Visual Data Mining Tools for Digital Soil Survey Updates

James E. Burt[1], A-Xing Zhu[1,2] , and Rongxun Wang[1]

1.Department of Geography, University of Wisconsin - Madison

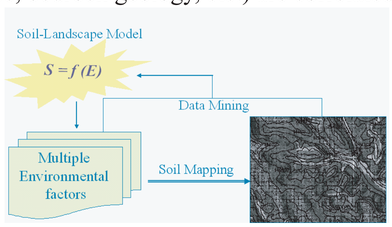
2.State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences

1. Overview

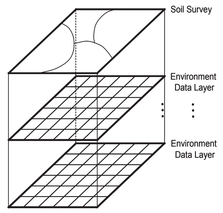
Most private-land soil survey work in the United States is directed at updating existing surveys. The soil-landscape models developed by the original soil scientist are implicit in their map products and---once recovered---can be used to jumpstart the revision process. This paper presents a variety of tools developed and under development for visual data mining that allow a user to recover explicit depictions of soil-landscape concepts, identify inconsistencies in application of those concepts, and develop new soil concepts.

2. Assumptions and Goals

Our fundamental assumption is that the existing survey map, though imperfect, reflects soil-landscape relations either knowingly or tacitly exploited by the original scientists in creating the survey. That is, soils have been placed in characteristic landscape positions whose combination of environmental conditions (e.g., slope, bedrock geology, etc.) are correlated with the appearance of those soils.



Our goal is to develop data mining tools for recovering the soil-landscape model. We do this by overlaying soil map polygons on raster data, as seen below:



The relative frequency of environmental conditions in each polygon of a given map unit reflects the distribution of conditions under which that unit was mapped. That is, the frequency distribution of pixel values contains systematic relations between environmental conditions and soils present in the original survey. Construction of empirical distributions allows discovery of expert knowledge employed in the survey.

3. Methods and Tools

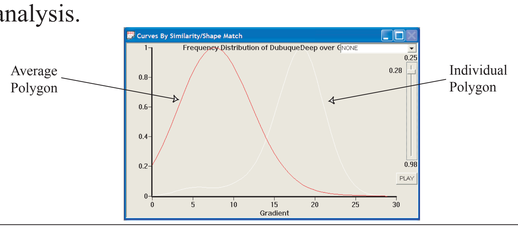
3.1 Univariate Tools

We have developed easy-to-use software for computing univariate distributions based on nonparametric kernel estimation. For a given soil map unit S_k and environmental condition x, the kernel estimate is:

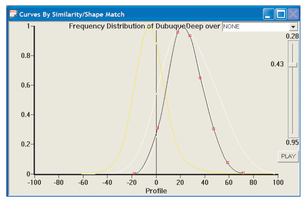
$$\hat{f}_k(x) = \frac{1}{n_k} \sum_{i \in S_k} K\left(\frac{x - x_i}{h}\right)$$

where $K(Z)$ is some suitably chosen kernel (we use the standard normal function) and h is a bandwidth controlling the dispersion of K.

An example from the 1962 Iowa County, WI survey appears below. Our software produces an average frequency curve (red) showing the distribution of a soil along a particular environmental covariate. Frequency curves for individual polygons (white) are also created and can be compared to the average polygon for consistency analysis.



The software also provides for graphical editing of frequency distributions as seen below:



After editing, this new knowledge can be saved and used directly in the SoLIM predictive soil mapping software. These data mining tools and the SoLIM software are available at no cost from:

http://solim.geography.wisc.edu/software

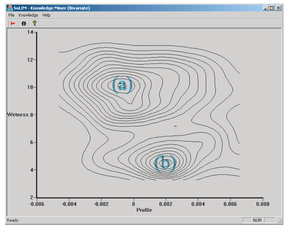
3.2 Bivariate Analysis

Currently under investigation are bivariate plots, where relative frequency is found via a 2-d kernel function:

$$\hat{f}_j(x, y) = \frac{1}{n_j} \sum_{i \in S_j} K\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}\right)$$

where x and y are environmental covariates and h_x and h_y are their corresponding bandwidths. Bivariate plots can potentially reveal features that would be missed in the univariate plots.

Consider, for example, the figure below showing relative frequency as a function of profile curvature and wetness index:



This soil has been mapped in two environments: (a) relatively wet, somewhat concave areas, and (b) drier convex slopes. Environments intermediate between the two are occupied by other soils. Note that the two soil instances would be much harder to detect using plots of either Wetness Index or Profile Curvature alone.

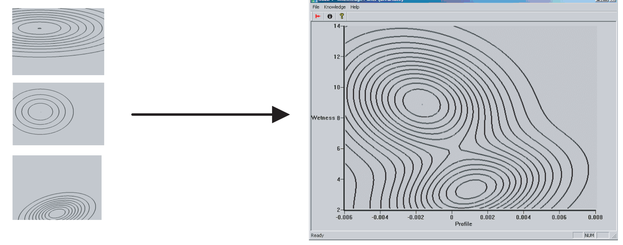
Bivariate functions are attractive because they are easy to visualize as contoured surface (above), or in pseudo-3d diagrams (not shown). However the full set of bivariate density functions is slow to compute owing to the fact that for N environmental covariates, there are $N\ (N\text{-}1)\ /2$ pairs of functions to estimate and contour for each soil type.

3.3 Constructing Bivariate Plots from Multivariate Models

As an alternative to pre-computing a very large number of functions, many of which likely won't be used, we are investigating Gaussian Mixture Models (GMM) as a way of concisely capturing the same information. Let $\mathbf{x}$ be the vector of all covariates and ϕ_j be a multivariate normal distribution with parameter vector θ (means, variances, and covariances). GMM estimates the relative frequency using a linear combination of M Gaussian functions, each with its own parameters:

$$\hat{f}_j(\mathbf{x}) = \sum_{k=1}^{M} \pi_k \phi_j\,(\mathbf{x}|\theta_k)$$

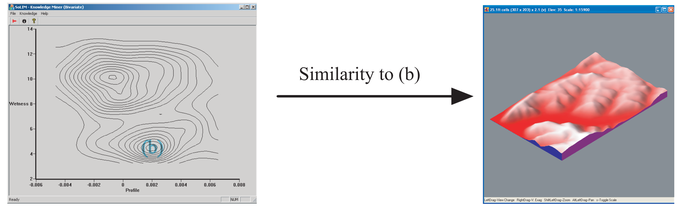
where the π_k are mixing weights. Standard methods are available for obtaining optimal π and θ given the observed $\mathbf{x}$ and their labels (soil type). Although these must be pre-computed, less time is needed than for estimation of all $\hat{f}(x,y)$, and the result is a small set of parameters (π and θ). When a bivariate plot is needed, it can be quickly found by integrating $\hat{f}_j(\mathbf{x})$ over all but whichever two variables are of interest, as seen in the following contour of a 3-component model:



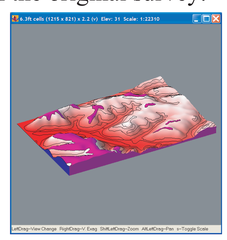
In addition to their computational advantages, GMMs are appealing because the individual mixture components can be interpreted as soil instances. That is, widely separated components amount to modes in the frequency distribution indicative of distinct environments. Research is needed to develop screening procedures that can find these for the soil scientist and also suggest combinations of variables for special scrutiny.

3.4 Linking Attribute and Geographic Spaces

Continuing with this example, a scientist will likely want to know where the instances appear in the landscape. To assess that, we can compute the similarity of landscape positions to a combination of attributes (x,y). For example, suppose the scientist clicks on the frequency mode in the lower right. Every point in geographic space can be scored for its similarity to the wetness, curvature pair selected. This is shown schematically in the figures below, where whiter is used for landscape positions highly similar to the selected point in attribute space.



An obvious followup would be to drape map polygons for the chosen soil on the landscape so that the analyst can see which highly similar places contain and which don't contain that soil in the original survey:



4. Summary

Visual spatial data mining tools deserve a place alongside automated methods for soil survey updates. Unlike purely quantitative approaches, visual techniques like those described here allow for discovery of new concepts not present in the original survey, and they provide for direct control over modification of existing concepts. Although initial steps have been taken, further work is needed to develop additional methods and intuitive software that implements those methods.

Acknowledgements

The support of USDA-NRCS National Geospatial Development Center and the Wisconsin State NRCS is gratefully acknowledged.

Contacts

James E. Burt
jeburt@wisc.edu
(608)263-4460

A-Xing Zhu
azhu@wisc.edu
(608)262-0272

Rongxun Wang
rongxunwang@wisc.edu
(608)262-1857