ASSISTING THE DEVELOPMENT OF KNOWLEDGE FOR PREDICTIVE MAPPING USING A FUZZY C-MEANS CLASSIFICATION

A-Xing ZHU1 2 Edward ENGLISH1
1 Department of Geography
University of Wisconsin
550 North Park Street
Madison, Wisconsin 53706
Tel: (608) 262-0272
Fax: (608) 265-3891
Email: enng@geography.wisc.edu

2 State Key Lab of Resources and Environmental Information System
Institute of Geography
Chinese Academy of Sciences
Beijing 100101, China

Keywords: Fuzzy Set, Geographic Information System, Soil Mapping, Expert System, Case-Based Reasoning

Abstract

Knowledge of relationships between a given geographic phenomenon and its observable environmental factors is needed for mapping geographic phenomena/features which cannot be directly observed, for example, soils, habitat potential. The knowledge is often developed through extensive fieldwork which is not only very labor intensive and slow, but also very costly. Methods are needed to assist local domain experts to acquire this knowledge efficiently. This paper presents an approach based on fuzzy c-means classification to assist the development of local domain experts’ knowledge of the relationships. The method is based on the assumption that the observable environmental factors have dominant impact on the distribution of the given geographic phenomenon and that unique environmental configurations reflect the unique status or properties of the geographic phenomenon. Under this assumption, clusters in the environmental space (parameter space) are directly related to different types (status) of the given phenomenon. We employed a fuzzy c-means classification to identify the natural clusters in the environmental space and use the centroids of these fuzzy clusters as a guide to allocate field investigation efforts for developing knowledge on relationships between environmental factors and the phenomenon to be mapped. Through a soil mapping case study, we found the approach is effective in helping local soil scientists to develop their understanding (knowledge) of soil-environmental relationships in areas the local soil experts are not familiar with the relationships. The soil map derived using the understandings achieved over 70% accuracy overall, compared to about 60% accuracy through the extensive fieldwork. For areas with a good environmental gradient, the accuracy is over 90% while the accuracy for area with little relief is about 63%.

1. Introduction

Predictive mapping, such as soil mapping and habitat potential mapping, uses observable environmental conditions to predict the spatial distribution of the geographic phenomenon to be mapped. Predictive mapping requires the knowledge of the relationships between the phenomenon and the observable environmental conditions. This knowledge is often developed through extensive fieldwork. For example, with conventional soil mapping, extensive field campaign is required for local soil scientists to develop the knowledge on how the soils in an area are related to the observable environmental conditions (such as elevation, slope gradient). Once the understanding of relationships has been developed through the campaign, the local soil scientists will then delineate the spatial distribution of soils based on the knowledge of the relationships and the observable environmental conditions.

Field campaigns for natural resource surveys are often very labor intensive, time consuming, and costly. Although field scientists may employ statistical sampling strategies (such as random sampling, regular grid sampling, or stratifying sampling), their field efforts may still not be effectively directed due to the repeated spatial patterns and the gradation of the phenomenon/feature. The layout of field investigation is often dependent on the experience of individual field experts. For example, some field scientists may be able to avoid transition areas and distill the major relationships very quickly while others may spend a lot of time in the transitional areas and was not able to quickly capture the major relationships. As a result, the development of knowledge on the relationships is often very slow and the accumulated knowledge is often very subjective.

Methods are needed to assist field scientists to quickly identify the key areas for developing the relationships. This paper investigates the use of an unsupervised fuzzy classification technique to identify areas of unique environmental niches from those areas of environmental transition. Using the results from the fuzzy classification field investigation efforts are directed towards areas of unique environmental niches to discover relationships between these niche and the geographic phenomena to be mapped.

The next section of this paper describes our method. Section 3 shows a case study using this method. Summaries are given in Section 4.

2. Methodology

2.1 The Basis:

The underlying assumption behind predictive mapping is that there is a relationship between the geographic phenomenon to be mapped and its observable environmental conditions. We further assume that the unique status of the given geographic phenomenon (such as different types of soils) is created under unique combinations of environmental conditions. Under this assumption, the spatial locations of the geographic phenomenon with unique status can be approximated by the locations of unique combination of environmental conditions. Thus, discovering the relationships between the geographic phenomenon and its environmental conditions is a matter of finding which unique combination of environmental conditions is related to which unique status (types) of the phenomenon.

Relating unique status of a given geographic phenomenon to unique combination of environmental conditions requires field investigation. Due to the repetition of spatial patterns and graduation of geographic phenomenon, locating the unique combination of environmental conditions is still very challenging. The efficiency of field investigation can be greatly improved if we can first identify the spatial locations of these
unique combinations of environmental conditions and discern areas of environmental transition. Unsupervised fuzzy classification allows us to identify and map the spatial locations of natural classes in environmental data. It can also be used to depict the areas of environmental transition. Thus it can facilitate the allocation of field investigation efforts and improve the efficiency of field investigation.

2.2 The Method

Our method of assisting knowledge development consists of the following four steps: (1) environmental database development; (2) environmental niches identification; (3) allocating field investigation efforts; (4) distilling relationships between the geographic phenomenon and environmental conditions.

2.2.1 Environmental database development. Environmental factors related to the geographic phenomenon are first identified. For example in soil mapping, the factors related to soil formation need to be identified. A GIS database on these environmental factors are then generated given that the source data is available and GIS data layers can be created for each of the environmental factors.

2.2.2 Identify environmental niches using a fuzzy c-means classifier (FCM). FCM is a classifier which first optimally partitions a dataset (such as the environmental dataset described in 2.2.1) into a given set of classes and computes the membership of each data element (such as the environmental conditions at a pixel) in each of the classes (Bezdék et al., 1984). It identifies the centroids of classes by minimizing the fuzzy partition error as given in Equation 1 (Bezdék et al., 1984):

\[ J_m(U, \mathbf{v}) = \sum_{i=1}^{c} \sum_{j=1}^{m} u_{ij}^m \left\| y_j - v_i \right\|^2 / A \]  

(1)

where \( y \) is the data, \( c \) is the number of clusters in \( Y \), \( m \) is a weighting exponent, \( u \) is a fuzzy partition of \( Y \), \( v \) is a vector of cluster centers, \( A \) is a weighting matrix, \( n \) is the number of objects in set \( y \); \( u_{ij} \) is the membership of the \( i \)-th object \( x_i \) belonging to the \( j \)-th cluster. \( J_m \), the fuzzy partition error, can be described as a weighted measure of the squared distance between pixels and class centroids, and so is a measure of the total squared errors as minimized with respect to each cluster (Ahn et al. 1999; Ross 1995). \( J_m \) decreases as the clustering improves (meaning that pixels tend to be closer to several representative centroids).

In most cases, one does not know the number of classes that best describe the structure in the data set. To judge the effectiveness of the clustering results generated using the above fuzzy c-means algorithm, two cluster validity measures (partition coefficient \( F \) and entropy \( H \)) are defined as (Bezdék et al., 1984):

\[ F_c(\hat{u}) = \sum_{i=1}^{c} \sum_{j=1}^{m} \frac{(\hat{u}_{ij})^2}{n} \]  

(2)

\[ H_c(\hat{u}) = -\sum_{i=1}^{c} \sum_{j=1}^{m} (\hat{u}_{ij} \log(\hat{u}_{ij})) / n \]  

(3)

Partition coefficient \( F \) will take the values of 1/c to 1, while entropy \( H \) ranges from zero to \( \log c \) (Ahn et al. 1999). \( F \) measures the amount of overlap between clusters, and is inversely proportional to the overall average overlap between pairs of fuzzy sets (Ahn et al. 1999). \( H \), conversely, is a scalar measure of the amount of fuzziness in a given fuzzy partition \( U \) (Bezdé3k 1981). The best fuzzy c-partition, e.g., the number of classes that best describe the structure in the data set, is thus the c-partition which realizes the highest \( F \) and the lowest \( H \) (Ward et al. 1992). Note that both \( H \) and \( F \) will reach maxima and minima at the same points, and in this sense they are essentially equivalent (Bezdé3k 1981).

It is often the case that \( F \) increases and \( H \) decreases as the number of classes decreases. To determine if a fuzzy clustering can be considered optimal, i.e. the number of clusters optimally describes the structure in the dataset, one should examine the improvement in entropy or partition coefficient over adjacent clusterings (Zhu 1989). If there is a significant improvement, one can consider the current clustering a better partition of the dataset.

2.2.3 Allocating field investigation efforts. Once the optimal clustering of the environmental data set is determined, membership maps for clusters can be produced. Spatial locations of environmental clusters and areas of environmental transition can be identified on these maps. For each membership map, the locations of its cluster centroid are in those areas with high membership values. Thus, field investigation efforts should be mostly allocated to these areas.

2.2.4 Distilling relationships between the phenomenon and environmental conditions. By investigating the status or property of the given phenomenon at the locations of environmental cluster, one can quickly establish the relationships between the phenomenon and its environmental conditions. Interpretation of membership maps and visualization of the clusters will allow one to develop an appreciation of how the phenomenon varies over space in response to variation in environmental conditions.

3. A Soil Mapping Case Study

Soil mapping is based on the classic concept that soil is the product of the interaction of its formative factors. Thus, the concept assumes that there is a relationship between soil and its formative environmental conditions. Soil mappers first obtain (establish) this relationship through extensive field work (survey) and then use this relationship with the observable environmental conditions to map the spatial distribution of soils.

Acquiring soil-environmental relationships (also referred to as soil-landscape models) through the conventional means (extensive field work) is very time-consuming, costly, and subjective. The method described above was applied to improve the efficiency of the field investigation.

3.1 Study Site

The study site is the Medina watershed, located in eastern Dane County, Wisconsin, approximately 30.5 kilometers east of Madison and 2.6 miles southeast of the town of Marshall. The Medina watershed is about 6,617,6 acres (103,4 km²) in area. The total relief of the Medina watershed is 50.3 meters (157 feet), and so is generally indicative of a gentle environmental gradient. The area is made of drumline and inter-drumline swales. Landuse in the Medina watershed at the present day is generally limited to corn and alfalfa farming. The soils in the area are formed on alluvial deposits of silt loam loess, which are underlain by sandy loam glacial till parent material.

The Medina watershed was chosen for this study as that the local soil experts have limited experience in this area, and so potentially stand to gain improvement to their knowledge through the use of a fuzzy c-means classification strategy.

3.2 Applying the Method

3.2.1 Developing the GIS data layers on environmental conditions. The following five environmental layers were considered to be primary importance to soil formation in the study area: elevation; slope percent; planform curvature; profile curvature; and upstream drainage area index. This decision was based on the literature, specifically McSwethy et al. (1994), which contended that these five environmental measures exerted the majority of influence on soil formation and development at the watershed scale. As no specific a-
priori information existed indicating the influence of other environmental factors, e.g., geology, vegetation, landuse, etc., the above landscape metrics were considered of primary importance in the Medina study area by default. GIS data layers on these five environmental variables were derived from a 3 by 3 meter resolution DEM using conventional digital terrain analysis (Zevenbergen and Thorne, 1987).

3.2.2 Identifying unique environmental combinations using FCM. An unsupervised FCM classification described in Section 2.2 was carried out on the environmental dataset compiled in Section 3.2.1. The FCM were applied on the environmental data across three different weighting exponents (m = 1.5, 2.2, and 2.5, after Bezdek (1981)). For each run (per m) the number of clusters examined range from 2 to 15. By examining the improvements in partition coefficient (P) and entropy (H) for all three sets, we observed the consistent improvement in both entropy and partition coefficient across all these three runs at ten clusters. No other cluster values improved consistently across all three runs, we argue that there are ten clusters (ten unique combinations of environmental conditions) within the data. Figure 1 shows the membership distributions of Class 7 and Class 9.

Figure 1: Membership distributions. (a): Class 7; (b): Class 9

3.2.3 Investigating the class centroids in the field. Over 20 field observations (2 to 3 per class) were made to determine the soil classes associated with each of the environmental classes. The field observations were guided by the membership values. For each class, the observation sites were at locations where the membership values for the classes are very high.

The association between the environmental classes and soil types observed in the field is summarized in Table 1. It is clear that there is a good association between the environmental classes and the soil types in the area. However, the association is not one-to-one. A soil type can occur under different unique environmental conditions. It is important to point out that we did not observe different soil types at a single environmental class. This validates the utility of fuzzy membership maps in allocating field investigation efforts since the membership distribution allowed us to avoid sampling in the transitional areas.

Table 1: Association between environmental clusters (for m=2.0) and soil types.

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Observed Soil Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Kidder</td>
</tr>
<tr>
<td>9</td>
<td>Kidder</td>
</tr>
<tr>
<td>1</td>
<td>McHenry</td>
</tr>
<tr>
<td>5</td>
<td>Transitional</td>
</tr>
<tr>
<td>6</td>
<td>ST Charles</td>
</tr>
<tr>
<td>4</td>
<td>ST Charles</td>
</tr>
<tr>
<td>2</td>
<td>ST Charles</td>
</tr>
<tr>
<td>3</td>
<td>Mayville</td>
</tr>
<tr>
<td>8</td>
<td>Virgil</td>
</tr>
<tr>
<td>10</td>
<td>Sable</td>
</tr>
</tbody>
</table>

3.2.4 Distilling the soil-landscape model for the area. By interpreting the cluster centroids in comparison with field observations and examining the membership distribution of each environmental cluster, the local soil scientist created an environmental description for each soil class. Examples of these descriptions are shown in Table 2. These environmental descriptions constitute the essence of the soil-landscape model of the area. Local soil scientists can map the distribution of each soil series using the environmental description for a given soil series.

Table 2: Descriptions of environmental conditions of Soil Series Kidder and McHenry in the study area.

<table>
<thead>
<tr>
<th>Environmental Variable</th>
<th>Environmental Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td>850 - 1050 feet</td>
</tr>
<tr>
<td>Gradient</td>
<td>15 - 35%</td>
</tr>
<tr>
<td>Profile Curvature</td>
<td>Slightly concave to convex</td>
</tr>
<tr>
<td>Platform Curvature</td>
<td>Slightly concave to convex</td>
</tr>
<tr>
<td>Upstream Drainage Area</td>
<td>Low</td>
</tr>
<tr>
<td>Landform Position</td>
<td>Drumlin tops and upslopes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Environmental Variable</th>
<th>Environmental Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td>850 - 940 feet</td>
</tr>
<tr>
<td>Gradient</td>
<td>6 - 15%</td>
</tr>
<tr>
<td>Profile Curvature</td>
<td>Slightly concave to convex</td>
</tr>
<tr>
<td>Platform Curvature</td>
<td>Slightly concave to slightly convex</td>
</tr>
<tr>
<td>Upstream Drainage Area</td>
<td>Moderately low</td>
</tr>
<tr>
<td>Landform Position</td>
<td>Strongly sloping backslopes and footslopes</td>
</tr>
</tbody>
</table>

3.3 Evaluating the soil derived soil-landscape model model. 3.3.1 Applying the soil-landscape model using Sollim. To assess the validity of the soil-landscape model constructed using the FCM approach, the soil-landscape model was used under the Sollim approach to generate soil maps for the area. The Sollim approach is a knowledge-based approach for soil mapping. It combines the knowledge of soil-environmental relationships (soil-landscape model) with the environmental conditions characterized in a GIS to infer the spatial distribution of soils (Zhu, 1997; Zhu 1999; Zhu et al., 2001). Case studies have demonstrated that the Sollim approach to soil mapping is successful (Zhu et al., 2001). However, the quality of the soil maps from Sollim largely depends on the quality of the soil-landscape model. Thus, the Sollim approach provides us with the opportunity to examine the quality of the soil-landscape model constructed with the use of FCM.

3.3.2 Evaluating the quality of the soil-landscape model. The soil map derived from Sollim with the use of the soil-landscape model above is shown in Figure 2. The soil map shows a catenary sequence of the soils in the area, soil series Kidder at the ridge (drumlin) tops and upper part of the slopes; McHenry at the backslopes; ST Charles at foot-slopes; Mayville at the toe-slopes; Virgil on the flat area of valleys and Sable at the wet areas besides the streams. This pattern matches field observation of catenary sequence in the area well.

To validate this soil map, observations of soils at 50 field sites were made. Soil type at each field site was identified at the series level. The field observed soil series at these sites were then compared with the soil series obtained from the inferred soil map at these locations. Soil series from the inferred soil map match field observed soil series at 38 of the 50 sites.
which accounts for 76% of accuracy. Conventional soil maps typically achieved 60-70% of accuracy. The 50 sites can be divided into two groups: one on the drumlin areas and the other on the inter-drumlin areas. The accuracy for the sites in the drumlin areas reaches 90% while the accuracy for the inter-drumlin areas is about 63%. The difference in accuracies between the two areas is due to the difference in environmental gradients and the fact that SoLM is sensitive to the ability of characterizing soil formative environmental conditions. The relief (gradient) in the drumlin areas is much stronger than that over the inter-drumlin areas. For the drumlin areas, the soil formative environment conditions for different soils can be easily distinguished (characterized) using GIS due to the stronger relief. As a result, the soil-landscape model is more "faithful" applied through SoLM. However, the soil formative environment conditions for different soils in the inter-drumlin areas is very difficult to distinguish due to low relief and it is difficult to distinguish soils over these areas using the environment conditions. Overall, we believe that the soil-landscape model developed through the use of FCM methodology is of good quality since the accuracy of the soil map exceeds that of common soil maps produced through extensive soil surveys.

![Figure 2: Soil map produced from SoLM using the soil-landscape model constructed using the FCM-based method.](image)

4. Summary

This paper presents a methodology to assist the development of knowledge of relationships between a given geographic phenomenon and its environmental conditions. The method employed a fuzzy c-means classification to identify unique combinations of environmental conditions and to discern locations of these unique combinations. The results (the unique combinations and the spatial locations of these unique combinations) were then used to direct field investigation efforts and to improve the efficiency of acquisition of knowledge on the relationships in the field.

Through a soil mapping case study it was found that the FCM assisted field investigation was effective in developing the knowledge of the soil-environmental relationships. First, the amount of field observations was reduced. Second, the acquired knowledge of the relationships was of high quality.

Acknowledgements

The research reported here was supported by a research grant from the Graduate School, University of Wisconsin-Madison, by the funding from Natural Resource Conservation Service, United States Department of Agriculture under Agreement No. 69-SF48-2-00186. The State Key Lab of Resources and Environmental Information System (LREIS), Institute of Geography, Chinese Academy of Sciences kindly provided travel support to the senior author to attend this meeting.

References


