretation of stereo photos. Subtle and gradual changes in environmental conditions are often difficult to discern via stereoscoping. It is easy to misplace the boundaries of soil polygons in the manual delineation process. Furthermore, due to the limitations of map scale, small patches of soil types may not be shown on a soil map, but they exist inside polygons of other soil types as soil inclusions. The misplacement of soil boundaries and the existence of soil inclusions result in some of the pixels being associated with incorrect environmental conditions. These pixels are considered systematic noise or outliers and will exert enormous influence on the accuracy of the extracted knowledge.

In our study, the major effort we make to reduce noise and effective size of the database is to sample only the pixels that are representative of the soil types. We assume that the misplacement of soil lines is not exorbitant, so that the majority of the polygon area is correctly categorized; thus the histogram mode(s) of a given environmental variable enclosed in the soil polygons for a given soil type represents the typical conditions under which the soil develops or is expected to occur. We thus believe that the representative pixels are those whose environmental conditions are at or close to the mode of environmental histograms. In implementation, for all pixels belonging to a single soil type, a histogram is constructed for each environmental variable, with the horizontal axis representing the intervals of the environmental variable, and the vertical axis representing the number of pixels whose environmental condition falls within the interval. The resulting histogram can be either unimodal or multimodal. A unimodal shape is the most common, with one single mode of the histogram indicating the central concept of the soil type and the low frequency tails representing map errors, inclusions, or transitional conditions. An example of such a histogram for soil type *Kickapoo* based on environmental variable *slope gradient* is illustrated in figure 2. This histogram has a modal *slope gradient* of approximately 5%, which demonstrates that *Kickapoo* develops mainly in areas with slope ranging from 4% to 6%. Multimodal shapes are possible when (1) the soil type occurs at more than one typical landscape positions, or (2) one soil mapping unit is used to represent more than one soil taxonomic units. An example of the first situation is that soil type *A* occurs both on narrow ridge tops and convex slope shoulders. The second situation happens when the map scale does not allow for detailed differentiation of two or more soil types thus a complex map unit is adopted.

A concern in the construction of a histogram is how to determine the number of intervals. In our study, the number of intervals is determined to be proportional to the number of pixels belonging to each soil type: $N_i = N_p/r$. Here N_i refers to the number of intervals of the histograms for a certain soil type, N_p refers to the total number of pixels contained in this soil type, and finally, r is the average number of pixels expected to fall within each interval, which is to be determined by the operator. With the increase of r, the number of intervals decreases, and the size of interval increases. In a case study discussed in later sections, different choices of r are experimented. This sampling specification allows the number of intervals for different soil types to be adjusted according to the number of pixels belonging to the soil type, so that the number of pixels falling within histogram mode is comparable across different soil types. This allows each soil type to be equally represented in the samples, thus preventing training bias and problematic performance evaluation (Gahegan 2000).

Once a set of such histograms is constructed for each soil type, sampling is conducted based on soil types. The individual sample set for each soil type is produced in two steps. The first step is to sample pixels one environmental variable at a time, that is, to sample just the pixels falling into the mode(s) of the histogram based on the given environmental variable. We investigated two options: one is to

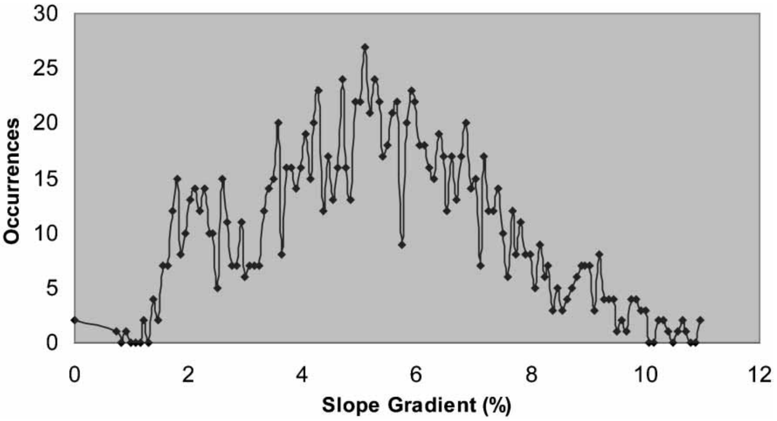


Figure 2. Histogram of slope gradient for area mapped as *Kickapoo*.

take the entire set of pixels within the mode(s), and the other is to sample randomly a fixed number of pixels (N_r) from each mode. The second step is to pool the samples from all environmental variables for the soil type, and then select samples from this pool. A single pixel may have more than one occurrence in this pool since it could be selected based on the modes of multiple environmental variables. Two approaches can be taken to generate non-redundant sample sets. The first approach is the union operation, which is done by retaining only one occurrence of repeats. The second approach is the intersection of the samples. This is accomplished by selecting only those pixels that show up in modes of more than one environmental histogram. With both approaches, the final set is constructed by simply combining the sample sets for all soil types.

As aforementioned, this sampling strategy aims to reduce the noise that is caused by errors in the original map. With the application of learning programs to only these rectified samples, the learned result is supposed to approximate the soil expert's knowledge rather than the error-prone map. This sampling strategy, therefore, is expected to improve the knowledge discovery performance.

3.3. *Data transformation and pattern extraction*

The central step of a knowledge discovery process is the extraction of patterns from data. To achieve this, various data mining algorithms have been developed (Fayyad 1996, Murray 1998, Koperski *et al.* 1999). The selection of suitable algorithms for a specific knowledge discovery task is determined by multiple factors, including the nature of the data source, appropriate knowledge representation, desired accuracy, and so on.

Among various data mining algorithms, inductive machine learning methods are widely used in knowledge discovery and in various classification practices because they offer significant advances in several ways. They usually make few assumptions about model parameters and data distributions. They are able to deal naturally with certain levels of noise. They scale well with the expansion of feature space, and many of them are computationally efficient. They also have their limitations in terms of generalization behaviour, are user-demanding in selection of operational parameters, and so on (Gahegan 2000). Different algorithms are suitable for different problem configurations.

In general, it is important to choose an appropriate knowledge representation for knowledge discovery tasks. Common knowledge representations include production rules, decision trees (Quinlan 1986), frames (Minsky 1975, Mennis *et al.* 2000), fuzzy membership functions (Zadeh 1965), semantic nets, regression and correlation analysis results, and adaptation knowledge. The knowledge representation extracted from soil maps could take the form of either rules/decision trees or fuzzy membership functions (Zhu 1999). We chose a decision tree to represent the extracted knowledge of soil-environment relationships in our study, because the conditional logic associated with a decision tree structure is compatible with how soil scientists understand the soil-landscape relationship, and thus it is an easily comprehensible form in this specific domain. Furthermore, a decision tree is easily transformed into rules and descriptions, which can be directly used with knowledge-based inference systems for automated soil inference.

Decision tree induction is one of the most established learning algorithms (Russell and Norvig 1995). With a decision tree, training data is partitioned to the children nodes using a splitting rule until all training samples can be categorized to a predefined class. For example, the soil-landscape relationships could be expressed as a decision tree, as shown in figure 3. One of its branches basically says that if the bedrock at a site is *Oneota*, the slope gradient is less than 12%, and if the planform curvature is almost linear on flat ridge tops, then the soil is expected to be '*Valton*'.

When constructing a decision tree from training data, choosing the right size for the tree is an important problem. A tree that classifies the training data perfectly may not be the tree with the best generalization performance when applied to real data since (1) there may be noise in the training data that the tree is fitting; and (2) the algorithm might be making some decisions about the leaves of the tree that are based on very little data. This phenomenon is called overfitting, and an overfitted tree may not reflect reliable trends in the data. To avoid overfitting, various efforts have been made to improve the decision tree algorithm itself, including various pruning algorithms (Esposito *et al.* 1997). Yet another effective way to avoid overfitting on noisy data is to reduce noise ahead of time, given prior knowledge in the specific application domain, as we tried to do in the data preparation stage.

In our study, we use the See5 algorithm (Quinlan 1993, 2001) to derive decision trees from training data. See5 is based on an information theoretic approach. To build a decision tree, See5 recursively grows a tree top-down through batch processing of the training data, using a greedy heuristic to search for a simple tree based on information gain (Quinlan 1993). It selects an attribute for each node with the most information gain. First, the entropy (or impurity or disorder) of a set of examples S is calculated as:

$$Entropy\ (S) = \sum_{i=1}^{c} -p_i \log_2(p_i),$$

where p_i is the proportion of category i samples in S, and c is the number of categories. The information gain of an attribute is the expected reduction in entropy

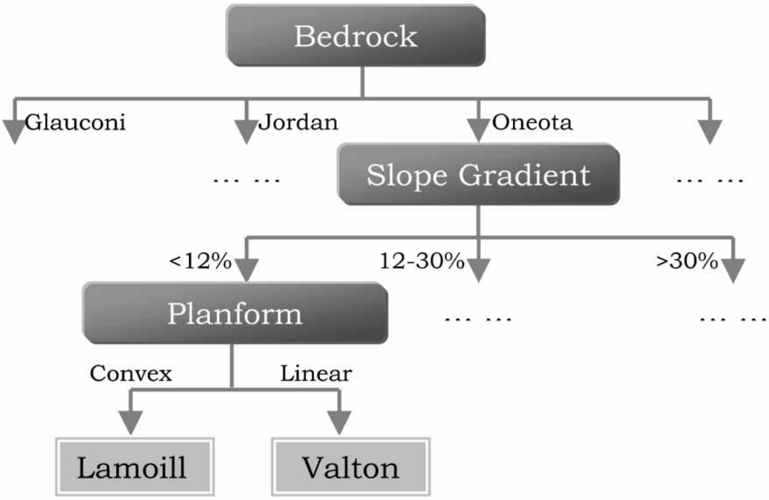


Figure 3. A decision tree representation of the soil-landscape model.

caused by partitioning on this attribute:

$$Gain\ (S,\ A) = Entropy\ (S) - \sum_{v} \frac{|S_v|}{|S|} Entropy\ (S_v),$$

where S_v is the subset of S for which attribute A has value v. The entropy of the partitioned data is calculated by weighting the entropy of each partition by its size relative to the original set.

To examine the algorithm's feasibility in terms of learning accuracy and result comprehensibility, we also experimented with two other algorithms. They are Naive Bayes and neural network backpropagation (Mitchell 1997). Naïve Bayes uses Bayes theorem to predict the value of a target field from evidence given by labeled examples. Under the assumption of conditional independence, it estimates the posterior probabilities of all possible classifications; and the classification with the highest posterior probability is chosen as the prediction. Although the independence assumption is essential for the optimality of the Naïve Bayesian classifier, empirical results have shown that it performs well in many domains containing clear attribute dependences and often outperforms more powerful classifiers. Domingos and Pazzani (1997) justify its optimality in other sufficient conditions, and suggest that the simple Bayesian classifier has a wide range of applicability. However, one limitation of this simple probability-based classification is that it does not explicitly reformulate the feature space; rather it classifies future samples on the fly. In other words, the learned concept is not interpretable.

Inductive learning with artificial neural networks uses an interconnected neural network structure to model interrelationships between features. A neural network usually consists of multiple layers of simple processing elements called neurons. Each neuron is linked to certain of its neighbors with varying coefficients of connectivity that represent the strengths of these connections. Learning is accomplished by adjusting these strengths to cause the overall network to output appropriate results. Back propagation (Mitchell 1997) is commonly used to train neural networks with labeled examples. It is known to be able to model a rich class of concepts through non-linear modeling. Although it provides the best predictive accuracy for many applications, problems arise when the learned network structure and weights need to be interpreted. Further efforts have to be made to understand the concept representations in a neural network (Craven and Shavlik 1997).

3.4. *Examining and consolidating the discovered knowledge*

In the last step of a complete knowledge discovery process, the extracted pattern is finally examined and interpreted to be of future use. In our study, the learned decision tree is tested using independent samples from the same map to obtain learning accuracy. Once the accuracy is satisfactory, it is considered to approximate the soil map sufficiently and is ready to be interpreted in terms of rules and descriptions, and to be incorporated into performance systems or simply documented and reported to interested parties.

To further validate this knowledge discovery methodology, more validation efforts were made in our case study. We chose a specific soil map from which to extract knowledge. We also conducted knowledge acquisition with the soil expert who created the soil map, through an interview to acquire his knowledge of the soil-landscape relationships over the same area. Assuming that the acquired knowledge is the exact knowledge that the expert applied in drawing the soil map, we compared our extracted knowledge with this acquired knowledge to test the accuracy. The details of this validation are described in §6.

Above we have outlined the basic knowledge discovery procedure. The actual process can contain iterations of some steps and even loops between any two steps. Most previous work on knowledge discovery from databases has focused on step 3—the specific algorithm taken. However, the other steps, especially the data preprocessing step, are of considerable importance for the successful application of knowledge discovery in practice. In this paper we pay attention to the importance of the data preprocessing step in knowledge discovery.

4. Study site

Our study site is the Raffelson watershed in the state of Wisconsin, USA, with a total area of approximately 4 km^2. Located on the edge of the 'driftless area' of southwestern Wisconsin that has remained free of direct impact from Pleistocene era continental glaciers, the watershed is of a typical ridge and valley terrain with relatively flat, narrow ridges. A 3-D view of the area with an orthophoto is shown in figure 4.

Zhu *et al.* (2001) have demonstrated the applicability of a soil-landscape model to mapping soils in this non-glaciated 'driftless' area. The soil map created from a recent soil survey indicates 16 different soil series in the area (figure 5). This area was chosen as the study site for knowledge discovery because the recently created soil map is suitable for testing our methodology. During the soil mapping process, expert knowledge of the local soil-landscape relationships was acquired through interviews with the soil expert and was documented in the form of rules and environmental descriptions. The same knowledge was directly applied to create the soil map. Because this study aims to reveal the knowledge that soil experts put into soil maps, it is essential to compare the derived knowledge with the knowledge that leads to the creation of the same map.

As §3.1 discussed, we started with five basic variables to characterize the formative environmental conditions of soils over the study area: elevation, slope gradient, planform curvature, profile curvature, and geology. Aspect was not

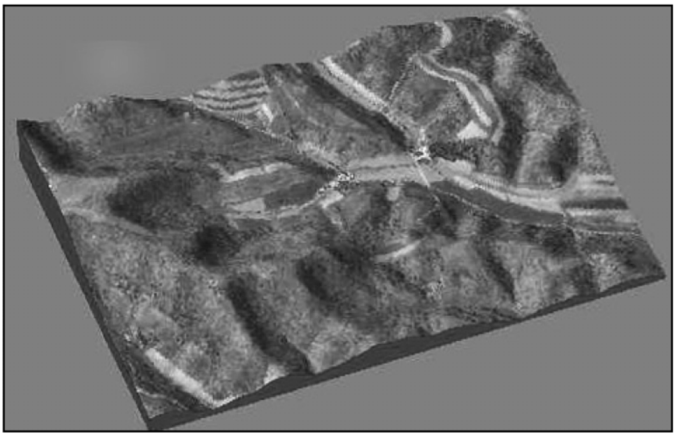


Figure 4. Raffelson watershed, Wisconsin, USA.

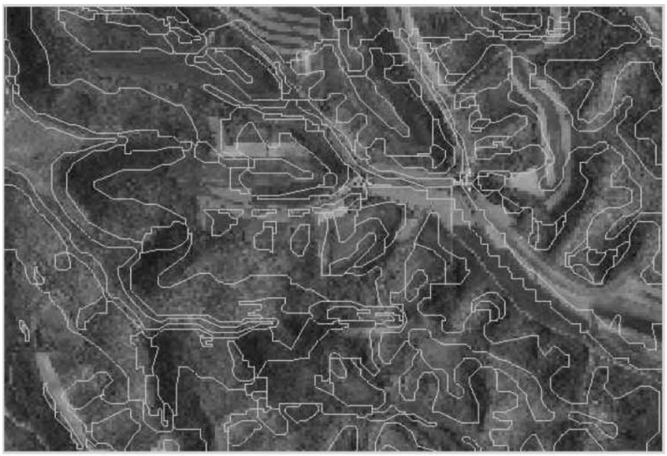
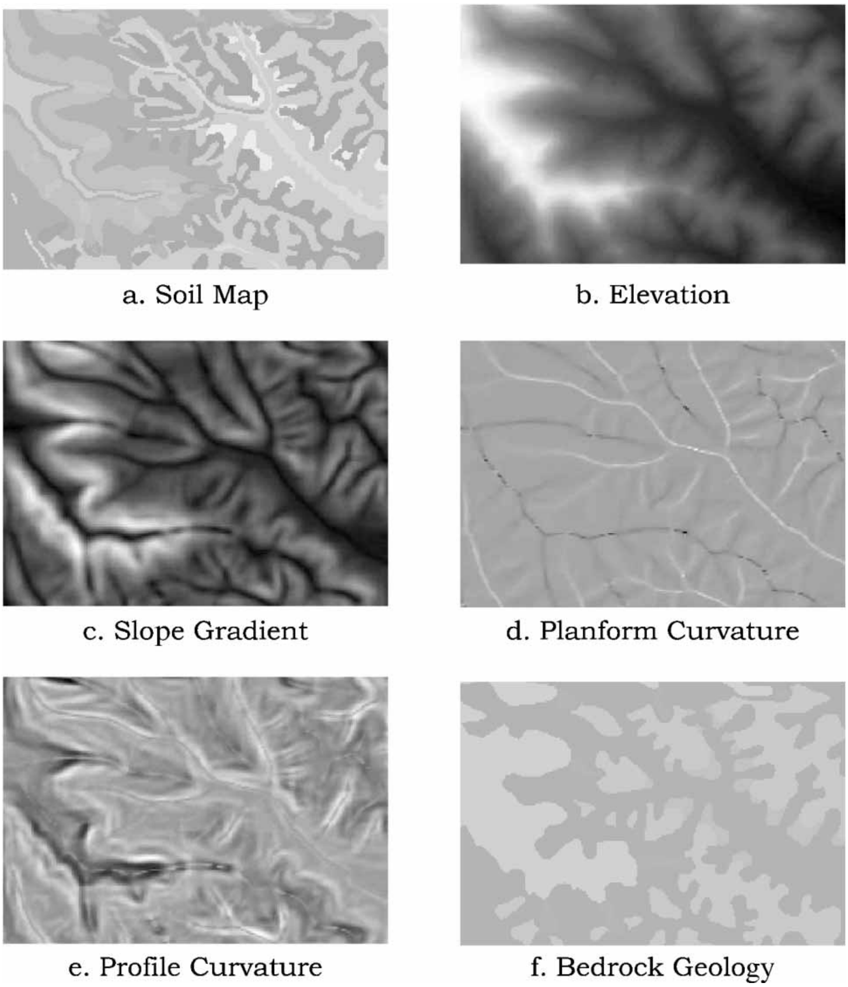


Figure 5. The soil map of Raffelson watershed.

chosen because there's almost no indication of soil type changes due to aspect. The lack of influence of aspect can be observed on the map, where soil polygons wrap around slopes of all aspects. Known from the soil survey report as part of the background knowledge, bedrock geology in this area is said to be complex and plays an important role in the soil formation. Therefore, a detailed bedrock geology layer is included in the database. The elevation, gradient, and curvature data are all derived from a 10-metre resolution DEM recently produced by the United States Geological Survey (USGS). Figure 6 illustrates the appearances of these data layers that constitute the preliminary soil-landscape database.

Considering the spatial nature of soil formation and distribution would add to the explicitness and extensiveness of the extracted soil-landscape model. However, the variables that describe spatial structure should be confined to those that soil experts are familiar with or can easily understand. Considering this, we included two kinds of variables in our case study to take into account spatial relations. One kind serves to capture the spatial relations of soil-formative environmental factors, and the other serves to describe spatial relations between soil types.

For the first kind, we experimented with three variables that might be related to soil formation, according to soil experts' suggestions. They are distance to streams, topographic wetness index, and percentage of colluvium from competing bedrocks. To obtain the distances to streams for given locations, stream channels are first derived using the approach described by O'Callaghan and Mark (1984). The spatial distance from the soil site to the closest stream is then measured. Topographic wetness index is used to combine connectivity information based on flow direction with slope dynamics to represent the hydrological topographic characteristics that influence soil formation (Moore *et al.* 1993, Band *et al.* 1993). Finally, since colluvium from different bedrocks tends to influence soil development, we calculate the percentage of colluvium from competing upslope bedrocks for footslope locations. Specifically, for a given footslope location, the relative amount of colluvium it received from a certain bedrock is approximated on the basis of the accumulated upstream drainage cells originating from the given bedrock polygon. The percentage of colluvium from multiple competing upslope bedrocks is then computed relatively.



a. Soil Map

b. Elevation

c. Slope Gradient

d. Planform Curvature

e. Profile Curvature

f. Bedrock Geology

Figure 6. The soil-landscape database for Raffelson watershed.

Unlike the above three spatial variables, the spatial relations between soil types are not soil-formative factors, but instead they serve as indicators of soil distributions over the area. In other words, by including the topological relations of the soil polygons in knowledge discovery, the extracted soil-landscape relationships can be enriched to describe spatial patterns of soil types. The next section will show that the extracted soil-landscape model can eventually be interpreted as catenary sequences of soil series over landscape. Ester *et al.* (2001) described three basic types of spatial relations that can be used in spatial data mining: topological, distance, and direction relations. As for the representation of soil-landscape models, both topological and direction relations are useful in generalizing soil distributions over landscape. Specifically, the topological relation of direct adjacency defines neighbors of a certain soil type. The direction relations of upslope and downslope define relative slope positions of soils over landscape. For example, the upslope neighbor and downslope neighbor of a certain soil type determine its position in a catenary sequence. For areas in which aspect plays a role in soil formation, directions other than up/down slope should also be considered. In our study area, where there's no indication of influence by aspect, we consider only the directions along slope.

Inductive learning algorithms require that the training dataset consist of labeled examples. A labeled example is a data record that contains values of all relevant variables (or 'features') and the classified category (or 'label') (Mitchell 1997). In our study, 'features' are the relevant environmental variables and spatial attributes, and 'label' is the soil type. Therefore, a labelled example is a pixel with the known soil category, a list of environmental conditions and spatial neighbors. Two kinds of labelled examples were prepared. The first kind was randomly drawn from the area and is for evaluating the different learning algorithms. This set is referred to as the unrectified sets. The second kind was drawn from the histogram modes of environment variables, as described in §3.2, and is for evaluating the extracted soil-landscape model. This kind is referred to as the 'rectified sets'.

5. Learning algorithms

We started by examining the learning accuracy of the See5 program with our preliminary soil-landscape database. Learning accuracy means how well the decision tree can approximate the soil map, regardless of the noise that might be present in the map. In order to measure the accuracy of the decision-tree learning algorithm, two other well-defined learning algorithms—Naïve Bayes and neural network backpropagation—were implemented on the same data for comparisons. A 10-fold cross validation was conducted for this purpose. Specifically, a set of 416 unrectified labeled examples was randomly drawn from the study area. These examples were then randomly divided into ten bins. Training was done ten times on nine of the bins, with one bin left out as a test set. The ten test set accuracies were measured for each of the three algorithms, and *t-test*s conducted between See5 and Naive Bayes, and between See5 and neural network.

Using the standard See5 algorithm, decision trees were constructed from the example sets. A typical output is shown in figure 7. The tree has a root attribute **Bedrock**, from which we can trace down to a final classification of soil type along the dotted lines. The number x/y in the parentheses after each soil type indicates the number of examples labelled as this class in the training set (x) and the number of examples that are incorrectly classified using this tree structure (y). If only one number is listed, all examples are correctly classified as that category. For example, the first path in the decision tree in figure 7 basically says: if **Bedrock** is *Oneota*, **slope gradient** is less than or equal to 13.17%, the soil type would be *Valton*, and there are 10 examples of this soil type in the training set, which are all correctly classified using this tree structure.

Figure 8 shows the accuracies of See5, Naive Bayes, and neural network algorithms from the ten-fold cross validation experiment. From the results, it is observable that the mean accuracy of the See5 algorithm is apparently higher than that of the Naive Bayes algorithm, but slightly lower than that of backpropagation. Furthermore, Paired Student *t-test*s are conducted under a 95% confidence level. The confidence interval of the relative performance for See5 versus neural network is 0.004 ± 0.052, or $(-0.048, 0.056)$. That of See5 versus Naïve Bayes is -0.1 ± 0.061 or $(-0.161, -0.039)$. From this result we can conclude that under a confidence level of 95%, the See5 algorithm is significantly better than Naive Bayes, while there's no

```
Bedrock = Oneota:
:...Slope <= 22.56:
:   :...Slope <= 13.17: Valton (10)
:   :   Slope > 13.17: Lamoille (10)
:   Slope > 22.56:
:   :...Planform <= -99.9: Dorerton (10)
:       Planform > -99.9: Elbaville (10)
Bedrock = Glauconite:
:...Slope > 28.28: Urne (10)
:   Slope <= 28.28:
:   :...Slope <= 13.08: Greenridge (10/1)
:       Slope > 13.08: Norden (9)
Bedrock = Wonewoc:
:...Planform <= -224.7: Boone (11/1)
:   Planform > -224.7:
:   :...Elevation <= 954.1: Elevasil (10/1)
:       Elevation > 954.1: Hixton (9/1)
Bedrock = Jordan:
:...Profile <= -53.2: Gaphill (5)
:   Profile > -53.2:
:   :...Slope <= 19.85: Rockbluff (9/1)
:       Slope > 19.85:
:       :...Profile <= 4.9: Rockbluff (3/1)
:           Profile > 4.9: Gaphill (3)
Bedrock = Alluvium:
:...Slope <= 5.59:
    :...Elevation <= 850: Orion (10)
    :   Elevation > 850: Kickapoo (10)
    Slope > 5.59:
    :...Elevation > 1004.1: Churchtown (7)
        Elevation <= 1004.1:
        :...Elevation <= 947.9: Council (7)
            Elevation > 947.9:
            :...Slope <= 19.96: Churchtown (2)
                Slope > 19.96: Council (4/1)
```

Figure 7. A decision tree built from training samples.

significant difference between the performance of See5 and that of neural network. Therefore, in terms of learning accuracy, the See5 and the neural network back-propagation algorithms exhibit similar performances. As mentioned in §3.3, a significant limitation of the neural network is that the representation is barely comprehensible. In order to interpret a neural network, Craven and Shavlik (1996)

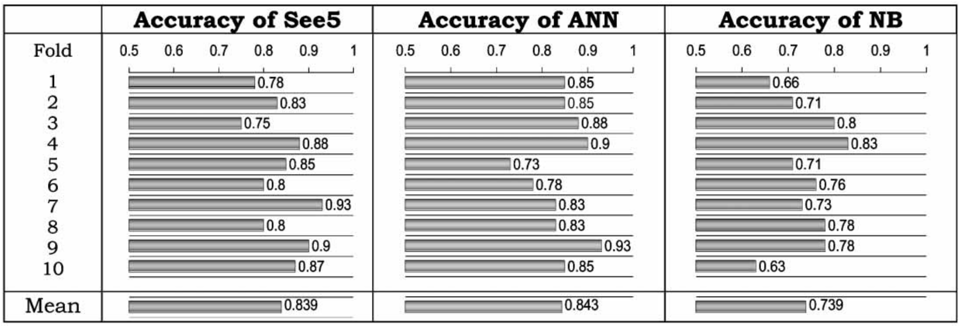


Figure 8. Accuracies of the See5, neural network, and Naive Bayes algorithms in a 10-fold cross validation.

used an additional decision tree learning program to re-represent the network structure and weights with a decision tree. Given that the See5 results are already easily comprehensible without extra effort, it is concluded to be a better approach to extracting knowledge of soil-landscape relationships.

6. Knowledge discovery results and discussion

In this section we examine the accuracy of the extracted knowledge and the effect of the noise reduction effort as outlined in §3.2. In order to measure accuracy, we obtained an expert-defined test set. Specifically, the soil expert who drew the map was provided with the orthophoto in 3-D view on a computer screen. He then digitized on screen a set of sample points that he believes are consistent with his understanding of the soil-landscape relationships over the area. These sample points were then recorded, and their values for the associated environmental variables were attached to generate a test set. Since the expert was asked to give only typical points that are consistent with his knowledge, this test set was expected to represent the expert knowledge instead of the map product.

6.1. *The impact of different sampling options*

Sample sets were constructed based on different sampling options in data preprocessing—either by taking the entire mode(s) or by randomly sampling from the mode(s), either using the union operation or using the intersection operation to pool modal samples. Each sample set was then used to derive a decision tree to investigate how the decision tree behaves in response to the different sampling options. Table 1 shows the sampling parameters of 32 different sample sets along with their resulting tree accuracies on the expert-defined test set. The parameter r determines the number of intervals in the histogram of an environmental variable (see §3.2), while the parameter N indicates the number of samples randomly extracted from each mode interval.

It is observable that the results from all intersection sample sets are apparently worse than those from the union sets. Actually the decision trees built from many of the intersection sets are not even complete, because for some of the soil types there are no training samples. Because the intersection sample set contains only pixels falling into multiple modes, the number of samples for different soil types differs. It often results in incomplete sample sets due to the fact that some soil types

Table 1. Rectified sample sets: sampling parameters and the resulting accuracies.

Sampling parameters		Pooled via INTERSECTION		Pooled via UNION	
		Sample size	Accuracy	Sample size	Accuracy
$r=3$	$N=5$	29	0.43	286	0.83
	Entire mode	67	0.40	1298	0.83
$r=5$	$N=5$	28	0.49	289	0.86
	$N=10$	103	0.52	517	0.83
	Entire mode	111	0.40	1777	0.83
$r=10$	$N=10$	67	0.65	564	0.83
	$N=20$	246	0.71	940	0.86
	Entire mode	284	0.60	2881	0.86
$r=20$	$N=20$	178	0.71	1051	0.86
	Entire mode	699	0.77	4728	0.86

Table 2. Accuracies of decision trees derived from rectified sample sets.

	Set1	Set2	Set3	Set4	Set5	Set6
Size	286	272	289	285	517	504
r	3	3	5	5	5	5
N	5	5	5	5	10	10
Accuracy	**0.83**	**0.80**	**0.86**	**0.86**	**0.83**	**0.80**

Table 3. Accuracies of decision trees derived from unrectified sample sets.

Fold	1	2	3	4	5	6	7	8	9	10	**Mean**
Accuracy	0.80	0.71	0.74	0.71	0.80	0.74	0.74	0.74	0.74	0.74	**0.75**

occupy only a very small portion of the area. Even if the sample sets are complete with bigger sample sizes, the severely uneven allocation of samples among the soil series can introduce bias into the training process, thus impairing the learning accuracy. Table 1 also shows that, although the sample sizes are usually much bigger if we take the entire mode rather than randomly sample from it, there's no significant difference between these two options in terms of test set accuracy under the union strategy. Given that the See5 algorithm is not computationally intensive, we say that taking the entire mode or sampling a few pixels from the mode basically has a negligible impact on the results.

6.2. *The effect of data preprocessing*

In order to examine the effect of our data preprocessing strategy, we first used six rectified sample sets generated through data preprocessing to derive decision trees, and then fed the expert-defined test set to determine the accuracy of each decision tree. The results of these tests are shown in table 2. The expert-derived sample set was also fed to the decision trees generated using the See5 algorithm based on the 10 fold unrectified example sets. The accuracy for each of the 10-fold decision trees is listed in table 3. The mean accuracy of the 10-folds is 0.75. Since the examples used to derive the 10 decision trees were not preprocessed using the strategy described in §3.2, we treat the above accuracy (0.75) as the knowledge discovery accuracy without data preprocessing.

Tables 2 and 3 show that the accuracy of each of the decision trees derived using the rectified sample sets is higher than the mean accuracy from the sample sets without data preprocessing (see figure 9). This provides one piece of evidence that the data-preprocessing step is effective in reducing noise and outliers from the original map.

6.3. *Comparison of the decision tree results with documented expert knowledge*

A decision tree representation can be easily converted into rule sets by traversing all paths of the tree. The rules can then be compared with the environmental descriptions directly obtained from the local soil expert for validation. Table 4 shows part of such a comparison between the result from a rectified sample set and the documented expert knowledge. The comparison reveals that consistency is high between the environmental descriptions for soil series constructed from the

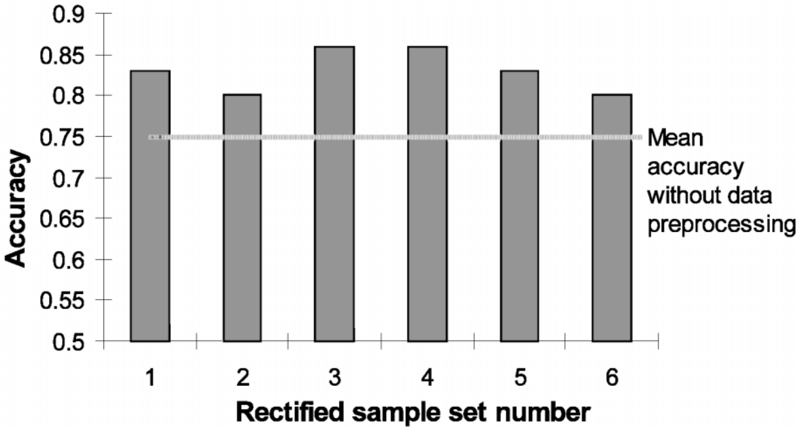


Figure 9. Accuracies measured on the expert-defined test set.

decision tree result and those acquired directly from the local soil expert. The best consistency is found for soil series developed on bedrock *Oneota* and *Glauconite*. Results for soil series developed on *Jordan* sandstone are not as good. The reason is that *Jordan* sandstone occupies only a very small area in the watershed, where, due to the limited spatial resolution, soil polygons on the soil map may not effectively capture the expert knowledge. For the continuous environmental variables, the decision tree program has generated breaks automatically. From the comparison in table 4, we see that most of these algorithm-derived breaks are fairly close to those provided by soil experts. Therefore, we can say that the knowledge discovered from the soil map has effectively approximated the knowledge that was used to create the soil map by soil experts.

We also compared the previous results, obtained without data preprocessing, to the documented expert knowledge. The breaks of the continuous features are less

Table 4. Comparisons of decision tree results with documented expert knowledge.

Soil series	Environmental Variable	Tree result	Expert Knowledge
Valton	**Geology**	Oneota	Oneota
	Elevation	>1298.68	>1300
	Gradient	<=12.57%	<12%
Dorerton	**Geology**	Oneota	Oneota
	Elevation	<1298.68	1150–1250
	Profile Curvature	Linear-convex	Linear-convex
	Planform Curvature	Convex	Convex
Elbaville	**Geology**	Oneota	Oneota
	Elevation	<1298.68	1150–1250
	Profile Curvature	Slightly convex-concave	Concave-slightly convex
	Planform Curvature	Slightly convex-concave	Concave-linear
Greenridge	**Geology**	Glauconite	Glauconite
	Gradient	<=14.37%	4–12%
Norden	**Geology**	Glauconite	Glauconite
	Gradient	14.37–28.84%	12–30%
Urne	**Geology**	Glauconite	Glauconite
	Gradient	>28.84%	>30%

Table 5. Comparisons of decision tree results with documented expert knowledge.

Soil series	Environmental Variable	Tree result (without data preprocessing)	Expert Knowledge	Tree result (mode sampling)
Valton	**Geology**	Oneota	Oneota	Oneota
	Elevation	>1194.95	>1300	>1298.68
	Gradient	<=9.86%	<12%	<=12.57
Lamoille	**Geology**	Oneota	Oneota	Oneota
	Elevation	>1194.95	>1250	>1298.68
	Gradient	>9.86%	12–20%	>12.57%
Dorerton	**Geology**	Oneota	Oneota	Oneota
	Elevation	N/A	1150–1250	<1298.68
	Gradient	>27.37%	>30%	N/A
	Profile Curvature	N/A	Linear-convex	Linear-convex

accurate than those from the rectified sample sets, due to the inclusion of noises (table 5).

6.4. *Effects of spatial information*

For our case study on the Raffelson watershed, the consideration of two kinds of spatial information is suggested by soil experts in addition to the basic environmental variables, as described in §4. In order to examine the effect of including spatial information in the knowledge discovery process, we use the same expert-defined test set to measure the changes in accuracies.

First, the three variables that portray spatial relations of soil-formative environments are added to the database. They are distance to streams, topographic wetness index, and percentage of colluvium from competing bedrocks. Ten rectified sample sets are generated to derive decision trees using See5. The test set accuracies of the ten resulting trees are compared to the accuracies obtained using only the basic environmental variables. Table 6 shows the result of this experiment, which indicates that the mean accuracy increases from 0.83 to 0.865 with the addition of the new variables. A paired *t-test* under a 95% confidence level gives the interval of relative performance (0.024, 0.056), indicating a significant improvement. In the

Table 6. Decision tree accuracies with and without using spatial information.

Run	Accuracy with preliminary database	Accuracy after spatial topographic variables added	Accuracy after spatial neighbor variables added
1	0.83	0.89	0.89
2	0.80	0.83	0.89
3	0.86	0.86	0.91
4	0.86	0.86	0.89
5	0.83	0.86	0.89
6	0.80	0.89	0.89
7	0.83	0.86	0.91
8	0.86	0.91	0.91
9	0.83	0.83	0.89
10	0.80	0.86	0.86
Mean	**0.83**	**0.865**	**0.893**

decision tree results, we see that two of the three newly added variables appear as tree nodes to separate soil series. They are wetness index and percentage of colluvium from competing bedrocks. It is the presence of these two features in the decision tree that accounts for the increase of accuracy. On the other hand, the other variable, distance to streams, does not affect the knowledge discovery performance in that it is found not to play a role in the soil-landscape relationships in this area. A merit of decision tree inductive learning is that it does not necessarily use all the given features to construct the decision tree result, but adaptively selects the most relevant variables. This is especially important when applying the process to other areas, where the actual soil-landscape model behind the soil map is unknown, and the environmental variables used by the soil experts who created the map are untraceable. It is then necessary to include a wide range of variables in the database to examine various potential factors in the soil-landscape relationships.

Last, we added another two variables into the above database to portray the topological relations between soil types. The two variables are the upslope neighbor and downslope neighbor of a given soil type. See5 was run on ten rectified sample sets to derive decision trees, and the test set accuracies are listed in table 6. The table shows that the mean accuracy has further increased from 0.865 to 0.893. A paired *t-test*, again, confirms the significance of the improvement, with a confidence interval (0.010, 0.036). An examination of the decision tree results shows that a considerable portion of tree nodes are now associated with these two variables. It is thus evident that the inclusion of spatial neighborhood information has led to a significant improvement of the knowledge discovery performance. Furthermore, the explicit spatial relationships between different soil types make it possible to create catenary sequences of soil series, which are commonly used in soil survey to illustrate soil-landscape models. For example, figure 10 shows part of a resulting decision tree. The tree branch on bedrock *Oneota* can eventually be generalized to the catenary sequence displayed in figure 11. Specifically, when the decision tree says the upslope neighbour of a certain soil type is 'None', it usually denotes that this soil type develops on ridge tops. Similarly, if one's downslope neighbor is 'None', this soil may be at the lowest drainage ways. When two soil types appear to be each other's downslope (or upslope) neighbours, they are most likely at similar

```
Bedrock = Oneota:
:...Elevation <= 1304.62:
:   :...downNeighbor = Dorerton: Elbaville (5)
:   :   downNeighbor = Elbaville: Dorerton (45/1)
:   :   downNeighbor = Churchtown: Elbaville (25)
:   Elevation > 1304.62:
:   :...upNeighbor = Valton: Lamoille (38)
:       downNeighbor = Elbaville: Lamoille (2)
:       upNeighbor = None: Valton (41/1)
Bedrock = Alluvium:
:...OffGlauconite <= 15.57835:
    :...Slope <= 0.327846: Churchtown (50/6)
    :   Slope > 0.327846: Elbaville (13)
    OffGlauconite > 15.57835:
    :...Elevation <= 860.18: Orion (47/5)
        Elevation > 860.18:
        :...Wetness <= 6.705073: Council (47/6)
            Wetness > 6.705073: Kickapoo (31/1)
```

Figure 10. Part of a decision tree result with spatial neighbour information.

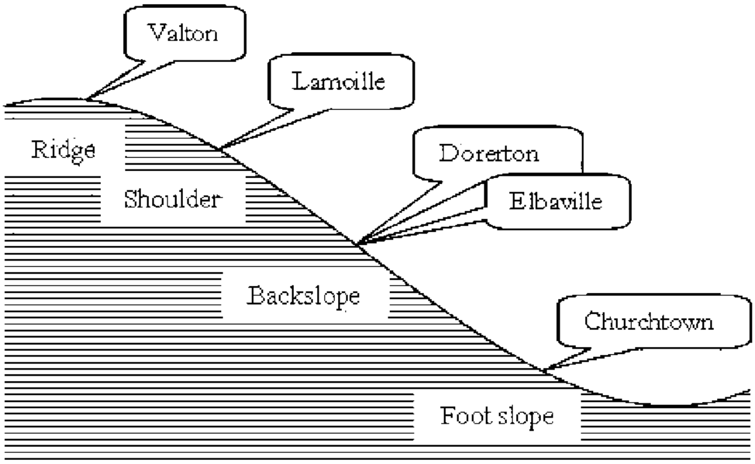


Figure 11. Catenary Sequence on *Oneota*.

slope positions and separable only by other environmental variables. Soil types *Elbaville* and *Dorerton* are examples of this case. By looking at previous decision tree results generated without using the spatial neighborhood information, we see that *Dorerton* develops on convex positions, while *Elbaville* is more related to linear and concave curvatures.

7. Conclusions and future efforts

This paper presented a knowledge discovery procedure to extract knowledge from soil maps. It shows that inductive machine learning algorithms can be applied to extract useful knowledge from previously underutilized soil maps. Previous research has demonstrated the success of decision tree learning algorithms in soil data modeling. Eklund *et al.* (1998) used decision tree induction in a knowledge-based system for secondary soil salinization analysis. Moran and Bui (2002) investigated the use of decision tree-generated rules for soil classification. Although that study showed that decision tree induction can be used to model existing soil maps, it is more desirable to recover true expert knowledge from the error-prone soil maps. Therefore, our study has paid specific attention to the reduction of representation noise in soil maps. Furthermore, although previous studies have used decision tree induction to model soils, the underlying soil-landscape model was not explicitly extracted and represented. We argue that the soil-landscape model is the key concept in soil survey practices. Once explicitly represented and documented, knowledge about the local soil-landscape model can be used by both inexperienced soil experts and automated soil inference systems for soil survey updates. In our study, the decision-tree learning algorithm See5 is found to be suitable for extracting descriptive knowledge of soil-landscape relationships from a soil map and the associated environmental database. The results are both comprehensible and accurate when compared to those obtained using other learning algorithms. The discovered soil-landscape model can be represented in three different ways: decision tree or production rules, soil descriptions, and catenary sequences.

In the knowledge discovery process, the data-preprocessing step, like the learning algorithm itself, is found to play a very important role. Since the decision-tree learning algorithm is a general algorithm that is suitable for processing data from various domains, it is important that data selection and data preprocessing are done with the aid of prior understanding of the application domain, so that data can be properly prepared to exclude noise and to make the samples more representative. The preprocessing method of sampling only modal pixels according to environmental histograms is found to be effective since it allows the selection of typical samples that represent the central concepts of soil types. It helps to reduce generalization bias of the algorithm and to avoid overfitting toward noisy data, thus significantly improving the knowledge discovery performance. When pooling samples from different environmental modes, the union operation proves to be more effective than intersection, since it maintains an even distribution of samples over different soil types to the greatest degree. This helps avoid training bias in the decision tree learning process.

We also showed that spatial relationships and other spatial variables can be incorporated into the proposed knowledge discovery procedure, and demonstrated that the incorporation of such spatial information further improves the accuracy of the extracted knowledge. Use of spatial neighborhood information also results in a more comprehensible knowledge representation in the form of catenary sequences. In geographical data mining, it is generally recommended to explicitly involve spatial dependency and heterogeneity (Miller and Han 2001). However, some spatial variables are not directly interpretable to soil experts. Therefore, our current study considers only the spatial information that can be used to represent the soil-landscape model in a way with which soil experts are most familiar. Yet we expect that the inclusion of other spatial variables may lead to the discovery of new insights into the soil-landscape relationships rather than strictly being limited to those with which soil experts are familiar.

Our case study shows that the proposed knowledge discovery procedure applies successfully in the 'driftless area' of Wisconsin, where the soil map was created based on the knowledge of a local soil-landscape model. Although the concept of the soil-landscape model is widely adopted in soil survey practices, particularly in the USA, it should also be noted that there are soil maps that were not produced using soil-landscape models. The knowledge discovery procedure reported in this paper may not work for these maps.

In this paper we have discussed the applicability of using a knowledge discovery procedure to extract expert knowledge from soil maps. Our aim is to approximate the knowledge that soil experts used to create the map. However, soil mapping is an inherently subjective process. Soil experts build the local soil-landscape model based purely on their own experience and understanding. It is thus not guaranteed that the soil-landscape model developed by individual soil experts represents accurately the real soil-landscape relationships of the local area. In other words, two experts may come up with different soil maps for the same area. Therefore, there are indeed two levels of approximation: how well the extracted model approximates the soil expert's knowledge, and how well the expert's knowledge represents the actual soil-landscape relationships. Our goal in the current study is to recover the subjective expert knowledge from the error-prone soil maps, and so the study concerns only the first level of approximation.

Although the knowledge discovery procedure described in this paper was developed in the context of soil mapping, it has broad relevance to knowledge discovery from other natural resource maps, particularly maps of those natural resources which cannot be directly observed using remote sensing techniques, such as wildlife habitats and potential natural hazards. The distribution of these natural resources cannot be directly observed due to obscuring overstories and the high cost of collecting information on these resources at many locations across the landscape. Therefore, their distributions are usually inferred (or indirectly mapped) from other easily observable environmental conditions (Mulder and Corns 1996, Zhu 1999). The procedure presented in this paper can thus be applied to extract the relationships between the mapped natural resource and its environment.

Our future plan includes exploring more realistic knowledge representations, incorporating the extracted knowledge into an automated inference system, modeling spatial autocorrelation, and developing an interactive knowledge discovery tool to allow synchronous integration of human expert knowledge with map information. Specifically, soil properties vary continuously over space. Soil-landscape relationships are more appropriately modelled when the natural fuzziness or uncertainties are considered. We are investigating the derivation of fuzzy membership values during the construction of decision trees under the See5 framework based on information theory. Furthermore, boosting can be used to capture the uncertainties that are ignored by constructing only one decision tree from the training data. Another potential knowledge representation of the soil-landscape model is the Bayesian network, which naturally models uncertainties through probability.

The extracted knowledge eventually can be used to infer soils for soil survey update. In automated soil inference, information on spatial autocorrelation of soil-formative factors can be incorporated. Soil maps and fuzzy representations of the soil distribution can be created by automated soil inference. The product can be validated using field data to measure the second level of approximation (See section 7). Since soil inference is virtually a knowledge based process, it is desirable to involve human experts in the knowledge discovery and soil inference process. An interactive data mining tool is under development to allow the expert to visualize the terrain in a 3-D view, direct the data preprocessing, choose knowledge representation, and control the use of different variables.

Acknowledgments

This study was funded by a grant from USDA-NRCS under Agreement No. 69-5F48-900186. Duane Simonson from USDA-NRCS Wisconsin Office served as the soil expert and provided the knowledge of soil-landscape relationships over the study area. His assistance in this project is greatly appreciated.

References

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

Bruin, S. D., Wielemaker, W. G., and Molenaar, M., 1999, Formalisation of soil-landscape knowledge through interactive hierarchical disaggregation. *Geoderma*, **91**, 151–172.

Craven, M., and Shavlik, J., 1997, Understanding time-series networks: a case study in rule extraction. *International Journal of Neural Systems*, **8**, 373–384.

Deka, B., Sawhney, J. S., Sharma, B. D., and Sidhu, P. S., 1995, Soil-landscape relationships in Siwalik hills of the semiarid tract of Punjab, India. *Arid Soil Research and Rehabilitation*, **10**, 149–159.

DOMINGOS, P., and PAZZANI, M. J., 1997, On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, **29**, 103–130.

EKLUND, P. W., KIRKBY, S. D., and SALIM, A., 1998, Data mining and soil salinity analysis. *International Journal of Geographical Information Science*, **12**, 247–268.

ESPOSITO, F., MALERBA, D., and SEMERARO, G., 1997, A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, 476–491.

ESTER, M., KRIEGEL, H. P., and SANDER, J., 2001, Algorithms and applications for spatial data mining. In *Geographic Data Mining and Knowledge Discovery*, edited by H. J. Miller and J. Han, (New York, NY: Taylor & Francis), pp. 160–187.

FASSNCHT, K. S., GOWE, S. T., MACKENZIE, M. D., NORDHEIM, E. V., and LILLESAND, T. M, 1997, Estimating the leaf area index of north central Wisconsin forests using the Landsat Thematic Mapper. *Remote Sensing of Environment*, **61**, 229–245.

FAYYAD, U., PIATETSKY-SHAPIRO, G., and SMYTH, P., 1996, From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining*, edited by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Menlo Park, CA: AAAI/MIT Press), pp. 1–34.

GAHEGAN, M., 2000, On the applications of inductive machine learning tools to geographical analysis. *Geographical Analysis*, **32**, 113–139.

GLINKA, K. D., 1927, *The Great Soil Groups of the World and their Development* (Ann Arbor, MI: Edwards Bros.).

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

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

JENNY, H., 1961, *E.W. Hilgard and the Birth of Modern Soil Science* (Berkeley, CA: Farallo Publication).

KOPERSKI, K., HAN, J., and ADHIKARY, J., 1999, Mining knowledge in geographic data. Accessed at URL: http://db.cs.sfu.ca/sections/publication/kdd/kdd.html.

MALERBA, D., ESPOSITO, A. L., and LISI, F. A., 2001, Machine learning for information extraction from topographic maps. In *Geographic Data Mining and Knowledge Discovery*, edited by H. J. Miller and J. Han, (New York, NY: Taylor & Francis), pp. 291–314.

MCLEOD, M., RIJKSE, W. C., and DYMOND, J. R., 1995, A soil-landscape model for close-jointed mudstone, Gisborne-East Cape, North Island, New Zealand. *Australian Journal of Soil Research*, **33**, 381–396.

MCSWEENEY, K., GESSLER, P. E., SLATER, B. K., PETERSEN, G. W., HAMMER, R. D., and BELL, J. C., 1994, Towards a new framework for modeling the soil-landscape continuum. *Factors of Soil Formation: A Fiftieth Anniversary Retrospective*. SSSA Special Publication, **33**, 127–143.

MENNIS, J. L., PEUQUET, D. J., and QIAN, L., 2000, A conceptual framework for incorporating cognitive principles into geographical database representation. *International Journal of Geographical Information Science*, **14**, 501–520.

MILLER, H. J., and HAN, J., 2001, Geographic data mining and knowledge discovery: an overview. In *Geographic Data Mining and Knowledge Discovery*, edited by H. J. Miller and J. Han, (New York, NY: Taylor & Francis), pp. 3–32.

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

MITCHELL, T. M., 1997, *Machine Learning* (New York: McGraw Hill).

MOORE, I. D., GESSLER, P. E., NIELSEN, G. A., and PETERSON, G. A., 1993, Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, **57**, 443–452.

MORAN, C. J., and BUI, E. N., 2002, Spatial data mining for enhanced soil map modeling. *International Journal of Geographical Information Science*, **16**, 533–549.

MULDER, J. A., and CORNS, I. G. W., 1996, Knowledge based ecosystem prediction: field testing and validation. In *GIS Applications in Natural Resources, 2*, edited by M. Heit, H. D. Parker and A. Shortreid (Fort Collins: GIS World, Inc.), pp. 392–398.

MURRAY, A. T., and ESTIVILL-CASTRO, V., 1998, Cluster discovery techniques for exploratory spatial data analysis. *International Journal of Geographical Information Science*, **12**, 431–443.

NEMANI, R. R., PIERCE, L. L., RUNNING, S. W., and BAND, L., 1993, Forest ecosystem processes at the watershed scale: Sensitivity to remotely sensed leaf area index estimates. *International Journal of Remote Sensing*, **14**, 2519–2534.

O'CALLAGHAN, J. F., and MARK, D. M., 1984, The extraction of drainage networks from digital elevation data. *Computer Vision, Graphics and Image Processing*, **28**, 323–344.

QUINLAN, J. R., 1986, Induction of Decision Trees. *Machine Learning*, **1**, 81–106.

QUINLAN, J. R., 1993, *C4.5 Programs for Machine Learning* (San Mateo, CA: Morgan Kaufmann).

QUINLAN, J. R., 2001, *See5: An Informal Tutorial.* Accessed at URL: http://www.rulequest.com.

RUSSELL, S., and NORVIG, P., 1995, *Artificial Intelligence: A Modern Approach*. (New York: Prentice Hall).

WRIGHT, R. L., 1996, An evaluation of soil variability over a single bedrock type in part of southeast Spain. *Catena*, **27**, 1–24.

ZADEH, L. A., 1965, Fuzzy Sets. *Information and Control*, **8**, 338–353.

ZHU, A. X., 1996, A similarity model for representing soil spatial information. *Geoderma*, **77**, 217–242.

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

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

ZHU, A. X., HUDSON, B., BURT, J. E., LUBICH, K., and SIMONSON, D., 2001, Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal*, **65**, 1463–1472.