CHAPTER

5

Digital Soil Mapping

By Suzann Kienast-Brown and Zamir Libohova, USDA-NRCS, and Janis Boettinger, Utah State University.

Principles and Concepts

Digital soil mapping is the generation of geographically referenced soil databases based on quantitative relationships between spatially explicit environmental data and measurements made in the field and laboratory (McBratney et al., 2003). The digital soil map is a raster composed of two-dimensional cells (pixels) organized into a grid in which each pixel has a specific geographic location and contains soil data. Digital soil maps illustrate the spatial distribution of soil classes or properties and can document the uncertainty of the soil prediction. Digital soil mapping can be used to create initial soil survey maps, refine or update existing soil surveys, generate specific soil interpretations, and assess risk (Carré et al., 2007). It can facilitate the rapid inventory, re-inventory, and project-based management of lands in a changing environment.[1]

SCORPAN Model

The scientific foundation of soil mapping is Hans Jenny's (1941) conceptual model that soils (S) on a landscape are a function of five environmental factors, namely climate (cl), organisms (o), relief (r), parent material (p), and time (t):

$$S = f(cl, o, r, p, t)$$

While this model, sometimes known as CLORPT, has been useful in conventional soil mapping, it is not quantitative nor spatially explicit.

[1] Trade or company names used in this chapter are for informational purposes only. This use does not constitute an endorsement by USDA–NRCS or the contributing authors of this chapter.

To represent soil and the related environmental factors in a spatial context and express these relationships quantitatively, McBratney et al. (2003) proposed the SCORPAN model, where soil (as either soil classes, S_c, or soil attributes, S_a) at a point in space and time is an empirical quantitative function of seven environmental covariates: soil (s), climate (c), organisms (o), relief (r), parent material (p), age (a), and spatial location (n):

$$S_{c,a} = f(s, c, o, r, p, a, n)$$

The important advances of the SCORPAN model for use in digital soil mapping are: (1) the recognition that the environmental factors are not necessarily independent of each other and are thus defined as environmental covariates, (2) the inclusion of soil as an environmental covariate, (3) the spatially explicit nature of the model, and (4) the quantitative nature of the functional relationships. In the SCORPAN model, soil, either as point observational data, existing soil maps, or remotely sensed spectral properties, can be used as input data. Environmental covariates are digital and spatially explicit data in a raster that is processed using a geographic information system (GIS). The SCORPAN model facilitates the quantification of the relationships between spatially explicit digital environmental covariates and the soil classes or attributes to be predicted in a spatial context. It also facilitates the estimation of error or uncertainty of the spatial prediction of soil classes or properties.

Digital vs. Conventional Soil Mapping

The availability and accessibility of geographic information systems (GIS), global positioning systems (GPS), remotely sensed spectral data, topographic data derived from digital elevation models (DEMs), predictive or inference models, and software for data analysis have greatly advanced the science and art of soil survey. Conventional soil mapping now incorporates point observations in the field that are geo-referenced with GPS and digital elevation models visualized in a GIS. However, the important distinction between digital soil mapping and conventional soil mapping is that digital soil mapping uses quantitative inference models to generate predictions of soil classes or soil properties in a geographic database (raster). Models based on data mining, statistical analysis, and machine learning organize vast amounts of geospatial data into meaningful clusters for recognizing spatial patterns.

Various digital soil mapping tools, methodologies, and inference models have been developed and tested in the U.S. and abroad to facilitate the rapid visualization and quantification of landscape patterns at multiple spatial scales (Lagacherie et al., 2007; Hartemink et al., 2008; Behrens et al., 2010; Minasny et al., 2012). A significant amount of the data used in digital soil mapping can be archived in a spatially explicit digital format in a GIS, so the expert knowledge used to predict soil distribution on the landscape is retained. Objective sampling plans can be implemented to statistically capture variability of the landscape, representing it by digital environmental covariates. Probably the most exciting aspect of digital soil mapping is the ability to generate spatially distributed information on soil classes and/or properties and the associated estimate of uncertainty (the probability that a particular soil type and/or property occurs at a specific point on the Earth's surface). There is a great demand globally for spatially distributed soil information. This is evidenced by the launch of GlobalSoilMap (Arrouyas et al., 2014), a project to make a digital soil map of the world using state-of-the-art technologies for soil mapping and predicting soil properties at 100-m resolution.

Maps that predict the spatial distribution of soil classes or properties are of interest in many countries because they inform soil use and management decisions. Digital soil mapping better captures observed spatial variability and reduces the need to aggregate soil types based on a set mapping scale (Zhu et al., 2001). An important component of digital soil mapping is the method of analysis used to define the relationship between soil observations and environmental covariates. Many types of methods have been investigated, including expert systems (Cole and Boettinger, 2007; Saunders and Boettinger, 2007; Zhu et al., 2001), unsupervised classification (Boruvka et al., 2008; Triantifilis et al., 2012), and machine learning or predictive modeling (Behrens et al. 2005; Behrens and Scholten, 2006; Bui and Moran, 2003; Stum et al., 2010; Brungard et al., 2015).

Discrete vs. Continuous Models

Discrete Models

A map of soil classes, such as soil map units, is a type of discrete, or crisp, model (Hole and Campbell, 1985; Burrough and McDonnell, 1998). Discrete models represent thematic or categorical data in which the values represent a predefined class with a finite number of possibilities. These models are typically nominal, ordinal, or binary and

therefore lack numerical meaning. When applied in a raster, each pixel value represents the class associated with the pixel (e.g., soil class A, soil class B, soil class C, etc.). Mathematical operations cannot be applied directly to discrete data because the values do not have true numerical meaning (e.g., soil class B is not twice as great as soil class A).

Soil mapping has traditionally used the discrete model to represent distinct soil types and groups of soil types on the landscape. In a raster environment, discrete models simplify the display of modeled classes and align conceptually with the conventional soil survey approach. However, discrete soil class models present the assumption that soils are constant across a class. Classes can be defined either narrowly or broadly for any soil landscape unit, similarly to how the traditional map unit can be categorized as a consociation or complex. Narrowly defined classes are best for providing site-specific interpretations and are most suitable in situations where sufficient field observations (training data) are available to adequately define the classes. Broadly defined soil classes may help bridge the gap from conventional (polygon, vector) to digital (raster) soil mapping and are most suitable in situations where field observations (training data) are limited.

Continuous Models

A map of soil properties is a type of continuous model. Continuous models represent data in which the values are measurements or calculations that have numerical meaning and represent a continuum. In a raster environment, each pixel value represents a real quantitative value (measured, calculated, or inferred) and can have various levels of precision (e.g., integer or floating point). Continuous models allow for any value over a continuous range, whereas discrete models have only a finite number of predefined outcomes.

Continuous soil models are designed to handle the continuous nature of soil properties more realistically than discrete models. In theory, continuous models eliminate the disadvantages of predefined classes and distinct boundaries in soil mapping. In practice, the continuity depends upon the cell size and the precision used. Predictions of soil properties are typically represented with a continuous data model.

The majority of the environmental covariates used in digital soil mapping are continuous data models. Terrain attributes derived from a digital elevation model (DEM), such as slope gradient, curvature, and area solar radiation, are continuous models. Spectral data, such as reflectance, derived from satellite or aircraft remote-sensing platforms are also continuous models.

Stages and Processes

Typically, each digital soil mapping project is unique. Many aspects of a project may vary (e.g., the objectives of the project, the biophysical properties of the study area, the availability of environmental covariates, the method of prediction applied). However, the stages and processes of digital soil mapping should be consistent in all projects. Each stage comprises a series of specific objectives that must be accomplished for the digital soil mapping project to progress. The digital soil mapping process is iterative and requires review and assessment at several points. The stages and processes of digital soil mapping projects are outlined in the following list and described in the following subsections.

Outline of Stages and Processes

Stage:

1. **Define area and project scope**
   a. Define and refine objective: soil classes or properties
2. **Identify physical features of interest**
   a. SCORPAN—important covariates and appropriate data
   b. Scale of processes and measurements
   c. Available measurements (field and remote sensing)
3. **Data sources and preprocessing**
   a. Identify and acquire data
   b. Assess data quality
   c. Organize data
   d. Preprocess data
4. **Data exploration and landform analysis**
   a. Derive terrain and spectral data products
   b. Select appropriate predictors
5. **Sample for training data**
   a. Case-based and *a priori* samples
   b. Field samples

   Review and assess:
   - Do the data layers represent the important environmental covariates?
     - o Yes—proceed to Stage 6
     - o No—return to Stages 2, 3, and 4
   - Are the training data adequate to predict the classes or properties of interest?
     - o Yes—proceed to Stage 6
     - o No—return to Stage 5

6. **Predict soil classes or properties**
   a. Choose and apply appropriate prediction method
      i. Soil classes – unsupervised or supervised classification, predictive modeling
      ii. Soil properties – predictive modeling, geostatistics

   Review and assess:
   - Are the prediction results reasonable?
     - Yes—proceed to Stage 7
     - No—apply a different prediction method, combination of predictors, or set of training data—return to Stages 4, 5, and 6
7. **Calculate accuracy and uncertainty of results**

   Review and assess:
   - Are accuracy and uncertainty results acceptable?
     - Yes—proceed to Stage 8
     - No—revisit prediction method, predictors, and training data—return to Stages 4, 5 and 6
8. **Apply digital soil mapping**
   a. Produce soil class or property maps
   b. Evaluate existing maps
   c. Create soil information products
   d. Apply to other disciplines

Defining the Area and Scope of the Project

Before beginning a digital soil mapping project, it is important to clearly define the project area and scope. For example:

- What is the specific objective of the project?
- Is the project intended to create initial soil survey information or to update existing soil mapping and data?
- Is the objective to produce a map for a specific purpose?
- What is the geographic extent of the project area?
- What are the biophysical characteristics of the area?
- How are the biophysical characteristics of the area related to the distribution of soils on the landscape?
- At what spatial scale is the expected variation in soil distribution expressed (local vs. regional)?
- Are soil classes and/or soil properties to be predicted?
- What is the scale of the final map product(s)?

Digital soil mapping can address a variety of questions. The key is to determine how digital soil mapping can be applied in *your* project area to achieve *your* objectives.

Identifying the Physical Features of Interest

Environmental Covariates and Appropriate Data

The first step after defining the area and scope is to determine which environmental covariates are most important to soil development and distribution in the project area. Once these are determined, the related specific terrain and spectral characteristics can be identified and appropriate digital data selected to allow the discrimination of those physical phenomena. Five environmental covariates in the SCORPAN model are commonly derived from digital data: soil properties (s), organisms (o), parent material (p), relief (r), and climate (c). How humans have altered the Earth's surface may also be considered, which in some cases can represent the time or age (a) covariate.

Soil (s).—Soil can be represented by covariates derived from: (1) georeferenced point data representing field and/or laboratory measurements, (2) remotely sensed spectral data, or (3) existing soil maps. Digital data may include point data such as soil taxonomic class, soil depth to bedrock, or soil chemical or physical properties by genetic horizon (e.g., soil laboratory data associated with a georeferenced sample location at the NRCS Kellogg Soil Survey Laboratory). Surface or near surface properties of the soil may have diagnostic spectral signatures distinguishable by remote sensing data. For example, Nield et al. (2007) used Landsat 7 ETM+ data to digitally map the occurrence of soils with surficial accumulations of gypsum, which was distinguished by a normalized difference ratio of the two shortwave-infrared (SWIR) bands (bands 5 and 7). Existing soil class data in the form of soil maps may also be useful, particularly in soil survey update projects or in disaggregating soil map unit associations into soil components (Nauman and Thompson, 2014).

Organisms (o).—Organisms are typically represented by vegetation or land cover digital data, including existing land cover data and remotely sensed spectral data. Existing land cover data can include maps of vegetation, land use, and species distribution, such as those available from the National Gap Analysis Program (USDI-USGS, 1999). Vegetation is commonly represented by remotely sensed spectral data because green vegetation reflects near infrared (NIR) and absorbs red electromagnetic radiation. The Normalized Difference Vegetation Index (NDVI) is a normalized difference band ratio of the NIR and red bands of a multispectral image. The values range from -1.0 to 1.0—higher values indicate higher vegetation density. NDVI can be quantified for any spectral data source that contains NIR and red bands, such as Landsat data. For example, NDVI was an important covariate in digitally

mapping the occurrence of badlands with very low vegetation cover in the Powder River Basin in Wyoming (Cole and Boettinger, 2007).

Parent material (p).—Parent material can be derived from a geology map or gamma radiometric data or by using remotely sensed spectral data to discriminate mineralogical correlates of parent material. Mineral assemblages in different parent materials (rocks and sediments) will vary in spectral response. Mineralogy is particularly responsive in the SWIR range of the electromagnetic spectrum, represented by Landsat TM or ETM bands 5 and 7, Landsat 8 OLI bands 6 and 7, and Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) bands 4 through 9. For example, the San Francisco Mountains in the Great Basin of southwestern Utah are characterized by mixed sedimentary rocks (mainly quartzite) intruded by igneous rocks (mainly andesite) with mixed basin fill. A principal components analysis of Landsat ETM+ bands 1 through 5 and 7 helped distinguish an andesite intrusion from sedimentary rocks and showed the influence of andesite on the composition of the alluvium downslope from the intrusion (Stum et al., 2010).

Relief (r).—The covariate representing relief can be derived from elevation data, such as Light Detection and Ranging (LiDAR), the National Elevation Dataset (NED), Interferometric Synthetic Aperture Radar (IFSAR), photogrammetric data, etc. These data derivatives are known as terrain attributes or elevation derivatives. Examples of terrain derivatives are slope gradient, slope length, slope curvature, wetness index, ruggedness index, slope aspect, landform, and relative elevation. Various combinations of terrain attributes can generate geomorphic surfaces and describe processes related to soil development.

Climate (c).—The climate covariate can be approximated in some areas by elevation, especially in landscapes subject to orographic effects (i.e., higher elevations are subject to cooler temperatures and greater amounts of precipitation). Regional climate models and data are also available, e.g., climate data at about 800-m resolution in the U.S. from the PRISM Climate Group (2016). Solar radiation is commonly an excellent proxy for climate, particularly in aspect-driven climate scenarios. Solar radiation models are widely available and can be calculated in various GIS software packages.

Age (a).—While not commonly considered a SCORPAN covariate, soil age has a major impact on the degree of profile development and soil properties. Humans, for example, play an important role in altering the landscape and/or land cover, thus changing soil properties (attributes), soil classes, and land use. Therefore, in some cases the human impact on the landscape can represent age. One example is the northern part

of the Las Vegas area, Nevada, where humans have urbanized the arid desert landscape and created green space via irrigation. In many areas, petrocalcic horizons have been destroyed, changing the habitat necessary for rare endemic plant species, and irrigation has altered soil properties and regional hydrology by leaching salts out of soils, raising water tables, and disrupting natural waterflow patterns. Human alterations of the landscape and land cover may also indicate soil properties. For example, the parts of a landscape converted into agriculture may indicate the location of soils that have desirable properties, such as lower contents of rock fragments or lower levels of salinity.

Scale of Processes and Measurements

The processes responsible for the development and distribution of soils on the landscape operate over a wide range of spatial scales, from continental (e.g., tectonic events and glaciation) to regional (e.g., deposition of alluvium and windblown sand) to hillslope (erosion and deposition) to pedon (addition, removal, transformation, and translocation of materials). These processes, their interactions, and their scale of spatial expression can create complex soil patterns. The processes must be understood and represented by the appropriate measurements for both environmental covariates and field observations. Digital data can be used to stratify landscapes into relatively homogenous geological and geomorphic units, which are helpful in understanding these processes and developing an appropriate design for collecting data in the field.

Field measurements.—Field measurements in digital soil mapping are derived from georeferenced points. They may be full or abbreviated pedon descriptions and associated laboratory data. The goal is to predict soil classes and properties beyond the location of field observations. Soil sample size and the area or volume of representation should be considered when determining the location of field sampling sites and timing of measurements (Bouma et al., 1989; Mohanty and Mousli, 2000).

Remote sensing measurements.—Remote sensing has been defined as the "art and science of deriving information from measurements made at a distance" (Colwell, 1997). Remote sensing measurements detect electromagnetic radiation from the Earth's surface in two different ways: passive and active. Passive remote sensing collects electromagnetic information produced as a result of the interaction between the sun's energy and surface materials, such as measurements collected by satellite sensors. Active remote sensing collects information returned from the Earth's surface as a result of an emitted signal, such as LiDAR (Light Detection and Ranging) or radar. (See chapter 6 for more information on remote sensing and other tools for proximal soil sensing.)

Remote sensing measurements that provide digital elevation and spectral response data are commonly used in digital soil mapping. The remote sensing of topography via passive sensors (e.g., aerial photographs) or active sensors (e.g., LiDAR) results in the generation of digital elevation models. The use of digital elevation models in soil mapping is extensive and well documented because variations in relief are closely linked to the distribution of soil properties and classes. Remote sensing of spectral data provides direct information about the surface properties of soils, vegetation, or other materials. Spectral properties remotely sensed at the surface can be related to environmental covariates that control soil development. The spectral properties can therefore potentially be used to infer other soil characteristics. Specifically, remote sensing data can be used to map the variations in relief, climate, organisms, parent material, and even time (indirectly).

When reviewing remotely sensed data sources, the data collection mechanism, the extent and consistency of the data, and the scale of the data compared to the scale of the physical phenomena need to be considered. The spatial detail, the spectral wavelengths of imagery, and even the season of the year or other temporal aspects of the physical environment that influence the timing of data acquisition should also be considered.

Because remote sensing measurements are collected at varying spatial and spectral resolutions, careful consideration should be given to selecting data at the appropriate spatial and spectral scale to represent the environmental covariates and processes in the project area. The focus should be the specific scope of a project, e.g., what spatial and spectral resolution is most appropriate for the question(s) being asked? These needs should then be compared against the range of data that is actually available given budget or other constraints.

Selecting Data Sources and Preprocessing

Identify and Acquire Data

One of the most critical steps in a digital soil mapping project is selection of the data. Incorporating data that match the question or problem being considered is essential to the success of the project. The properties of the data should be directly related to the physical attributes and soil-forming processes in the area of interest. For example, in mountainous areas a 30-m DEM might adequately characterize the significant features on the landscape. In low-relief areas where soil formation is driven by very subtle changes in topography, a much higher resolution DEM may

be necessary to adequately characterize the terrain features. Several studies have shown that soil-landscape relationships exist over a range of scales (Thompson et al., 2001; Smith et al., 2006; Park et al., 2009). Spatial information commonly has to be down-scaled or up-scaled to match other environmental covariates.

A project may require a mix of data to adequately represent the multiple SCORPAN covariates that influence soil development in a particular area. Elevation derivatives and spectral derivatives are a powerful combination for predicting soil classes or properties in most areas. However, depending on the question being considered and the physical features of the area, a project may require only one of these data sources.

In the United States, there are multiple sources for both DEMs and remote sensing images. One of the largest archives of remote sensing imagery is the USGS EarthExplorer site (USDI-USGS, 2016a). The USGS National Elevation Dataset provides DEMs for most locations (USDI-USGS, 2016b). Many States have archives available for DEMs (USGS and LiDAR), Landsat, and ASTER imagery and should be investigated as potential data sources. The NRCS Geospatial Data Gateway also provides many different types of data layers (USDA-NRCS, 2016a).

Assess Data Quality

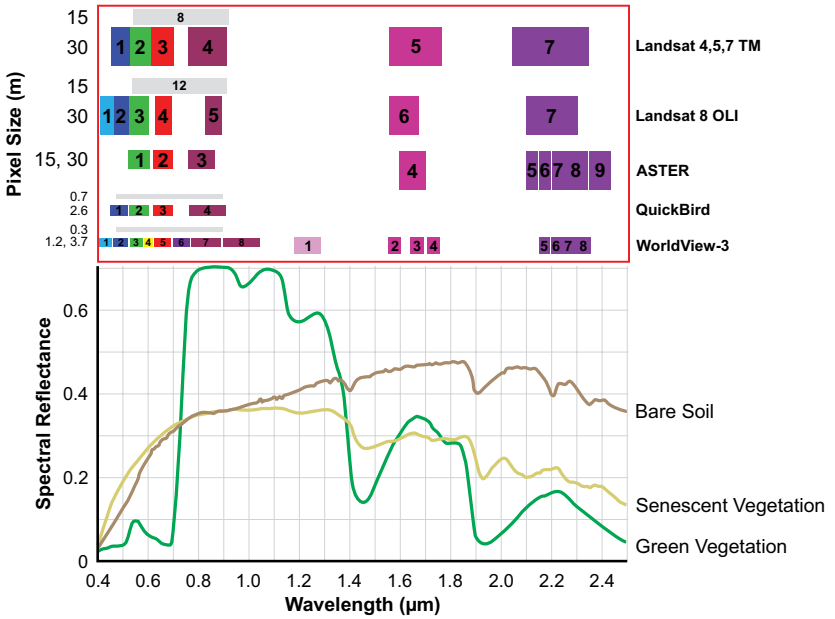
Once data sources have been identified, the quality of the data should be assessed to ensure the best data available are being used for model development. Data attributes to be considered include resolution, spatial projection, units, and source.

Resolution.—Resolution of the data is one of the most important attributes to consider when selecting data. Many high-resolution data sources are currently available, but they may not address the problem being considered. High-resolution data can provide "too much information" and add undesirable noise and/or excess data storage and processing time to analysis and modeling. The scale of physical features or properties on the landscape should be considered when choosing the most appropriate resolution.

The types of resolution—spatial, spectral, temporal, and radiometric—must be considered. Spatial resolution applies to all data sources and equates to grid cell size. In deciding the appropriate spatial resolution, the features of interest on the landscape must be considered; the grid cell size must be able to adequately capture the desired features. One rule of thumb is that the smallest object recognized should be equivalent to four grid cells of a DEM (Rossiter, 2003).

When considering spectral data derived from remote sensing sources, spectral resolution may be the most important attribute. Spectral resolution refers to the number of bands of data a sensor provides and which part of the electromagnetic spectrum they capture. Generally, the red and NIR part of the spectrum is most important if the focus is vegetation and the SWIR part of the spectrum is most important if the focus is minerals, parent materials, or bare soils (fig. 5-1).

Figure 5-1



Comparison of spectral bands of common sensors to the reflectance spectra of common materials.

Temporal resolution indicates the time of year and frequency of image acquisition. Seasonality or repetition of image acquisition over several years may be an important variable. In addition, noting the date of acquisition is important if several images are mosaicked together. Ideally, the images for a mosaic should be acquired on or near the same date to minimize differences in atmospheric and Earth surface conditions. If data meeting those criteria are not available, and data from different years are used, the data used should at least be from the same time of

year. Image acquisition frequency typically ranges from every day (e.g., MODIS and AVHRR) to every 16 days (e.g., Landsat).

Radiometric resolution is an important, though rarely considered, spectral sensor attribute. It refers to the number of gray levels the sensor can potentially differentiate. Gray levels describe the brightness values (BV) or the digital number (DN) values that are recorded for an image. Because these quantization values are integers, they are only whole numbers. Therefore, there is a direct correlation between the range of numbers that are used in describing an image and the level of detail in the brightness variation.

Spatial projection.—It is important to ensure that all digital data are the same spatial projection (geographic vs. projected datum, etc.) for ease of processing. There are many software packages that can be used to define the projection (if data comes without a projection file but the projection is known) and re-project the data. The georeferencing of the data should be checked by comparing key features in a data source with the same key features in a reliable image source, such as the National Agriculture Imagery Program (NAIP). If georeferencing needs to be corrected, many software packages offer this functionality.

Units and data type.—Understanding the units of the data and how to interpret them is important. If units between data sources are not compatible (e.g., feet vs. meters for a DEM), values may need to be converted. Data ranges should be noted as they will impact certain classification methods.

The data type and how the data are stored should also be noted, such as whether the data is a floating point (contains decimal places) or an integer and the number of bits of the data. For integers, 1-bit data are binary and store 2 values (0 and 1), 8-bit data can store 256 values, and 16-bit data can store 65,536 values. Floating point numbers are either single (32 bit) or double (64 bit). Another aspect to consider is the file type: thematic (discrete, categorical) or continuous. Typically, thematic data are integer and continuous data are floating point. The numbers in a continuous dataset have intrinsic meaning and represent real physical measurements (elevation or reflectance) or are the result of a calculation that has been performed on the data (e.g., wetness index from elevation, spectral band ratio from reflectance). In contrast, thematic (categorical) data typically represent an interpreted class. All of the information needed to properly understand the data typically can be found in the file's metadata.

Data issues.—Several issues may occur with data, but most can be resolved during preprocessing. For spectral imagery, issues include clouds, smoke, sun glint, data loss, and calibration. When possible,

another image of better quality should be used. Images without clouds and smoke are preferable since these issues cannot be resolved through preprocessing. If an alternate image is not available, data preprocessing techniques should be tried to reduce the impact of sun glint, data loss, or calibration issues on analysis.

Elevation data are developed to model the bare earth terrain features from a number of sources, and each source has a unique set of issues. The most frequently used form of elevation data in digital soil mapping is a raster surface comprised of a matrix of cells arranged in rows and columns. The elevation values in the cells can be interpolated from points or contour lines. The accuracy of the elevation values themselves is commonly reported for data sources and indicated in the metadata. Accuracy typically is expressed using root mean squared error (RMSE) as related to the absolute error of the elevation surface. Smaller RMSE values more closely match the absolute elevations of the modeled surface. The spatial resolution of a cell determines the level of characterization detail that can be attained for the analysis of the bare earth terrain features. The cell size used should not exceed the accuracy level of the source data.

DEMs derived from hypsography (digital contour data), also called HypsoDEMs, will have a characteristic contour-line bias, which is expressed as an artificial, terraced landscape. DEMs produced from hypsography may also have flat-topped ridges, peaks, and indistinct junctures between footslopes and toeslopes. The contour line interval is a critical factor when considering the use of DEM-derived data for terrain analysis, especially in areas of low relief. Derivative products created from the HypsoDEM in which the contour line interval exceeds the change in relief will portray features that reflect the locations of the contour lines and not the features of the terrain surface. No satisfactory solutions are available to correct this problem.

DEMs produced from LiDAR may have areas of uncertainty associated with vegetation and the presence of water. Areas with dense vegetation may have few to no returns of the emitted signal from the actual ground surface. The LiDAR sensor rarely receives a return when the pulse makes contact with any surface water. Areas with very shallow water will have either no data collected or will have points where the pulse contacted vegetation above the water surface. In these cases, the DEM will have elevations that are greater than the actual ground surface. If there are isolated patches of dense vegetation, artificial “spikes” may occur in the DEM. In areas with mixed land cover, such as cultivated cropland and small woodlands, the effect of the wooded areas may be

pronounced. Performing a minimum focal filter in association with iterative focal smoothing operations can help minimize these problems.

DEMs produced from radar (such as x-band radar, e.g., IFSAR) will not represent the bare earth surface where vegetation is present unless augmented with elevation data from another source. Unlike LiDAR, IFSAR produced with x-band radar will not adequately penetrate vegetation to model the bare earth surface. In addition, IFSAR is not sensitive to features with abrupt changes in slope, such as narrow, convex ridges or concave, closed depressions. Because DEMs from this data source may mute the expression of such features, modifications should be made to accurately reflect terrain derivatives such as slope or curvature, or the less defined representation of the features should be acknowledged.

Artifacts derived from the data management scheme used in the source data may be apparent with DEMs developed from LiDAR or radar data. Tiling is an effective method of managing and processing the large volumes of source data. It organizes the data into small, systematic, rectangular grids. The juncture between adjacent tiles may introduce inadvertent artifacts. One or several smoothing (Gaussian or focal) operations may be able to adequately blend away these artifacts. However, the best practice is to consult the original data source (if available).

Organize Data

A data management plan is needed at the onset of a project. It should include a common directory structure, file naming convention, minimum metadata standard or other means of documentation, and a data backup process. This plan is particularly important if the project will include multiple members of a team accessing and utilizing the same data. It should be simple enough for the members to effortlessly implement.

One approach is to keep the original data sources separate from the processed data. The folder structure should represent the steps in the process, and the names of folders and files should reflect their content. The processing and analysis steps and the file naming convention should be kept in a separate document or in the metadata. Regardless of the folder structure and naming conventions, the processing steps of the project could be used as a guide to organizing the data.

Preprocess Data

Data rarely are in an immediately usable format. Some degree of preprocessing typically is needed before the data can be incorporated

into analysis or modeling. Some basic guidelines for data preprocessing are:

- Ensure that all data are in the same projection and have the same extent.
    - Select natural boundaries when possible for the project area and include a buffer around the perimeter of the area when clipping or subsetting data for processing. This minimizes or eliminates potential edge-effect from processing along the margins of a dataset.
    - Use a snap raster to maintain consistency in grid cell alignment.
- Validate georeferencing with a reliable image source (e.g., NAIP).
- Normalize spatial resolution (grid cell size) between layers.
    - If multiple datasets are being combined, it may be best if they share a common spatial resolution.
- For elevation data, include in DEM preparation:
    - Filling sinks and trimming peaks;
    - Removing linear, human-made artifacts (e.g., roads, railroads, channelized waterways);
    - Applying a low-pass filter or other smoothing algorithm; and
    - Ensuring that derivatives based on hydrology (e.g., flow accumulation, upslope contributing area, topographic wetness index, stream power index) encompass entire watersheds for consistent interpretation and application of values across the entire project area.
- For spectral data, apply image standardization or atmospheric correction to calculate surface reflectance when:
    - Mosaicking images for classification (if images were not acquired on the same day/time and under the same atmospheric conditions);
    - Calculating band ratios;
    - Extracting biophysical information from the image (biomass, NDVI); and
    - Extending class signatures across multiple images, particularly if images were acquired on a different date or location.

  Landsat 4, 5, 7, 8 surface reflectance products are available from USGS EarthExplorer (USDI-USGS, 2016a).

- If a mosaic is required, apply all the preprocessing prior to mosaicking.
- Stratify the area to reduce variability for analysis, modeling, or classification.
    - Choose a stratification that applies to the overall goal of the project and is based on natural boundaries, such as geology, elevation, physiographic areas, etc.

Data Exploration and Landscape/Landform Analyses

The process of digital soil mapping requires exploring the data available for a project and linking it to key SCORPAN covariates and pedological knowledge. With the soil processes and end goal of the project in mind, an exploration analysis should be used to determine if the data will provide adequate information on the variability and distribution of key covariates across the area of interest. Commonly, unexpected variation in the data is discovered and an evaluation is needed to determine if real information or noise is represented. In most cases, the development of terrain or spectral data derivatives is necessary to exploit the data to its full potential for predicting soil classes or properties.

Deriving Terrain Attributes

Terrain attributes are derived from DEMs and are typically represented using the raster data format. Elevation can also be represented as points (e.g., LiDAR returns) or triangulated irregular networks (TIN), but the raster format is typically preferred due to its greater flexibility. Elevation data are typically developed from contours, topographic surveys, or LiDAR data. Terrain attributes may be broadly grouped into two categories: (1) primary attributes, which are computed directly from a DEM; and (2) compound attributes, which are combinations of primary attributes (Moore et al., 1991). The field of geomorphometry (Hengl and Reuter, 2008) has advanced with the technology of GIS and is contributing to the evolving list of terrain attributes. Table 5-1 lists some terrain attributes commonly used in digital soil mapping. An exhaustive list is available in Wilson and Gallant (2000). All these terrain attributes can be calculated using commonly available GIS and statistical software packages (e.g., ArcGIS, SAGA, R).

A critical variable to consider when calculating terrain derivatives is the neighborhood size used. The typical raster GIS operation uses a roving window of 3 x 3 cells when calculating first and second derivatives, such as slope gradient and slope curvatures, respectively. This small window can be problematic if the source DEM has a high resolution (e.g.,

Table 5-1

Selected Primary and Compound Terrain Attributes Used in Digital Soil Mapping

Attribute	Measures	Biophysical property
Primary		
Curvature	Second derivative of slope	Flow characterization, i.e., runoff or run-on
Relief, a.k.a. Topographic Ruggedness (Riley et al., 1999)	ABS(Zmax − Zmin) for specified neighborhood	Broad characterization of terrain (infers parent material)
Normalized Slope Height, a.k.a. Relative Elevation or Relative Position	(Z − Zmin)/ (Zmax − Zmin) where Z = elevation of center cell for specified neighborhood	Relative landform position, catenary sequence, vegetation distribution
Compound		
Solar Radiation (Hofierka and Suri, 2002)	Estimates potential or actual incoming solar radiation for specified time interval	Solar energy incidence on surface, a means of modeling aspect
Wetness Index, i.e., Topographic Wetness Index (Moore et al., 1991)	W = (A/S) where A = upslope contributing area for a cell and S = the tangent of slope gradient	Spatial distribution of zones of saturation for runoff (assumes uniform soil transmissivity within the catchment)
Potential Drainage Density (Dobos and Daroussin, 2005)	Cell count of stream segments within specified neighborhood	A measure of landscape dissection
Morphometric Protection Index (Olaya and Conrad, 2009)	A measure of topographic openness	Plant communities, soil development, impact of wind
Multi-Resolution Valley Bottom Flatness Index and Ridge Top Flatness Index (Gallant and Dowling, 2003)	Process to differentiate valley floor and ridgetop positions	Landscape position
Geomorphon (Jasiewicz and Stepinski, 2013)	Landform classification based on line-of-sight	Crisp landform classes, catenary sequence

< 10 meters) or contains substantial noise. For example, calculating a slope gradient from a 3-m resolution NED DEM for an area in the Midwestern United States using the typical 3 x 3 neighborhood yields a noisy surface, whereas a larger neighborhood yields a smoother surface that better represents the slope patterns that govern soil distribution.

The expanding neighborhood size over which the DEM derivatives are calculated allows flexibility in depicting local or regional features. The larger the neighborhood, the greater the emphasis on broad trends and large features. The most suitable neighborhood size for the modeling target(s) under investigation should be determined. Neighborhood sizes should vary according to the terrain attribute being calculated. For example, an attribute like topographic ruggedness is commonly calculated using a larger neighborhood to characterize a regional trend (e.g., geomorphic/physiographic region) but slope gradient is typically modeled as a localized attribute (e.g., hillslope).

Terrain attributes based on hydrology must be calculated using extents that include intact, complete watersheds. Terrain attributes such as upslope contributing area (flow accumulation), wetness index, stream power index, and downslope distance gradient will have consistent, uniform output values when calculated for complete watersheds, and the output values will have the same meaning when compared across different watersheds.

Another factor related to hydrologically based attributes is the manner in which flow direction is determined. One of the first algorithms developed limited flow to one of the eight directions in a 3 x 3 neighborhood. It is known as the deterministic 8 (D8) algorithm (O'Callaghan and Mark, 1984). The D8 algorithm works well if flow paths are confined to areas of concentrated flow and there is only one cell of lower elevation to route flow toward. Problems occur if the flow is diffuse. More recent algorithms, such as the multiple flow direction method (MFD) (Quinn et al., 1991) or the deterministic infinity (Dinf) method (Tarboton, 1997), allocate flow to multiple directions and so render a flow path that better represents the diffuse nature of water flow.

Several terrain attributes listed in table 5-1 or in Wilson and Gallant (2000) are appropriate for stratifying study areas or defining broad, regional areas. They include Topographic Ruggedness (Riley et al., 1999), Roughness by Relief and Aspect (Behrens, 2003), Hammond's Landforms (1954, 1964), Iwahashi and Pike's Topographic Classification (2007), Fuzzy Landform Elements (Schmidt and Hewitt, 2004), and Geomorphons (Jasiewicz and Stepinski, 2013). Since most of these attributes are based on a large neighborhood, they can be used to describe regional characteristics. Creating combinations of these attributes may also be useful. For example,

a crisp class (i.e., Geomorphon landform elements) in combination with relative elevation would be useful for investigations of the relationship between upper, mid, and lower backslopes (Libohova et al., 2016).

Spectral Data Derivatives

Spectral data commonly is transformed (not just used as raw spectral bands) in order to emphasize useful spectral signatures. A spectral data derivative is simply the conversion of spectral data, either digital numbers or surface reflectance, into a new composite spectral variable. Typically, these transformations involve some combination of the spectral values in two or more bands. The original bands represent a measure of radiance for a specific spectral band, whereas the derivative transforms the data and typically represents some information useful for subsequent analysis.

Spectral derivatives are useful for several purposes, including: (1) indices of biophysical properties, commonly related to environmental covariates (SCORPAN); (2) data reduction, by concentrating information into a small number of new bands; and (3) suppression of topographically related illumination variation (considered noise, not information). Of these spectral transformations, the conversion of spectral data into indices of biophysical properties is probably the most important for digital soil mapping. The most effective and widely used biophysical indices relate to vegetation abundance, in part because vegetation has such a distinctive spectral reflectance pattern. However, any physical property, including soil mineralogy and moisture, can potentially be the focus of a transformation if the property has a measurable effect on the spectral reflectance that can differentiate it from other surface materials in an image. Three of the most widely used spectral transformations are band ratios, principal components analysis, and the Tasseled Cap (Kauth-Thomas) transformation.

Band ratios.—Ratios of spectral bands can be used to accentuate the differences between reflectance and absorption features (Jensen, 2005). The two kinds of ratios commonly used are simple and normalized. *Simple ratios* simply divide the digital number (DN) or surface reflectance value (%) of one sensor band by another (e.g., band 1/band 2). *Normalized ratios* divide the difference between two bands by the sum of the two bands. Because ratios are not scene-dependent, ratios from different images potentially can be compared. Table 5-2 lists commonly used band ratios. The information in the ratio image must be validated with *a priori* knowledge of the area or other measured data. Specialized ratios can be developed based on a surface feature that reflects highly in one band and absorbs greatly in another, such as gypsum (Nield et al., 2007). Ratios must be calculated on atmospherically corrected or standardized images (images converted to surface reflectance).

Table 5-2

Spectral Band Ratios Used in Digital Soil Mapping*

Ratio name	Equation	Sensor/bands	Biophysical property
NDVI[1] (Normalized Difference Vegetation Index)	$\frac{NIR - Red}{NIR + Red}$ Values range from -1 to 1	Red and Near Infrared bands; Landsat 5, 7—bands 3, 4; Landsat 8—bands 4, 5	Healthy green vegetation
Soil Enhancement Ratios[2]	1) Red/Green: carbonates 2) Red/SWIR(a): iron 3) SWIR(a)/SWIR(b): hydroxyls (clay)	See band combinations for carbonate, iron, hydroxyls (clay) ratios below	Three simple ratios for carbonate, iron, and hydroxyls (clay) are combined into one three-layer image
Carbonate Normalized Ratio[3]	$\frac{Red - Green}{Red + Green}$ Values range from -1 to 1	Red and Green bands; Landsat 5, 7—bands 3, 2; Landsat 8—bands 4, 3	Calcium carbonate-bearing minerals
Iron Normalized Ratio[4]	$\frac{Red - SWIR(a)}{Red + SWIR(b)}$ Values range from -1 to 1	Red and SWIR bands; Landsat 5, 7—bands 3, 7; Landsat 8—bands 4, 7	Iron-bearing minerals
Clay (hydroxyls) Normalized Ratio[5]	$\frac{SWIR(a) - SWIR(b)}{SWIR(a) + SWIR(b)}$ Values range from -1 to 1	SWIR bands; Landsat 5, 7—bands 5, 7; Landsat 8—bands 6, 7	Clay or hydroxyl-bearing minerals
Rock Outcrop Normalized Ratio[6]	$\frac{SWIR(a) - Green}{SWIR(b) + Green}$ Values range from -1 to 1	SWIR and Green; Landsat 5, 7—bands 5, 2; Landsat 8—bands 6, 3	Sedimentary (bright pixels) vs. igneous (dark pixels) parent material
Ferrous Normalized Ratio	$\frac{SWIR(a) - NIR}{SWIR(a) + NIR}$ Values range from -1 to 1	SWIR bands; Landsat 5, 7—bands 5, 4; Landsat 8—bands 6, 5	Ferrous iron-bearing minerals

* For documentation and ERDAS Imagine models available for most ratios listed, see USDA-NRCS (2016b).

[1] Jensen, 2005

[2] Developed by U.S. Bureau of Land Management

[3] The carbonate band from the Soil Enhancement Ratio (see above) as a normalized index

[4] The iron band from the Soil Enhancement Ratio (see above) as a normalized index

[5] The clay band from the Soil Enhancement Ratio (see above) as a normalized index

[6] Bodily, 2005; Stum et al., 2010

Principal components analysis.—In applications for remote sensing, principal components analysis (PCA) is an image-dependent data transformation and varies depending on the spectral properties of pixels in the image. Because each PCA transformation is unique, the results of a PCA transformation from one image cannot be compared directly to that from another image. This condition is both a strength and a weakness: a strength because the transformation will adapt to highlight the information present in the particular image, and a weakness because interpreting the results of a PCA transformation can be difficult (i.e., each scene's PCA is different and needs to be interpreted based on its specific transformation).

A PCA transformation is the rotation and translation of the *n* bands of original image data to produce *n* bands of new data, which are orthogonal or mutually perpendicular in spectral feature (n-dimensional) space and uncorrelated. The consequence of this method of arranging the new bands is that most of the variance will be concentrated in a subset of the PC bands (Jensen, 2005). PCA reduces variance of the data in the new PC bands, a reduction which may be desirable. The resulting PC bands should be examined closely to determine which new PC bands contain the most information and could potentially be most useful in subsequent analysis and modeling. The most useful are typically PC 1, 2, 3, but they should be evaluated for each individual image). PCA transformation is available in many software packages and does not require an atmospherically corrected (surface reflectance) image.

Tasseled Cap (Kauth-Thomas) transformation.—The Tasseled Cap transformation is similar to PCA in that it is an orthogonal, multiband transformation. Unlike PCA, the rotations are directed to capture specific biophysical properties and are not scene specific. The original Tasseled Cap transformation was developed for Landsat MSS data and then extended for Landsat TM data. It was based on an analysis of agricultural data from the U.S. Midwest but since has been used globally and for non-agricultural areas (including forestry and urban applications).

The Tasseled Cap transformation is based on the observation that most of the variability in Landsat TM data can be explained by three properties: (1) *brightness*, which is similar to the average DN value across all bands; (2) *greenness*, which is a measure of vegetation abundance, similar to a vegetation index, but which incorporates all the bands and not just red and NIR; and (3) *wetness*, which tends to be correlated with the amount of water present. It is available in image processing software packages, such as ERDAS Imagine, and requires an atmospherically corrected (surface reflectance) image (Jensen, 2005).

Selection of Appropriate Predictors

After data has been explored and appropriate terrain and/or spectral derivatives established, but before the process of model building is started, an optimal set of predictor variables (i.e., covariates) needs to be selected. Digital soil mapping requires spatially exhaustive environmental covariates (SCORPAN) related to the soil class or property of interest. Generating 10s to 100s of covariates is inexpensive and relatively easy (Brungard et al., 2015; Miller et al., 2015; Xiong et al., 2014), particularly when multi-resolution digital elevation models are used (Behrens et al., 2010; Roecker and Thompson, 2010; Smith et al., 2006). Although it is possible to use all available covariates as predictor variables in modeling, it is best to select an optimal subset. Inclusion of non-informative covariates increases model uncertainty, particularly for linear models. Covariate reduction (also known as feature selection) is also important because as the number of covariates increases so does the chance of model overfitting and the amount of computation time. Moreover, simpler models are easier to interpret.

Pedological knowledge should be integrated in the covariate selection process (as described earlier in this chapter) because digital soil mapping is most accurate when fundamentally driven by an expert with significant knowledge of the soil system (Kuhn and Johnson, 2013). If pedological knowledge is lacking or uncertain (particularly regarding scale) and/or if multiple data layers represent the same SCORPAN covariate, these methods should be used. In some cases, semi-automated covariate selection methods can identify a subset of covariates from the larger set of all available covariates so that prediction accuracy is optimized with the fewest number of covariates (Nilsson et al., 2007; Xiong et al., 2014). Pedological knowledge and semi-automated covariate selection methods should be used together (Kempen et al., 2009; Kuhn and Johnson, 2013).

Semi-automated covariate selection methods can be grouped into two broad categories: unsupervised and supervised (Kuhn and Johnson, 2013). Unsupervised methods evaluate covariate relevance outside of a predictive model by selecting covariates that pass some criterion (Kuhn and Johnson, 2013). Supervised methods select optimal covariates by identifying the covariate set that maximizes model predictive ability (Kuhn and Johnson, 2013).

Unsupervised covariate selection methods include correlation analysis, Optimal Index Factor (OIF), and principal components analysis (PCA). Correlation analysis retains or removes covariates that meet a pre-determined correlation threshold. OIF ranks any covariate combinations of three bands so that the within-covariate variance is maximized and the between-covariate correlation is minimized (Kienast-Brown and

Boettinger, 2010; Nield et al., 2007). Combinations with the highest OIF are assumed to contain the most information. PCA transforms covariates so that they fall along the multivariate axes of greatest variance (Fox and Metla, 2005; Levi and Rasmussen, 2014). It eliminates between-covariate correlations, but because it transforms covariates, the results can be difficult to interpret. Unsupervised methods are likely to be most useful when covariates are highly correlated.

Supervised covariate selection methods include forward and backward selection, simulated annealing, genetic algorithms, and the Boruta algorithm. Forward and backward selection iteratively adds (forward selection) covariates or removes (backward selection) covariates to determine which covariates are not significant. Forward and backward selection is particularly useful for linear regression when combined with Akaike's Information Criterion (AIC). Recursive feature elimination is a variant of backward selection that avoids fitting multiple models at each step (Guyon et al., 2002; Kuhn and Johnson, 2013). Simulated annealing modifies an initial random subset of covariates based on a slowly decreasing probability, so that over a number of iterations it becomes very unlikely that a suboptimal covariate set will be selected (Kuhn and Johnson, 2013). Genetic algorithms randomly change multiple covariate sets until a covariate set that produces the most accurate model is identified. The Boruta algorithm scores each covariate against a set of random covariates. Covariates that have importance scores significantly larger than the random covariates are deemed relevant (Kursa and Rudnicki, 2010). Additionally, several tree- and rule-based statistical models (i.e., random forests, cubist models, multivariate adaptive regression splines, and lasso models) conduct intrinsic covariate selection. Because each supervised method has a different approach to covariate selection, different methods identify different optimal covariate sets. Generally, it is useful to compare multiple supervised covariate selection approaches. Implementations of these methods can be found in the caret (Kuhn et al., 2015) and Boruta (Kursa and Rudnicki, 2010) packages for the R software for statistical computing (R Core Team, 2013).

Unsupervised and supervised covariate reduction methods can be used together. For example, in a digital soil mapping study of soil depth in southeastern Utah, correlation analysis was initially used to identify and remove highly correlated covariates from a set of 94 potential covariates. Next, both the Boruta algorithm and simulated annealing were used to identify a final set of 7 covariates. The final covariate set provided equal or better predictive accuracy than larger covariate sets (Brungard, unpublished data).

Qualitative visual inspection of spatial predictions should also be used to assess selected covariates. Covariates which are pedologically and statistically plausible but produce visually incorrect predictions, such as sharp linear boundaries where none exist, should be removed (Padarian et al., 2014).

In summary, optimal covariate selection begins with using existing pedologic knowledge to identify data layers that represent relevant SCORPAN covariates. The result may be a relatively large number of covariates since it is likely that multiple data layers, at multiple scales, can represent each SCORPAN covariate. Supervised and unsupervised techniques can be used to further refine these covariates. The optimal predictor set should be the covariate set that is pedologically and statistically plausible, results in the most accurate model, and produces visually correct predictions. A guide to covariate selection is presented in figure 5-2.

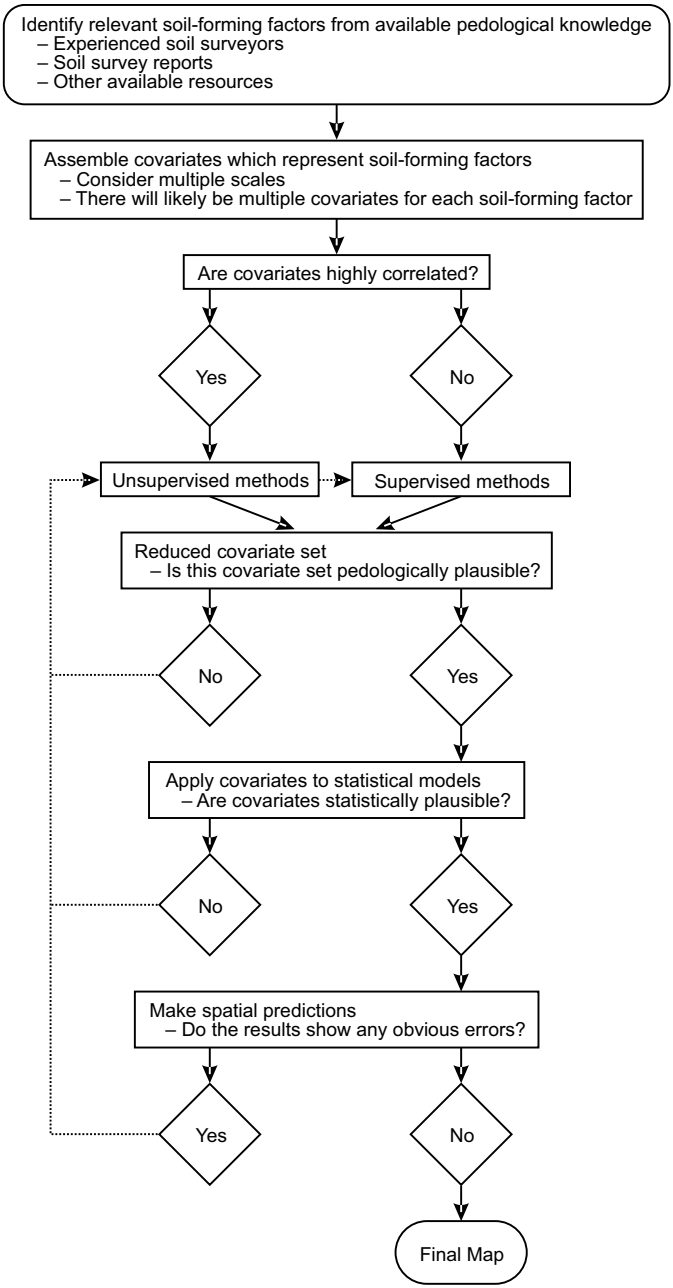
Sampling for Training Data

The digital soil mapping process is dependent on the relationship between predictor variables (i.e., covariates) and the target soil feature (soil class or property) of the model. This relationship applies to both knowledge-driven and data-driven modeling methods. It is important to select samples of covariates that are representative of the distribution of the target soil feature. These samples, known as training data, provide the data that will be used to train the model to predict similar occurrences. Prediction of soil classes or properties is most successful when precise, observed locations of typical soil members are available or when experts can provide precise tacit points. Directed (purposive) field investigations may be used in support of a knowledge-based modeling approach but should not be used exclusively. Random or stratified sampling is more robust and less prone to bias. Training data can be collected with case-based or *a priori* sampling if existing data or knowledge is utilized or by *in situ* sampling if new data are collected specifically for the purpose of training a model.

Case-Based and A Priori *Knowledge Sampling*

Case-based sampling for training data uses prior mapped locations of classes or properties to train a model to map the same classes or properties in unmapped locations. The empirical relationship between the outcome (class or property) and the covariates at known locations (previously mapped) can be used to predict an outcome in unknown areas with similar biophysical characteristics. The known and unknown areas must

Figure 5-2



Flow chart illustrating the general steps in selecting environmental covariates.

have similar soil-landscape relationships. Knowledge of soil-landscape relationships, along with model performance measures (discussed under "Validation and Uncertainty"), should be used to determine how reliable and applicable the empirical relationships will be in unmapped areas.

A priori sampling for training data uses previous knowledge of an area to sample a training data location from the covariate data. It is best applied to classes that are very distinct and whose location is easily determined using high-resolution imagery, such as a rock outcrop or water class. It should not be used for classes that contain more variability, like a soil class, to avoid introducing bias into the sampling process. It is best to use case-based or field sampling for more variable and complex classes.

Field Sampling

Collecting training data in the field is an essential part of the digital soil mapping process. Data must be collected in the field using the selected set of covariates and a sampling design amenable to the modeling objectives. Sample point selection typically is determined using GIS software. Generally, a GPS receiver is used to navigate to sample locations in the field.

The positional accuracy of GPS receivers varies dynamically according to satellite configuration, atmospheric and solar conditions, terrain, and type of GPS receiver in use. If possible, comparable GPS receivers should be used for all data collection activities for a given project. All GPS receivers provide a dynamic display of positional accuracy. A minimally acceptable standard of positional accuracy should be determined for the data collection activities.

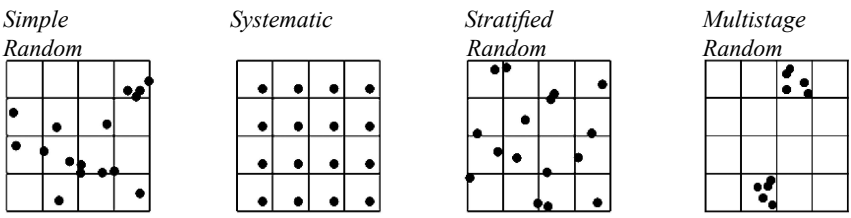
It is important for field personnel to know what the sample is intended to represent. Field computers that display spatial data against the GPS position and sample location are ideal for ensuring that the field location is close to the sample location. In remote areas where computers cannot be used, corroborating information should be supplied to help better reference the site location for field staff. For example, if the sample is located near the juncture of a side slope and footslope, but clearly on the side slope, this information should be given to the field crew. The information should be on a hard-copy field collection sheet or database form. Data collection forms, either digital or hard copy, should be standardized throughout a given project and include all variables needed to satisfy the target modeling objective(s). Including a data field item for GPS accuracy may be helpful and provide a reference throughout the course of a project.

Sampling Design

The choice of sampling design depends on the size and accessibility of the project area, modeling objectives, desired level of confidence and precision, expected variability of the soil feature(s), and the cost of obtaining samples. The selected design needs to satisfy the statistical rigor of randomness as well as remain within the limits of time, money, and staff available for sampling.

Simple random.—Simple random sampling is the most straightforward way to select independent and unbiased samples. Sample locations each have an equally probable chance of being selected (fig. 5-3). This design has the primary advantage of being unbiased and satisfying the statistical requirements of randomness. It gives every location the same probability (i.e., chance) of being selected for sampling. However, this design may result in irregular and/or clustered spacing of samples. In addition, detecting systematic variation may be difficult using this sampling method. This design is most useful for study areas that are small and homogeneous and have few explanatory variables.

Figure 5-3



Simplistic representation of sampling locations as determined by simple random, systematic, stratified random, and multistage random sampling designs.

Systematic.—A sample is taken according to a regularized pattern (fig. 5-3). This approach ensures even spatial coverage. Patterns may be rectilinear, triangular, or hexagonal. This design can be problematic with data that vary cyclically or vary at an interval smaller than the sample spacing. It is important to ensure that selected samples do not coincide with a particular cycle (e.g., the microhighs of hummocks) but fall on the complete spectrum of the population.

Stratified random.—The sampling region is spatially subset into different strata, and random sampling is applied to each strata (fig. 5-3). Strata are typically geographic, such as land cover type, landform, slope gradient, slope aspect, or parent material. It is assumed that these strata

are strongly related to the target soil feature(s). Strata may be sampled equally or in proportion to area. However, if the target is rare in the population, it may be preferable to sample the strata equally (Franklin and Miller, 2009; Kuhn and Johnson, 2013). Stratified random sampling offers higher accuracy at lower cost. These benefits are dependent on the suitability of the defined strata, which is dependent on adequate prior knowledge of the target soil feature(s).

Cluster.—A cluster or group of points is selected at one or more sites, and only a portion of the available strata or primary sampling units (such as geographic strata, fields, or other separations) are sampled. If strata are an important determinant of the target soil feature(s) being evaluated, it is better to use a stratified random sample and sample all strata. Ideally, each cluster in a cluster sampling design represents the full variability of the area in question and the within-cluster variability is greater than the between-cluster variability (Lohr, 2009). When the costs of getting to a primary sampling unit are high (e.g., when sampling areas are far from a road) and the cost of individual sampling units is low, cluster sampling is highly efficient. However, it can introduce bias if clusters are not representative of the population as a whole (e.g., if a cluster is on an odd highly disturbed area) and a loss of precision if the between-cluster variability is high.

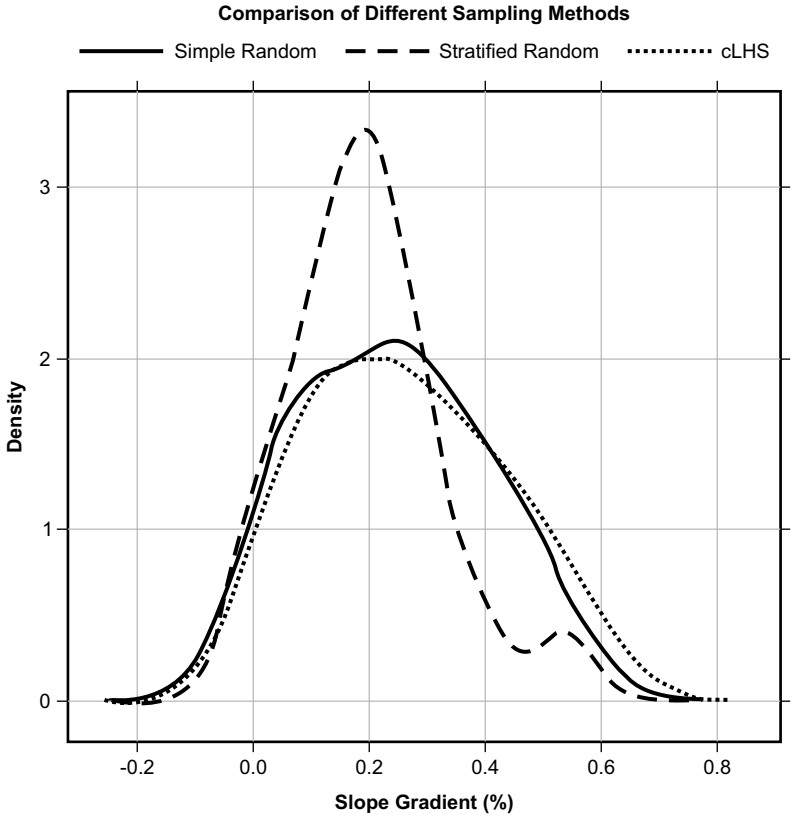
Multistage random.—Multistage random sampling is a complex form of stratification and cluster sampling. In this sampling design, only a subset of individual sampling units (such as pedons) within each cluster are selected for sampling. The individual sampling units can be arranged in order to maximize the variability, or arranged randomly, within the primary sampling unit. For example, as shown in figure 5-3, a two-stage random sampling design may stratify an area into a standard grid and randomly select a subset of strata units (first stage), then randomly select individual sample locations from within each strata unit (second stage) (Schaeffer et al., 1990; de Gruijter et al., 2006). This design offers the advantage of efficiency at reduced costs. The drawbacks include the potential for lower accuracy and precision. Successful multistage random sampling depends greatly on proper selection of strata.

Conditioned Latin hypercube.—Conditioned Latin hypercube sampling (cLHS) is a special type of stratified random sampling that uses the principle of Latin hypercube sampling conditioned with ancillary data (covariates). This sampling method selects sample locations that maximize the variability represented by multiple covariates and works on both continuous and categorical data (Minasny and McBratney, 2006). It differs from other sampling strategies, which focus on sampling geographic space, by focusing on sampling covariate feature

(n-dimensional) space. This type of sampling design is efficient because it can represent the multivariate distribution of input covariates with relatively small sample sizes (Brungard and Boettinger, 2010).

This robust sampling method has been favored in digital soil mapping because it provides a representative sample based on the distribution of covariate data. Without a technique such as cLHS, obtaining a sample that is representative of the feature (n-dimensional) space becomes increasingly difficult as the number of covariates increases. Figure 5-4 compares the distribution of different sampling methods over the data range of a covariate layer.

Figure 5-4



A comparison of the distribution of simple random, stratified random, and cLHS sampling methods over the data range of a slope gradient covariate.

Conditioned Latin hypercube sampling is appropriate for any digital soil mapping project for which multiple independent covariates related to the target soil feature(s) are known or can be inferred. If soil-covariate relationships are unknown or highly uncertain, another sampling design should be used. For areas with access constraints, constrained cLHS (Roudier et al., 2012) or cLHS with fuzzy k-means clustering (Kidd et al., 2015) can be used.

The information needed to run cLHS includes: (1) covariates covering the entire project area, (2) the number of desired samples, and (3) the number of iterations needed to reach a satisfactory sampling scheme. Conditioned Latin hypercube sampling can be performed in Matlab software (MathWorks, Inc.); the R software for statistical computing (Roudier, 2011); and the USFS (U.S. Forest Service) TEUI (Terrestrial Ecological Unit Inventory) Geospatial Toolkit (Vaughan and Megown, 2015).

Predicting Soil Classes and Properties

After the optimal set of SCORPAN covariates (predictor variables) has been selected and training data have been collected, a method may be applied to the data to predict soil classes or properties. Many prediction methods are available and applicable in digital soil mapping. Considerations in choosing a prediction method include:

- Are discrete soil classes or continuous properties the goal?
- Are the training data adequate to support the desired prediction method and/or number of desired classes?
- Are the data parametric (normally distributed) or nonparametric?
- At what step in the soil survey process is the prediction being applied: pre-mapping, initial mapping, update mapping, or secondary product?
- What are the time restrictions for completing the prediction?

Classification is the process of predicting discrete classes. It can be described as the process of sorting pixels into a finite number of classes, based on their data values and distribution in feature (n-dimensional) space. Simply stated, if a pixel satisfies the criteria defining a class, the pixel is assigned to that class. This process is executed according to a classification algorithm. Depending on the type of information one wants to extract from the predictor data, classes may simply represent clusters that look statistically different to the computer (exploratory) or that are associated with known features on the ground (definitive) (refer to ERDAS Field Guide, Intergraph Corp., 2013).

Regression and interpolation methods predict continuous values rather than discrete classes. Interpolation methods model spatial patterns based on values at known locations and the assumption that locations that are closer to one another are more similar than those that are farther apart. Geostatistical approaches are forms of interpolation that rely on statistical functions rather than mathematical functions. Regression approaches use some statistical function to model the relationship between soil observations and a set of predictor variables.

Unsupervised Classification

Unsupervised classification is the prediction method most reliant on computer automation presented in this chapter, and it is the only method that does not require soil observations (i.e., training data) covering the area. The algorithm uncovers statistical patterns inherent in the data and groups pixels with similar characteristics into unique clusters (classes) based on statistically determined criteria (Duda et al., 2001). The resulting class definitions are only dependent upon the predictor data representing the SCORPAN covariates and a few parameters defined at the time the classification is executed. The resulting classes must be interpreted to determine if they are meaningful in terms of soil-landscape relationships. Classes can be merged, disregarded, or manipulated based on evaluation of the class signature or definition in feature (n-dimensional) space.

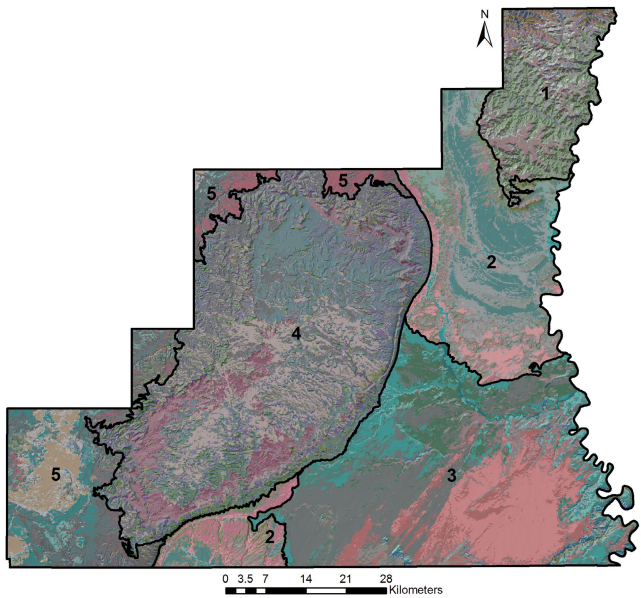
Iterative Self-Organizing Data Analysis Technique (ISODATA) (Tou and Gonzalez, 1974) and k-means (MacQueen, 1967) are the most commonly used unsupervised classification algorithms and are available in many software packages. ISODATA is a modification of the k-means algorithm (Schowengerdt, 1997). Both algorithms are parametric (assuming a normally distributed dataset). They employ an iterative process that creates clusters and classifies pixels until the change in class assignment at each pixel location is small, at which point final classes are defined. The main difference between the two algorithms is that k-means requires the number of classes to be set *a priori* while ISODATA allows a range for the number of final classes to be set. ISODATA can split, merge, and delete clusters during the classification process but k-means cannot. For this reason, ISODATA is considered more computationally robust and flexible than k-means and is commonly preferred.

Unsupervised classification provides a non-subjective, data-driven method for exploring the inherent clustering of data and determining how many classes the data (predictor variables) can support. Because no prior knowledge of the area is required, unsupervised classification is a useful exploratory tool that can help direct field sampling and develop map unit concepts. However, because there is very little control over how

the clusters are defined, the results may be difficult to interpret. Using an appropriate selection of predictor data based on SCORPAN covariates helps to produce the most useful results of an unsupervised classification.

Unsupervised classification is most applicable in the exploratory or pre-mapping stage of soil survey (fig. 5-5). It can help target initial field sampling and be useful in comparing mapped and unmapped areas. Unsupervised classification can be beneficial in the initial phase of digital soil mapping in determining the number of classes the predictor data can support or in determining potential classes in areas with inadequate training data. These determinations prevent using more target classes than the data can separate or support.

Figure 5-5



ISODATA unsupervised classification of both terrain and spectral data derivatives in eastern Emery County, Utah, showing natural clustering in the data and how potential classes may be distributed across the landscape. The area was divided into five subsets based on geology to minimize variability for the classification, which was run on each subset independently (10-m grid resolution). Different colors represent different classes within each subset of the survey area.

Supervised Classification

Supervised classification differs from unsupervised classification in that it requires soil observations covering the area and the target classes.

Soil observations, or training data, must be carefully chosen in order to adequately represent the target classes and produce a meaningful classification. Class definitions from training data are combined with carefully selected predictor data representing SCORPAN covariates, and the applied algorithm determines the class in which each pixel belongs.

There are multiple algorithms for supervised classification that are frequently applied in digital soil mapping. This section discusses minimum distance to means, maximum likelihood (discriminant analysis), fuzzy classification, knowledge-based classification, and predictive modeling (machine learning or statistical modeling).

Minimum distance to means.—Using this classification algorithm, candidate pixels can be classed according to the closest training class mean. This method, by definition, does not include information on the class variability. Therefore, if there are large differences in the variance of each class, the method will likely be unreliable. This method is computationally very rapid.

Maximum likelihood (discriminant analysis).—This classification is one of the most widely used standard supervised classification methods and is based on probability. Maximum likelihood uses the training class means and covariance matrices to classify candidate pixels. The probability of a candidate pixel belonging to each of the classes is calculated, and the class for which the probability is highest is assigned to the pixel. In addition, maximum likelihood allows the prior probability for the class (if known) to be specified across the dataset.

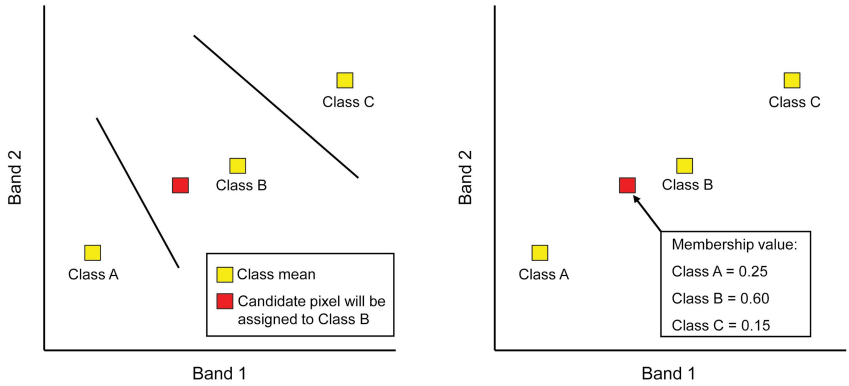
Minimum distance to means and maximum likelihood are both parametric classifiers and assume a normally distributed dataset. Therefore, training data sites and class definitions must be homogenous. These approaches to supervised classification can be useful in areas that have large extents of homogenous soils whose properties do not vary over short distances. This kind of soil landscape allows very clean class definitions and a successful classification, if training and predictor data are properly selected.

Fuzzy classification.—Homogenous soil landscapes are more simplistic for digital soil mapping. However, natural environments are more likely to contain subtle variation over short distances and non-distinct boundaries between soil types. Commonly, a candidate pixel may be mixed and have properties that overlap multiple classes.

Fuzzy set theory provides tools for working with imprecise data (Zadeh, 1965; Wang, 1990). Fuzzy classification allows information from multiple classes to contribute to the classification of a candidate pixel through the use of fuzzy logic and membership functions. In figure

5-6, for example, a candidate pixel may have a membership value of 0.25 for Class A, 0.60 for Class B, and 0.15 for Class C. The pixel is most like Class B, but information about Classes A and C is still obtained. The major difference between fuzzy classification and traditional hard classification (like minimum distance to means and maximum likelihood) is the ability to obtain information about constituent classes occurring in a mixed pixel (Foody, 2000). Due to this characteristic, fuzzy classification can accommodate nonparametric datasets.

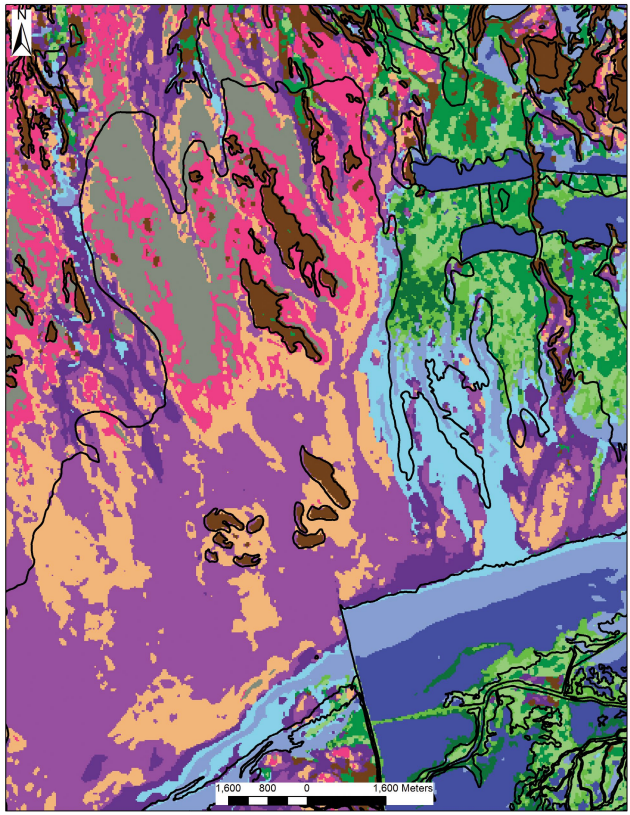
Figure 5-6



Simplistic representation of hard classification (left) and fuzzy classification (right). Hard classification requires a candidate pixel to be assigned to only one class, whichever class mean is closest. Fuzzy classification uses class means but allows candidate pixels to express properties of several classes instead of just one. (Image based on Jensen, 2005.)

Fuzzy classification has the same starting point as the other supervised classification methods, i.e., training and predictor data. However, because of its ability to handle mixed pixels, training data for fuzzy classification can represent both homogenous and heterogeneous classes (Jensen, 2005). Fuzzy classification is most useful in heterogeneous areas where variations in soil type result in mixed pixels or classes (common for soil landscapes). In the fuzzy classification process, it is possible to assign a single class to a pixel, also described as "hardening" (Zhu et al., 2001). However, information regarding constituent classes is still retained and can be used to understand the relationships in the data, refine class definitions or sort out confusion in the classification, and understand soil-landscape relationships (fig. 5-7).

Figure 5-7



Supervised fuzzy classification of Landsat imagery for an area along the east shore of the Great Salt Lake, Utah, showing a "hardened" version of the fuzzy classification (i.e., one class assigned per pixel). Results from the fuzzy classification were used to disaggregate broad map unit concepts in areas with wet and saline soils in an update soil survey project. Original Soil Survey Geographic Database (SSURGO) line work is shown in black; land cover classes representing clusters defined by soil-vegetation-moisture relationships are shown in color.

Knowledge-based classification.—Knowledge-based classification uses expert systems to represent an expert's knowledge as rules and data within a computer (Jensen, 2005). It is not only applicable to predicting soil classes but also very useful in documenting a soil scientist's knowledge about soil-landscape relationships (Zhu et al., 2001). A knowledge-based expert system consists of the following:

- Source (expert, training data, predictor data)
- Knowledge base (rule-based domain)

- Inference engine
- User

The knowledge base or rule set is constructed using the predictor data and the expert's knowledge about soil-landscape relationships and how they are expressed through the data (fig. 5-8). Specific knowledge that defines soil-landscape relationships, and subsequent soil classes, is required. An example is "badland soil complexes occur on steep eroded slopes." This knowledge can be converted into specific rules, such as "badland soil complexes occur on slope % >= 8 and have Fe band ratio value >= 67," and integrated into a knowledge base to predict the desired class (e.g., badland soil complex). In this example, the predictor data (a DEM-derived slope layer and the Fe band ratio layer derived from spectral data) are applied to the expert's knowledge (the rule) about the badland soil-landscape relationship.

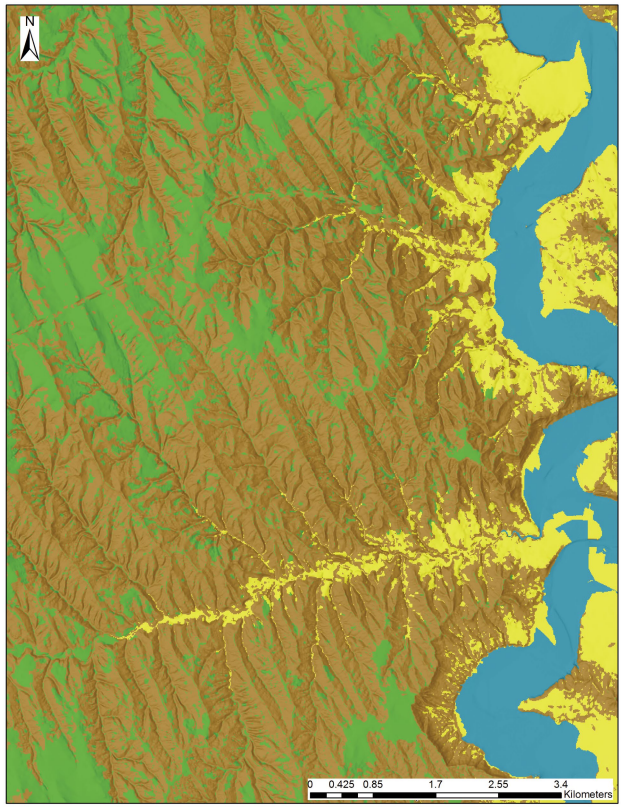
Knowledge-based classification requires the most *a priori* knowledge about soil-landscape relationships of all the classification methods presented in this chapter. It can be successful in areas where a lot of fieldwork and documentation have been completed and soil-landscape relationships are well documented and understood. Also needed are adequate predictor data to support and discriminate the specific rules defined in the knowledge base.

Knowledge-based classification is a very time-intensive approach. It requires field observations to understand the soil-landscape relationships well enough to develop specific rules for each class as well as to refine the rules in an iterative manner (as more knowledge is acquired or needed). If the resources are available, knowledge-based classification can be worth the investment, especially in terms of its ability to capture the tacit knowledge of a soil scientist.

Several software packages offer knowledge-based classifications. Some provide a hierarchical decision-tree classifier (ERDAS Imagine Knowledge Classifier) while others employ a fuzzy classification approach (SoLIM, ArcSIE). Most expert systems have the flexibility of using both continuous and categorical predictor data.

Supervised classification methods are best applied once preliminary field documentation has been collected and map unit concepts are in development. Supervised classification can be effectively applied in both initial and update soil survey projects. Since *a priori* knowledge and class definitions in the form of class signatures (or rules) are required, the methods of supervised classification discussed above can be more time intensive to initiate than classification options that are more data driven and do not require as much input initially, such as unsupervised classification and predictive modeling.

Figure 5-8



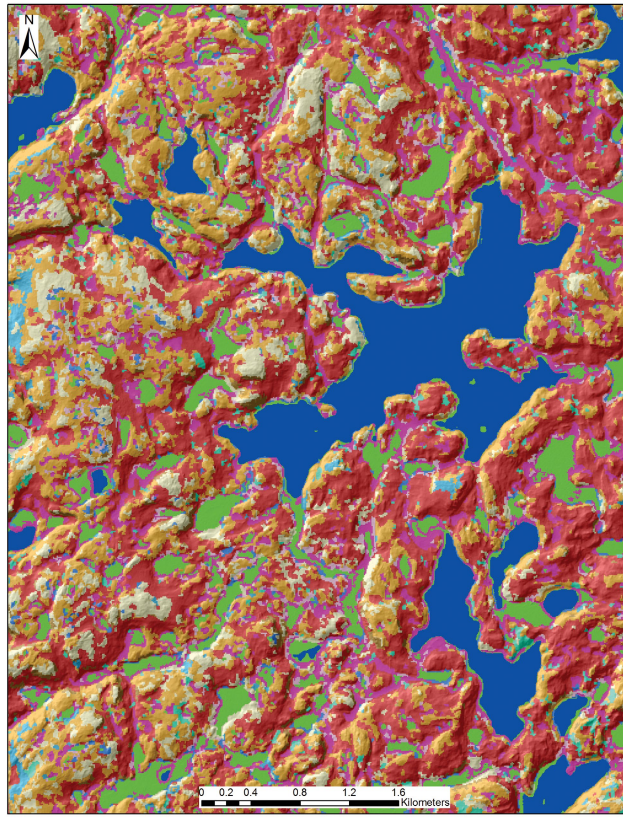
Output from a hierarchical decision-tree knowledge-based classification for four classes—fluvial soils, badland soils, uplands, and alluvial fans (shown in different colors)—in an area near the Powder River Breaks, Wyoming (Cole and Boettinger, 2007). Predictor data included both terrain and spectral data derivatives (10-m grid resolution).

Predictive modeling.—Predictive modeling (commonly referred to as statistical modeling or machine learning) for digital soil mapping is the process of developing a mathematical model that approximates the true relationship between soil properties or classes and environmental covariates in order to produce an accurate prediction. It involves choosing the necessary predictor data representing SCORPAN covariates and an appropriate model or algorithm.

Predictive models can be conceptually divided into two broad groups: classification and regression. Classification methods are used

for predictions of a soil class, and regression methods are used for predictions of a continuous soil property. Within these broad groups, predictive models can be further divided based on the type of model: linear, non-linear, or tree- and rule-based. Examples of linear methods are simple linear regression and discriminant analysis. Examples of non-linear methods are multivariate adaptive regression splines and neural networks. Examples of tree- and rule-based methods are random forests (fig. 5-9) and gradient boosting machines. Kuhn and Johnson (2013) and James et al. (2014) discuss each model algorithm in depth as well as the overall process of predictive modeling.

Figure 5-9



Classification using random forests method for parent material classes in the Boundary Waters Canoe Area Wilderness, Minnesota. Predictor data included both terrain and spectral data derivatives and training data points from field data collection (5-m grid resolution).

Although many potential predictive models are available, a model that can always produce the most accurate predictions for any digital soil mapping project is difficult to find. This is because model predictive ability depends upon the structure of individual datasets and the methods used for covariate selection. The best approach is to apply several predictive models and pick the model that produces the most accurate prediction. One could start with a complex model (e.g., random forests or neural networks), then compare it to simpler models (e.g., linear regression or classification trees). If the accuracy of the simpler model is comparable to the more complex model, the simpler model can be selected. Simple models are favored for their ease of interpretation.

Overfitting can occur when applying predictive modeling for digital soil mapping. The term "overfitting" indicates that the statistical model over-emphasizes random noise instead of the underlying function. Overfit models will not produce accurate predictions. Cross-validation (a model validation method for assessing how the results will generalize to an independent data set) should be used during the model building process to avoid overfitting. Cross-validation is inherent in, or at least an option for, many algorithms.

Predictive modeling should be applied after preliminary fieldwork is complete and there is adequate training data to satisfy the model and produce an accurate prediction. It can be useful for initial or update soil survey and for soil property mapping. Depending on the model, parametric and non-parametric datasets as well as continuous and categorical data can be used in the modeling process. As a result, predictive modeling is one of the more flexible approaches to digital soil mapping prediction.

Predictive modeling provides a non-subjective, quantitative alternative to conventional soil survey and returns an estimate of prediction uncertainty based on cross-validation. However, accurate predictive modeling may require more pedon observations than are available or can be collected given project constraints. Predictive modeling works best if observations are collected using a probabilistic sampling design and if it is driven by an expert with significant knowledge of the soil system (Kuhn and Johnson, 2013).

Geostatistics

The field of geostatistics encompasses a range of techniques for modeling spatial patterns that satisfy the basic assumption that nearby objects are more related to each other than distant objects. Central to this assumption is the concept of regionalized variable theory, or the description of spatial patterns as an additive mixture of trend, spatially correlated variation, and noise. Typically, geostatistical methods are

used to estimate values at unsampled locations (interpolation) based on a limited set of sampled continuous (and to a lesser extent, categorical) properties, such as A horizon pH, depth to root-restricting layer, or presence of a duripan. Geostatistics is closely related to a number of other spatial interpolation methods, such as Voronoi polygons, triangulation, natural neighbors, inverse distance weighting, trend surfaces, and splines. Geostatistical methods, however, are commonly preferred (when sufficient data are available and critical assumptions met) because they provide unbiased estimates of uncertainty.

Once the appropriate data have been collected, the typical steps involved in geostatistical analysis (Webster and Oliver, 2007; Isaaks and Srivastava, 1989) are as follows:

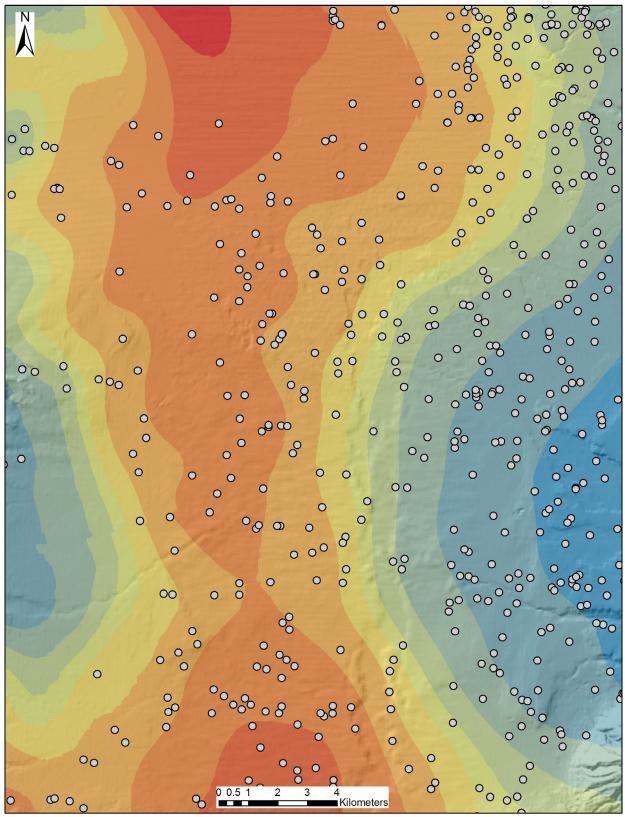
1. Check data for outliers, extreme deviance from a normal distribution, and any spatial trend.
2. In the presence of a strong trend (e.g., elevation gradient), de-trend or use hybrid approaches such as regression-kriging (Hengl et al., 2007).
3. Transform data as needed (log transformation, normal-score transformation, and logit transformation are commonly used).
4. Compute the empirical variogram (a description of how the data are correlated with distance), and check for the influence of any unusual values.
5. Fit a model to the empirical variogram, and verify that the parameters make sense.
6. Use some form of kriging to make predictions for unvisited locations.

The greatest limitation of geostatistics for soil survey is that the reliability of the variogram (and thus subsequent spatial predictions) is dependent upon both sample size and design. Typical soil survey sampling methods are commonly inadequate for reliable variogram estimation. However, geostatistics may be used for new soil products, provided that sampling design is given special attention and sufficiently large numbers of observations are collected (fig. 5-10). At least 150 samples are needed for robust variogram estimation (Webster and Oliver, 2007). The mean sampling interval (i.e., distance between samples) should be at least one order of magnitude less than the variogram range (Olea, 2009). Additionally, the application of geostatistical methods requires special consideration of anisotropy, i.e., existing trends or gradients that exhibit some form of directionality (such as the orographic effect on climate or the complex pattern of a braided stream system). It is possible to incorporate external information on such trends into the kriging process

using methods such as universal kriging, kriging with external drift, or regression-kriging (Odeh et al., 1994, 1995).

Basic geostatistical methods have been implemented in the gstat package (Pebesma, 2004) for the R statistical software (R Core Team, 2013). Other commonly available software packages, such as ArcGIS, include geostatistical analysis functionality.

Figure 5-10



Interpolation using ordinary kriging of soil K concentration in the Salt Lake City Valley, Utah. Points represent locations of soil K measurements collected in the field. Concentration of K ranges from low (blue) to high (orange).

Validation and Uncertainty

Qualitative (conventional soil survey) and quantitative (digital soil mapping) soil survey methods rely on conceptual or mathematical models

to describe soil spatial distribution. These models are approximations of reality and are thus subject to uncertainty. Due to the quantitative nature of digital soil mapping, predictions of soil classes or properties lend themselves to quantitative assessments of accuracy and uncertainty. Communicating the accuracy and uncertainty associated with soil spatial predictions is imperative and should be an integral part of any digital soil mapping project, particularly given that soil information is used in decision making and risk assessment.

Accuracy

All soil maps are approximations of reality, such that the values depicted on a map will deviate to some extent from true values. Accuracy estimates are therefore necessary to quantify prediction quality. Prediction accuracy is the difference between the predicted value at a location and the measured value at the same location (Brus et al., 2011). Desirable predictive models have high prediction accuracy (i.e., small differences between predicted and observed values).

Prediction accuracy is quantified differently depending on whether soil classes or soil properties are being modeled. Soil class prediction accuracy is quantified using overall accuracy, user's accuracy, and producer's accuracy. These metrics are best understood by reviewing a confusion matrix (table 5-3) that compares the number of correctly and incorrectly predicted observations for each class. Overall accuracy is the proportion of correctly classified observations in the entire dataset. User's accuracy (also known as "errors of commission" or precision) is the proportion of a predicted class that matches the observed class. Producer's accuracy ("errors of omission" or specificity) is the proportion of an observed class that matches the predicted class (Congalton, 1991; Kuhn and Johnson, 2013).

Table 5-3 shows a confusion matrix of three modeled soil subgroup classes, modified from data presented in Brungard et al. (2015). Observation numbers were 26 Ustic Haplargids, 2 Ustic Paleargids, and 21 Ustic Torriorthents. Overall accuracy was calculated by summing the correctly predicted observations (matrix diagonal; 11) and dividing by the total number of observations (49). User's accuracy for each class was calculated by dividing the correctly predicted observations for each class by the row totals. Producer's accuracy for each class was calculated by dividing the correctly predicted observations for each class by the column total. Overall accuracy was relatively low because the Ustic Paleargid class was never modeled correctly (an effect of low numbers of training observations). Low overall accuracy masks the relatively high accuracy of the other two classes.

Table 5-3

Confusion Matrix of Three Modeled Soil Subgroup Classes

Predicted soil class	Observed soil class			Total correctly predicted	User's accuracy
	Ustic Hapl-argid	Ustic Pale-argid	Ustic Torri-orthent		
Ustic Haplargid	6	1	1		0.75
Ustic Paleargid	0	0	0		0.00
Ustic Torriorthent	1	0	5		0.83
				11	
Producer's accuracy	0.86	0.00	0.83		Overall accuracy: 0.22

It is important to note that the above accuracy metrics are all threshold-dependent, i.e., they depend upon a cutoff threshold above which observations are classified as belonging to a particular soil class. All predictive models output probability or membership values, which are then classified as belonging to a particular soil class if they are above some threshold (commonly 0.5 by default). However, if this threshold is changed, then validation observations may be included or excluded from a particular class and the confusion matrix and resulting accuracy metrics altered. Though most commonly used for two class predictions, threshold-independent metrics, such as the area-under-the-curve (AUC), provide an estimate of prediction accuracy over all threshold values (Kuhn and Johnson, 2013).

Accuracy of soil property predictions is typically quantified using mean square error (MSE), root mean square error (RMSE), and coefficient of determination (R^2). Mean square error is the average squared difference between predicted and measured values. Because MSE is a squared difference, the square root of MSE (RMSE) commonly is used to report accuracy in the same units as the original measurements (Kuhn and Johnson, 2013). Smaller RMSE indicates a more accurate model. The coefficient of determination (R^2) is a measure of the correlation between observed and predicted values and commonly is interpreted as the proportion of the data explained by a model. Caution is needed when using R^2 because it is a measure of correlation, not accuracy, and is dependent upon the variation in the test set (Kuhn and Johnson, 2013).

Validation observations (also known as reference observations) necessary to calculate prediction accuracy metrics can be derived from independent validation data, internal model performance measures, or data-splitting methods. Independent validation data are observations gathered independently from data used for model building (the training data set). Independent validation is the best way to assess prediction accuracy because it is the only way to determine true prediction accuracy. Independent validation data should be gathered using probabilistic sampling methods to avoid bias. Sampling schemes for validation can be found in Brus et al. (2011) and de Gruijter et al. (2006), and methods for calculating the required number of observations can be found in Congalton (1991).

Although independent validation data are preferable for accuracy assessment, in some cases it is not possible to collect such data (such as with legacy data) and other methods are required. Internal model performance measures (also termed calibration accuracy) are used for model tuning. They indicate how well the model matches the data. Examples of internal model performance measures include the out-of-bag error (OOB) used in the random forests tree-based model and the mean squared error commonly used in many regression models (James et al., 2014). Internal model performance measures are useful for assessing model parameters, but such measures commonly overestimate actual prediction accuracy because statistical models are designed to minimize (or maximize) these internal accuracy measures. Prediction accuracy should not be inferred solely from internal model performance measures.

Related to internal model performance measures are data-splitting methods. Data-splitting methods involve reserving a portion (commonly 10 to 30 percent) of the available training data to use only for validation. Using an observation for both model training and validation is redundant and strictly prohibited. In data splitting, the reserved portion of the data is only used in model validation and not in model training/building. While data-splitting practices are common, there is no guarantee that a different subset of the training data would result in the same accuracy estimates. A better alternative is to use cross-validation, which repeatedly divides the training data into n (commonly 5 or 10) training and validation subsets and thus evaluates many alternate versions of the data (Kuhn and Johnson, 2013). Cross-validation results in prediction accuracy estimates with associated variability (e.g., standard deviations). If the initial field-sampling method was biased, cross-validation accuracy estimates may not adequately capture true prediction accuracy because cross-validation relies strictly on the data used in modeling.

Estimates of prediction accuracy are necessary for quantifying digital soil mapping prediction quality and should be included as a vital component of any digital soil mapping project. Measures for accuracy calculation are available in many software packages and commonly are included in the execution of prediction models.

Uncertainty

Uncertainty in traditional soil survey results from the scale of mapping (e.g., order 1 vs. order 3), the placement of map unit lines, and the inclusion of similar soils. This uncertainty is quantified using map unit composition (e.g., Map unit 1 is 55% soil A, 30% soil B, and 15% soil C).

Uncertainty in digital soil mapping results from several sources: (1) positional accuracy of the pedon location (particularly for legacy pedon observations); (2) covariate accuracy (e.g., vertical uncertainty of a digital elevation model); (3) soil class or property measurement (e.g., taxonomic classification or laboratory analysis); and (4) model structure (e.g., using a linear model for curvilinear data).

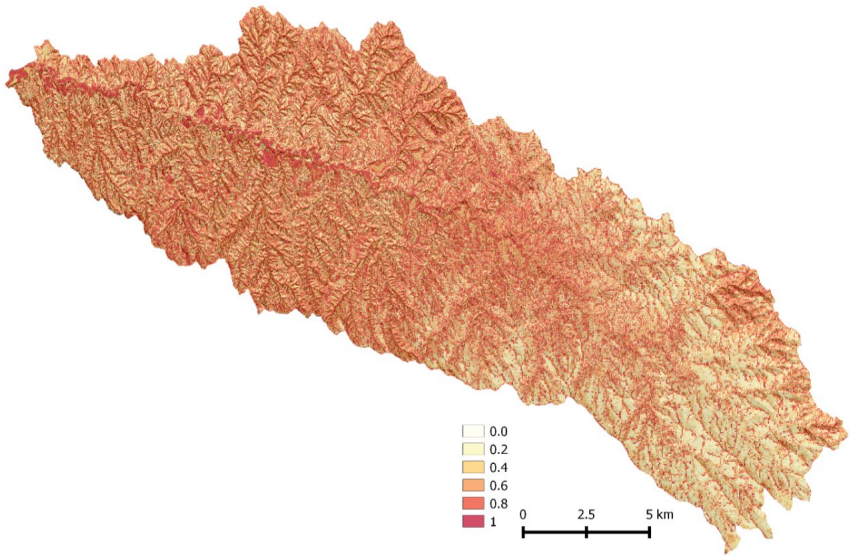
Digital soil mapping uses memberships or probabilities to quantify prediction uncertainty when modeling soil classes. Soil class memberships/probabilities indicate the similarity of soil class occurrence in each grid cell. Digital soil mapping produces a membership/probability grid for each modeled soil class. Confusion between soil class predictions is quantified with the confusion index (CI):

$$CI = [1-(\mu_{max} - \mu_{(max-1)})]$$

where μ_{max} is the membership/probability value of the class with the maximum membership/probability and $\mu_{(max-1)}$ is the second-largest membership/probability value. If the memberships/probabilities of the two most likely classes are similar (e.g., 0.3 and 0.2) then the CI will approach 1, indicating high confusion about the class to which the prediction should belong (fig. 5-11). If the memberships/probabilities of the two most likely classes are dissimilar (e.g., 0.7 vs. 0.1) then the CI will approach 0, indicating little confusion between classes (Burrough et al., 1997; Odgers et al., 2014).

Digital soil mapping uses prediction intervals to quantify uncertainty in soil property predictions (fig. 5-12). Prediction intervals (not confidence intervals, which measure uncertainty about the mean) indicate the range in values within which the true value is likely to occur (Malone et al., 2011). Digital soil mapping most commonly uses 90% prediction intervals, which indicate the range in values in which a new measurement will be found 9 times out of 10. Prediction intervals are most commonly shown as companion maps, where the lower prediction

Figure 5-11



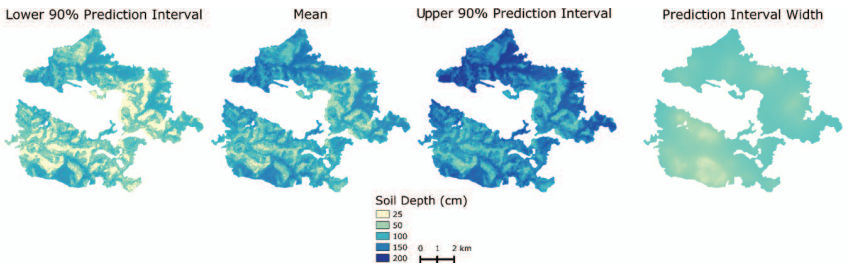
Example of the confusion index for soil class prediction over approximately 300 km^2 in the Powder River Basin, Wyoming. Confusion index values near 1 indicate areas of uncertainty in soil class spatial predictions. Figure adapted from Brungard et al. (2015).

interval, mean, and upper prediction interval are shown side by side (fig. 5-12). In some cases, the prediction interval width is also provided to indicate the spatial variability of uncertainty (fig. 5-12). Although less common, another option for displaying soil property prediction uncertainty is through "whitening" (Hengl, 2003, 2007), i.e., predictions whiten/pale based on the uncertainty so that highly uncertain predictions approach the color white. Methods for calculating prediction uncertainty are readily available in many software packages.

Applications of Digital Soil Mapping

Digital soil mapping is widely used to predict soil classes and properties and produce a soil map. However, the process of generating spatially explicit predictions of natural phenomena using quantitative relationships between training data and predictor variables can be applied to create a broad spectrum of information products. The following

Figure 5-12



Example of prediction intervals and prediction interval width for soil depth to a restricting layer over approximately 50 km² in San Juan County, Utah. Wider prediction intervals indicate greater uncertainty.

paragraphs discuss examples of the application of digital soil mapping in pedology and related fields to produce information products other than soil maps.

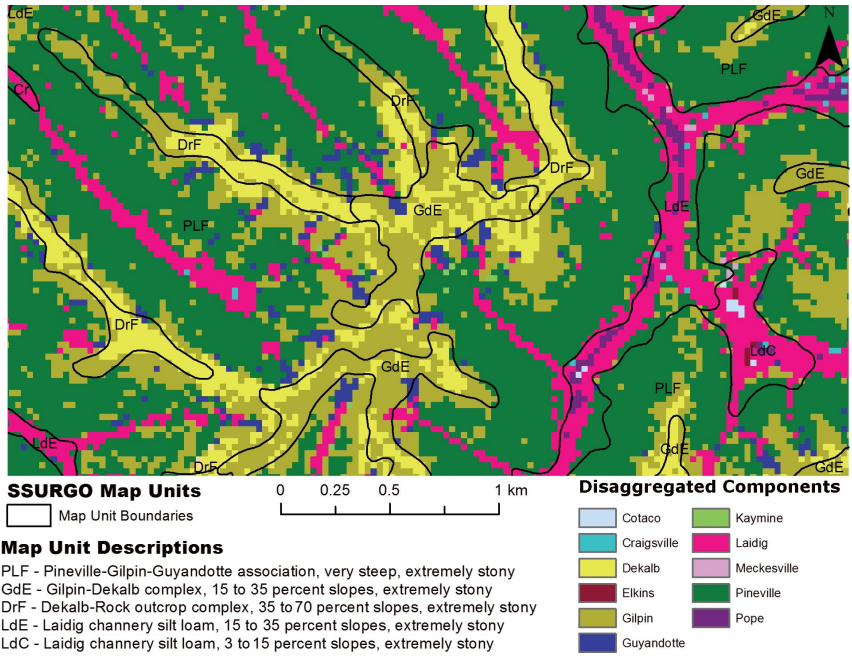
Raster vs. Polygon, Disaggregation, and Evaluation of Existing Maps

A fuzzy classification of Landsat 7 spectral data was applied in an update soil survey of wet and saline map units along the east shore of the Great Salt Lake, Utah, specifically for disaggregation of a few very broad map units. The disaggregated product showed the distribution of soil components (tied to land cover type) with an overall map accuracy of 88%. It highlighted the additional information a raster product can convey that a vector product cannot. The disaggregated raster product allowed for refinement of map unit concepts and line work, particularly in areas previously lumped into a miscellaneous "Playa" map unit, which had no soil information to support it. This survey area is important for wetland preservation and migratory habitat for large populations of birds and is experiencing pressure from encroaching development (Kienast-Brown and Boettinger, 2007).

Disaggregation of the Soil Survey Geographic Database (SSURGO) legacy data into maps at soil component level was completed for two West Virginia counties using soil-landscape knowledge, data mining, and predictive modeling (Nauman and Thompson, 2014). Descriptions of the soil-landscape relationships stored in the SSURGO database for the two survey areas were used, along with elevation data and derived

geomorphic indices, to build a set of representative training areas for all soil components. The training areas were used in classification tree ensemble models with additional environmental covariates to predict soil series extent (fig. 5-13). Underlying prediction frequency surfaces were also generated from the models and used to create continuous soil property maps. Model predictions agreed with validation pedons 22 to 44% of the time. This study demonstrates how disaggregation techniques may be used to update soil surveys.

Figure 5-13



Example of a disaggregation of SSURGO in West Virginia (modified from Nauman and Thompson, 2014) showing the hardened classification of soil series components with an overlay of the original map unit boundaries.

Predicting Biological Soil Crusts

Biological soil crusts are communities of cyanobacteria, algae, microfungi, mosses, liverworts, and lichens at the soil surface (Soilcrust.org, 2016). They stabilize soil, minimize wind and water erosion, and are important sources of soil N and organic C in arid and

semiarid ecosystems (Belnap et al., 2001). Biological soil crust level-of-development (LOD) classes represent a development sequence from low to high, with higher classes indicating greater cyanobacteria development (Belnap et al., 2008). Spatial predictions of low, moderate, and high LOD classes were completed for an area surrounding and including Canyonlands National Park, Utah, to assist in management of this important resource (Brungard and Boettinger, unpublished data).

Spatial predictions of the presence or absence of biological soil crust LOD class were derived using unweighted model averaging (Malone et al., 2014) of five statistical models: stochastic gradient boosting, random forests, maximum entropy, generalized linear models, and generalized additive models. Observations of biological soil crust used in model development were obtained during a 2006-2009 soil survey update of Canyonlands National Park, Utah.

Prediction uncertainty was calculated as the standard deviation of the combined probability predictions from each model. Lower prediction uncertainty indicates more robust predictions. Prediction quality was assessed using concordance. Concordance is the number of models predicting class occurrence in each raster cell. High concordance values (e.g., 5) indicate areas where all models predict biological soil crust presence and thus identify areas where greater confidence may be placed in presence predictions. Conversely, low concordance values (e.g., 1) indicate areas where only a few models predict biological soil crust presence and thus identify areas where less confidence may be placed in spatial predictions.

Predicting Ecological Sites

Correlating ecological sites with soil map units is an important component of soil mapping in the United States. It provides an understanding of how biotic and abiotic factors in the environment interact and influence one another. (Appendix 4 discusses ecological site descriptions.) Ecological sites are considered a vital part of many land management decisions (USDA-NRCS, 2008). Several studies focused on predicting distribution of vegetation types, to assist in understanding spatial relationships of ecological sites, have been conducted in Rich County, Utah. A selected set of elevation (DEM) and spectral (Landsat) data derivatives were used as input to logistic regression models to produce predictions of vegetation types that play a key role in ecological site identification (Peterson, 2009). An accuracy of 71% was reported based on an independent validation data set.

A subsequent study in Rich County, Utah, used a combination of elevation and spectral derivatives and random forests classification to predict five dominant vegetation types (Stam, 2012). Reported overall accuracies were between 81% and 98%. Prediction of ecological sites and states was also explored in this same study using Landsat spectral data derivatives and supervised classification, specifically the maximum likelihood classifier. A similarity index was calculated, based on the Mahalanobis distance generated during the classification, and related to various states (6 total) of the ecological site. The similarity index was successful in defining where different states of a given ecological site occur on the landscape, with a reported accuracy of 65%.

Predicting Rare Plant Habitat

Shrubby reed-mustard *(Schoenocrambe suffrutescens),* a U.S. federally listed endangered shrub endemic to the Uinta Basin, Utah, faces habitat loss due to fossil fuel energy development and extraction. Random forests models and digital environmental covariates were used to identify potential shrubby reed-mustard (SRM) habitat (Baker et al., 2016). A three-step approach was used to create the final predictive map. First, soil properties measured in the field were used to predict SRM presence or absence (out-of-bag [OOB] error of 10%). Second, these soil properties were correlated to elevation and spectral data, including a DEM, DEM derivatives, and Landsat 5 TM imagery, to predict SRM habitat onto a spatial extent and generate training data points for a final model (OOB error of 28%). Calcium carbonate equivalent, silt content, and dry color value were strongly correlated with yellowness from the Tasseled Cap transformation, 3/2 normalized difference ratio, and 3/1 normalized difference ratio (spectral band ratios typically associated with geology and carbonate content). Third, the spectral and elevation data were used to create a final predictive raster of potential SRM habitat with OOB error of 23%, validated by an independent dataset of SRM locations. Variable importance plots were used in all models to indicate the mean decrease in accuracy for each predictor variable. The most important predictor variables were selected and reduced to a subset by manual stepwise elimination to obtain the best model fit with the fewest variables. The final model can be used to identify potential habitat across a large area, especially where remote or rugged terrain make access difficult and time- and labor-intensive. Once soil and site data are located for potential habitat areas, they can be used to verify SRM habitat suitability and focus conservation or restoration efforts.

Summary

Digital soil mapping uses field and laboratory observations coupled with spatially explicit environmental covariates (SCORPAN) and modern computer technology to predict soil classes or properties. It complements and builds upon the collective knowledge and expertise accumulated over many decades of conventional soil survey work. Major advantages of digital soil mapping include:

- The most accurate model that resources can support through the iterative process of development and testing can be used to create the final soil map. Models can be refined until the resulting soil map meets accuracy and uncertainty standards.
- The uniform application of the model across the project area results in a consistent soil map.
- The degree of accuracy and uncertainty associated with the soil map can be expressed quantitatively.
- Soil information is captured for each grid cell rather than aggregated for entire polygons. As a result, there is a more detailed portrayal of the short-range soil variability over the landscape.
- The models developed to predict soil classes or properties are an effective way to capture and preserve expert knowledge about soil and landscape relationships.

References

Arrouyas, D., N. McKenzie, J. Hempel, A.R. de Forges, and A.B. McBratney. 2014. GlobalSoilMap: Basis of the global spatial soil information system. Taylor and Francis, London, UK.

Baker, J.B., B.B. Fonnesbeck, and J.L. Boettinger. 2016. Modeling rare endemic shrub habitat in the Uinta Basin using soil, spectra, and topographic data. Soil Science Society of America 80:395–408.

Behrens, T. 2003. Digitale Reliefanalyse als Basis von Boden-Landschafts-Modellen-am Beispiel der Modellierung periglaziarer Lagen im Ostharz. Boden und Landschaft 42:189.

Behrens, T., and T. Scholten. 2006. A comparison of data-mining techniques in predictive soil mapping. *In* P. Lagacherie, A.B. McBratney, and M. Voltz (eds.) Digital soil mapping: An introductory perspective, Elsevier, Amsterdam, pp. 353–617.

Behrens, T., H. Förster, T. Scholten, U. Steinrücken, E.-D. Spies, and M. Goldschmitt. 2005. Digital soil mapping using artificial neural networks. Journal of Plant Nutrition and Soil Science 168:21–33.

Behrens, T., A.X. Zhu, K. Schmidt, and T. Scholten. 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155(3–4):175–185.

Belnap, J., J.H. Kaltenecker, R. Rosentreter, J. Williams, S. Leonard, and D. Eldridge. 2001. Biological soil crusts: Ecology and management. Technical Reference No. 1730-2, USDI Geological Survey, Forest and Rangeland Ecosystem Science Center, Denver, CO.

Belnap, J., S.L. Phillips, D.L. Witwicki, and M.E. Miller. 2008. Visually assessing the level of development and soil surface stability of cyanobacterially dominated biological soil crusts. Journal of Arid Environments 72(7):1257–1264.

Bodily, J.M. 2005. Developing a digital soil survey update protocol at the Golden Spike National Historic site. M.S. thesis, Utah State University, Logan.

Boruvka, L., L. Pavlu, R. Vasat, V. Penizek, and O. Drabek. 2008. Delineating acidified soils in the Jizera Mountains region using fuzzy classification. *In* A.E. Hartemink, A. McBratney, and M.L. Mendonça-Santos (eds.) Digital soil mapping with limited data, Springer, The Netherlands, pp. 303–309.

Bouma, J., A.G. Jongmans, A. Stein, and G. Peek. 1989. Characterizing spatially variable hydraulic properties of a boulder clay deposit in The Netherlands. Geoderma 45:19-29.

Brungard, C.W., and J.L. Boettinger. 2010. Conditioned Latin hypercube sampling: Optimal sample size for digital soil mapping of arid rangelands in Utah, USA. *In* J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink, and S. Kienast-Brown (eds.) Digital soil mapping: Bridging research, environmental application, and operation, Springer, Dordrecht, The Netherlands, pp. 67–75.

Brungard, C.W., J.L. Boettinger, M.C. Duniway, S.A. Wills, and T.C. Edwards. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239-240:68-83.

Brus, D.J., B. Kempen, and G.B.M. Heuvelink. 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62(3):394–407.

Bui, E.N., and C.J. Moran. 2003. A strategy to fill gaps in soil survey over large spatial extents: An example from the Murray-Darling Basin of Australia. Geoderma 111:21–44.

Burrough, P.A., and R. McDonnell. 1998. Principles of geographical information systems. Oxford University Press.

Burrough, P.A., P.F.M. van Gaans, and R. Hootsmans. 1997. Continuous classification in soil survey: Spatial correlation, confusion and boundaries. Geoderma 77(2–4):115–135.

Carré, F., A.B. McBratney, T. Mayr, and L. Montanarella. 2007. Digital soil assessments: Beyond DSM. Geoderma 142:69-79.

Cole, N.J., and J.L. Boettinger. 2007. A pedogenic understanding raster classification methodology for mapping soils, Powder River Basin, Wyoming, USA. *In* P. Lagacherie, A.B. McBratney, and M. Voltz (eds.) Digital soil mapping: An introductory perspective. Developments in Soil Science, Vol. 31, Elsevier, Amsterdam, pp. 377-388.

Colwell, R.N. 1997. History and place of photographic interpretation. *In* Manual of Photographic Interpretation, second ed., American Society for Photogrammetry and Remote Sensing.

Congalton, R. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37(1):35–46.

de Gruijter, J., D.J. Brus, M.F.P. Bierkens, and M. Knotters. 2006. Sampling for natural resource monitoring. Springer, Berlin.

Dobos, E., and J. Daroussin. 2005. The derivation of the Potential Drainage Density Index (PDD). *In* E. Dobos, J. Daroussin, and L. Montanarella (eds.) An SRTM-based procedure to delineate SOTER terrain units on 1:1 and 1:5 million scales.

Duda, R.O., P.E. Hart, and D.G. Stork. 2001. Pattern classification. John Wiley & Sons, New York, NY.

Foody, G.M. 2000. Estimation of sub-pixel land cover composition in the presence of untrained classes. Computers and Geosciences 26:469-478.

Fox, G.A., and R. Metla. 2005. Soil property analysis using principal components analysis, soil line, and regression models. Soil Science Society of America Journal 69(6):1782.

Franklin, J., and J.A. Miller. 2009. Mapping species distributions: Spatial inference and prediction. Cambridge University Press.

Gallant, J.C., and T.I. Dowling. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research 39(12).

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik. 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46(1-3):389–422.

Hammond, E.H. 1954. Small scale continental landform maps. Annals of the Association of American Geographers 44:32-42.

Hammond, E.H. 1964. Analysis of properties in land form geography: An application to broad-scale land form mapping. Annals of the Association of American Geographers 54(1):11-19.

Hartemink, A.E., A. McBratney, and M.L. Mendonça-Santos (editors). 2008. Digital soil mapping with limited data. Springer, Heidelberg.

Hengl, T. 2003. Visualisation of uncertainty using the HSI colour model: Computations with colours. Proceedings of the 7th International Conference on GeoComputation, pp. 1–12.

Hengl, T. 2007. A practical guide to geostatistical mapping of environmental variables. Geoderma 140(4):417–427.

Hengl, T., and H.I. Reuter. 2008. Geomorphometry. Developments in Soil Science, Vol. 33, Elsevier, Amsterdam.

Hengl, T., G.B.M. Heuvelink, and D.G. Rossiter. 2007. About regression-kriging: From equations to case studies. Computers and Geosciences 33(10):1301–1315.

Hofierka, J., and M. Suri. 2002. The solar radiation model for Open Source GIS: Implementation and applications. Proceedings of the Open Source GIS-GRASS User's Conference, pp. 1-19.

Hole, F.D., and J.B. Campbell. 1985. Soil landscape analysis. Rowman & Littlefield.

Intergraph Corporation. 2013. ERDAS field guide. Available at http://e2b.erdas.com/Libraries/Misc_Docs/ERDAS_FieldGuide_PDF_Intergraph_brand.sflb.ashx. [Accessed 20 September 2016]

Issaks, E.H., and R.M. Srivastava. 1989. An introduction to applied geostatistics. Oxford University Press, New York, NY.

Iwahashi, J., and R.J. Pike. 2007. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology 86(3):409-440.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2014. An introduction to statistical learning: With applications in R. Springer, New York, NY.

Jasiewicz, J., and T.F. Stepinski. 2013. Geomorphons—A pattern recognition approach to classification and mapping of landforms. Geomorphology 182:147-156.

Jenny, H. 1941. The factors of soil formation. McGraw Hill, New York, NY.

Jensen, J.R. 2005. Introductory digital image processing: A remote sensing perspective, 3rd edition. Pearson Prentice Hall, pp. 296-300, 301-321, 315-316.

Kempen, B., D.J. Brus, G.B.M. Heuvelink, and J.J. Stoorvogel. 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A

multinomial logistic regression approach. Geoderma 151(3-4):311–326.

Kidd, D., B. Malone, A. McBratney, B. Minasny, and M. Webb. 2015. Operational sampling challenges to digital soil mapping in Tasmania, Australia. Geoderma Regional 4:1–10.

Kienast-Brown, S., and J.L. Boettinger. 2007. Land cover classification from Landsat imagery for mapping dynamic wet and saline soils. *In* P. Lagacherie, A.B. McBratney, and M. Voltz (eds.) Digital soil mapping: An introductory perspective. Developments in Soil Science, Vol. 31, Elsevier, Amsterdam, pp. 235–244.

Kienast-Brown, S., and J.L. Boettinger. 2010. Applying the optimum index factor to multiple data types in soil survey. *In* J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink, and S. Kienast-Brown (eds.) Digital soil mapping: Bridging research, environmental application, and operation, Springer, Dordrecht, The Netherlands, pp. 385–398.

Kuhn, M., and K. Johnson. 2013. Applied predictive modeling. Springer, New York, NY.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, T.R.C. Team, M. Benesty, R. Lescarbeau, A. Ziem, and L. Scrucca. 2015. caret: Classification and regression training.

Kursa, M.B., and W.R. Rudnicki. 2010. Feature selection with the Boruta package. Journal of Statistical Software 36(11):1–13.

Lagacherie, P., A.B. McBratney, and M. Voltz (editors). 2007. Digital soil mapping: An introductory perspective. Developments in Soil Science, Vol. 31, Elsevier, Amsterdam.

Levi, M.R., and C. Rasmussen. 2014. Covariate selection with iterative principal component analysis for predicting physical soil properties. Geoderma 219–220:46–57.

Libohova, Z., H.E. Winzeler, B. Lee, P.J. Schoeneberger, J. Datta, and P.R. Owens. 2016. Geomorphons: Landform and property predictions in a glacial moraine in Indiana landscapes. Catena 142:66-76.

Lohr, S.L. 2009. Sampling: Design and analysis, 2nd edition. Nelson Education.

MacQueen, J.B. 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Symposium on Math, Statistics, and Probability, University of California Press, Berkeley, CA.

Malone, B.P., A.B. McBratney, and B. Minasny. 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160(3–4):614–626.

Malone, B.P., B. Minasny, N.P. Odgers, and A.B. McBratney. 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232-234:34–44.

MathWorks, Inc. MATLAB 8.0 and Statistics Toolbox 8.1. Natick, MA.

McBratney, A.B., M.L. Mendonça-Santos, and B. Minasny. 2003. On digital soil mapping. Geoderma 117:3-52.

Miller, B.A., S. Koszinski, M. Wehrhan, and M. Sommer. 2015. Impact of multi-scale predictor selection for modeling soil properties. Geoderma 239-240:97–106.

Minasny, B., and A.B. McBratney. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32:1378–1388.

Minasny, B., B.P. Malone, and A.B. McBratney. 2012. Digital soil assessments and beyond. Proceedings of the 5th Global Workshop on Digital Soil Mapping. CRC Press, Boca Raton, FL.

Mohanty, B.P., and Z. Mousli. 2000. Saturated hydraulic conductivity and soil water retention properties across a soil-slope transition. Water Resources Research 43(11):3311-3324.

Moore, I.D., R.B. Grayson, and A.R. Ladson. 1991. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrological Processes 5(1):3-30.

Nauman, T.W., and J.A. Thompson. 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213:385-399.

Nield, S.J., J.L. Boettinger, and R.D. Ramsey. 2007. Digitally mapping gypsic and natric soil areas using Landsat ETM data. Soil Science Society of America Journal 71:245-252.

Nilsson, R., J.M. Peña, J. Björkegren, and J. Tegnér. 2007. Consistent feature selection for pattern recognition in polynomial time. Journal of Machine Learning Research 8:589–612.

O'Callaghan, J., and D. Mark. 1984. The extraction of drainage networks from digital elevation data. Computer Vision, Graphics, and Image Processing 28:323–344.

Odeh, I.O.A., A.B. McBratney, and D.J. Chittleborough. 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63:197–214.

Odeh, I.O.A., A.B. McBratney, and D.J. Chittleborough. 1995. Further results on prediction of soil properties from terrain attributes—Heterotrophic cokriging and regression kriging. Geoderma 67:215–226.

Odgers, N.P., W. Sun, A.B. McBratney, B. Minasny, and D. Clifford. 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215:91–100.

Olaya, V., and O. Conrad. 2009. Geomorphometry in SAGA. Developments in Soil Science, Vol. 33, Elsevier, Amsterdam, pp. 293-308.

Olea, R.A. 2009. A practical primer on geostatistics. USGS Open-File Report 2009-1103, USDI Geological Survey, Reston, Virginia.

Padarian, J., B. Minasny, A.B. McBratney, and N. Dalgliesh. 2014. Predicting and mapping the soil available water capacity of Australian wheatbelt. Geoderma Regional 2–3:110–118.

Park, S.J., G.R. Ruecker, W.A. Agyare, A. Akramhanov, D. Kim, and P.L.G. Vlek. 2009. Influence of grid cell size and flow routing algorithm on soil-landform modeling. Journal of the Korean Geographical Society 44:122–45.

Pebesma, E.J. 2004. Multivariable geostatistics in S: The gstat package. Computers and Geosciences 30:683–691.

Peterson, K.A. 2009. Modeling potential native plant species distributions in Rich County, Utah. All Graduate Theses and Dissertations Paper 649, Utah State University.

PRISM Climate Group. 2016. PRISM climate data. Northwest Alliance for Computational Science & Engineering (NACSE) at Oregon State University. Available at http://prism.oregonstate.edu. [Accessed 16 September 2016]

Quinn, P.F., K.J. Beven, P. Chevallier, and O. Planchon. 1991. The prediction of hillslope flowpaths for distributed modelling using digital terrain models. Hydrological Processes 5:59-80.

R Core Team. 2013. R: A language and environment for statistical computing. Vienna, Austria.

Riley, S.J., S.D. DeGloria, and R. Elliot. 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Intermountain Journal of Sciences 5:23-27.

Roecker, S.M., and J.A. Thompson. 2010. Scale effects on terrain attribute calculation and their use as environmental covariates for digital soil mapping. *In* J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink, and S. Kienast-Brown (eds.) Digital soil mapping: Bridging research, environmental application, and operation, Springer, Dordrecht, The Netherlands, pp. 55–66.

Rossiter, D.G. 2003. Methodology for soil resource inventories, 3rd edition. ITC Lecture Notes SOL.27, International Institute for Aerospace Survey and Earth Sciences, Enschede, The Netherlands.

Roudier, P. 2011. clhs: An R package for conditioned Latin hypercube sampling.

Roudier, P., D.E. Beaudette, and A.E. Hewitt. 2012. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. *In* B. Minasny, B.P. Malone, and A. McBratney (eds.) Digital soil assessments and beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping, CRC Press, Sydney, Australia, pp. 227–231.

Saunders, A.M., and J.L. Boettinger. 2007. Incorporating classification trees into a pedogenic understanding raster classification methodology, Green River Basin, Wyoming, USA. *In* P. Lagacherie, A.B. McBratney, and M. Voltz (eds.) Digital soil mapping: An introductory perspective. Developments in Soil Science, Vol. 31, Elsevier, Amsterdam, pp. 389-399.

Schaeffer, R.L., W. Mendenhall, and L. Ott. 1990. Elementary survey sampling, 4th edition. PWS-Kent Publishing Company.

Schmidt, J., and A. Hewitt. 2004. Fuzzy land element classification from DTMs based on geometry and terrain position. Geoderma 121(3):243-256.

Schowengerdt, R.A. 1997. Remote sensing: Models and methods for image processing, 2nd edition. Academic Press.

Smith, M.P., A.X. Zhu, J.E. Burt, and C. Stiles. 2006. The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma 137(1–2):58–69.

Soilcrust.org. 2016. Biological soil crusts. http://www.soilcrust.org/index.htm [Accessed 20 September 2016]

Stam, C.A. 2012. Using biophysical geospatial and remotely sensed data to classify ecological sites and states. All Theses and Dissertations Paper 1389, Utah State University.

Stum, A.K., J.L. Boettinger, M.A. White, and R.D. Ramsey. 2010. Random forests applied as a soil spatial predictive model in arid Utah. *In* J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink, and S. Kienast-Brown (eds.) Digital soil mapping: Bridging research, environmental application, and operation, Springer, Dordrecht, The Netherlands, pp. 179–190.

Tarboton, D.G. 1997. A new method for the determination of flow directions and contributing areas in grid digital elevation models. Water Resources Research 33(2):309-319.

Thompson, J.A., J.C. Bell, and C.A. Butler. 2001. Digital elevation model resolution: Effects on terrain attribute calculation and quantitative soil-landscape modeling. Geoderma 100(1-2):67–89.

Tou, J.T., and R.C. Gonzalez. 1974. Pattern recognition principles. Addison-Wesley.

Triantifilis, J., N.Y. Earl, and I.D. Gibbs. 2012. Digital soil-class mapping across the Edgeroi district using numerical clustering and gamma-ray

spectrometry data. *In* B. Minasny, B.P. Malone, and A. McBratney (eds.) Digital soil assessments and beyond. Proceedings of the 5th Global Workshop on Digital Soil Mapping, CRC Press, Sydney, pp. 187–191.

U.S. Department of Agriculture, Natural Resources Conservation Service. 2008. Ecological sites: Understanding the landscape fact sheet. Available at http://nitcnrcsbase-www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1043492.pdf. [Accessed 20 September 2016]

U.S. Department of Agriculture, Natural Resources Conservation Service. 2016a. Geospatial Data Gateway. https://gdg.sc.egov.usda.gov [Accessed September 20, 2016]

U.S. Department of Agriculture, Natural Resources Conservation Service. 2016b. Job Aids (Soil Databases, GIS). http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/edu/ncss/?cid=nrcs142p2_054322 [Accessed 20 September 2016]

U.S. Department of the Interior, Geological Survey. 1999. National GAP analysis program land cover data, version 2.

U.S Department of the Interior, Geological Survey. 2016a. EarthExplorer. http://earthexplorer.usgs.gov [Accessed 20 September 2016]

U.S Department of the Interior, Geological Survey. 2016b. The National Map. http://ned.usgs.gov/ [Accessed 20 September 2016]

Vaughan, R., and K. Megown. 2015. The Terrestrial Ecological Unit Inventory (TEUI) Geospatial Toolkit: User guide v5.2. RSAC-10117-MAN1. USDA Forest Service, Remote Sensing Applications Center, Salt Lake City, UT.

Wang, F. 1990. Improving remote sensing image analysis through fuzzy information representation. Photogrammetric Engineering and Remote Sensing 56:1163-1169.

Webster, R., and M.A. Oliver. 2007. Geostatistics for environmental scientists, 2nd edition. John Wiley & Sons.

Wilson, J.P., and J.C. Gallant. 2000. Terrain analysis: Principles and applications. John Wiley & Sons.

Xiong, X., S. Grunwald, D.B. Myers, J. Kim, W.G. Harris, and N.B. Comerford. 2014. Holistic environmental soil-landscape modeling of soil organic carbon. Environmental Modelling and Software 57:202–215.

Zadeh, L.A. 1965. Fuzzy sets. Information and Control 8:338-353.

Zhu, A.X., B. Hudson, J. Burt, K. Lubich, and D. Simonson. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal 65:1463-1472.