

Knowledge Discovery from 'Area-class' Resource Maps: Data Preprocessing for Noise Reduction

Feng Qi

Department of Geography,
University of Wisconsin-Madison,
550 North Park Street,
Madison, WI 53706, USA
fqi@wisc.edu

1 Introduction

Researches on data mining or knowledge discovery from databases have been receiving continuous attention in the past decade. Studies have extended the scope of data mining from traditional databases to spatial databases. With the distinct power of discovering previously unclear knowledge in spatial data, geographic data mining not only improves our spatial data analysis abilities; knowledge discovery from previously underutilized data sources (i.e. image data, map data, etc.) also provides an alternative to knowledge construction for knowledge-based systems where traditional knowledge acquisition is difficult. There have been a variety of data mining methods investigated and reported by GIScientists (see Ester *et al.* 2001 and Miller and Han 2001 for a review). Recent research has either demonstrated the success of many well-established data mining algorithms in the geographic domain, or aimed to advancing existing AI techniques for spatial applications.

While the searching for efficient spatial data mining algorithms continues in various fields, less work has been reported on an important component of the knowledge discovery process: preprocessing of data. Data preprocessing tasks in data mining usually include data cleaning, data reduction, data transformation, and feature selection. Several researchers have examined the general data transformation and feature selection techniques in knowledge discovery from traditional databases (Lu *et al.* 1996, Pyle 1999, Bernstein and Provost 2001, Kietz *et al.* 2001). The data cleaning task is usually more application specific and remains least investigated in the AI community. This paper studies how data preprocessing can improve the performance of knowledge discovery from 'area-class' natural resource maps; and our focus is on the preprocessing method for noise reduction.

Natural resource maps, particularly maps of those natural resources that cannot be directly observed using remote sensing techniques, are usually created by domain experts through a modeling process. Examples include soil maps, maps of wildlife habitats, and potential natural hazards, among others. The distributions of these natural resources are usually inferred from other easily observable environmental conditions based on a relationship model. Our hypothesis is that the expert knowledge of the particular model is implicitly embedded in the map product and can be extracted using spatial data mining techniques. The extracted knowledge can then be used to facilitate map updates in natural resource inventory mapping.

Traditional natural resource maps created through manual surveys are often prone to errors. The map classes thus contain noise that does not reflect the experts' true knowledge of the relationship model. In this study, particular attention is paid to minimize the impact of these unavoidable errors on knowledge discovery. We designed a sampling strategy in the data preprocessing step to obtain samples that are representative of the central concept of individual natural resource classes. This effort aims to reduce noise in the original map and improve the accuracy of the extracted knowledge.

We applied the knowledge discovery procedure in a case study to extract knowledge of soil-landscape models from a soil map. Previous research has demonstrated the success of inductive machine learning algorithms in modeling soil maps (Moran and Bui 2002). Qi and Zhu

(2003) compared three inductive learning algorithms in knowledge discovery from soil maps and found that the decision tree algorithm is the most suitable to extract and represent knowledge of soil-landscape models. This paper reports the experiments using the See5 decision tree learning algorithm (Quinlan 2001) and focuses on the effectiveness of the noise reduction method.

The remainder of the paper is organized as follows. Section 2 gives a brief introduction to natural resource inventory mapping and the errors common to 'area-class' resource maps to provide the background for the knowledge discovery and specifically the noise reduction method. The sampling strategy for noise reduction and the knowledge discovery procedure are then presented in the next Section. The experiments in a case study are described in Section 4. Section 5 reports the results. And Section 6 concludes the paper with discussions and future efforts.

2 Background

2.1 Natural resource inventory mapping based on modeling

Detailed natural resource maps provide essential information for environmental management and decision making. However, it is often impossible to directly map natural resources based on field observations in resource inventory mapping due to obscuring overstories and the high cost of collecting information on these resources at many locations across the landscape. While remote sensing techniques can be used to map some natural resources, many are not directly observable through remote sensing. Natural resources like soils, wildlife habitats, and potential environmental hazards are usually mapped through environmental modeling. The distributions of these resources are inferred from other easily observable environmental conditions based on expert knowledge of the relationships between the resources and their environmental conditions (as referred to as knowledge driven environmental modeling, Skidmore 2002). For example, wildlife habitat maps can be modeled based on species-environmental relationships, where the environmental factors include climate, vegetation, landscape characteristics, and human factors. Wildfire assessment map can be modeled based on three main environmental factors: the fuel (biomass type, moisture, etc.), the weather, and topography (Skidmore 2002). Under these models, the natural resources are classified and represented as 'area-class' maps (Mark and Csillag 1989).

This paper uses soil map as an example to illustrate noise reduction for knowledge discovery. 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 Dokuchaev (Glinka 1927) and Hilgard (Jenny 1961). The autocorrelation of soil formative environmental factors results in natural entities of soil series developed on unique landscape units (Hudson 1990). Hudson (1992) generalized this concept 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 work out first the relationships between soil and its landscape conditions and then draw polygons of soil types based on the perceived landscape units. 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 of a soil-landscape model that is embedded in the map and can be extracted using spatial data mining.

2.2 Errors in the 'area-class' resource maps

Errors associated with 'area-class' maps have been well studied by GIScientists. Goodchild (1992) examined the nature of errors in 'area-class' maps and recognized two major types of errors: inclusions and generalization of transition zones. Ehlschlaeger and Goodchild (1996) summarized three major forms of errors in these maps: mislabeling, inclusion, and in accurate class boundaries. In our study to extract knowledge from 'area-class' maps, we

differentiate two different kinds of errors based on the error source. The first kind of errors comes from the model used to create the map, and is referred to as modeling errors. Modeling errors can include: 1) generalization of the continuous feature as discrete categories, thus the over-simplification of transitional zones as lines; and 2) inclusions that cannot be eliminated by increasing map scale, that is, the model cannot discriminate between two classes using the available environmental factors. The second kind of errors is introduced in the mapping process, and is referred to as mapping errors. Major mapping errors include: 1) misplacement of class boundaries, 2) mislabeling of polygons, and 3) inclusions that can be avoided by increasing map scale, that is, the missing of small patches of classes due to the limitation of map scale.

This paper explicitly considers the mapping errors existing in an 'area-class' resource map. The objective of our study is to extract the knowledge of the particular model that a domain expert used to create the 'area-class' map. The knowledge is usually based on the expert's experience and is by all means subjective. It is thus not guaranteed that the model developed by individual expert represents accurately the real resource-environment relationships of the area. Therefore, there are indeed two levels of approximation: how well the extracted model approximates the expert knowledge, and how well the expert knowledge represents the reality. Our goal in the current stage is to recover the subjective expert knowledge from the error-prone 'area-class' maps; and the second level of approximation (thus the modeling errors) is not explicitly considered.

As aforementioned, there are three major types of mapping errors: misplacement of boundaries, mislabeling of polygons, and inclusions. Our assumption is that in well-controlled natural resource inventory mapping, these errors are not exorbitant in the map product. Mislabeling is usually minimized under strict quality control. Maps are created by domain experts to best represent their knowledge of the relationship model at the given map scale. Polygons of each class are carefully delineated to enclose well the central concept of the class, where inclusions and unintentional misplacement of polygon boundaries exist as minor disturbances. In soil survey, the central concept of a soil class is the soil unit that has the modal or typical properties of the soil class. Since the key hypothesis of soil survey is that specific combinations of soil-forming factors lead to specific soil properties, the central concept of a soil class is identified at typical landscape positions where the certain combinations of environmental factors are typical (Rossiter 2000). When the environmental factors associated with a soil class in the soil map are projected to the parameter domain with frequency distribution, the central concept of this soil class occupies the highest frequency and the errors are the relatively low frequency parts. In our study, this frequency distribution is approximated with histograms of the environmental factors associated with each soil class. By sampling only the modal area, we obtain samples that are representative of the central concept of the soil class and exclude the low frequency noises caused by mapping errors. Next Section will outline the knowledge discovery procedure we applied to extract knowledge from 'area-class' maps and will describe this sampling strategy in more detail.

3 Knowledge discovery

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 preprocessing, pattern extraction, and knowledge examination and interpretation.

The data preparation step is to construct a GIS database containing relevant environmental variables and spatial relationships. The choice of proper variables is domain-specific. For detailed soil survey, the underlying model is usually a soil-landscape model at the

watershed scale of meso-scale or smaller sizes. The practical environmental variables include bedrock geology, topographical characteristics, and vegetation conditions. Spatial relationships can be captured with 1) spatial relations of primitive environmental variables, and 2) topological and directional relations of map polygons.

The data preprocessing step is designed to reduce noise contained in the original map and to obtain representative samples. As aforementioned, 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 mapped classes. Under the assumption that the majority of the area mapped as a certain class is correctly categorized, the histogram mode(s) of a given environmental variable enclosed in the polygons of the same class represents the typical environmental conditions associated with the class. Based on this heuristic, we obtain representative samples whose environmental conditions are at or close to the mode of environmental histograms.

In implementation, for all pixels belonging to the same class, 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 class under concern and the low frequency tails representing mapping errors and transitional conditions. In soil maps, multimodal shapes are occasionally present when 1) the soil class 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 co-dominant 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 categorized as each class: $N_i = N_p / r$. Here N_i refers to the number of intervals of the histograms for a certain class, N_p refers to the total number of pixels categorized as this class, and r is the average number of pixels expected to fall within each interval. 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 class to be adjusted according to its area, so that the number of modal pixels is comparable across classes. This allows each class 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 class, sampling is conducted based on the stratification of map classes. The individual sample set for each class is produced in two steps. The first step is to obtain samples one environmental variable at a time, that is, to sample just the modal pixels based on the given environmental histogram. We investigated two options: one is to take the entire set of modal samples, and the other is to obtain randomly a fixed number of samples (N_i) from each mode. The second step is to pool the samples from all environmental variables for the class, and then select samples from this pool. A single pixel may have more than one occurrence in the pool since it could be selected based on 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 multiple environmental histograms. With both approaches, the final set is constructed by simply combining the sample sets for all classes.

The so obtained representative samples are then fed to a data mining program for pattern extraction. Previous research has established the success of decision tree algorithms in learning and representing descriptive soil-landscape models (Qi and Zhu 2003). In this study, we experiment with a decision tree algorithm called See5 (Quinlan 2001). When constructing a

decision tree from training data, choosing the right size for the tree is an important issue. 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 toward the leaves of the tree that are based on very little data (this is known as small disjuncts). 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 preprocessing stage.

In the last step of a complete knowledge discovery process, the extracted pattern is examined and interpreted 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 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 test the knowledge discovery method, especially the data preprocessing strategy, more validation efforts were made in our case study, for which we chose a specific soil map to extract knowledge from. We conducted knowledge acquisition with the soil expert who created the soil map and compared our extracted knowledge with the acquired expert knowledge to test our accuracy. The details of this validation will be described in Section 5.

4 Case study

We implemented our knowledge discovery procedure to extract soil-landscape models from an existing soil map. The study area of this research is a small watershed 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. Complex soils from many epochs of soil formation and movement can be found. The soil map created from a recent soil survey indicates there is a total of 16 different soil series in the area (Figure 1).



Figure 5. Soil map of the Raffelson watershed.

The GIS database contains five primitive variables to characterize the formative environmental conditions of soils over the study area: elevation, slope gradient, planform curvature, profile curvature, and geology. Three derived variables are used to capture the spatial relations of soil-formative environmental factors. They are distance to streams, topographic wetness index, and percentage of colluvium from competing bedrocks. Additionally, we attach another two attributes to training samples to capture topological and direction relations between soil types. They are upslope and downslope neighbors. Unlike the other variables, the spatial relations between soil types are not soil-formative factors, instead they serve as indicators of soil distributions over the area. In other words, by including the spatial relations of the soil polygons in knowledge discovery, the extracted soil-landscape relationships can be enriched to describe spatial patterns of soil types.

Expert knowledge of the local soil-landscape relationships was acquired through interviews with the soil expert who created the map and was used to validate our extracted

knowledge. We also obtained an expert-defined test set to quantitatively test our training results. The soil expert was provided with the orthophoto in 3D view on a computer screen. He then digitized on screen typical points that he believes are consistent with his understanding of the soil-landscape model over the area. These sample points were 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.

5 Results and discussions

5.1 The effect of different sampling options

Sample sets were constructed based on different sampling options in data preprocessing—either taking the entire mode(s) or randomly sampling from the mode(s), either pooling with the union operation or the intersection operation, and different settings of r and N_r (see Section 3). 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.

Table 1. Accuracies of decision trees derived from rectified sample sets using an expert-defined sample set as the test set.

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

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 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. However, we noticed from the decision tree output that the bigger the size of the training set, the more detailed the resulting decision tree is. We experimented with sample sets of different sizes and compared the decision tree results with documented expert knowledge. It is found that with a sample size less than 150 (that is, the sample size for an individual class is less than 10), the learned structure is too coarse to capture the expert knowledge of the soil-landscape model to its actual level of details. When the sample size increases to over 500 (that is, the sample size for an individual class exceeds 30), the learned structure is overly detailed so as to overfitting many small disjuncts. In this case,

pruning has to be done to reduce overfitting. However, the selection of pruning parameters is often subjective. The preferred sample size should thus be between 200 and 500, with which the learned decision tree is detailed enough but requires the least subjective pruning efforts to approximate well the structure of the expert knowledge.

When constructing histograms, the number of intervals (and thus the interval size) is determined by the selection of number r , which directly controls the shape of the histogram. Several choices of r ($r = 5, 10, 15$ or 20) were used to examine the stability of the mode when the histogram shape changes. The histograms of soil *Lamoille* based on elevation are shown in Figure 2, from which one can observe the mode of the histogram is fairly stable with regards to changes of r . It is also shown in Table 1 that the accuracies from sample sets with different r do not show apparent differences. This indicates that although the shape of the histogram may change when the histogram parameters change, the mode(s) of the histogram would always consist of a set of samples that exclude the low frequency errors and represent the relatively typical environmental conditions for a soil type under the given configuration. Using samples obtained from the histogram mode(s) to train decision trees would thus lead to results that represent the central concepts of the mapped soil types.

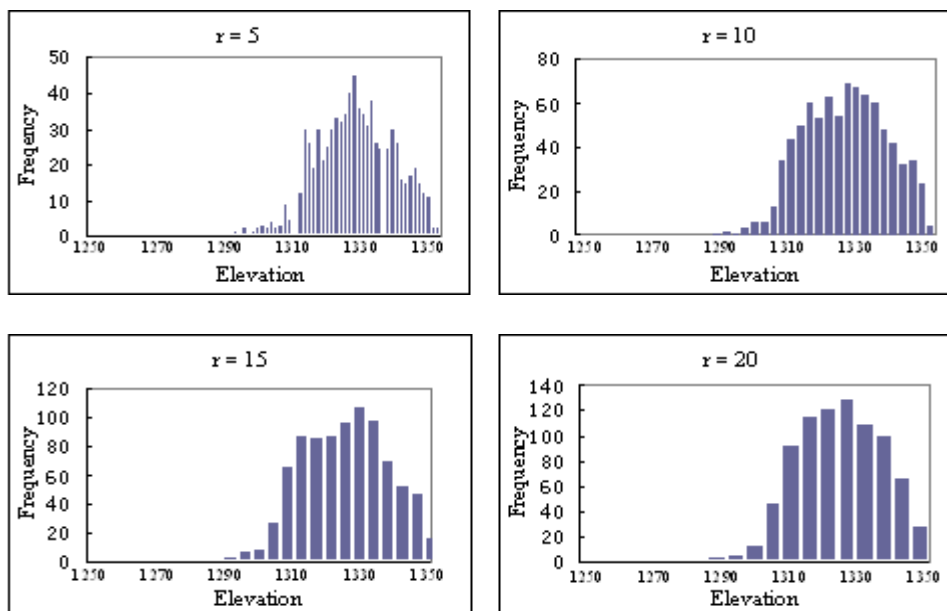


Figure 2. Elevation histograms for *Lamoille* with $r = 5, 10, 15$ and 20

5.2 The impact of data preprocessing

In order to examine the effect of our data preprocessing strategy, we generated ten sample sets by randomly drawing 416 (an arbitrary number that is less than 500, the effective sample size determined in Section 5.1) samples from the map area. These are referred to as unrectified sample sets. Six rectified sample sets were also obtained by preprocessing the map using the histogram sampling heuristic. These sample sets were trained to derive decision trees, and then tested using the expert defined test set. The results of these tests are shown in Table 2 and Table 3. Table 2 shows that the mean accuracy without data preprocessing is 0.75. Table 3 shows 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. This provides one piece of evidence that the data-preprocessing step is effective in reducing noise and outliers from the original map.

Table 2. Accuracy on expert-defined test set: sample sets without data preprocessing.

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

Table 3. Accuracies of decision trees derived from the samples without data preprocessing.

	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

A decision tree representation can be easily converted into rule sets by traversing the tree paths. 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 among the results from a rectified sample set, the results from an unrectified sample set, and the documented expert knowledge. The comparison reveals that consistency is high between the results with data preprocessing and those acquired directly from the local soil expert. For the continuous environmental variables, the decision tree program has generated breaks automatically. From the comparison in Table 4, we see that the environmental breaks generated from the rectified sample set are less accurate than those from the unrectified sample set, due to the inclusion of noises. Therefore, we can say that the data preprocessing method has effectively reduced the impact of mapping errors and improved the performance of knowledge discovery.

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

Soil series	Environmental Variable	Tree result (without data preprocessing)	Expert Knowledge	Tree result (mode sampling)
<i>Valton</i>	Geology	Oneota	Oneota	Oneota
	Elevation	>1194.95	>1300	>1298.68
	Gradient	<=9.86%	<12%	<=12.57
<i>Lamoille</i>	Geology	Oneota	Oneota	Oneota
	Elevation	>1194.95	>1250	>1298.68
	Gradient	>9.86%	12-20%	>12.57%
<i>Dorerton</i>	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

6 Conclusions and future efforts

Most previous work on knowledge discovery has focused on the pattern extraction step—the specific algorithm taken. However, we conclude that the other steps, especially the data preprocessing step, are of considerable importance for the successful application of knowledge discovery in practice, especially when applying general-purpose data mining algorithms to geographic knowledge discoveries. It is important that 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 lead to a better accuracy.

In this study we designed a sampling method for data preprocessing in order to reduce the impact of noises contained in 'area-class' resource maps. Environmental histograms are used to approximate the frequency distributions of the environmental conditions related to individual map classes. Although the shape of the histogram may change when the number of intervals changes, the mode(s) of the histogram is relatively stable and the modal samples represent the typical environmental conditions for a map class. The preprocessing method of sampling only modal pixels based on environmental histograms is found to be effective since it allows the selection of samples that represent the central concepts of the mapped classes. It helps to reduce generalization bias of the algorithm and to avoid overfitting toward noisy data, thus significantly improving the knowledge discovery performance.

When selecting modal samples, still appropriate sampling strategy needs to be chosen both to keep a reasonably proper sample size and to balance between different map classes. In the case of constructing decision tree of a soil-landscape model, it is found that the best sample size for individual soil type is from 10 to 30 samples in order to gain the best generalization performance while minimizing subjective pruning efforts. 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.

With the application of data mining program to preprocessed data, the extracted knowledge is supposed to approximate the expert's knowledge rather than the error-prone 'area-class' map. A good approximation of the true expert knowledge, the extracted knowledge is valuable in at least two ways. First, it has the potential to facilitate traditional natural resource map updates. In traditional natural resource inventory mapping, the relationship models often exist as experts' tacit knowledge and are not documented. Since the map update cycle is usually longer than the career span of a domain expert, new experts would have to develop their own model from scratch, which would involve a tremendous amount of fieldwork. The extracted model in our study allows new domain experts to build upon, thus facilitate the update of resource survey. Second, the knowledge of the relationship model, once properly formulated, could also be used for automated natural resource mapping, modeling, and classification.

Our future plan includes considering explicitly the modeling errors associated with 'area-class' resource maps, modeling the fuzzy transitions of natural resource categories, and developing an interactive visualization tool for data preprocessing. Specifically, in the current study, the frequency distributions of the environmental conditions related to a mapped class are approximated by the environmental histograms, where the modal area of the histograms represents the central concept of the class and the low frequency tails are the errors and transitional conditions. The preprocessing method presented in this paper selects only the modal samples to exclude noise. One side effect of this approach is that the low frequency transitional conditions are also excluded from the training samples. This may help to reduce the modeling error of generalizing the continuous natural resource feature to discrete categories thus the arbitrary placement of class boundaries. However, the inductive learning algorithm used in this study derives rules that also separate different classes with crisp lines, which overlooks the natural fuzziness of natural resource categories and the corresponding resource-environment models. 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 generating only a single output using the modal data. At last, since the natural resource modeling discussed in this paper is virtually a knowledge based process, it is desirable to involve human experts in the knowledge discovery process. An interactive data preprocessing tool is under development to allow the expert to visualize the spatial distributions of samples grouped using histograms in the parameter space. The tool will allow domain experts to direct the data preprocessing, control feature selection and manage iterations.

References

- Bernstein, A. and Provost, F. 2001, An Intelligent Assistant for the Knowledge Discovery Process. In *Proc. of the IJCAI-01 Workshop on Wrappers for Performance Enhancement in KDD*, Seattle, WA.
- Ehlschlaeger, C. R., and Goodchild, M. F., 1996, Dealing with uncertainty in categorical coverage maps: Defining, visualizing, and managing errors. *Proceedings, Workshop on Geographical Information Systems at the Conference on Information and Knowledge Management*, Gaithersburg, MD, pp. 86-91.
- 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**(5), 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.
- 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**(1), 113-139.
- Glinka, K. D., 1927, *The Great Soil Groups of the World and their Development*. (Ann Arbor, MI: Edwards Bros.).
- Goodchild, M. F., Sun, G., and Yang, S., 1992, Development and test of an error model for categorical data. *International Journal of Geographical Information Systems*, **6**, 87-104.
- 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).
- Kietz J. U., Vaduva A., Zücker R., 2001, MiningMart: Metadata-Driven Preprocessing. In *Proceedings of the ECML/PKDD Workshop on Database Support for KDD*.
- Lu, H., Sung, S.Y., and Lu, Y., 1996, On Preprocessing Data For Effective Classification. *ACM SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, Canada.
- Mark, D. M., and Csillag, F., 1989, The nature of boundaries on 'area-class' maps. *Cartographica*, **26**, 65-78.
- 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.
- Moran, C. J., and Bui, E. N., 2002, Spatial data mining for enhanced soil map modeling. *International Journal of Geographical Information Science*, **16**(6), 533-549.
- Pyle, D., 1999, *Data Preparation for Data Mining Morgan Kaufmann*, California.
- Qi, F., and Zhu, A. X., 2003, Knowledge discovery from soil maps using inductive learning. *International Journal of Geographical Information Science*, accepted.
- 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>.
- Rossiter, D. G., 2000, *Methodology for Soil Resource Inventories*. Accessed at URL: http://www.itc.nl/~rossiter/teach/ssm/SSM_LectureNotes2.pdf.
- Skidmore, A., 2002, *Environmental Modelling with GIS and Remote Sensing*. Taylor & Rancis, London.