

Machine Learning and Knowledge Acquisition Techniques for Natural Resources Inventory

A-Xing Zhu, Ph.D.
Department of Geography
University of Wisconsin-Madison
Madison, Wisconsin 53706
axing@geography.wisc.edu

Overview

The knowledge on relationships between natural resources and their environmental conditions is needed for predictive mapping of natural resources (such as soils) and for mapping the susceptibility of natural hazards (such as landslides). This knowledge often exists in the form of human expertise or in raw data (point observations and map) forms. This workshop introduces the techniques for extracting this knowledge from these sources using knowledge elicitation and machine learning techniques. The following sections cover personal construct-based interview techniques, neural networks, case-based reasoning, and data mining techniques.

1. Knowledge Acquisition: the Personal Construct-Based Approach

1.1 References:

- KELLY, G.A., 1955, *The Psychology of Personal Constructs* (New York: Norton).
- KELLY, G.A., 1970, A brief introduction to personal construct theory. In *Perspectives in Personal Construct Theory*, edited by D. Bannister (London: Academic Press), pp. 1-29.
- Zhu, A.X., 1999. A personal construct-based knowledge acquisition process for natural resource mapping using GIS. *International Journal of Geographic Information Science*, Vol. 13, No. 2, pp. 119-141.

1.2 Personal Construct Theory

In general, personal construct theory (Kelly, 1955; Kelly, 1970) assumes that people typically use cognitive dimensions (termed “constructs”) to learn and evaluate their experience. Each construct, by definition, represents a single bipolar distinction. For example, a bird ecologist might use the construct “hot/cold” to describe the temperature requirement of birds; a person may use the construct “wet/dry,” among others, to distinguish climate types. The underlying relation between the alternative poles of any construct is contrariety and no construct can be understood fully without considering the meaning of both poles. A construct is a basis for making a distinction and is a dichotomous reference axis in a person’s psychological space.

Under Kelly's personal construct theory, the accumulation of a resource expert's knowledge about the relationships between a resource and its environment can be considered as a process of constructing that person's psychological space about the relationships. For example, a soil scientist may accumulate knowledge about the relationship between a specific soil type and its environment by formulating constructs (such as "steep/flat," "south facing/north facing") and locating the intersections of these constructs to define the environmental niche or niches under which the soil exists. Under this notion, acquiring knowledge about relationships between a resource and its environment becomes a process of defining relevant (proper) constructs (both poles and intervals) and locating the intersections of these constructs for various classes of that resource.

1.3. The Knowledge Acquisition Process

Based on the assumption that resource experts acquire and organize knowledge in a way theorized by personal construct theory, a knowledge acquisition process consisting of two phases was developed (Fig. 1). The first, or *iteration phase*, consists of many knowledge acquisition iterations. Each iteration consists of five sessions: **1**) the preparation session, **2**) the resource-environment key development session, **3**) the resource-environment description session, **4**) the key and description comparison session, and **5**) the optimality curve definition session. These five sessions are tightly connected and form a structured interview. In Session 1, constructs are defined. Session 2, 3 and 4 are used to define the intersections of these constructs. Session 5 is used to extract *fuzzy membership functions* describing the relationships. Results from different iterations are then compared and analyzed in the second phase called the *knowledge verification phase*.

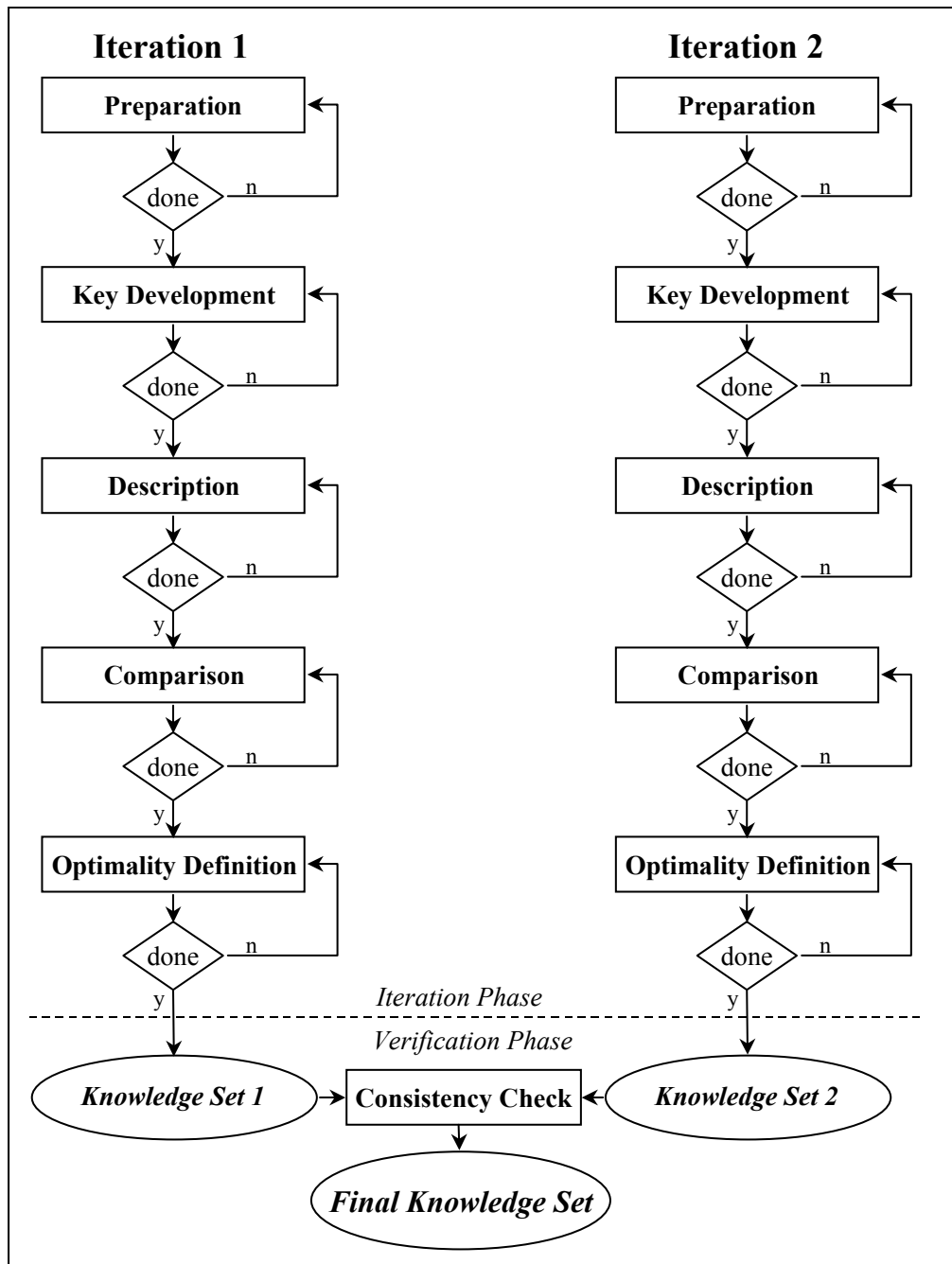


Figure 1: A personal construct-based knowledge acquisition process

2. Soil Inference Under Fuzzy Logic: A Neural Network Approach

2.1 References:

Masters, Timothy, 1993. *Practical Neural Network Recipes in C++*, Academic Press, pp. 77-116.

Zhu, A.X., 2000. Mapping soil landscape as spatial continua: the neural network approach. *Water Resources Research*, 36, 663-677.

2.2 Basic Concepts and Issues:

Artificial neural networks (ANN) are a form of computing motivated by the functioning of biological neural networks. An ANN solves a problem by first developing a memory associating a large number of input patterns with a set of resulting outputs through training on examples, and then by applying this association to produce an output when given an input pattern. There are many forms of ANN. One of them is called multilayer feedforward networks (MFN). This workshop focuses on this type of ANNs.

2.2.1 Structure: node and links

A MFN is made of many processing elements (neurons or nodes). These neurons are usually arranged in layers: an input layer, an output layer, and one or more layers in between called hidden layers (Figure 2). The neurons in one layer are connected to the neurons in the next layer with different strengths of connection. The strength of connection is referred to as a weight.

The structure of a MFN (the numbers of layers and the number of neurons in each layer) is problem specific and is one of the important issues in applying ANN techniques.

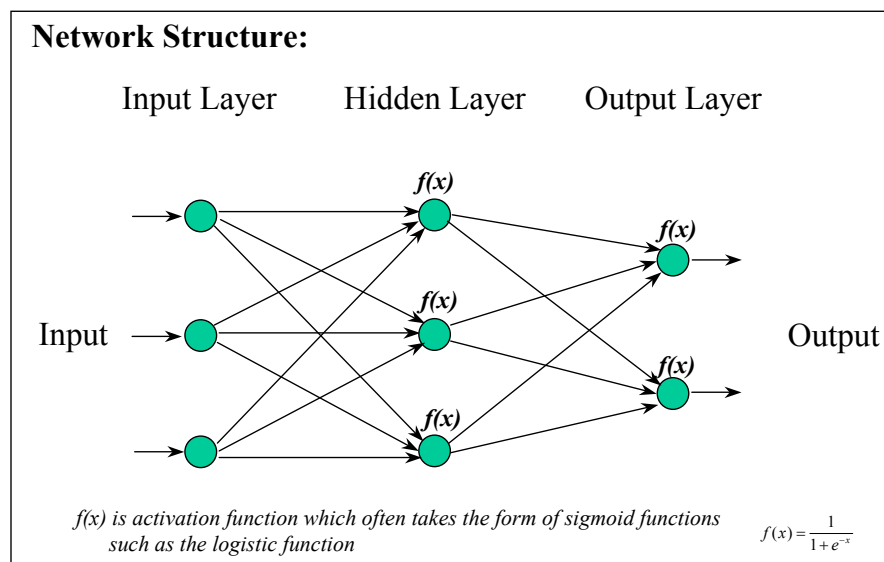


Figure 2: A three layer feedforward network

2.2.2 Operation

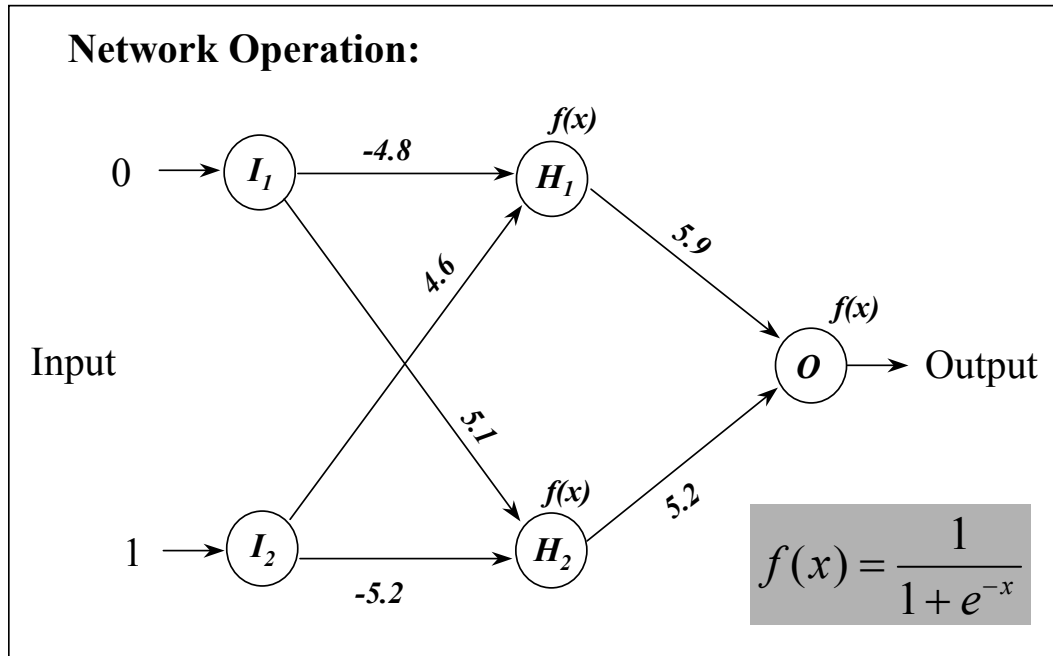


Figure 3: Example of ANN operation

2.3 Using NN

2.3.1 The size of the network:

number of inputs (m) = number of variables

number of outputs (n) = number of classes

initial number of hidden nodes = $\text{sqrt}(m \cdot n)$

2.3.2 Training the network:

Network training (network learning) is a process of determining a set of weights that will produce the best possible input/output mapping. Most network training employs a supervised approach under which the network is presented with a set of input patterns and a set of corresponding desired outputs (together referred to as training data). The training process starts by initializing all weights to small non-zero values. Then, training samples are presented to the network one at a time to produce corresponding results. A measure of error between the network outputs and the desired outputs is computed and weights are updated to reduce error. Many iterations or epochs (from presenting training samples to measuring error and to updating weights) may be required before a network reaches a given level of accuracy.

1) Purpose of training

a) determining the structure

b) determining the weights

2) Training data

a) training sample size:

minimum number of training cases = $2(m+1)n$

reasonable number of training cases = $4(m+1)n$

b) training cases should be representative

c) testing cases should be different from training

testing cases should be independent of training

if the test fails, a new set of testing cases should be used for further testing

3) How to determine the optimal structure and weight configuration

The number of hidden neurons should be the number at which the network performed well (with a high accuracy and low training error). There are several possible numbers of hidden neurons as shown in Figure 4. The number of hidden neurons could be 6 or 10. The final decision among these possible numbers was made based on the test accuracy for individual soil categories over the validation set. This is justified that the networks were trained to learn from the training data set. It is possible that with some structures (defined by the number of hidden neurons in this case) the network could be tuned to the training data set too much (over-fitting). As a result, the network would not generalize well to a different data set. Thus, the selection of number of hidden neurons should be based on more heavily on the test accuracy from the validation data set while still considering the training error.

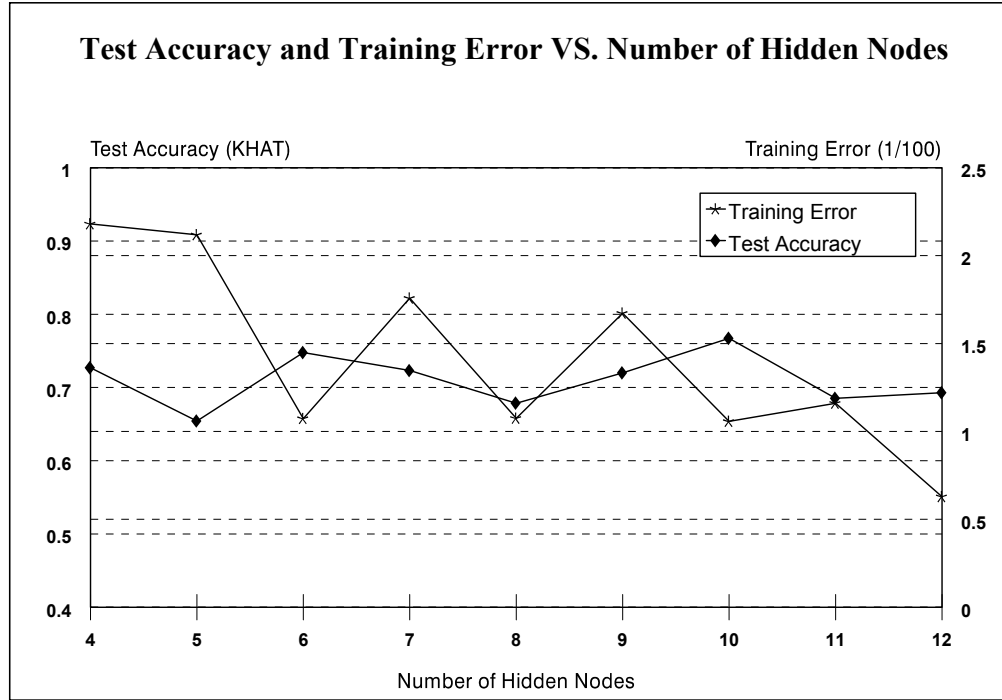


Figure 4: No of hidden nodes vs. test accuracy and training error

3. Case-Based Reasoning Approach to Soil Mapping

3.1 References:

- Kolodner, J. 1993. Case-Based Reasoning. Morgan Kaufmann Publishers, San Mateo, CA.
- Shi, X., A.X. Zhu, J.E. Burt, F. Qi, and D. Simonson, 2003. A case-based reasoning approach to fuzzy soil mapping. *Soil Science Society of America Journal*, In press.

3.2 Basic Idea:

Case-based reasoning (CBR) refers to a concept and the corresponding technology in the knowledge-based system discipline. It uses the knowledge represented in specific cases to solve a new problem. A case in CBR contains two basic parts: the description of the problem and the solution of the problem. The description part is for evaluating the similarity between the case and a new problem. If the case and the new problem are similar enough, then the solution part of the case is used to solve the new problem.

The applicability of CBR in soil mapping can be justified through examining the two assumptions of CBR. The first assumption is that cases are capable of representing domain experts' knowledge. It has been documented that a large part of soil scientists' knowledge can be subsumed to *tacit knowledge*, which is learned from practical work, especially from field experiences. The tacit knowledge of soil scientists is often the most important knowledge in soil mapping, yet is the most difficult knowledge to learn by a new soil scientist and by the computer, because it is usually hard to articulate and generalize. The main reason for this is that the soil-forming process can be highly complicated and has not been fully understood. As a result, a large part of the knowledge of soil-environment relationship is empirical and exists in the form of "cases" or is associated with particularly locations ("tacit locations" or "tacit points").

The second assumption of CBR is that a new problem can be solved by referring to similar cases. The concept of *landscape unit* in traditional soil survey and mapping provides a basis for using the similarity-based method to conduct soil inference. Particularly, two basic characteristics of *landscape unit* are most relevant to applying CBR to this field: "Generally, the more similar two units are, the more similar their associated soils tend to be; conversely, dissimilar units tend to have dissimilar soils"; and "Same or similar units can occur again and again in space". This means for determining what soil at a given location one just needs to compare the local landscape conditions with the conditions of the typical landscape units for a set of prescribed soil classes.

3.3.The Methodology:

The goal of soil inference under fuzzy logic is to derive, for every location over the mapping area, the fuzzy membership values of all the soils found in the area. With the CBR method, these fuzzy membership values will be computed based on the similarity between the environmental configuration of the given location and that of each *tacit*

point. For example, suppose that the soil scientist knows that there are three soils, A, B, and C, in the mapping area, and that the soil scientist has created a group of *tacit points* (with full fuzzy memberships) for each of these three soils, the process for deriving fuzzy membership is as follows: first, for a location X in the mapping area, the inference engine obtains the data on environmental conditions, such as elevation, slope gradient, slope aspect, etc., from the GIS database; next, the inference engine calculates the similarities between the environmental configuration at X and those of the *tacit points* for soil A; and then, these similarity values will be integrated to derive the fuzzy membership for soil A at X . This process is repeated for soils B and C to obtain their fuzzy membership values at X .

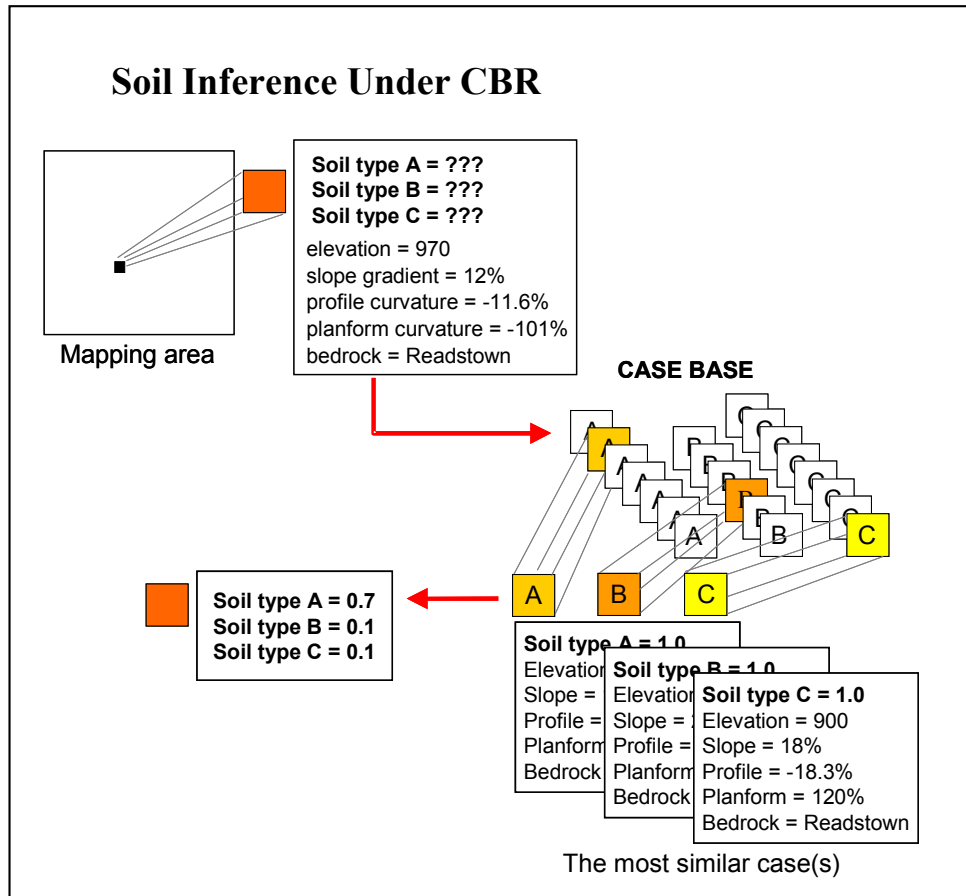


Figure 5: The process of soil inference using CBR

The technical details of computing the fuzzy membership for a certain soil at a specific location can be represented with a general equation:

$$s_{ij}^k = T_{ij}^k \left\{ P_{ij}^t [E_{ij}^{v,t} (e_{ij}^v, e^{v,t})] \right\} \quad [2]$$

where s_{ij}^k is the fuzzy membership at location (i, j) for soil k ; m is the number of environmental variables taken into account, and n is the number of *tacit points* for soil k ; e_{ij}^v is the value of the v th environmental variable at location (i, j) , and $e^{v,t}$ is the environmental condition for the v th variable associated with the t th *tacit point* for soil k ;

E is the function for evaluating the similarity on the v th variable, and this function can be specific for variable v , *tacit point* t , and location (i, j) ; P is the function for evaluating the similarity at the case level (based on all the environmental variables, i.e., the configuration of environmental conditions), and can be specific for *tacit point* t and location (i, j) ; T is the function for deriving the final fuzzy membership value based on the similarities between site (i, j) and all the *tacit points* for soil k , and can be specific for soil k and location (i, j) .

There can be various choices for functions T , P , and E in Equation 2. In this research, the maximum operator is used for function T , which is the simplest possible form for T under the nearest neighbor principle. The maximum operator selects, among all the *tacit points* for soil k , the single *tacit point* that is most similar to the given location, and uses this similarity value as the fuzzy membership for soil k at the given location. For function P , the minimum operator is used. This follows Zhu and Band (1994) and is based on the limiting factor principle in ecology. The limiting factor principle assumes that the limiting factor controls the development of soil formation, thus no additional information about the relative importance of each factor at a local point is needed. While the limiting factor method is probably the easiest and simplest choice for function P , more research, nevertheless, is needed to find out the most reasonable way to integrate the influences of different environmental variables on soil formation. The choice for function E should be based on the data type of the environmental variable. For a variable whose values are categorical, Boolean operators can be used. For the variables whose values are continuous, the soil scientist can choose from the functions illustrated in Fig. 5. In summary, in this research Equation 2 takes the form of:

$$s_{ij}^k = \max_{t=1}^n \{ \min_{v=1}^m [E^{v,t}(e_{ij}^v, e^{v,t})] \} \quad [3]$$

The environmental variables used in this research for soil inference include parent material (from geological data), elevation, slope gradient, surface curvatures (profile and planform curvatures), and wetness index. The selection of these variables is based on the knowledge of the local soil scientist. The issue of weight for each variable has been considered. However, the soil scientist finds it difficult to quantify the importance of each variable in this area. Therefore, a simplified strategy is adopted: all environmental variables are treated equally important.

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

4. Extracting soil-landscape model Using spatial data mining

Feng Qi and A-Xing Zhu
 Department of Geography
 University of Wisconsin-Madison
 Madison, Wisconsin 53706

4.1 References:

Miller, H. J., and J. Han, 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.

Qi, F. and A.X. Zhu, 2003. Knowledge discovery from soil maps using inductive learning, *International Journal of Geographic Information Science*, In press.

4.2 Basic Idea:

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 stereoscopic. 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 analyzing soil maps together with the landscape characteristics captured using GIS (Figure 6).

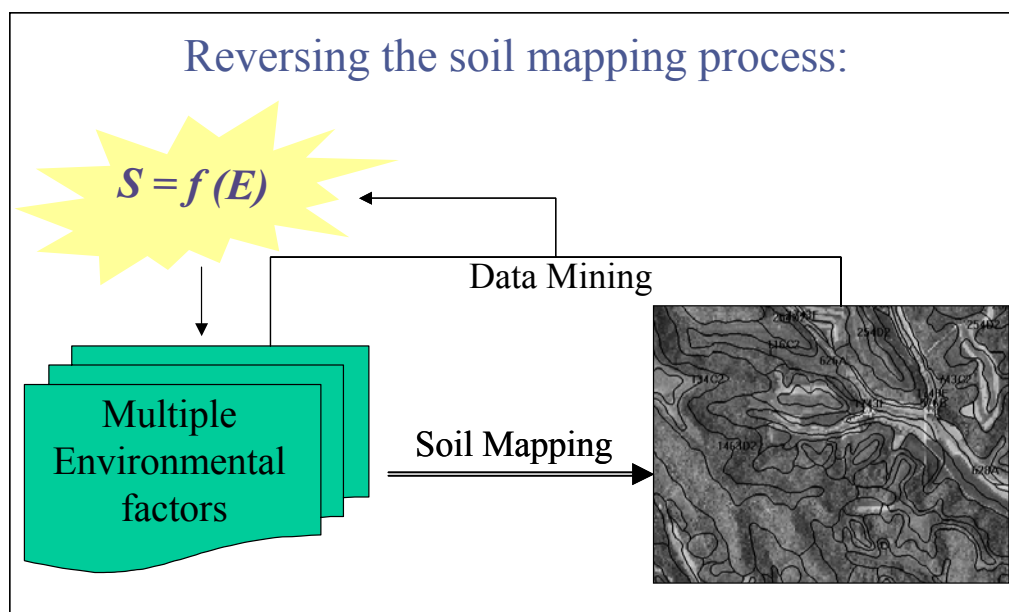


Figure 6: Data mining of soil maps

4.3 The Process:

4.3.1 Data preparation. The first step is to construct a GIS database that contains the soil map, relevant environmental variables, and spatial relationships.

- 1) **Soil map.** The soil maps that can be used for data mining are detailed survey maps at the county level, i.e. the SSURGO maps. Soil series are the basic taxonomic unit involved in the data mining process. SSURGO maps can be downloaded from USDA NRCS website in digital form (ARC/INFO coverage). We rasterize the soil map with the same resolution as the other environmental data layers.
- 2) **Environmental variables.** Generally, we start with the following environmental layers: elevation, slope gradient, slope aspect, surface curvature (planform and profile), local drainage conditions, geology, and vegetation conditions. The elevation, slope, aspect, and curvature data layers can be derived directly from a DEM. If applicable, drainage factors such as upper stream drainage area and wetness index can also be calculated from the DEM using flow accumulation algorithms. When necessary, geology map needs to be digitized and rasterized in consistency with other environmental data layers. Vegetation information can be obtained from vegetation maps or from remote sensing images (such as tree canopy coverage and leaf area index). The usefulness of vegetation information very much depends on the ability of these data layers in relating to soil formation in the given area.
- 3) **Spatial relationships.** Two kinds of spatial relations can be considered in data mining: a) spatial relations of primitive environmental variables that are related to soil formation, and b) topological and directional relations of soil polygons. Examples for the first kind include distance to streams, topographic wetness index, and percentage of colluvium from competing bedrocks. These can be derived using existing GIS packages. For the second kind, the upslope neighbor and downslope neighbor of a certain soil type determine its position in a catenary sequence, and will help the construction of a comprehensive soil-landscape model.

4.3.2 Data preprocessing. This step is designed to reduce noise contained in the original map and to obtain representative samples. This involves sampling pixels that are at or close to the histogram modes of environmental variables for individual soil types (Figure 7). Experiments have determined preferred histogram parameters and sample sizes for the purpose of extracting soil-landscape model from soil maps. We provide a program to do the sampling and generate sample sets.

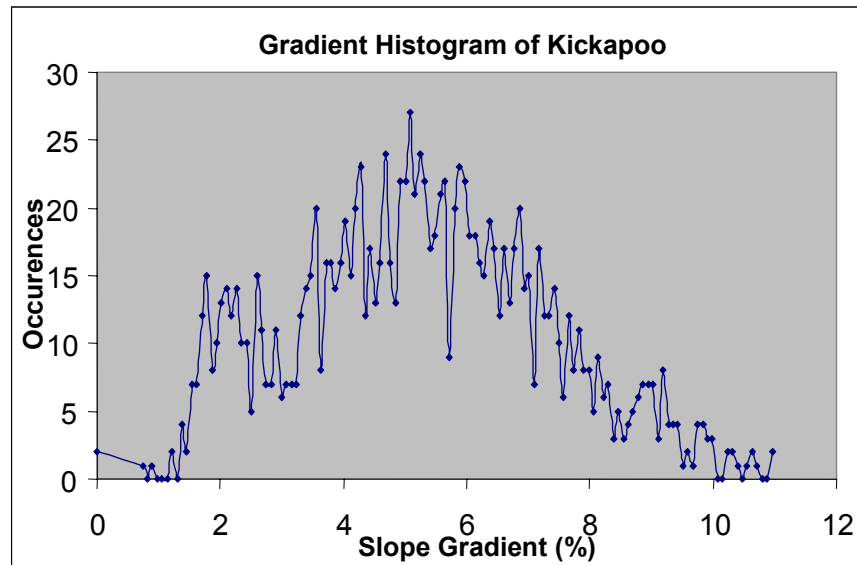


Figure 7: Histogram of gradient values within the Kickapoo soil polygons

- 1) **Input.** The preprocessing program reads in the environmental data layers used.
- 2) **Output.** The output is a file that contains the representative samples as labeled examples. The file format is in consistency with the requirements of the data mining program (See5). One labeled example is a data record that contains values of all environmental variables as the “features” and the soil type as the “label”.

4.3.3 Decision tree construction. The representative samples are then fed to a data mining program for pattern extraction. We use the commercial software See5 to construct decision trees from training samples. The program requires two essential inputs. One is a descriptive file that defines the data type of each environmental data layer and a list of soil types. The second essential input is the sample set, which was generated in the preprocessing step.

4.3.3 Knowledge examination and interpretation. The last step is to examine and interpret the decision tree to be of future use. The learned decision tree can be 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 can be transformed to rules and soil descriptions. The See5 program provides function to transform a decision tree to rule sets.