Algorithms for Quantitative Pedology: A Toolkit for Soil Scientists

D.E. Beaudette

Dept. Land, Air and Water Resources

University of California, Davis

May 19, 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 other parameters within the "clorpt" framework are: *cl*
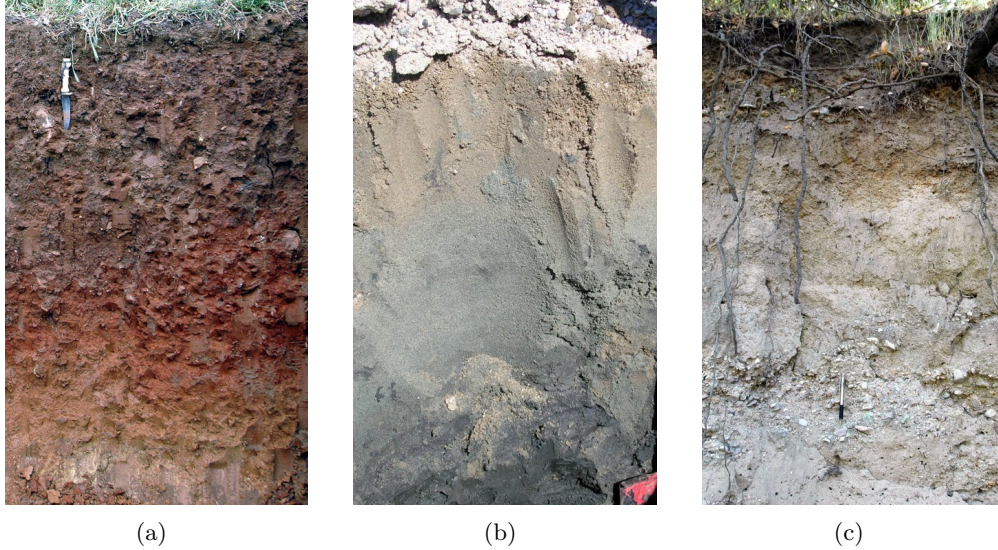
Figure 1: Examples of soil profiles illustrating how horizons change with depth. Color, texture, structure and root abundance are common visual indicators of near surface processes in soil.

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), and several authors have suggested the potential merit to such an approach (Webster, 1968; Arkley, 1976; Minasny and McBratney, 2007; Carré and Jacobson, 2009). However, these methods are rarely employed outside of case studies presented within scientific journals.

2   The aqp Package

The **aqp** (Algorithms for Quantitative Pedology) package for **R** 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, and most functions use basic dataframes as input, where rows represent soil horizons and columns define properties of those horizons. Common to most functions are the requirements that horizon boundaries are defined as depth from 0, and that profiles are uniquely defined by an *id* column. The **aqp** package defines two 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.
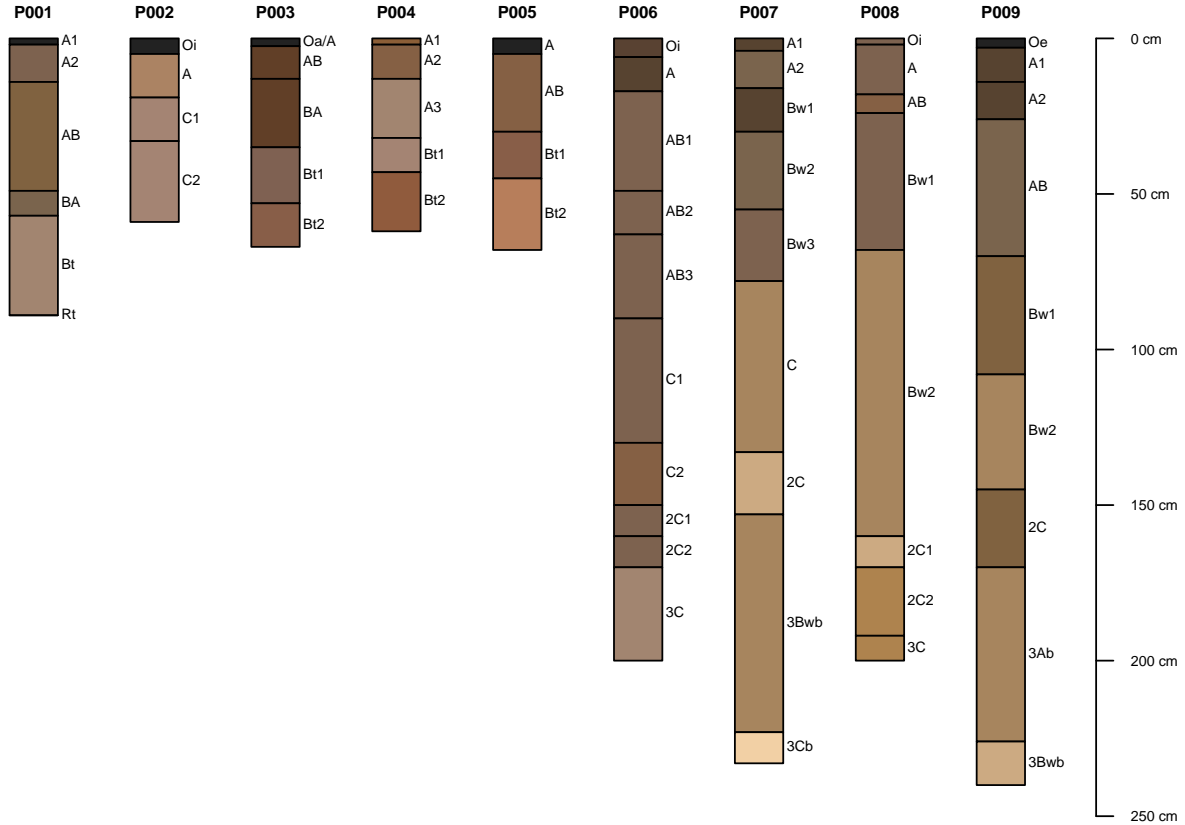

Figure 2: Visualization of nine soil profiles, colored by RGB representations of field-described, dry colors.

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 §3).

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[1] and color conversion equations[2]. The 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 D65[3] 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).
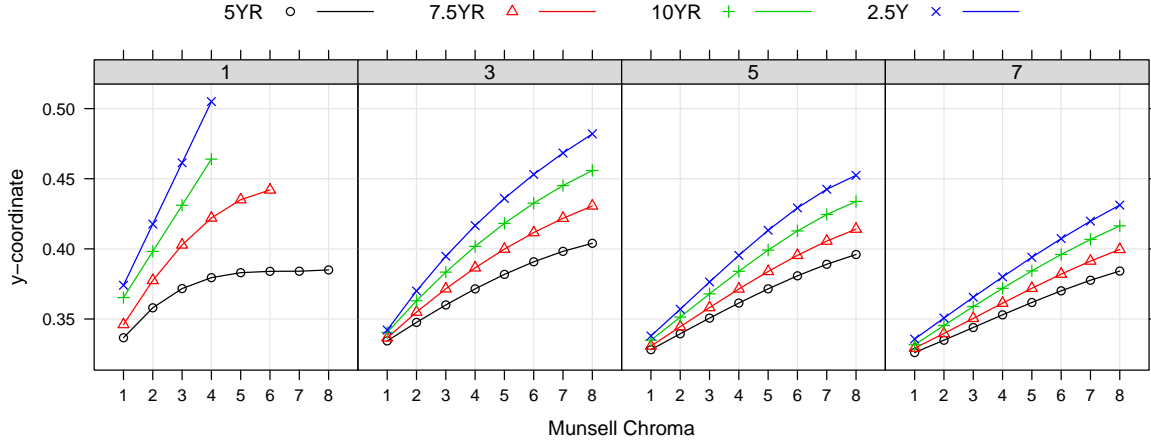

Figure 3: Relationship between Munsell chroma and y-coordinate (xyY colorspace), for selected Munsell hue (defined by line color and point symbol) and Munsell value (panels). Points at even-numbered Munsell chroma values were derived from the MCSL spectral database. Points at odd-numbered Munsell chroma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

$$X = \frac{xY}{y}$$
$$Y = Y$$
$$Z = \frac{Y(1 - x - y)}{y}$$
(2)

---

[1]http://www.cis.rit.edu/mcsl/online/munsell.php

[2]http://www.brucelindbloom.com/index.html?Math.html

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

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

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

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

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: 1) require single estimates of a given soil property at sampling locations in space, or, 2) 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).
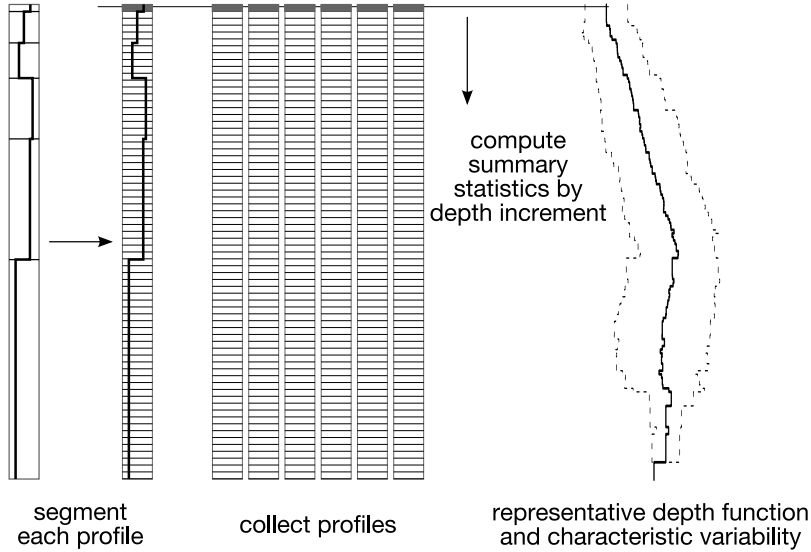

Figure 4: Profile aggregation algorithm.

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 (i.e. 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 (mean, median, SD, IQR, etc.) are computed along segments within the collection of profiles. The resulting estimate of central tendency and spread around that tendency for each segment 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 (i.e. presence/absence of diagnostic feature), and soil depth probability (i.e. probability that the soil profile ends at a given depth).

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.

2.3 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 a pair-wise dissimilarity metric 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 (i.e. every slice, every other slice, or every n slices), forming an $n_{profiles} \times n_{properties}$ dissimilarity matrix for each depth slice (Figure 5(a)).

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). Internally the `daisy()` function (**cluster** package) is used to compute Gower's generalized dissimilarity metric between samples (Maechler et al., 2005). This metric accepts both continuous and categorical variables, and can accommodate missing observations (Kaufman and Rousseeuw, 2005). A `max_d` parameter limits the maximum depth to which between-profile dissimilarity is computed.

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: $w = e^{-k \times d}$. 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 30cm, 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
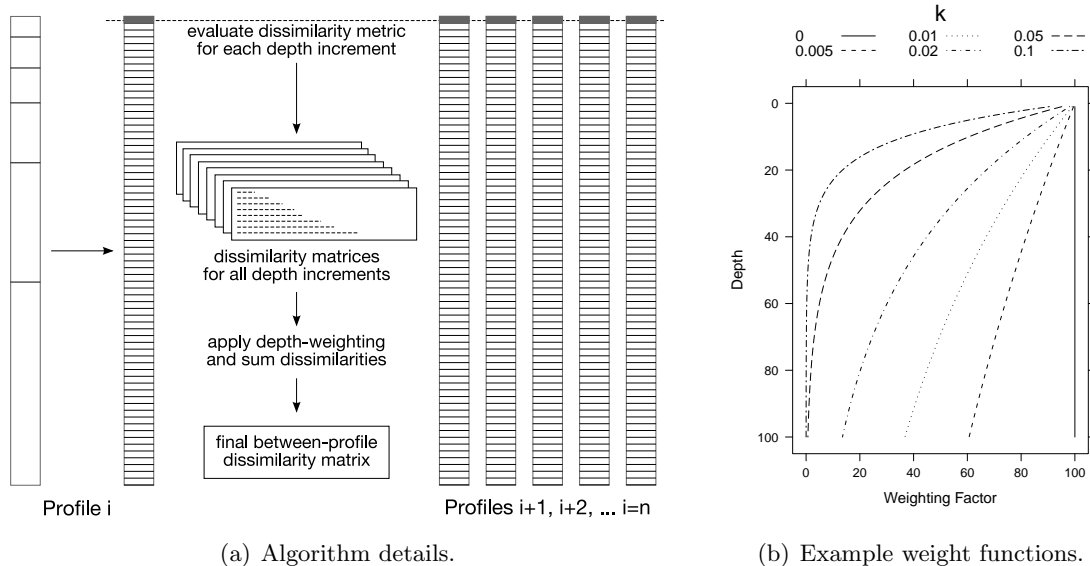
(a) Algorithm details.

(b) Example weight functions.

Figure 5: Calculation of pair-wise profile dissimilarity.

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).
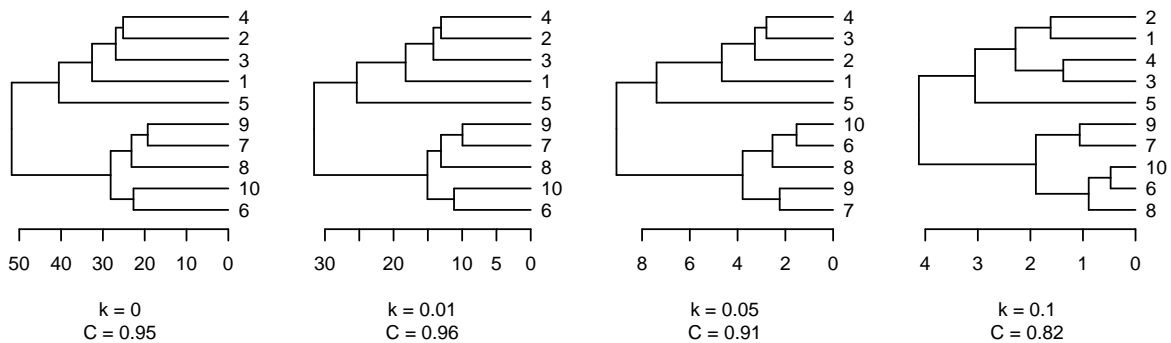

Figure 6: Effect of adjusting the depth weighting (k) parameter from 0 to 0.1 on soil profile grouping. Cophenetic correlation coefficients are printed below k values. "Average linkage" agglomerative clustering was used to build dendrograms.

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 7). However, the specific threshold defining a reasonable trade-off between
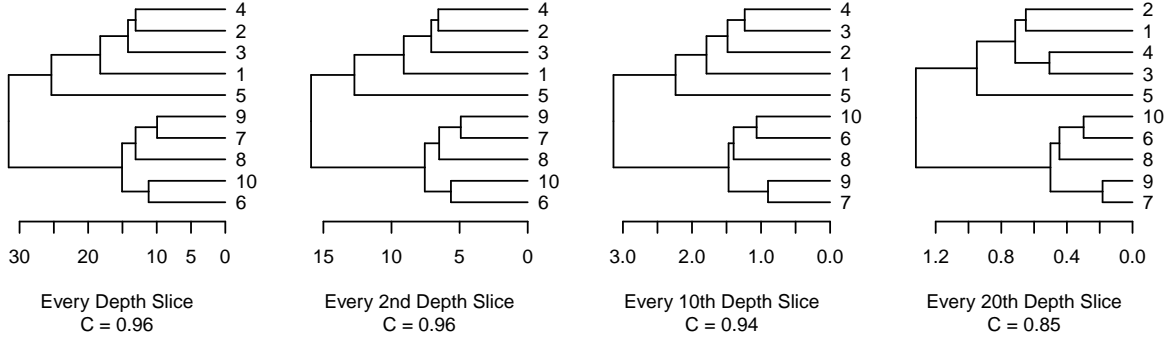
Figure 7: Effect of adjusting the `sample_interval` parameter from 1 to 20 on soil profile grouping, when k = 0.01, and `max_d` = 100 cm. "Average linkage" agglomerative clustering was used to build dendrograms.

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 $n_{profiles} \times n_{properties} \times n_{depth\ slices}$ list of dissimilarity matrices is currently in development.

3   Case Study

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).

```
# 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))
```

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

8

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.

```
# 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 8). 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 8). 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.

```
par(mar=c(1,1,1,1))
p.plot <- plot(p, cex=0.8, label.offset=-3, direction='down', y.lim=c(80,0),
x.lim=c(1,sp3.list$num_profiles+1), show.tip.label=FALSE)
tiplabels(col=c(1,2,4)[cutree(as.hclust(p), 3)],
pch=c(15,15,15,16)[cutree(as.hclust(p), 4)], cex=2)
# get the last plot geometry
lastPP <- get("last_plot.phylo", envir = .PlotPhyloEnv)
# the original labels, and new (indexed) order of pedons in dendrogram
d.labels <- attr(d, 'Labels')
new_order <- sapply(1:lastPP$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=1.2)
```

The results of this numerical classification (Figure 8) 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
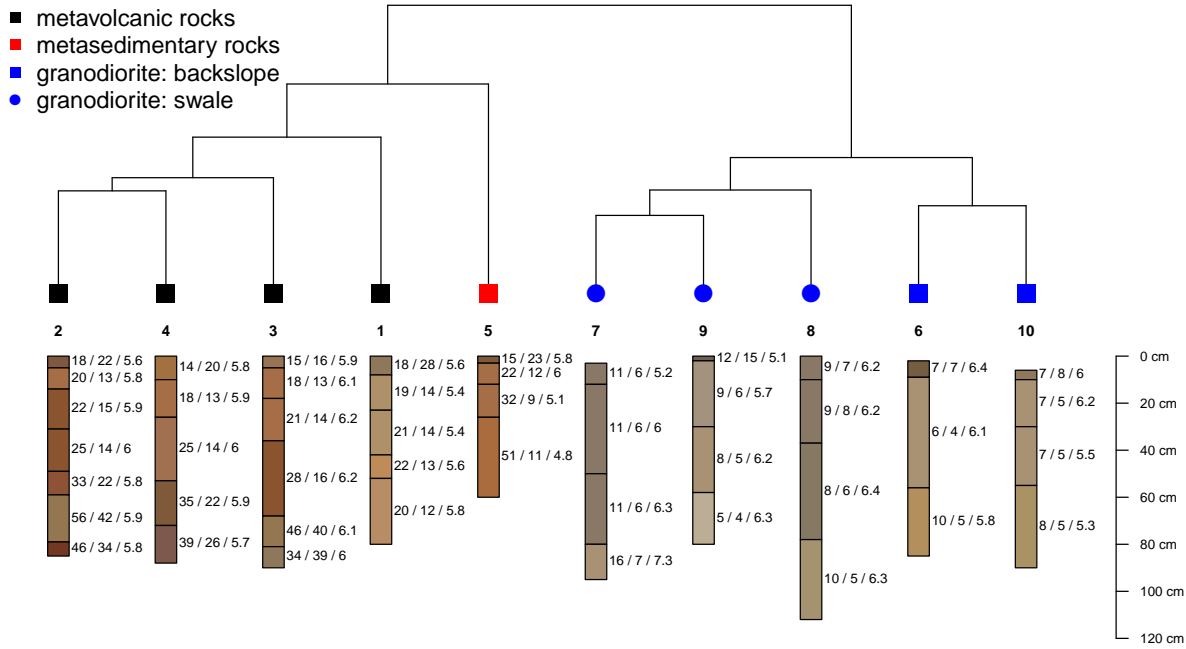
Figure 8: Divisive hierarchical clustering of soil profiles from the `sp3` sample dataset. Tip colors represent group membership defined by cutting the dendrogram into three classes, and labeled according to underlying rock type. Horizon names have been substituted with: "clay / CEC / pH".

(Figure 8).

    According to branching within the dendrogram (Figure 8), 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.

```
# 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
```

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 8). 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 trends and uncertainty (+/- 1 standard deviation) is performed with the custom lattice panel function `panel.depth_function()` (Figure 9). Note that the following code listing corresponds to Figure 9(b).

```
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)
```

```
# 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))
)
```
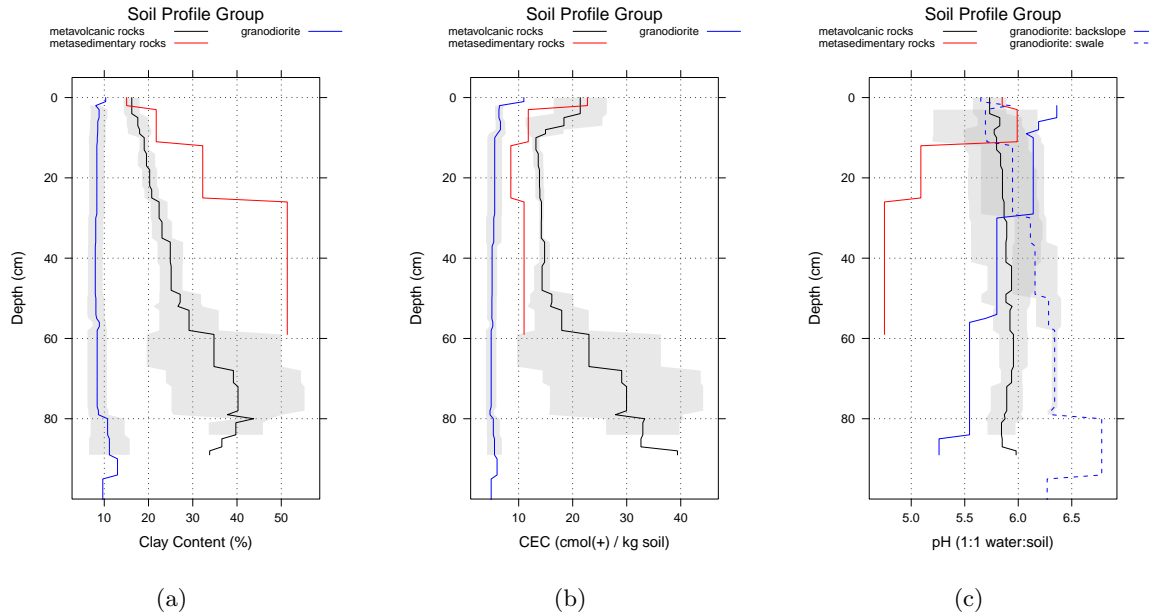

Figure 9: Depth-slice aggregation of clay content (a), cation exchange capacity (b) and pH (c) based groups identified via cluster analysis. Lines are mean values, shaded area represents the mean ± 1 standard deviation.

Aggregation of soil profile information gives an indication of group-wise central tendency and an empirical estimate of variability (Figure 9). Clay content (Figures 9(a)) and CEC values (Figures 9(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 9(a) and 9(b)). Diverging pH depth trends (Figure 9(c)) differentiate the two sub-groups identified within the granitic-soils (backslope vs. swale hillslope position).

4   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 paper 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. Pending submission to CRAN, the active development version of the **aqp** package will be hosted on R-Forge (http://aqp.r-forge.r-project.org/).

5   Acknowledgments

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.

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.

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

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

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.

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.

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

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.

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.