New Digital Soil Survey Products to Quantify Soil Variability Over Multiple Scales

D.E. Beaudette¹ and A.T. O'Geen¹*

Objectives

1. Create a framework for the rapid digitization and proofreading of historic paper-based soil characterization data. Digitized data will be stored in PedLogic, an existing soil database designed to store a variety of data associated with soil analysis and description, developed in the University of California, Davis, Soil Resource Lab.

2. Produce raster-based depictions of soil landscape relationships. The resulting grid-based data will be used to up-scale and extrapolate point observations, as well as refine and down-scale existing soil survey polygons. Models will be constructed from point data (current and historical sources) and selected environmental data that reflect soil-forming factors (e.g., microclimate, landscape features). An assessment of uncertainty at each grid cell will be derived from regression diagnostics, such as standard error of the estimate. Models will be evaluated using bootstrapping-style tests for over fitting, and an independently collected verification data set.

Approach and Procedures

Several programs, written in the Python programming language, were developed to automate the conversion of paper data sheets into digital representations, that were then stored in a relational database management system (Postgresql). It was possible to digitally convert the scanned images into text (via optical character recognition or OCR) at a very high level of accuracy by performing several image processing and segmentation steps. An extension to an already existing Web-based application (PedLogic) was created to facilitate proofreading of the digitized data by an experienced operator (fig. 1).

A solar radiation modeling approach was used to quantify landscape-scale variation in microclimate. We used the ESRA solar model (Rigollier et al. 2000) to estimate an energy budget at each grid cell of a 10-meter resolution DEM from Pinnacles National Monument, Calif. Parameterization of the clear-sky version of this model was accomplished with daily estimates of the Linke turbidity factor, using local pyranometer measurements (11-year record) (Rapti 2000). Logistic regression, a type of generalized linear model (GLM), was used to couple annual solar radiation load values with geologic class to predict the spatial distribution of upland Mollisols within Pinnacles National Monument, Calif. (fig. 2). The resulting probability values were used to predict surface carbon and nitrogen concentrations as well.

¹University of California, Davis, Department of Land Air and Water Resources
*Principal Investigator
Figure 1. Example screen shot from the Web-based proofreading application. Note that the small images "chunks" above text boxes are from the scanned paper surveys, and the contents of the text entry boxes are from the OCR (optical character recognition) process. Color coding is used to help the operator detect subtle mistakes, such as transposition of '0' and 'O' or blank fields.

Results

Each entry from the three-volume published Soil-Veg database was scanned, converted into text via OCR (optical character recognition), inserted into a relational database, and finally proofread for mistakes (fig. 1). These data consisted of up to 40 chemical assays for each sampled horizon, a full morphologic description for each horizon, and a brief description of the sample site's geology and geomorphology. Now that flexible framework for working with historic sources of pedon data is ready, we will begin digitizing data from other sources. The 557 pedon records from the Soil-Veg database are now ready to be updated to modern conventions (e.g., horizon designations, units). This framework was initially tested and populated with data from the Pinnacles National Monument soil survey.

Modeled daily irradiance values matched the local weather station data with an R2 of 0.965 (n=365, p < 0.001). Solar radiation values coupled with a local geologic map were used as predictor variables in a logistic regression model constructed to predict the spatial distribution of upland Mollisols. A total of 185 field observations were used to build the model, which had an 83% PCC (percent correctly classified) rate and a receiver operating characteristic (ROC) area of 0.78 to 0.96. A 100-fold cross-validation (repeated re-fitting of the model with a subset of observations) procedure indicated a mean classification error rate of 23%. Field validation sites...
(n=35) showed a 77% PCC rate, and an ROC area of 0.90. New soil survey products including an estimated probability map of Mollisol distribution were produced (fig. 2).

**Figure 2.** Predicted occurrence (probability) of upland Mollisols within Pinnacles National Monument, Calif. A logistic regression model was constructed from point observations, modeled annual solar radiation load and a geologic map. The black dots are the original observation points, and white lines are the soil survey map unit delineations (as determined by NRCS soil mapping staff). Note that the Mollisol probability map depicts variation at a much finer scale than the soil survey, even though the two products were created from the same observation points.

**Discussion**

The framework of applications developed to support digitization of historic, paper-based soil characterization data has added considerable capability to our existing soil database tool called PedLogic. The approximately 90 hours of software development time spent resulted in an efficiency slightly higher than manual entry of each data element. It was estimated that each record (chemical and morphologic data) would have taken an experienced operator about 30 minutes of time to manually type in. Including software development time, the 557 records from the Soil-Veg database required approximately 25 minutes per record. The next phase of digitization will use the existing framework, and should require about 15 minutes per record.

The quantification of upland microclimate variability via solar radiation modeling proved to be a versatile tool for predicting soil properties. Methods developed for the calibration of the ESRA radiation model, through local weather data, provide a simple mechanism for extending this approach to other locations (i.e., the benchmark catenas outlined in the original proposal).
Logistic regression and related ordinal response methods are flexible and efficient tools for linking \textit{categorical} soil properties observed at point locations (pedon scale) to remotely-sensed or DEM-based (raster) datasets. This "link" (i.e., the statistical model) represents a digital analogue to the traditional mental models that have been successfully used by field mappers for decades (Hartung et al. 1991). Model predictions, in the form of probabilities, deviate from the traditional map unit approach (i.e., well-defined polygons), however, their interpretation are not incompatible with soil survey (fig. 2). On the contrary, raster-based probability maps of soil features can address two long-standing issues with soil survey: 1) explicit depiction of map unit components, and 2) an assessment of uncertainty. In a general sense, digital soil mapping techniques given rise to a transparent, repeatable, and compact media for conveying the essentials of the mental models of soil formation created by experienced field mappers. Successfully preserving these mental models would fill a long standing gap in Soil Survey (Arnold 2006).

Statistical models that encapsulate the relationship between soil properties observed at point observations and as grid cells (process-based or factorial data layers) are a proven first step in the scaling process. Once suitably accurate models are created, prediction at unsampled locations, defined by a grid size and extent, represent an up-scaling operation (fig. 2). Suitable interpretation of the resulting raster data, with respect to existing soil survey polygons, would represent a down-scaling operation (fig. 2).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Overview of the relationship between point observations (used to build statistical models), existing soil survey map unit delineations, and predicted grids of soil parameters.}
\end{figure}
References


This research was funded by the Kearney Foundation of Soil Science: Understanding and Managing Soil-Ecosystem Functions Across Spatial and Temporal Scales, 2006-2011 Mission (http://kearney.ucdavis.edu). The Kearney Foundation is an endowed research program created to encourage and support research in the fields of soil, plant nutrition, and water science within the Division of Agriculture and Natural Resources of the University of California.