Algorithms for Quantitative Pedology: A Toolkit for Soil Scientists

D.E. Beaudette
Dept. Land, Air and Water Resources
University of California, Davis

September 15, 2010

Abstract

Soils are routinely sampled and characterized according to genetic horizons (layers), resulting in data that are associated with principal dimensions: location (x,y), depth (z), and property space ($\mathbf{p}$). The high dimensionality and grouped nature of this type of data can complicate standard analysis, summarization, and visualization. The **aqp** package was developed to address some of these issues, as well as provide a useful framework for the advancement of quantitative studies in soil genesis, geography, and classification.

1 Background

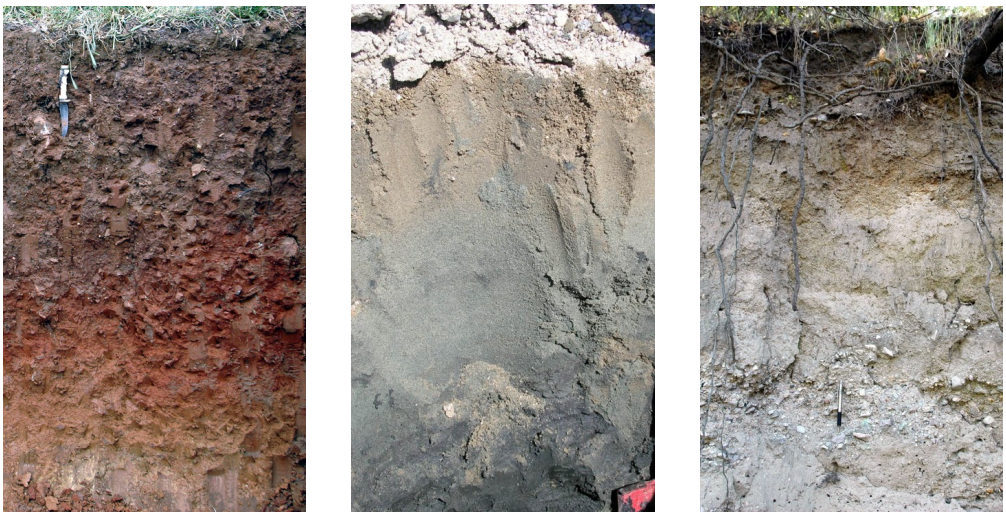
The soils of the world support a wide range of natural ecosystems, agricultural production, industrial processes, and the largest surficial carbon pool (Schlesinger, 1997). The rise and fall of past civilizations can be directly linked to the use and misuse of the soil resource (Hillel, 1998). A staggering quantity of soils information has been collected over the last 100 years, yet these data are often underutilized due to the sheer volume and complex structure. We have developed an **R** package that supports the interpretation of massive soils databases through numerical extensions to traditional methods of visualizing, aggregating, and classifying soils information. Further development of these numerical analogues will provide a new set of *quantitative* tools that soil scientists and surveyors can use in conjunction with well-established, *qualitative* methods.

Soil science is an integrative approach to understanding surficial processes that includes concepts from several disciplines (Buol et al., 2003). Pedology, one of several branches of soil science, is the study of the genesis, morphology, classification, and geography of soils. Soil profiles are usually described, sampled, and characterized by genetic horizons ("layers" defined by morphology and usually associated with an inferred process), extending from the surface to a lower boundary determined by bedrock contact or to a depth of 150-200 cm (Soil Survey Division Staff, 1993). The stratigraphy and morphology of soil horizons are usually the first data that the soil scientist uses to qualitatively classify a soil: i.e. degree of alteration relative to the parent material (Figure 1(a)), expression of oxidized or reduced forms of iron (Figure 1(b)), accumulation of organic matter, or evidence of cyclical deposition of new material (Figure 1(c)).

Hans Jenny was one of the first researchers to advocate a semi-quantitative theory of soil genesis; in which he described the "factors of soil formation" concept (Jenny, 1941). This novel approach is based on the expression:

$$S = f(cl, o, r, p, t) \tag{1}$$

where S represents a branch within a soil classification system, a collection of soil properties associated with a soil profile or a single layer (horizon). The parameters within the "clorpt" framework are: cl



(a) (b) (c)

Figure 1: Examples of soil profiles illustrating how horizons change with depth. Color, texture, structure and root abundance are common visual indicators of near surface processes in soil.

representing a climate factor, o representing an organic factor, r representing a relief factor, p representing a parent material factor, and t representing time. The S term in equation 1 can be modeled as matrix of soil properties (columns) associated with either genetic horizons or regular depth-slices (rows), occurring at some location in space. While the "clorpt" model is a useful construct for understanding how soil genesis might proceed, quantitative evaluation is usually not possible because of complex interaction and possible feedback mechanisms between terms on the right-hand side of the expression (Huggett, 1975). The left-hand side of the expression, S, is especially difficult to quantitatively define when it describes a collection of soil horizons and properties. The magnitude of measured properties, correlation between properties, and trends with depth are all critical elements of how a soil profile is interpreted as a whole (Arkley, 1976)

Several mature systems exist for the classification of soil profiles; Soil Taxonomy, World Reference Base, Australian Soil Classification, etc. (Buol et al., 2003). Each system is based on current knowledge of soil genesis, manifestation of specific processes in the form of field or lab measured properties, and region-specific land use limitations. Most soil classification systems seek to accommodate the (potential) global variability of soils (including Soil Taxonomy and World Reference Base), while others are tailored to region-specific soil variability. Soil Taxonomy (Soil Survey Staff, 1999) provides a rich vocabulary for grouping soils into several levels of a hierarchy based on established land-use limitations and our current knowledge of soil genesis. However, Soil Taxonomy does not currently define an approach for *numerically* describing the difference between soils. There has been limited work on purely numerical systems of soil classification (Rayner, 1966; Moore and Russell, 1967; Moore et al., 1972; Little and Ross, 1985; Dale et al., 1989), and several authors have suggested the potential merit to such an approach (Webster, 1968; Arkley, 1976; Minasny and McBratney, 2007; Carré and Jacobson, 2009). In particular, Young and Hammer (2000) suggested that fine-scale soil variability is more adequately captured by numerical classification as opposed to Soil Taxonomy. To date, most numerical soil classification methods are rarely employed outside of case studies presented within scientific journals. A numerical classification system could potentially be used to bridge national taxonomic systems (i.e. Soil Taxonomy and the Australian Soil Classification system) based solely on soil physical and chemical processes. Alternative classification schemes could be generated from the same underlying data, but directed towards specific goals, by selecting which variables and dissimilarity

metric are used. For example, soil texture, organic matter content, and aggregate stability information (weighted such that near-surface horizons contribute more to the final classification) could be used to generate a classification scheme supporting erosion prevention.

2 The aqp Package

The **aqp** (Algorithms for Quantitative Pedology) package for **R** (R Development Core Team, 2006) was developed to address some of the difficulties associated with processing soils information, specifically related to visualization, aggregation, and classification of soil profile data. This package is based on S3-style functions and classes. Most functions use basic dataframes (rectangular data tables) as input, where rows represent soil horizons and columns define properties of those horizons. Rows associated with each profile are "stacked" (i.e. long format), with collections of rows corresponding a single soil profile identified by an appropriate ID. Functions within the AQP package assume that horizon boundaries are defined as depth from the soil surface, and that the lower boundary of the deepest horizon represents a logical "end" to the soil profile– either contact with a root restricting layer or to a conventionally used lower boundary (e.g. 150 cm). The **aqp** package defines two specialized classes, 'SoilProfile' and 'SoilProfileList', for storage of profile-level metadata, as well as summary, print, and plotting methods that have been customized for common tasks related to soils data.

2.1 Visualizing Soil Profile Data

Visualization of key soil morphologic properties (i.e. color) is the first step in the interpretation of soil profile information. Therefore, a simple diagram (Figure 2) illustrating horizon depths, colors, and names (for a collection of soil profiles) represents an ideal starting point for presenting the soils within an area of interest. The `profile_plot()` function provides an approach for rendering soil profiles, based on basic stratigraphic parameters: horizon top boundary, bottom boundary, horizon name, and optionally horizon color (specified in a format that **R** understands). Combined with the base graphics plotting and layout capabilities, the `profile_plot()` function can be used to quickly organize and depict soils information.

The **aqp** package has several other functions for visualizing soils information: 1) `plot_slices()` for generating maps of soil properties by depth slice, 2) `panel.soil_profile()` for plotting grouped soil properties vs. depth as step functions, and 3) `panel.depth_function()` for plotting grouped depth functions, accompanied by upper and lower confidence limits, vs. depth. In addition, there are several examples within the manual pages that describe how to integrate these functions into calls to base and lattice graphics commands for the production of complex diagrams (See §4.1).

2.1.1 Color Conversion

Since soil colors are measured in Munsell notation (hue, value, chroma), conversion to the RGB colorspace is required for digital reproduction. The `munsell2rgb()` function uses a look-up table of common soil colors, and can directly convert (hue, value, chroma) coordinates into (R, G, B) triplets or hexadecimal-encoded colors. The `munsell` look-up table was generated from the MCSL spectral database of Munsell chips (Munsell Color Science Laboratory, 2010) and color conversion equations (Lindbloom, 2010). The Munsell Color Science Laboratory (MCSL) spectral database contains xyY colorspace coordinates for a range of commonly used Munsell colors, defined at even-numbered chroma values. Colors at odd-numbered chroma values were derived by estimating xyY colorspace coordinates along the entire range of chroma defined for each Munsell hue and value, via spline interpolation (Figure 3).

The conversion from xyY coordinates to RGB coordinates was performed with the following 4 steps: conversion from xyY to XYZ coordinates (Equation 2), chromatic adaption transformation from the C to

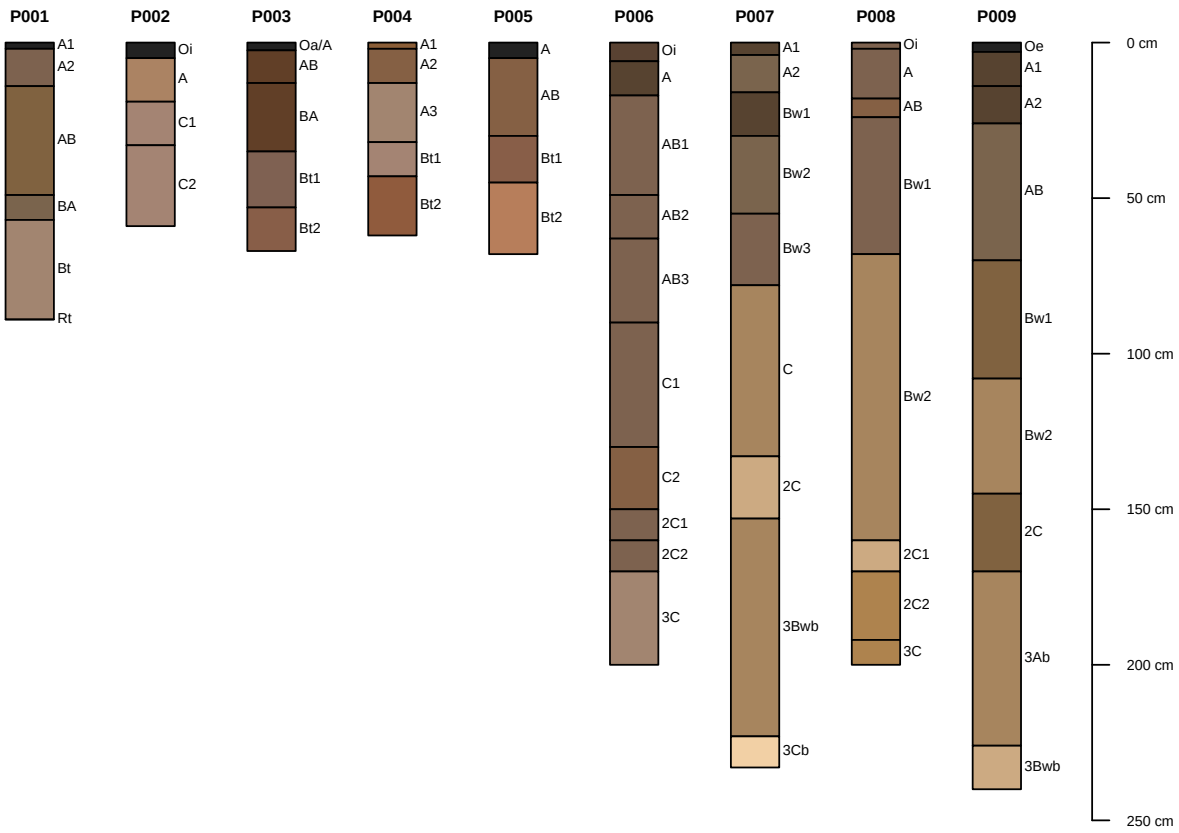


Figure 2: Visualization of nine soil profiles from Pinnacles National Monument, CA; colored by RGB representations of field-described dry colors.

D65[1] illuminant (Equation 3), conversion from XYZ (D65 illuminant) to rgb (Equation 4), scaling of rgb values to conform to a specific gamma value (Equation 5).

$$\begin{aligned} X &= \frac{xY}{y} \\ Y &= Y \\ Z &= \frac{Y(1-x-y)}{y} \end{aligned} \tag{2}$$

$$\begin{bmatrix} X_{D65} & Y_{D65} & Z_{D65} \end{bmatrix} = \begin{bmatrix} X & Y & Z \end{bmatrix} \begin{bmatrix} 0.990448 & -0.012371 & -0.003564 \\ -0.007168 & 1.015594 & 0.006770 \\ -0.011615 & -0.002928 & 0.918157 \end{bmatrix} \tag{3}$$

$$\begin{bmatrix} r & g & b \end{bmatrix} = \begin{bmatrix} X_{D65} & Y_{D65} & Z_{D65} \end{bmatrix} \begin{bmatrix} 3.24071 & -0.969258 & 0.0556352 \\ -1.53726 & 1.87599 & -0.203996 \\ -0.498571 & 0.0415557 & 1.05707 \end{bmatrix} \tag{4}$$

$$R, G, B = \begin{cases} 12.92 \times \{r, g, b\} & : r, g, b \leq 0.0031308 \\ 1.055 \times \{r, g, b\}^{(1.0/2.4)} - 0.055 & : r, g, b > 0.0031308 \end{cases} \tag{5}$$

[1]Most **R** plotting functions, and computer monitors in general, use the sRGB color profile which assumes a D65 illuminant.

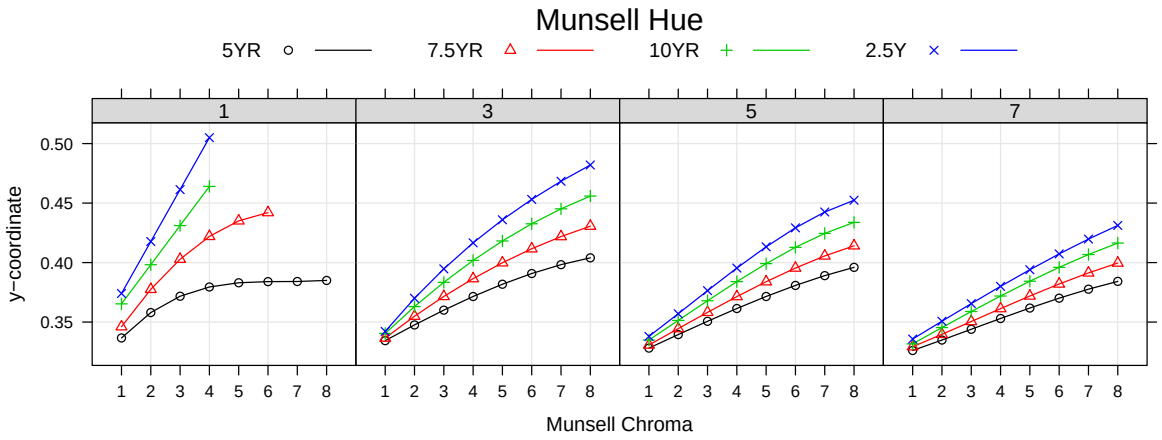


Figure 3: Relationship between Munsell chroma and y-coordinate (xyY colorspace), for selected Munsell hue (defined by line color and point symbol) and Munsell value (panels). Points at even-numbered Munsell chroma values were derived from the MCSL spectral database. Points at odd-numbered Munsell chroma values were estimated by spline interpolation.

2.2 Re-Alignment of Soil Horizons into Depth Slices

2.2.1 Aggregation by Depth Slice

Standard aggregation or summarization of soils information usually involves properties summed across all horizons (carbon quantity), mean values that are weighted by profile thickness (clay content), or depth to a diagnostic feature (bedrock contact). These approaches generalize well to tasks that require either: single estimates of a given soil property at sampling locations in space, or, group-wise estimates of a given soil property. These approaches do not generalize well to cases in which vertical variation in a given soil property is of interest, and needs to be summarized for a group of soil profiles. For example, the change in clay content with depth is used as a diagnostic element in US Soil Taxonomy, and is an important criterion for several land use interpretations. A collection of soils within a given region is likely to include a wide range in horizon designation (A, B, C, etc.), depth, thickness, and horizon sequences. Therefore, summarization by major horizon type is confounded by variable thickness of major horizon types (A, B, C, etc.), and potential absence of major horizon types at some locations. We present an alternative approach, where soil properties are summarized along a set of depth slices, despite being collected by genetic horizon (Figure 4).

The algorithm (implemented in the `soil.slot()` function) is based on the assumption that a *representative depth function* for some soil property (i.e. clay content) can be generated from a collection of soil profiles by summarizing this property along depth slices. Depth slices are defined by a segmenting vector: either regularly spaced (1 cm) intervals, or a user-defined vector of segment boundaries (e.g. 0-10, 10-25, 25-50, 50-150). Each profile in the collection is first segmented according to the specified segmenting vector. Then, summary statistics are computed along slices within the collection of profiles. If the segments (s) associated with some property are represented as a matrix, rows would represent a slice of values across the collection of profiles ($\mathbf{s}_i$) at depth interval i, and columns would represent the sequence of values ($\mathbf{s}_j$) associated with profile j. Therefore, the computation of the aggregate value by slice ($\bar{s}_i$) can be symbolized as:

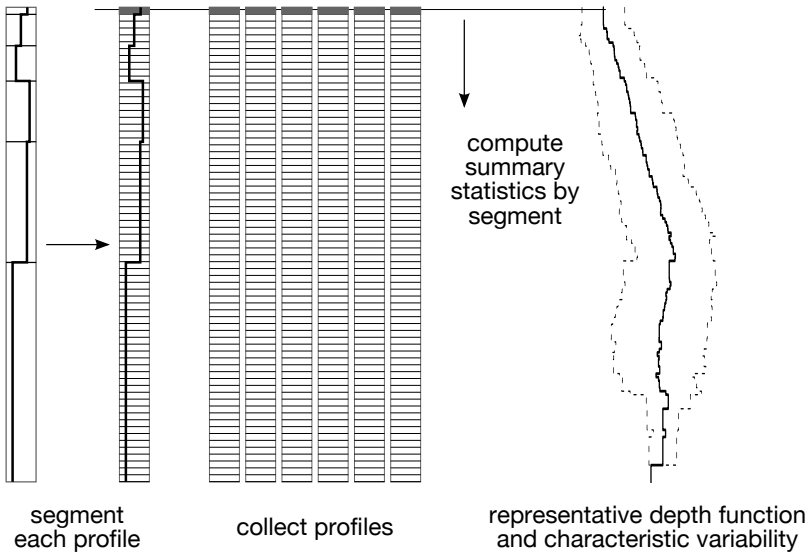


Figure 4: Demonstration of the soil profile aggregation algorithm, for a single soil property, aggregated along a regular sequences of depth slices.

$$\begin{vmatrix} s_{1,1} & s_{1,2} & s_{1,3} & s_{1,4} & \dots & s_{1,j} \\ s_{2,1} & s_{2,2} & s_{2,3} & s_{2,4} & \dots & s_{2,j} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{i,1} & s_{i,2} & s_{i,3} & s_{i,4} & \dots & s_{i,j} \end{vmatrix} \xrightarrow{\bar{s}_i = f(\mathbf{s}_i)} \begin{vmatrix} \bar{s}_1 \\ \bar{s}_2 \\ \vdots \\ \bar{s}_i \end{vmatrix} \tag{6}$$

where $f(\ )$ is a scalar-returning function applied row-wise, resulting in a new column vector. The algorithm currently supports calculation of mean, standard deviation, quantiles, or a user-defined function. The resulting estimate of central tendency and spread around that tendency for each depth slice are reconstituted into a single *representative depth function* (Figure 4). When available, weights (i.e. area fractions) can be supplied for each profile, resulting in weighted versions of most summary statistics. Representative depth functions can be computed for continuous variables (i.e. clay content), categorical variables, and soil depth probability (i.e. probability that the soil profile ends at a given depth). Probabilities for each level k of a categorical variable are computed by slice:

$$\hat{s}_{k.i} = freq(\mathbf{s}_{k.i})/j \tag{7}$$

where $freq(\mathbf{s}_{k.i})$ is the frequency of level k along slice i, and j is the number of profiles in the collection. Depth-slice probabilities generated from major horizon types can reveal site-wide patterns in soil morphology, such as A horizon thickness, presence or absense of transitional horizons (i.e. AB or BA horizons), and depth to paralithic or bedrock contact. In addition, a collection of horizon types can be combined into a new, synthetic soil profile from depth-slice probabilities; representative of the collection of soils used within the aggregation.

2.2.2 Investigation of Spatial Patterns by Depth Slice

Representation of spatial patterns in soil properties, either at points or interpolated along a regular grid, is confounded by irregular horizon depths, variation in naming conventions used by different workers, and the absence of certain horizon types at certain locations. The depth slice aggregation methods presented

in the previous section can be extended to re-align soils data (collected by genetic horizon) onto a common depth-basis. The `format_slices()` function is provided to re-format the resulting "sliced" data into a list of `sp` class elements, suitable for mapping or modeling tasks.

2.2.3 Limitations of the Algorithm

Care should be taken when applying this algorithm in two major cases, where either 1) several heterogeneous groups of soil profiles are aggregated together, and, or 2) soil depth varies greatly within a collection of soil profiles. In the first case, aggregation will result in depth functions that are *numerically* correct, but in no way representative of any of the original soils in the original collection. In the second case, aggregation will result in depth functions that are reasonable near the surface, but quite unreasonable beyond the average depth of the shallowest profiles within the collection. An important diagnostic that is returned by this algorithm (the "contributing fraction") describes what fraction of the original profile collection was used to compute an aggregate value at each depth slice. Inspection of this value, along with an evaluation of spread relative to central tendency (i.e. coefficient of variation) can help determine when aggregate depth functions may not be sufficiently representative.

2.3 Numerical Classification of Soil Profiles

2.3.1 Pair-Wise Comparison by Depth Slice

One approach to a purely numerical extension to soil classification requires the calculation of pair-wise dissimilarity between soil profiles. Since soil profiles are defined by an ordered (in depth) set of horizons, a numerical comparison must account for variation in horizon thickness and associated properties between profiles (Webster and Oliver, 1990). Our approach builds on work of Moore et al. (1972) and the previously mentioned depth-slicing algorithm. Between-profile dissimilarity is evaluated along regular depth slices: every slice, every other slice, or every *nth* slice (Figure 5(a)). The final between-profile dissimilarity is computed by summing the collection of slice-wise dissimilarity matrices. Internally, this is accomplished by forming a series of soil property matrices $\mathbf{P}_j$ representing "sliced" soil properties from profile j, where a single cell in this matrix $x_{i,p}$ represents a soil property p from depth slice i. The vector of properties defined by a single slice $\mathbf{x}_{ij}$ from the collection of property matrices are accumulated, row-wise, forming a new matrix $\mathbf{X}_i$. In this matrix, rows represent profiles and columns represent properties. Pair-wise dissimilarity $\mathbf{D}$ between profiles is computed as the sum of slice-wise dissimilarity:

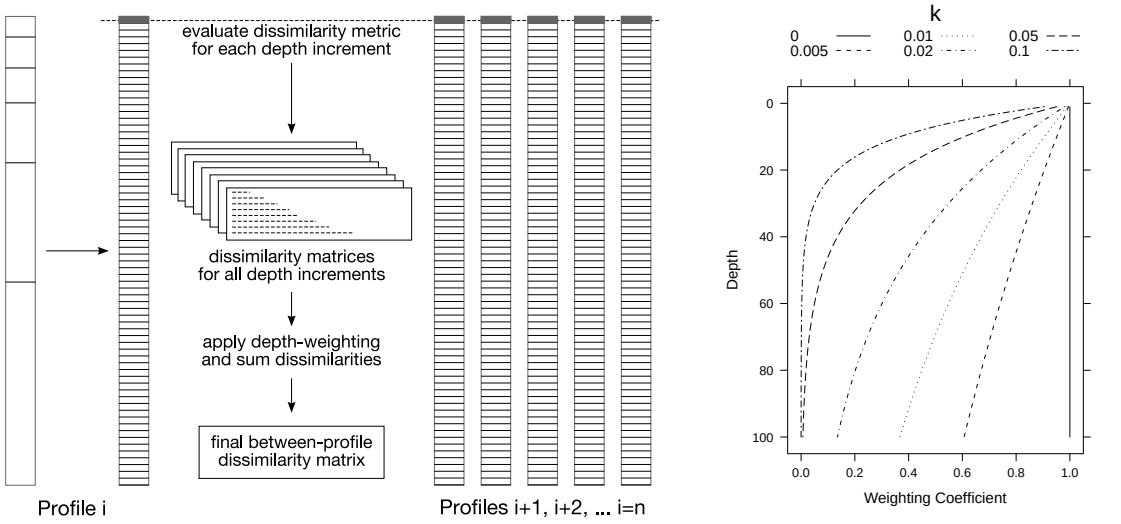
$$\mathbf{D} = \sum_{i=1}^{n} w_i G(\mathbf{X}_i) \tag{8}$$

where n is the number of slices, w_i is an optional weighting coefficient, and $G(\ )$ is Gower's generalized dissimilarity metric (Gower, 1971).

The algorithm (implemented in the `profile_compare()` function) represents a compromise between the way soils are commonly described and sampled (by genetic horizon type) and a normalized basis for the comparison of measured properties (depth slice). Gower's generalized dissimilarity metric is available via the `daisy()` function (from the **cluster** package). This metric can be evaluated from binary, categorical, and continuous variables; and can accommodate limited occurrences of missing observations (Kaufman and Rousseeuw, 2005). Between-profile dissimilarities are computed to a user-specified maximum depth.

2.3.2 Tuning Parameters

Before summation of dissimilarities across depth slices, the matrix of between-profile dissimilarities can be weighted according to the depth of a given slice (d) via an exponential decay function: $w = e^{-k \times d}$, where



(a) Algorithm details.

(b) Example weight functions.

Figure 5: Calculation of pair-wise profile dissimilarity.

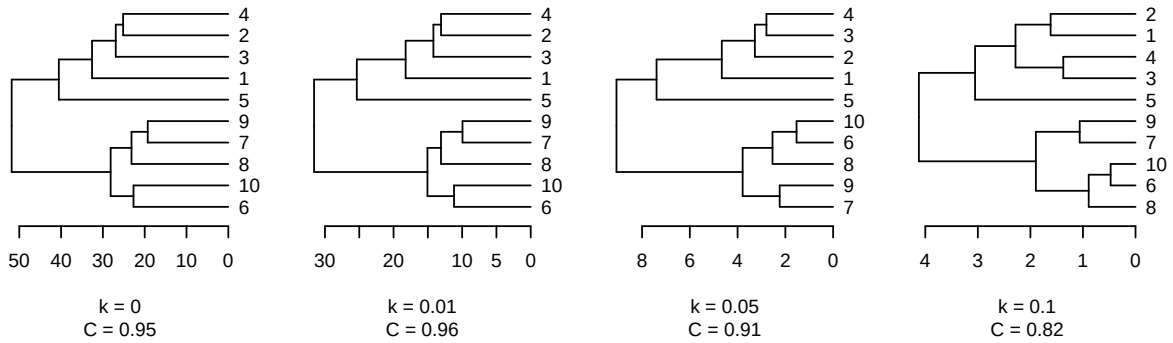


Figure 6: Effect of adjusting the depth weighting (k) parameter from 0 to 0.1 on soil profile grouping. Cophenetic correlation coefficients are printed below k values. "Average linkage" agglomerative clustering was used to build dendrograms.

k and d are > 0. The decay rate parameter (k) determines how rapidly a slice's dissimilarity value is down-weighed with depth: a value of 0.1 would effectively remove any influence of dissimilarities computed below about 30 cm, and a value of 1 would weight all slices equally (Figure 5(b)). The actual value for k should be determined as objectively as possible; i.e with a combination of knowledge about expected vertical anisotropy and a metric such as the cophenetic correlation coefficient (Sneath and Sokal, 1973). Within the sample dataset `sp3`, incrementing k from 0 to 0.1, with respect to resulting agglomerative ("average" method) clustering is demonstrated in Figure 6. For this dataset, lower levels of k result in better agreement (larger cophenetic correlation coefficients) between the dissimilarity matrix and grouping defined by agglomerative clustering. The highest cophenetic correlation coefficient is encountered when k = 0.01, close to the depth weighting values suggested by Russell and Moore (1968).

Variable soil depth can interfere significantly with the calculation of between-profile dissimilarity. For example, how should the dissimilarity between two profiles at a given depth, when one of the profiles is shallower than that depth? When soil depths are not well defined (e.g. alluvial soils excavated with different tools) a lower-limit to the depth-wise dissimilarity calculation should be sufficient. When the soils

in question have been described down to a natural lower boundary (e.g. bedrock, root-restricting layer, etc.) the dissimilarity between soil and non-soil material should be incorporated into the final dissimilarity between profiles. As currently implemented in the function `daisy()` (Kaufman and Rousseeuw, 2005), Gower's dissimilarity metric is undefined when one of the two inputs is missing. Therefore, when a 25 cm deep profile is compared with a 50 cm deep profile, pair-wise dissimilarities are only accumulated for the first 25 cm of soil (dissimilarities from 26 - 50 cm are NULL). When summed, the total dissimilarity between these profiles will generally be much lower than if the profiles had been of equal depth.

Our algorithm has an option (setting `replace_na=TRUE`) to replace NULL dissimilarity with the maximum dissimilarity between any pair of profiles for the current depth slice (Figure 7). In this way, the dissimilarity between a slice of soil and a corresponding slice of non-soil reflects the fact that these two materials should be treated very differently (i.e. maximum dissimilarity). This alternative calculation of dissimilarity between soil and non-soil slices solves the problem of comparing shallow profiles with deeper profiles. However, it can result in a new problem: dissimilarity calculated between two shallow profiles will be erroneously inflated beyond the extent of either profile's depth when deeper profiles exist in the collection. Our algorithm has an additional option (setting `add_soil_flag=TRUE`) that will preserve NULL distances between slices when both slices represent non-soil material (Figure 7). With this option enabled, shallow profiles will only accumulate mutual dissimilarity to the depth of the deeper profile.

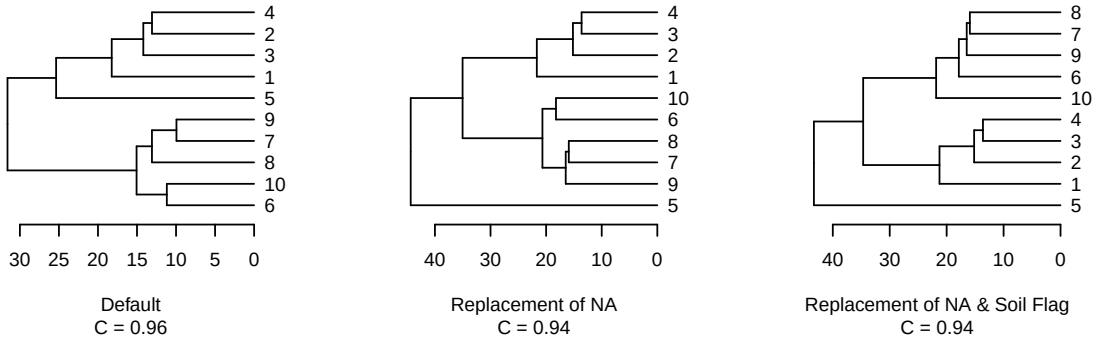


Figure 7: Effect of enabling the `replace_na` and `add_soil_flag` options on between-sample dissimilarity when $k = 0.01$, and `max_d` = 100 cm. Cophenetic correlation coefficients are printed below options. "Average linkage" agglomerative clustering was used to build dendrograms.

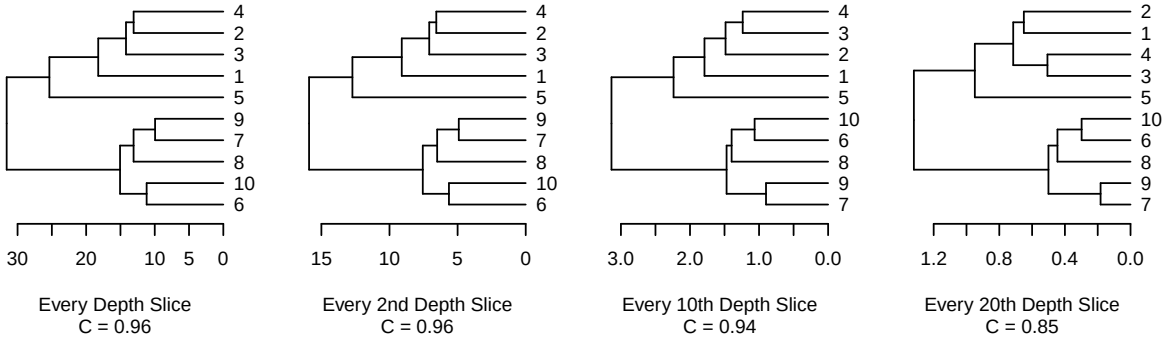


Figure 8: Effect of adjusting the `sample_interval` parameter from 1 to 20 on soil profile grouping, when $k = 0.01$, and `max_d` = 100 cm. Cophenetic correlation coefficients are printed below options. "Average linkage" agglomerative clustering was used to build dendrograms.

For massive collections of soil profiles the `sample_interval` argument to `profile_compare()` can be

used to reduce memory consumption by computing pair-wise dissimilarities every n slices. For example, the comparison of 1,000 soil profiles, each with 5 variables, to a maximum depth of 50 cm requires 192.3 Mb of RAM for the storage of the entire dissimilarity matrix (all depth slices) and takes about 70 seconds to perform (1.3 Ghz Intel CPU). Computing dissimilarity values every 5 slices reduces memory consumption to 1 fifth the original size (38.5 Mb) and processing time by a factor of about 3 (22 seconds). Within the sample dataset `sp3`, larger `sample_interval` values result in lower total dissimilarity values, minor differences in grouping structure, and minor reduction in cophenetic correlation coefficients; up to a sampling interval of every 10th slice (Figure 8). However, the specific threshold defining a reasonable trade-off between computational efficiency and preservation of detail will depend on the input dataset, available computing resources, and the purpose of the analysis. An optimized version of `profile_compare()` that uses file-based storage for the collection of dissimilarity matrices is currently in development.

Selection of variables included in the dissimilarity calculation, dissimilarity metric, depth-weighting coefficient, replacement of NULL distances, and grouping criteria all affect the output of this algorithm– and require further evaluation. Ideally, variables should be selected to accomodate the type of grouping that is most appropriate for the task at hand. For example, a classification reflecting basic parameters of soil development could be built from physical and chemical properties (particle size, pH, CEC, %BS, soil color, etc.) whereas a classification geared towards soil management decisions could be built from other properties (horizon morphology, bulk density, soil depth, etc.).

3 XRD Full-Pattern Matching

X-Ray diffraction (XRD) has contributed more to the current understanding of phyllosilicate minerals in soil systems than any other single method of analysis (Whittig and Allardice, 1986). Crystal structure and composition for common species of clay minerals have largely been derived from extrapolated XRD studies, as clay minerals are usually too small for single-crystal methods (Hughes et al., 1994). The platy structure and common stacking along with c-dimension make it possible to identify clay minerals via changes in measured (00l) diffraction peaks according to a battery of treatments that affect interlayer spacing (Hughes et al., 1994). Peak position, intensity and width are generally modified by a combination of sequential ion-exchange reactions, possibly followed with heat treatments. For example, it is possible to differentiate between clay minerals at the subgroup level (i.e. smectites from serpentines from kaolins) through changes in interlayer spacing associated with 1) solvation with Mg, 2) solvation with Mg and glycerol, 3) solvation with K, and 5) solvation with K followed by heating to 550 degrees C for 2 hours (Whittig and Allardice, 1986). Differentiation between dioctahedral vs. trioctahedral species is generally much more difficult, and requires alternative solvation/heat treatments such as Green-Kelly test (Greene-Kelly, 1953), interpretation of (060) diffraction peaks, or full-pattern quantitative XRD (QXRD) methods.

There are several methods that have been used for quantitative interpretation of X-ray diffraction patterns, and can be split into two major categories: 1) peak-matching, and 2) full pattern matching (Bish, 1994). The reference intensity ratio (RIR) method is based on normalization of one or more peaks of a mineral phase with respect to the intensity (or integrated intensity) of an internal standard. A known quantity of corundum is typically added to the sample, in a 1:1 ratio, and used as the internal standard. Replicated analysis and the use of several peaks can increase the accuracy of the RIR method (Bish, 1994). Full pattern matching methods have become more common with the advent of digital X-ray diffractometers and high-speed computers, and can overcome many of the problems associated with peak-based identification (Hughes et al., 1994). Of these methods, Rietveld refinement and Observed Patterns are two of the most widely used. Rietveld refinement is based on fundamental physical laws that describe the interaction of radiation with crystal structures, and relies on a detailed crystallographic data on all phases present in a sample. Therefore application of this method is much more difficult when working

with substances containing large quantities of amorphous, poorly crystalline, or micro-crystalline (i.e. clay mineral) phases (Bish, 1994).

The Observed Patterns method is based on empirical mixture modeling, rather than fundamental principals, and thus requires less detailed characterization of the sample (Bish, 1994). A simulated pattern is fit to the unknown pattern (all previously normalized to an internal standard) based on iteratively estimating the fractions of phases that are present within the unknown sample. Proportions are estimated by minimizing the sum of absolute differences between the unknown pattern and the composite pattern at each 2θ value:

$$D(2\theta) = \sum |I_u(2\theta) - W_p * I_p(2\theta)| \tag{9}$$

where $I_u(2\theta)$ is the intensity of the unknown sample at each 2θ value, W_p is the proportion of phase p, and $I_p(2\theta)$ is the intensity of phase p at each 2θ value. This approach requires knowledge of the main mineral phases within the unknown, pure standards of those phases, and that all analysis be conducted under the same operating conditions. If pure standards cannot be obtained, it is possible to simulate patterns for those phases from the Powder Diffraction File. The accuracy of this method can be greatly improved when pure standards are available that closely match the crystal size and composition of the phases present in the unknown sample.

4 Example Usage of the aqp Package

4.1 Profile Visualization, Classification, and Aggregation

The **aqp** package ships with several example soil profile datasets, collected from the Sierra Foothill and Sacramento Valley regions. The `sp3` dataset is based on a collection of 10 soil profiles from 3 major geologic groups (metavolcanic rocks, metasedimentary rocks, and granodiorite), representative of the Sierra Foothill Region. These data contain field-measured color values (Munsell notation) along with lab-measured clay content, cation exchange capacity (CEC), pH, total carbon, and analytically-measured color values (Soil Survey Staff, 2004). The variability of soil properties in this region is largely controlled by the type of underlying bedrock. Finer-textured, redder soils are usually found on metavolcanic rocks, whereas coarser-textured, yellow to gray colored soils are found on granitic rocks. Soils formed on metasedimentary rocks generally resemble soils formed on metavolcanic rocks, but variability in the type of rock and degree of metamorphism can result in drastically different soil properties. The following case study demonstrates how functions from the **aqp** package can be used to *numerically* describe: 1) differences between soil profiles, 2) dissimilarity-based group membership, and 3) aggregated soil property information defined by these groups.

Surrogate horizon names based on the clay content, cation exchange capacity (CEC), and pH of each horizon, are generated to facilitate interpretation of profile classification. Next field-measured colors are converted to RGB triplets for visualization with the `munsell2rgb()` function. Missing values, illogical combinations, or Munsell values not matched by rows in the look-up table result in no data (NA).

Between-profile dissimilarity is computed with the `profile_compare()` function using clay content, cation exchange capacity (CEC), and pH values, to a maximum depth of 100 cm, and using a depth-weighting coefficient of 0.01. Divisive hierarchical clustering (`diana()` function from the **cluster** package) is used to group soil profiles into a dendrogram for visualization (Kaufman and Rousseeuw, 2005; Maechler et al., 2005). The output from `diana()` is converted into an **ape** class object, and ladderized (Paradis et al., 2004). Divisive clustering was used as it most closely resembles the top-down approach that a soil scientist would (usually) take when sorting soils: i.e. splitting an initial collection of individuals into subsequently smaller and smaller groups. Finally, the `sp3` dataframe is converted into a `SoilProfileList` class object.

Listing 1: Setup the environment and load an example dataset.

```
# setup environment
library(aqp) ; data(sp3)
# generate surrogate horizon names from clay / CEC / pH
sp3$name <- paste(round(sp3$clay), '/', round(sp3$cec), '/', round(sp3$ph,1))
# color conversion
sp3$soil_color <- with(sp3, munsell2rgb(hue, value, chroma))
```

Listing 2: Compute between-profile dissimilarity and build a dendrogram from the result using divisive hierarchical clustering.

```
# load required libraries
require(ape) ; require(cluster)
# perform comparison of profiles
d <- profile_compare(sp3, vars=c('clay','cec','ph'), max_d=100, k=0.01)
h <- diana(d)
p <- ladderize(as.phylo(as.hclust(h)))
# convert soil data into ProfileList object for plotting
sp3.list <- initProfileList(sp3)
```

A new plot of the dendrogram is generated with the standard plot method for **ape** class objects; adjustments are made in order to accommodate sketches of the soil profiles below (Figure 9). Information on the ordering of soil profiles is extracted from the special `last_plot.phylo` object, and used to position profile sketches below corresponding terminal nodes of the dendrogram. Finally, soil profile sketches are generated by the `profile_plot()` function, applied to a `SoilProfileList` class object (Figure 9). If so desired, alternative depth-function plots could be inserted below their corresponding "leaves" of the dendrogram; i.e. particle size information, principal component scores, etc.

The results of this numerical classification (Figure 9) match field observation of soil properties, and expected differences between major lithologic types. Profiles 1-4 were collected from soils formed on metavolcanic rocks of varying iron content; with higher clay and pH values found on rocks with the highest iron content (profiles 2-4). Profile 5 was collected from a soil formed on metasedimentary rock, with high clay content and much lower pH values. Profiles 6-10 were collected from soils with low clay content and slightly higher pH values formed on granodiorite. Slightly higher clay contents and an increasing pH depth-function differentiate profiles 7-9 (swale position) from profiles 6 & 10 (backslope position). General patterns in soil color mirror the 3 groups identified within the clustering: deep red colors found in group 1 (high-iron soils from metavolcanic rocks) and group 2 (metasedimentary rocks), gray to brown colors found in the swale position of group 3, and the lighter, more yellow colors found on the backslope position (Figure 9).

According to branching within the dendrogram (Figure 9), the metasedimentary-soil appears to be most similar to the metavolcanic-soil group. Inspection of the dissimilarity matrix reveals that this soil is approximately 31% similar to the soils of the metavolcanic group and only 9% similar to the soils of the granodiorite group.

Next, depth-slice aggregation of cec and clay values is performed by calling the `soil.slot()` function for each of the three major groups identified via cluster analysis. Depth-slice aggregation of pH values is applied to groups defined by cutting the dendrogram at a lower level, such that the granodiorite group is split according to hillslope position (Figure 9). The `ddply()` function (**plyr** package) is simplest to use, however the `by()` and `do.call()` functions could have been used as well. Visualization of the depth-wise

Listing 3: Plot the dendrogram with soil profile sketches below.

```
par(mar=c(1,1,1,1))
p.plot <- plot(p, cex=0.8, label.offset=-3, direction='up', y.lim=c(60,-2),
x.lim=c(1,sp3.list$num_profiles+1), show.tip.label=FALSE)
tiplabels(col=c(1,2,4)[cutree(as.hclust(p), 3)],
pch=c(15,15,15,16)[cutree(as.hclust(p), 4)], cex=2)
# get the last plot geometry
lastPP <- get("last_plot.phylo", envir = .PlotPhyloEnv)
# vector of indices for plotting soil profiles below leaves of dendrogram
new_order <- sapply(1:lastPP$Ntip,
function(i) which(as.integer(lastPP$xx[1:lastPP$Ntip]) == i))
# plot the profiles, in the ordering defined by the dendrogram
# with a couple fudge factors to make them fit
profile_plot(sp3.list, color="soil_color", plot.order=new_order,
scaling.factor=0.3, width=0.1, cex.names=0.65,
y.offset=max(lastPP$yy)+8, add=TRUE)
# add a legend
legend(0.4, -2, legend=c('metavolcanic rocks', 'metasedimentary rocks',
'granodiorite: backslope', 'granodiorite: swale'),
col=c(1,2,4,4), pch=c(15,15,15,16), bty='n', cex=1)
```

Listing 4: Compute a normalized similarity between a single profile and all others within the collection.

```
# get groups from above and leave out soil number 5
groups <- factor(cutree(as.hclust(p), 3)[-5],
labels=c('metavolcanic','granodiorite'))
# using dissimilarity matrix from above,
# subset soil number 5 vs. all others
d.5 <- as.matrix(d)[5, -5]
# normalized similarity = 1 - ( dissimilarity / max(dissimilarity) )
1 - round(tapply(d.5, groups, mean) / max(d), 2)
# metavolcanic granodiorite
# 0.31 0.09
```

trends and uncertainty (+/- 1 standard deviation) is performed with the custom lattice panel function `panel.depth_function()` (Figure 10).

Aggregation of soil profile information gives an indication of group-wise central tendency and an empirical estimate of variability (Figure 10). Clay content (Figures 10(a)) and CEC values (Figures 10(b)) are highest within the metavolcanic-soils, with a marked but highly variable increase at 60-80 cm in depth. CEC values are lowest in the granitic-soils and show very low variability with depth. The metasedimentary-soil group lies closer to the metavolcanic-soils, and additional observations (required to compute depth-wise variability) would assist with further, interpretation. Visualization of aggregate soils information can also aid interpretation of the results from the previous classification. Of the three characteristics supplied to the `profile_compare()` function (clay content, CEC, and pH), the distribution of cec values and clay content with depth appears to be the most important factor contributing to differences between groups (Figures 10(a) and 10(b)). Diverging pH depth trends (Figure 10(c)) differentiate the two sub-groups identified within the granitic-soils (backslope vs. swale hillslope position).

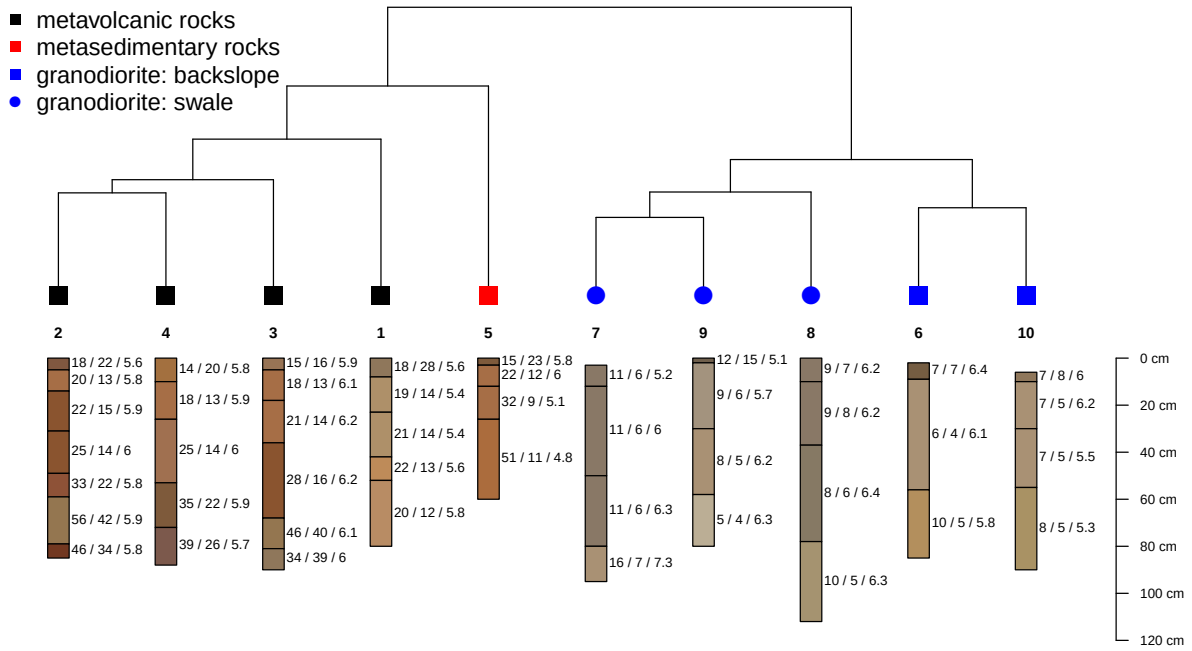


Figure 9: Divisive hierarchical clustering of soil profiles from the `sp3` sample dataset. Tip colors represent group membership defined by cutting the dendrogram into three classes, and labeled according to underlying rock type. Horizon names have been substituted with: "clay / CEC / pH".

4.2 Estimation of Proportions in a Composite XRD Pattern

pending

5 Concluding Remarks

The examples presented in the previous sections represent only a handful of the functions within the **aqp** package. Several additional functions are included that can be used to format and display depth slices of soils information according to spatial coordinates. A `random_profile()` function is included to simulate soil profile data, for the development and testing of aggregation and classification algorithms. The bundled documentation includes extensive, annotated examples based on three sample soils datasets. Examples presented in this chapter were based on a small number of soil profiles for clarity. However, functions in the **aqp** package have been successfully applied to studies involving several thousand soil profiles. Stable versions of the **aqp** are hosted on CRAN[2], and the active development version of the **aqp** package will continue to be hosted on R-Forge[3].

6 Acknowledgments

Several portions of this research was funded by the Kearney Foundation of Soil Science. I would like to thank Dr. Brent Myers for providing thoughtful commentary on several of the ideas presented in this

[2]http://cran.r-project.org/web/packages/aqp/
[3]http://aqp.r-forge.r-project.org/

Listing 5: Aggregate soil property information according to groups identified through divisive hierarchical clustering.

```
require(plyr) ; require(lattice)
# note that this example only illustrates a single iteration of the steps outlined above
# split data into 3 major classes (following rock type)
g <- factor(cutree(as.hclust(p), 3), labels=c('metavolcanic rocks',
'metasedimentary rocks', 'granodiorite'))
g <- data.frame(group=g, id=factor(names(g)))
# combine groups with original dataframe
sp3.new <- merge(sp3, g, by='id')
sp3.new$prop <- sp3.new$cec
# perform aggregation, by group
a <- ddply(sp3.new, .(group), .fun=soil.slot)
# manually add mean +/- SD to the result
a$upper <- with(a, p.mean+p.sd)
a$lower <- with(a, p.mean-p.sd)
```

Listing 6: Plot aggregate soil property data. Note that the following code listing corresponds to Figure 10(b)

```
# use custom plotting function for uncertainty viz.
xyplot(
top ~ p.mean, data=a, groups=group, subscripts=TRUE,
lower=a$lower, upper=a$upper, ylim=c(100,-5), alpha=0.3,
ylab='Depth (cm)', xlab='CEC (cmol(+) / kg soil)',
panel=panel.depth_function,
auto.key=list(lines=TRUE, points=FALSE, columns=2,
title='Soil Profile Group', cex=0.75, size=4, between=1),
par.settings=list(superpose.line=list(col=c(1,2,4), lty=1))
)
```

Listing 7: Estimation of proportions used to generate a composite XRD pattern.

```
# setup environment
library(aqp) ; data(rruff.sample)
# get number of measurements
n <- nrow(rruff.sample)
# number of components
n.components <- 6
# mineral fractions, normally we don't know these
w <- c(0.387, 0.232, 0.153, 0.096, 0.049, 0.065)
# make synthetic combined pattern
rruff.sample$synthetic_pat <- apply(sweep(rruff.sample[,2:7], 2, w, '*'), 1, sum)
# estimate proportions, f.noise is the objective function
o <- optim(par=c(0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1), f.noise,
method='CG', pure.patterns=rruff.sample[,2:7],
sample.pattern=rruff.sample$synthetic_pat)
```

paper.

References

Arkley, R. J., 1976. Statistical methods in soil classification research. p. 37–69. *In* N. C. Brady (ed.) Advances in Agronomy. Academic Press, New York, NY.

Bish, D., 1994. Quantitative x-ray diffraction analysis of soil. sssa miscellaneous publication 9, p. 267–295. *In* J. Amonette and L. Zelazny (ed.) Quantitative Methods in Soil Mineralogy. Soil Science Society of America, Inc.

Buol, S. W., R. C. Graham, P. A. McDaniel, and R. J. Southard, 2003. Soil Genesis and Classification. 5th ed. Iowa State Press, Ames, IA.

Carré, F. and M. Jacobson, 2009. Numerical classification of soil profile data using distance metrics. Geoderma 148:336–345.

Dale, M., A. McBratney, and J. Russell, 1989. On the role of expert systems and numerical taxonomy in soil classification. European Journal of Soil Science 40:223–234.

Gower, J. C., 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–871. ISSN 0006341X.

Greene-Kelly, R., 1953. The identification of montmorillonite in clays. Journal o f Soil Science 4:233–237.

Hillel, D., 1998. Environmental Soil Physics. Academic Press.

Huggett, R., 1975. Soil landscape systems: A model of soil genesis. Geoderma 13:1 – 22. ISSN 0016-7061.

Hughes, R., D. More, and H. Glass, 1994. Qualitative and quantitative analysis of clay minerals in soils. p. 330–359. *In* J. Amonette and L. Zelazny (ed.) Quantitative X-Ray Diffraction Analysis of Soil. Soil Science Society of America.

Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill, New York.

Kaufman, L. and P. J. Rousseeuw, 2005. Finding Groups in Data An Introduction to Cluster Analysis. Wiley-Interscience.

Lindbloom, B. J., 2010. Useful color equations.

Little, I. and D. Ross, 1985. The levenshtein metric, a new means for soil classification tested by data from a sand-podzol chronosequence and evaluated by discriminant function analysis. Australian Journal of Soil Research 23:115–130.

Maechler, M., P. Rousseeuw, A. Struyf, and M. Hubert, 2005. Cluster analysis basics and extensions. See the 'Changelog' file (in the package source).

Minasny, B. and A. B. McBratney, 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142:285–293.

Moore, A. and J. Russell, 1967. Comparison of coefficients and grouping procedures in numerical analysis of soil trace element data. Geoderma 1:139–158.

Moore, A., J. Russell, and W. Ward, 1972. Numerical analysis of soils: A comparison of three soil profile models with field classification. Journal of Soil Science 23:194–209.

Munsell Color Science Laboratory, 2010. Munsell renotation data.

Paradis, E., J. Claude, and K. Strimmer, 2004. Ape: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rayner, J., 1966. Classification of soils by numerical methods. Journal of Soil Science 17:79–92.

Russell, J. and A. Moore, 1968. Comparison of different depth weightings in the numerical analysis of anisotropic soil profile data. *In* Proceedings of the 9th International Congress of Soil Science, volume 4, p. 205–213.

Schlesinger, W. H., 1997. Biogeochemistry: An Analysis of Global Change. 2nd ed. Academic Press, New York, NY.

Sneath, P. H. A. and R. R. Sokal, 1973. Numerical Taxonomy. W.H. Freeman and Company. Examples related to soils on page 439.

Soil Survey Division Staff, 1993. Soil Survey Manual. U.S. Department of Agriculture Handbook 18. United States Department of Agriculture.

Soil Survey Staff, 1999. Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys. Number 436 *In* Agricultural Handbook. USDA.

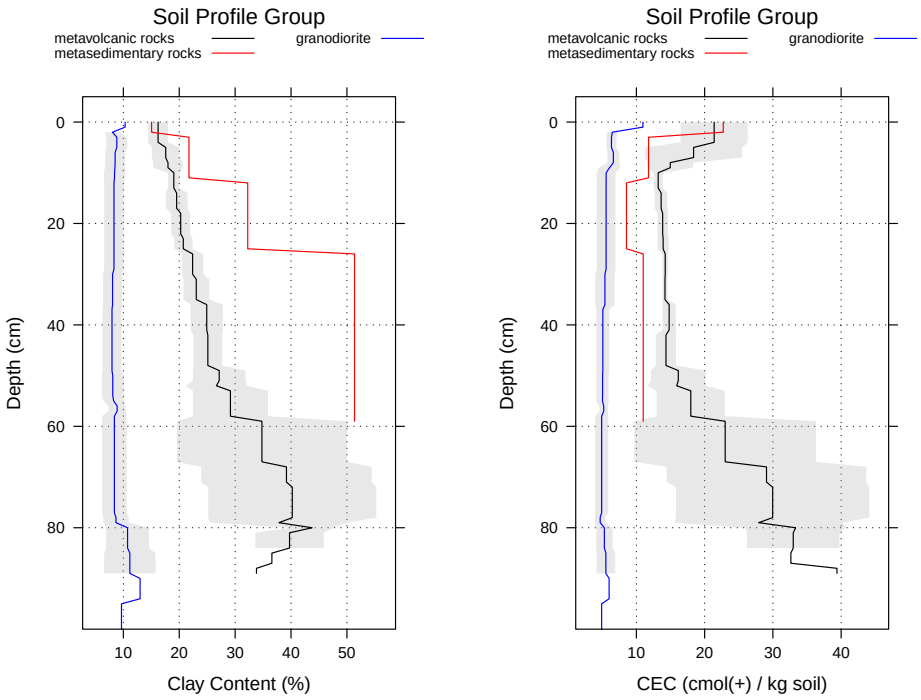
Soil Survey Staff, 2004. Soil Survey Laboratory Methods Manual. 4th ed. Number 42 *In* Soil Survey Investigations Report. USDA-NRCS, Washington, D.C.

Webster, R., 1968. Fundamental objections to the 7th approximation. Journal of Soil Science 19:354–366.

Webster, R. and M. Oliver, 1990. Statistical Methods in Soil and Land Resource Survey. Oxford University Press.

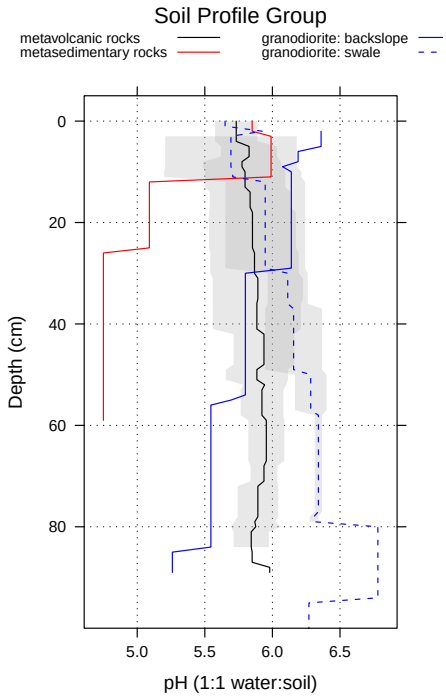
Whittig, L. D. and W. R. Allardice, 1986. Principles of x-ray diffraction. *In* Methods of Soil Analysis, Part 1. Physical and Mineralogical Methods. American Society of Agronomy-Soil Science Society of America, 2nd ed.

Young, F. and R. Hammer, 2000. Defining Geographic Soil Bodies by Landscape Position, Soil Taxonomy, and Cluster Analysis. Soil Sci Soc Am J 64:989–998.



(a)

(b)



(c)

Figure 10: Depth-slice aggregation of clay content (a), cation exchange capacity (b) and pH (c) based groups identified via cluster analysis. Lines are mean values, shaded area represents the mean $\pm$ 1 standard deviation.