Algorithms for quantitative pedology: A toolkit for soil scientists

D.E. Beaudette a,*, P. Roudier c, A.T. O’Geen b

a USDA-NRCS, 19777 Greenley Road Sonora, CA 95222, USA
b Department of Land, Air and Water Resources, University of California, One Shields Ave Davis, CA 95616, USA
c Landscape Research, Private Bag 11052, Manawatu Mail Centre, Palmerston North 4442, New Zealand

ABSTRACT

Soils are routinely sampled and characterized according to genetic horizons, resulting in data that are associated with principle dimensions: location (x, y), depth (z), and property space (p). The high dimensionality and grouped nature of this type of data can complicate standard analysis, summarization, and visualization. The “aqp” (algorithms for quantitative pedology) package was designed to support data-driven approaches to common soils-related tasks such as visualization, aggregation, and classification of soil profile collections. In addition, we sought to advance the study of numerical soil classification by building on previously published methods within an extensible and open source framework. Functions in the aqp package have been successfully applied to studies involving several thousand soil profiles. The stable version of the aqp package is hosted by CRAN (http://cran.r-project.org/web/packages/aqp), and the development version is hosted by R-Forge (http://aqp.r-forge.r-project.org).

1. Introduction

A staggering quantity of soils information has been collected to support soil survey operations, natural resource inventories, and research over the last 100 years. Integrated analysis of large soil profile collections, however, is hampered by the multivariate nature of these data, changes in soil classification over time, and difficulties associated with processing horizon data that vary widely in depth and thickness. In addition, the magnitude of properties, correlation between properties, and trends with depth each affect how a soil profile is interpreted as a whole (Arkley, 1976). Many new users of soil survey data (climate change, hydrologic, ecologic, and atmospheric modelers) have noted that there is a need for basic statistical measures (mean, variance, confidence intervals, etc.) on reported soil properties (Webster and Beckett, 1968; Brown, 1988; Brown and Huddleston, 1991). Also, these users typically require soil property information at several levels of generalization, from landscape to regional scales (Heuvelink and Pebesma, 1999; Rasmussen, 2006; Schoorl and Veldkamp, 2006). We have developed a package for the general purpose analytic software, R (R Development Core Team, 2011), that supports the analysis of massive soil databases. The “aqp” (algorithms for quantitative pedology) package was designed to support data-driven approaches to common soils-related tasks such as visualization, aggregation, and classification of soil profile collections. In addition, we sought to advance the study of numerical soil classification by building on previously published methods within an extensible and open source framework. This paper describes the design and implementation of algorithms used within the package.

Stratigraphy and morphology of horizons are usually the first data used to infer dominant pedogenic processes within the profile: i.e. degree of alteration relative to the parent material, expression of oxidized or reduced forms of iron, accumulation of organic matter, or evidence of cyclical deposition of new material. Soil profiles are usually described, sampled, and characterized by genetic horizons 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). While the investigation of soil profile characteristics and horizon-level morphology are strongly based on visual and tactile cues, communication of these data are typically delivered via written narrative or tabular form—complicating integration or further meta-analysis. Soil profile sketches aligned with landscape, parent material, or vegetation gradients are commonly used as the foundation for the development of soil–landscape models, and ultimately take the form of block diagrams in soil survey documents (Fig. 1). Despite artistic merits, the reliance on hand-drawn soil profile sketches highlights a missing component in the soil scientist’s digital toolkit. An automated approach to the creation, alignment (e.g. along environmental...
designations, sequences, boundaries, and soil texture classes are usually sufficient for describing complex soil–landscape relationships.

Significant research has been conducted on numerical systems of approach for numerically genesis. However, these systems do not currently define an established land-use limitations and our current knowledge of soil cation systems (Buol et al., 2003) provide a rich vocabulary for tations of categorical soil properties (e.g. horizon designation or etc.) or absence of horizons at some locations. In addition, there range in horizon designations (A, B, C, etc.), depths, thicknesses, properties within a collection of soils. For example, the collection property. However, these approaches generally cannot adequately sampling locations in space or site-wide estimates of a given soil that require either: single estimates of a given soil property at division of sampling locations in space or site-wide estimates of a given soil property. However, these approaches generally cannot adequately provide aggregate representations of vertical variation in soil properties within a collection of soils. For example, the collection of soils associated with a given region is likely to include a wide range in horizon designations (A, B, C, etc.), depths, thicknesses, and horizon sequences. Consequently, computing an aggregate representation of some property with depth (e.g. clay content) can be confounded by variable thickness of horizon types (A, B, C, etc.) or absence of horizons at some locations. In addition, there are no established methods for describing aggregate representations of categorical soil properties (e.g. horizon designation or structure) and how they vary with depth.

Soil Taxonomy (Soil Survey Staff, 1999) and other soil classification systems (Buol et al., 2003) provide 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, these systems do not currently define an approach for numerically describing the difference between soils. Significant research has been conducted on numerical systems of soil classification over the last 40 years (Rayner, 1966; Fitzpatrick, 1967; Moore and Russell, 1967; Moore et al., 1972; Little and Ross, 1985; Carré and Jacobson, 2009); however, most of these methods are rarely employed outside of case studies within scientific literature. Despite this fact, several authors have suggested the potential merit of numerical soil classification (Webster, 1968; Arkley, 1976; Dale et al., 1989; Minasny and McBratney, 2007). In particular, Young and Hammer (2000) suggested that fine-scale soil variability is more adequately captured by numerical classification as opposed to Soil Taxonomy. Digital soil mapping efforts may be improved through an incorporation of between-sample taxonomic proximity (via numerical classification) into the development of statistical models (Minasny and McBratney, 2007).

2. Design

Methods described in this paper are available in the aqp (Algorithms for Quantitative Pedology) package for R (R Development Core Team, 2011). This package is free of charge and distributed with source code for full customization. Bundled documentation includes extensive, annotated examples based on several (included) sample soils datasets that are typical of both field and laboratory characterization. Functions provided by the aqp package expect rectangular data tables, where rows represent soil horizons and columns define properties of those horizons. Rows associated with each profile must be identified by a unique profile ID. Most functions 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 bedrock contact or to a conventionally used lower boundary (e.g. 150 cm). Specialized data types (classes) are included to support the multivariate hierarchy of linked spatial data (e.g. coordinates), site data (e.g. landscape position), diagnostic horizon data (e.g. argillic horizon), and horizon data (e.g. clay content). These data types facilitate a variety of analysis such as, “what is the site-wide median change in clay content from 10 cm to 30 cm?” or “what is the range of spatial auto-correlation for carbon content, at a depth of 5 cm?”.

In addition, these structures make it possible to rapidly generate maps of soil properties along regular depth intervals. A detailed description of the SoilProfileCollection class and associated methods can be found in the aqp package manual.

Examples presented in this paper are based on a small number of soil profiles for clarity: 15 soils formed on granodiorite and 72 soils formed on metavolcanic rocks from the Sierra Foothill Region of California. Functions in the aqp package have been successfully applied to studies involving several thousand soil profiles. The stable version of the aqp package is hosted by CRAN (http://cran.r-project.org/web/packages/aqp), and the active development version of the aqp package is hosted by R-Forge (http://aqp.r-forge.r-project.org). Users interested in the most well-tested and documented code are advised to use the stable version of the package, while users interested in the most recently implemented functions can try them out in the development version. A vignette is included with the package source code, containing a mini-tutorial for all of the major functionality within the aqp package.

2.1. Visualizing soil profile data

2.1.1. Color conversion

Soil colors are typically measured with the Munsell system (hue, value, chroma), therefore conversion to an RGB (red, green, blue) colorspace is required for display on most computer monitors. We developed a function to convert soil colors in
Munsell notation (hue, value, chroma) into RGB triplets, based on a look-up table of common soil colors. The look-up table was generated from the Munsell Color Science Laboratory (MCSL) spectral database of Munsell chips (Munsell Color Science Laboratory, 2010) and color conversion equations (Lindbloom, 2010). The MCSL spectral database contains xyY color space 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 color space coordinates along the entire range of chroma defined for each Munsell hue and value, via spline interpolation. The conversion from xyY coordinates to RGB coordinates was performed with the following four steps: conversion from xyY to XYZ coordinates (Eq. (1)), chromatic adaption transformation from the C to D65\(^{1}\) illuminant (Eq. (2)), conversion from XYZ (D65 illuminant) to rgb (Eq. (3)), and scaling of rgb values to conform to a specific \(\gamma\)-correction\(^{2}\) (Eq. (4)).

\[
X = \frac{xY}{y} \\
Y = Y \\
Z = \frac{Y(1-x-y)}{y} \\
\]

(1)

\[
[X_{D65}, Y_{D65}, Z_{D65}] = [X, Y, Z] \times \begin{bmatrix} 0.990448 & 0.0012371 & 0.003564 \\ -0.007168 & 1.015594 & 0.006770 \\ -0.011615 & -0.002928 & 0.918157 \end{bmatrix} \\
\]

(2)

\[
[r, g, b] = [X_{D65}, Y_{D65}, Z_{D65}] \times \begin{bmatrix} 3.24071 & -0.969258 & 0.0556352 \\ -1.53726 & 1.87599 & -0.203996 \\ -0.498571 & 0.0415557 & 1.05707 \end{bmatrix} \\
\]

(3)

\[
R, G, B = \begin{cases} 
12.92 \times (r, g, b) & : r, g, b \leq 0.0031308 \\
1.055 \times (r, g, b)^{1/2.4} - 0.055 & : r, g, b > 0.0031308 
\end{cases} \\
\]

(4)

\(^{1}\) Most R plotting functions, and computer monitors in general, use the sRGB color profile which assumes a D65 illuminant.

\(^{2}\) Gamma (\(\gamma\)) is the scaling parameter used in a non-linear transformation of color values into an approximately linear representation used by digital devices.

An \textit{rgb} equivalent for soil colors described in the Munsell system makes it possible to graphically depict as a function of horizon designation, depth, and color within printed documents, integrated into online resources, and smartphone devices (Beaudette and O’Geen, 2009). Since ambient lighting, monitor calibration, and different printing methods (liquid ink, laser, offset, etc.) all affect the perception of color, \textit{rgb} representation of soil color should only be used to illustrate relative differences in color.

### 2.1.2. Profile sketches

The \textit{aqp} package provides a function for rendering soil profiles, based on basic stratigraphic parameters: horizon boundaries, horizon designation, and soil color. Horizon level information including designation, texture, structure, pH, etc. can be displayed beside horizons. A convenient coordinate system is used for referring to soil profile sketches, where \(x\) coordinates are based on an integer indexing of profiles, and \(y\) coordinates are referenced to depth from the soil surface (Fig. 2). Profile sorting can be performed using an integer index reflecting ranks derived from aggregate soils information (e.g. change in clay content with depth), landscape parameters (e.g. landscape position or surface curvature), taxonomic relationships, etc. This index describes the order in which soil profile sketches are added to the figure. This functionality combined with \textit{R}’s plotting and layout capabilities make it possible to quickly generate complex diagrams that describe entire collections of soil profiles.

Fig. 2 depicts 15 soils formed on granodiorite ordered according to an index of relative hillslope position. From this figure it is possible to rapidly extract several patterns: summit positions were generally shallow and had redder soil colors (indicative of iron oxide accumulation) as compared to darker colors (indicative of organic matter accumulation) found in swales (Fig. 2). Soils “08” and “11” were sampled within dioritic pockets and were distinctly deeper, redder, and finer in texture than other summit and backslope soils formed on granodiorite. Soil “06” was distinctly redder and finer in texture than other soils found in the swale position, and likely formed from dioritic material transported from upslope. Soils “13” and “14”, sampled on an especially silica-rich granite vein, were considerably shallower and coarser in texture. A similar diagram was generated from 15 soils...
formed on metavolcanic rocks, with profiles sorted according to change in clay content with depth (Fig. 3). From this diagram it is possible to explore the magnitude of abrupt textural changes that frequently occur just above bedrock contact in these landscapes.

2.2. Realignment of soil horizons into depth slices

Soil property data sampled from genetic horizons are difficult to process due to variable horization. This problem can be solved by normalizing a collection of horizons, irrespective of the horizon type, according to a common system of "slices". Essentially, each soil property (from each soil profile) is aligned to a common depth basis. With this new data structure it is possible to plot, aggregate, map, or compute numerical measures of similarity by slice. This approach relies on the assumption that sample depth, not provenance (i.e. genetic horizon type), is a logical basis for between-sample comparisons.

2.2.1. Aggregation by depth slice

The slice-wise aggregation algorithm in the aqp package is based on the premise that a "representative depth function" for some soil property (e.g. clay content) can be generated from a collection of soil profiles by summarizing this property along depth slices (Fig. 4). Depth slices are based on 1 unit intervals ranging from the minimum to maximum profile depth. Alternatively, slices can be defined by regularly spaced (e.g. 10-unit) intervals, or by customized boundaries (e.g. 0, 10, 25, 50, 150). Each profile in the collection is first segmented into 1-unit slices. Then, summary statistics are computed along slices across the collection of profiles.

Slices (s) associated with a given soil property are grouped into a matrix, where rows represent a slice of values across the collection of profiles (si) at depth interval i, and columns represent the sequence of values (sj) associated with profile j. Therefore, computation of the aggregate value by slice (\( \tau_i \)) can be symbolized as

\[
\begin{array}{cccccc}
S_{1,1} & S_{1,2} & S_{1,3} & \ldots & S_{1,j} \\
S_{2,1} & S_{2,2} & S_{2,3} & \ldots & S_{2,j} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
S_{i,1} & S_{i,2} & S_{i,3} & \ldots & S_{i,j} \\
\end{array}
\]

\[ \tau_i = f(s_i) \]

where \( f \) is a scalar-returning function applied row-wise, resulting in a new column vector of slice-wise summaries. The algorithm currently supports calculation of mean, standard deviation, quantiles, or evaluation of a user-defined function. An estimate of central tendency and spread around that tendency for each depth slice is reconstituted as a single "representative depth function" (Fig. 5). This approach is similar to the methods described in Lentz and Simonson (1987a,b) for defining "conceptual soil profiles".

Studies on the vertical variation in soil properties are common for cases in which the property in question is continuous in nature (e.g. clay content). However, there are few cases where the vertical variability in categorical properties (e.g. horizonation) has been described (Vanwalleghem et al., 2010). Aggregate depth functions for categorical properties can be modeled with probabilities, one for each class of the categorical variable, and along each depth slice. Probabilities for each class \( k \) of a categorical variable are computed by slice:

\[
\hat{s}_{k,i} = \frac{\text{freq}(s_{k,i})}{j}
\]

where freq(s_{k,i}) is the frequency of class \( k \) along slice \( i \), and \( j \) is either the number of profiles contributing to the calculation at slice \( i \) or optionally, the total number of profiles within the collection. Depth-slice probabilities generated from major horizon types can quantitatively describe site-wide patterns in soil
morphology, such as A horizon thickness, presence or absence of transitional horizons, and depth to paralithic or bedrock contact (Fig. 6).

Profile weights can be included in the estimation of weighted mean, standard deviation, and quantiles. In this case, weights should be assigned according to some logical criteria: e.g. area associated with each profile, field insight suggesting the relative dominance of a given profile, or perhaps number of supplementary auger observations that are similar to a given profile. Weights need not be rescaled to any particular range, nor sum to any given value; weights are automatically normalized within the algorithm. An area-weighted aggregation of soil profile information (e.g. from a soil survey) by major land-use, physiographic province, or along a regular grid could serve as an effective means of communicating regional-scale patterns of soil properties and their intrinsic variability.

The algorithm automatically determines an appropriate type of aggregation (continuous or categorical) based on the data type of its input. Statistical measures for central tendency and spread (mean, standard deviation, median, quantiles) are computed for continuous variables, and group probabilities are computed for categorical variables. In addition, a "contributing fraction" value is returned for each depth slice, describing the number of soil profiles contributing to computations.
profiles that were used in the computation. The “contributing fraction” value can be interpreted as an aggregate measure of soil depth probability—if soil profiles were described consistently to either a constant depth or limiting horizon.

Fig. 5(a) depicts clay content and total carbon values (measured from samples collected by genetic horizon) summarized along 1-cm depth slices to demonstrate how the slice-wise aggregation algorithm can be used. From this data display it is possible to interpret several key characteristics of these soil profiles. Clay content ranges from about 15% to 25% near the surface, increases linearly until about 60 cm, and approaches a collection-wide mean value of about 38% (Fig. 5(a)). Variability in clay content is smallest near the surface and generally increases with depth.

The aggregate trend in total carbon with depth follows the typical exponential decay-shaped curve; however, a zone of higher variability is apparent between 10 and 30 cm (Fig. 5(b)). This zone tracks the described range in depth of transitional (BA) horizons at this site (Fig. 6(b)), and is probably related to a combination of bioturbation and colluvial processes. Total carbon values outside of the range defined by the slice-wise weighted mean ± 1 standard deviation (on a log scale) indicate samples that were contaminated by either ground rock material (e.g. values lower than expected) or by organic material within root channels (e.g. values higher than expected).

Fig. 6 demonstrates how this slice-wise aggregation algorithm can be applied to categorical data types. In this case, horizon designations from soil profiles collected at two different sites were aggregated along 1-cm depth slices to produce site-wide depictions of horizonation. Contributing fraction values (percentages along right-hand y-axis) can be interpreted as aggregate measures of soil depth probability. Lower contributing fraction values suggest that there is a lower probability of encountering soil material at that depth, as long as soils within the collection have been consistently described to either bedrock contact or to a common depth. Likewise, aggregate descriptions of horizon designation can be interpreted as the probability of encountering a specific horizon at some depth; in this case, the largest class-wise probability (x-axis) at any depth is the most likely horizon type (A, AB, Bt, etc.).

Several features can be efficiently communicated from this procedure and resulting graphic. For example, on soils formed from granodiorite: (1) A horizons typically occurred from 0–14 cm, sometimes followed by transitional (AB) horizons from 14–30 cm; (2) Bw and Bt horizons typically occurred from 30–80 cm; and, (3) C horizons and Cr contact occurred over 80–120 cm (Fig. 6(a)). Similarly, soils formed from metavolcanic rocks typically had A horizons from 0–8 cm, BA horizons from 8–20 cm, Bt horizons from 20–85+ cm, and Cr or R contact near 85 cm (Fig. 6(b)). The larger number of profiles sampled from metavolcanic terrain resulted in smoother estimates of slice-wise horizon probability estimates as compared to the probabilities computed from soils sampled on granodiorite.

2.2.2. Limitations of the algorithm

Care should be taken when interpreting the output from the profile aggregation algorithm described above where either: (1) several heterogeneous groups of soil profiles have been aggregated together, or, (2) soil depth varies greatly within the collection of soil profiles. In the first case, aggregation will result in depth functions that are numerically correct, but in no way representative of the individual soils from 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 any given depth slice (Fig. 6). Inspection of this value, along with an evaluation of spread relative to central tendency (e.g. coefficient of variation) can help determine when aggregate depth functions may not be sufficiently representative. As with any interpretation of data summaries, expert judgment should always be applied when drawing inference from a sample.

2.3. Numerical classification of soil profiles

A numerical comparison of soil profiles is complicated by several factors: (1) soil profiles represent a collection of horizons that must be treated as a group; (2) soils are generally sampled according to genetic horizons, but horizon type and thickness can vary widely; and (3) the dissimilarity metric and selection of attributes used to compute between-sample dissimilarity will affect the classification. Rayner (1966) presented an approach for computing average dissimilarity between pairs of profiles where calculation of between-profile dissimilarity was based on those horizons that were most similar and occurred at or below similar depths. While this method integrates the concept of a soil profile as a collection of horizons into the dissimilarity calculation, it does not account for differences in horizon thickness or soil depth. Three alternate approaches were later evaluated by Moore et al. (1972) where between-profile distances were derived from either: (1) coefficients of orthogonal polynomials fit to soil property depth functions; (2) profile sums of depth-wise dissimilarity; or, (3) global clustering of horizon samples (into a small number of “horizon types”) followed by comparison of transition matrices. Little and Ross (1985) suggested a more efficient approach for comparing profiles, after horizon samples were clustered into nominal “horizon types”, using the Levenshtein metric. Considerations on the number and weighting of variables used within dissimilarity calculations, along with a correlation-based criteria for variable reduction, were presented by Sarkar et al. (1966). Recent work by Carré and Jacobson (2009) demonstrated a fundamentally different approach to numerical soil classification based on the k-means partitioning algorithm (Venables and Ripley, 2002) using either Euclidean or Manhattan distance metrics. Many of the previously proposed numerical soil classification algorithms have two major limitations in common: (1) no attempt to account for comparison between soil and non-soil material (i.e. between a deep and shallow profile), and (2) lack of support for binary and nominal scale variables.

2.3.1. Pair-wise comparison by depth slice

One approach to numerical soil classification requires the calculation of a pair-wise dissimilarity between soil profiles. In the simplest case, horizon samples can be compared using one or more properties (e.g. clay content, pH, carbon content, etc.) according to the pair-wise Euclidean distance between observations in property space. Larger distance values are interpreted as greater dissimilarity. Since soil properties are rarely collected on comparable scales, each property would need to be centered (e.g. variable-wise subtraction of its mean value) and scaled (e.g. variable-wise division by its standard deviation) before an unbiased distance could be calculated. Euclidean distance metrics require scaling and standardization and do not readily accommodate binary or nominal scale variables. Gower’s generalized dissimilarity metric (Gower, 1971) is an ideal alternate metric for the comparison of soils data as variable standardization is automatic, and the metric accounts for any combination of binary, categorical, or continuous variables. In addition, this metric can
The vector of properties defined by a single slice forming a new matrix collection of property matrices ($w_i$) is an exponential decay function:

$$w_i = e^{-k \cdot i}$$

where $k$ and $i$ are $>0$. The decay rate parameter ($k$) determines how rapidly a slice's dissimilarity value is down-weighted with depth: a value of 0.1 would effectively remove any influence of dissimilarities computed below about 30 slices, and a value of 0 would weight all slices equally (Fig. 7(b)). The actual value for $k$ should be determined as objectively as possible; ideally with a combination of expert knowledge on vertical anisotropy in properties used within the classification and the depths over which relevant processes occur. The determination of an optimal $k$ value can be estimated by iterating over the range of $k$ and evaluating the quality (e.g. using a priori knowledge of groupings within the profile collection) of the resulting classification at each iteration. In the absence of expert knowledge, a $k$ value of 0 would be a good starting point.

Our numerical soil profile classification algorithm was repeatedly applied to a set of similar soils (backslope position on granodiorite terrain), using a depth-weighting coefficient ($k$) ranging from 0 (no depth weighting) to 0.1 (Fig. 8). Between-profile comparisons were based on clay content, silt content, very coarse sand content, in-transformed total carbon content, CEC, and Munsell chroma. Larger values of $k$ result in between-profile comparisons that are more greatly influenced by near-surface differences. The overall structure of between-profile differences was fairly consistent as $k$ was increased from 0 (no depth-weighting) to 0.01 (moderate depth-weighting), with minor changes as $k$ was increased to 0.05 (high depth-weighting) (Fig. 8). As $k$ was adjusted from 0.05 to 0.1, between-profile differences were dominated by those soil properties that vary near the soil surface. In this case, the final grouping structure ($k=0.1$) follows total carbon content. Profiles “01”, “09”, and “11” were sampled on south-facing backslopes (lowest carbon content); profiles “03”, “07”, and “15” were sampled on north-facing backslopes (higher carbon content); and profile “05” was sampled under an oak canopy (highest carbon content) (Fig. 8).

Variable soil depth can significantly interfere with the calculation of between-profile dissimilarity. For example, how should dissimilarity, evaluated between two profiles at a given depth, account for one of the profiles being shallower than the other? When soil depths have been arbitrarily defined (e.g. sampling to a common depth) a user-defined 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 physical difference between soil and non-soil material should be incorporated into final pair-wise dissimilarities. Any dissimilarity metric is accommodate limited occurrences of missing observations (Kaufman and Rousseeuw, 2005).

Since soil profiles are defined by an ordered (with depth) set of horizons, any numerical comparison must account for variation in horizon thickness and vertical changes in soil properties with depth (Webster and Oliver, 1990). Our approach builds on the 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 (Fig. 7(a)). The final between-profile dissimilarity is computed by summing the collection of slice-wise dissimilarity matrices. This is accomplished by forming a series of soil property matrices $P_j$:

$$P_j = \begin{bmatrix} x_{1,\text{clay}} & x_{1,\text{pH}} & x_{1,\text{TC}} & \cdots & x_{1,p} \\ x_{2,\text{clay}} & x_{2,\text{pH}} & x_{2,\text{TC}} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,\text{clay}} & x_{n,\text{pH}} & x_{n,\text{TC}} & \cdots & x_{n,p} \end{bmatrix}$$

representing “sliced” soil properties from profile $j$, where a single cell in this matrix $x_{ij}$ represents soil property $p$ from depth slice $i$. The vector of properties defined by a single slice $x_i$ taken from the collection of property matrices ($P_j$) are accumulated, row-wise, forming a new matrix $X_i$. In this matrix, rows represent profiles and columns represent properties. Pair-wise dissimilarity $D$ between profiles is computed as the sum of slice-wise dissimilarity:

$$D = \sum_{i=1}^{n} w_i G(X_i)$$

where $i$ is the slice index, $n$ is the total number of slices, $w_i$ is an optional weighting coefficient, and $G()$ is Gower's generalized dissimilarity metric. In this way, a collection of horizons with different vertical boundaries can be compared along a normalized basis (depth), with proper weighting according to horizon thickness.

### 2.3.2. Tuning parameters

In certain circumstances, specific sections of a profile (e.g. rooting zone) may be more relevant to a classification than others. In these cases, the matrix of between-profile dissimilarities can be weighted according to the depth of a given slice ($i$) via an exponential decay function:

$$w_i = e^{-k \cdot i}$$

Fig. 7. Calculation of pair-wise profile dissimilarity (a) and demonstration of the depth-weighting coefficient (b).
Effect of adjusting the depth weighting \((k)\) parameter from 0 to 0.1 on soil profile grouping, with comparisons limited to a depth of 100 cm. Dissimilarity matrices were each normalized to the range of \([0,100]\). Divisive hierarchical clustering of each normalized dissimilarity matrix was used to generate dendrograms. Numbers correspond to the IDs of soils collected from backslope positions on granodiorite terrain (Fig. 2).

![Dendrogram](image)

Fig. 8. Effect of adjusting the depth weighting \((k)\) parameter from 0 to 0.1 on soil profile grouping, with comparisons limited to a depth of 100 cm. Dissimilarity matrices were each normalized to the range of \([0,100]\). Divisive hierarchical clustering of each normalized dissimilarity matrix was used to generate dendrograms. Numbers correspond to the IDs of soils collected from backslope positions on granodiorite terrain (Fig. 2).

Effect of replacing undefined distances and preserving undefined distances between mutually shallow profiles on soil profile grouping. No depth weighting was used and comparisons were limited to 100 cm depth. Dissimilarity matrices were each normalized to the range of \([0,100]\). Divisive hierarchical clustering of each normalized dissimilarity matrix was used to generate dendrograms. Numbers correspond to the IDs of soils collected from backslope and summit positions from granodiorite terrain (Fig. 2). Shallow soils are marked with filled circle symbols.

![Dendrogram](image)

Fig. 9. Effect of replacing undefined distances and preserving undefined distances between mutually shallow profiles on soil profile grouping. No depth weighting was used and comparisons were limited to 100 cm depth. Dissimilarity matrices were each normalized to the range of \([0,100]\). Divisive hierarchical clustering of each normalized dissimilarity matrix was used to generate dendrograms. Numbers correspond to the IDs of soils collected from backslope and summit positions from granodiorite terrain (Fig. 2). Shallow soils are marked with filled circle symbols.

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 undefined). When summed, the total dissimilarity between these profiles will generally be much lower than if the profiles had been of equal depth. An example of this phenomena is demonstrated in the left panel of Fig. 9. Two shallow soils (profiles “13” and “14”) from summit positions (see Fig. 2) were incorporated into the previously described numerical classification performed without depth weighting (Fig. 8). The presence of undefined distances at depths below the lower boundaries of profiles “13” and “14” (approximately 40 cm) causes pair-wise dissimilarities between deeper profiles in the collection to be underestimated (Fig. 9(a)). Although these three profiles have similar properties near the surface, profile “11” is erroneously grouped with profiles “13” and “14” because it is much deeper, contains finer soil textures, and has redder soil colors (Fig. 2).

Our algorithm addresses this problem via replacement of undefined dissimilarities with the maximum between-slice dissimilarity (within the collection of profiles). 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). Replacement of undefined dissimilarities in this manner clearly separates the shallow summit soils (“13” and “14”) from the deeper backslope soils (Figs. 2 and 9(b)). However, this approach can result in a new problem: dissimilarities calculated between two shallow profiles will be erroneously inflated beyond the extent of either profile’s depth when deeper profiles exist in the collection. For example, shallow summit soils, although split from deeper backslope soils, are erroneously distant from each other (note high level of branching between profiles “13” and “14” in Fig. 9(b)). In order to address this problem, undefined distances are preserved between slices, when both slices represent non-soil material (Fig. 9). With this option enabled, shallow profiles only accumulate mutual dissimilarity to the depth of the deeper profile. When undefined distances are conditionally replaced, shallow soils are split from deeper soils without artificially inflating between-profile distances among shallow soils (Fig. 9(c)).

Selection of variables included in the dissimilarity calculation, dissimilarity metric, depth-weighting coefficient, replacement of undefined distances, and grouping criteria all affect the output of this algorithm—and require further evaluation. Ideally, variables should be selected to accommodate the type of grouping that is most appropriate for the task at hand. For example, a classification reflecting quantifiable parameters of soil development (such as those used by Soil Taxonomy) could be built from physical and chemical properties (particle size, pH, cation exchange capacity, base saturation, etc.) whereas a classification tailored for soil management decisions could be built from other properties (horizon thickness, bulk density, soil depth, consistency, redox, etc.).

2.3.3. Application to large collections of soil data

For massive collections of soil profiles, the sampling interval of slices can be used to reduce memory consumption by computing pair-wise dissimilarities every \(n\) slices. For example, the comparison of 1000 soil profiles, each with five variables, to a maximum depth of 100 cm requires 385 Mb of RAM for the storage of the entire dissimilarity matrix (all depth slices) and takes about 80 s to perform (3 GHz Intel (7 CPU). Computing dissimilarity values every five slices reduces memory consumption to 20% of the original size (77 Mb) and processing time to 27% of the original (22 s). 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 the algorithm that uses parallel computation (distributed across CPU cores) and
file-based storage for dissimilarity matrices is currently in development.

2.3.4. Comparison with soil taxonomy

Subgroup level classification (Soil Taxonomy) and our numerical soil classification algorithm (described in Section 2.3.1) were qualitatively compared using a combination of 15 soils sampled from the granodiorite landscapes of the Sierra Foothill Region of California and four soil series mapped in the area. These soil series included: Ahwahnee (Mollic Haploxeralfs) on stable summit positions, Tollhouse (Entic Haploxerolls) on summit positions with shallow depth to bedrock, Vista (Typic Haploxercepts) on the sideslope positions, and Chualar (Typic Argixerolls) in swales. Between-profile dissimilarities were computed using clay content, silt content, very coarse sand fraction, natural log transform of total carbon concentration, cation exchange capacity (CEC), and Munsell chroma (dry). Depth-weighting of dissimilarities was performed ($k=0.01$), slice-wise comparisons were made to a depth of 150 cm, and differences in soil depth were accounted for using conditional replacement of undefined dissimilarity. Dendrograms were created via divisive hierarchical clustering (Kaufman and Rousseeuw, 2005) (as opposed to alternatives such as agglomerative hierarchical clustering) because it most closely resembles the way in which most soil classification systems operate: soils are first organized into large initial groups which are subsequently split into smaller and smaller groups.

Taxonomic distances computed from subgroup membership were based on the number of matches at order, suborder, great group, and subgroup levels (per Gower's metric). This approach allows for the derivation of a quasi-numerical classification system from Soil Taxonomy, but is severely limited by the fact that each split in the hierarchy is given equal weight (Fig. 10(a)). In other words, the quasi-numerical dissimilarity associated with divergence at the soil order level is identical to that associated with divergence at the subgroup level. At the subgroup level, these soils (labeled “01”–“15”) fall into roughly four groups: Typic Haploxercepts (similar to the Vista series), Mollic Haploxeralfs (similar to the Ahwahnee series), Typic/Ultic Argixerolls (similar to the Chualar series), and Typic/Lithic Haploxercepts (similar to the Tollhouse series) (Fig. 10(a)).

Looking closer at the results from the numerical classification, it is possible to identify four major groupings (starting from the left-hand side of Fig. 11). Soils in group 1 were moderately deep, coarse-textured soils on backslopes, with the exception of soils “08” (summit position) and “02” (swale position) (Fig. 11). Soils “02” and “04” (collected from the swale position) were erroneously included in group 1, since they had not been described to either Cr or R contact. This fact highlights one of the most important drawbacks to a numerical classification system: between-profile dissimilarities can be easily skewed by missing data or inconsistent profile description style. Soils in group 2 were shallow to moderately deep, with predominately loamy coarse sand to coarse sandy loam textures, and found on summit positions (Fig. 11). The Tollhouse series was included in this group, which roughly follows taxonomic classes within this group (Typic Haploxeroloxes, Entic Haploxerolls, and Lithic Haploxerolls). The numerical dissimilarity between soils “13” and “14” (sampled on adjacent interfluve landscape positions) mirrors their similar morphology. Soils in group 3 were deep to very deep, moderately fine-textured soils sampled on north-facing backslope positions (“03”) and swale positions (“10” and “06”) (Fig. 11). The Chualar and Vista soil series were included in this group; however, within-group dissimilarity values were large as indicated by the height at which branches merge within this group. Soils in group 4 were moderately deep, with well-expressed argillic horizons (Fig. 11). The Ahwahnee soil series was included in this group even though it differs from soil “11” at the subgroup taxonomic level (Mollic Haploxeralfs vs Ultic Haploxerolls).

3. Concluding remarks

The aqp software package could be a valuable tool for the National Cooperative Soil Survey’s data collection and data delivery efforts. The visualization capabilities included in our package were designed specifically for summary of field mapping exercises. An automated approach to summarizing soil profile photos and profile sketches could serve as a useful tool for soil survey outreach and education. The use of aggregate depth functions could support a fundamental shift in how soil survey is presented: from the concept of a “modal profile” to a collection of “representative depth functions”. Representative soil property depth functions would give users a continuous estimate of soil properties (e.g. with depth) and fulfill a long-standing criticism of soil survey regarding the current lack of uncertainty estimates for soil property data. Also, the use of aggregate characterization (instead of a single pedon) should lead to more reliable series-
level classification within Soil Taxonomy (Allardice and Southard, 1991). The numerical soil classification system described in this paper could potentially be used to bridge editions of Soil Taxonomy or national soil taxonomic systems—based solely on soil physical and chemical properties. It would allow for a more flexible integration of “tax adjuncts” into existing classification systems by identifying possible taxa that are functionally similar. Alternative classification schemes could be generated from the same underlying data, but directed towards specific goals, by selecting which variables and dissimilarity metrics are used. In addition, numerical soil classification could be used to automate the process of normalizing soil series differences (i.e. map unit harmonization) that frequently occur at the boundaries of soil surveys of different vintages. The algorithms implemented in app provide a new suite of tools for modern challenges facing soil survey and scientists working digital soil profile databases.

Acknowledgments

Several portions of this research were funded by the Kearney Foundation of Soil Science. We would like to thank Dr. Brent Myers, Jay Skovlin, Dr. Randy Dahlgren, and Stephen Roecker for providing thoughtful commentary on several of the ideas presented in this paper.

References