Chapter 8
Soil Material Classes

Nathan P. Odgers and Alex. B. McBratney

"When you get close to the raw materials and taste them at the moment they let go of the soil, you learn to respect them".

Rene Redzepi

Soil classification is really about answering the question *what makes a soil?* Or, perhaps, *what makes one soil different from another?* To answer questions like these, soil classifiers create taxonomic rules to separate one kind of soil from another and categorise and make sense of the diverse pattern of the soil continuum. Traditionally a great deal of consideration has been given to characterising and classifying the whole soil profile in a top-down fashion. Pedometric methods allow us to answer the same questions in a bottom-up trajectory. Thus, the starting point is not the whole soil profile or even its major constituents, the soil horizons. Rather we start by classifying the actual, tangible, skeleton of soil itself: the *soil material*.

8.1 Soil Material Classes

We classify soil because we think the soil varies sufficiently in its properties from place to place for distinct kinds to be recognised. These various kinds of soil ought to be identifiable in the physical space (the natural landscape) and, hopefully, in the

N.P. Odgers (✉)
Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia

Soils and Landscapes Team, Manaaki Whenua – Landcare Research, PO Box 69040, Lincoln 7640, New Zealand
e-mail: OdgersN@landcareresearch.co.nz

A.B. McBratney
Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia
e-mail: alex.mcbratney@sydney.edu.au

A.B. McBratney et al. (eds.), *Pedometrics*, Progress in Soil Science,
DOI 10.1007/978-3-319-63439-5_8

attribute space—the *n*-dimensional space defined by *n* measurable soil attributes (MacVicar 1969). In classifying soils, we aim to use soil information gathered from the physical space to define soil classes that are as coherent as possible in the attribute space.

Taxonomists classify objects we may call *individuals*. An individual is an object that is "complete in itself" (Cline 1949). Individuals can be allocated to (can become members of) classes, and a *universe* is a superclass that contains all individuals. A *particulate* universe is one in which discrete objects can readily be counted. Biological universes are frequently particulate universes: for example, in the universe of birds, individual birds are readily discernible. On the other hand, a universe of soils does not seem to fit the description of a particulate universe since its characteristics frequently change from place to place more gradually than sharply (Simonson and Gardiner 1960). Discrete, definitive bodies of soil are not as discernible as individual organisms often are in a biological universe. Soil is, in other words, a continuum and therefore occupies a *continuous* universe in which individuals cannot readily be counted; individuals in a continuous universe must be created arbitrarily (Knox 1965).

Because soils occupy a continuous universe, the identification of soil individuals has been the source of some contention (Buol 2003). Yet as a practical matter, some notion of a soil individual has traditionally been necessary in order to discretise the soil continuum and simplify its classification.

8.1.1 Point Representation of Soil

The pedon is probably the most well-known soil individual. The pedon and the polypedon were devised as three-dimensional soil individuals for the purpose of soil sampling and classification, respectively (Soil Survey Staff 1993) and are fundamental concepts in the Soil Taxonomy classification system (Soil Survey Staff 1999). A polypedon is a collection of contiguous pedons of like taxa and is considered internally homogenous for taxonomic purposes.

Delineation of pedons in the field is complicated by the fact that soil occupies a continuous universe. For example, there is effectively an infinite number of pedons since their boundaries in the landscape must be established arbitrarily (Knox 1965). This means that their dimensions are also arbitrary, although guidelines dictate their minimum and maximum limits (Soil Survey Staff 1993). Several methods of setting the dimensions of a pedon have been proposed that are based on, for example, the volume of soil occupied by plant roots or the minimum volume that can be sampled by a particular instrument or the examination of lateral soil horizon variability (Simonson and Gardiner 1960; Soil Survey Staff 1993).

In reality what we tend to examine and classify in the field are two-dimensional soil profiles rather than three-dimensional soil bodies (Webster 1968). We often assume, implicitly, that the variation in the third dimension is irrelevant for taxonomic purposes because it is impractical in most circumstances to excavate and describe a three-dimensional body of sufficient size and in sufficient detail.

Since soil is a continuum, it seems appropriate that in our attempts to classify it we should not ignore the lateral variation when sampling it; however, it appears that little impetus is traditionally provided to collect such information. Since sampling and classification are linked, it suggests that a revised philosophy of the soil individual is necessary.

There is precedent in the literature. According to Holmgren (1986), the conceptualisation of a soil individual as an arbitrary volume is unsuitable because a soil property measurement is actually made on a volume of material originating *from a specific location on Earth*. Location is therefore a definitive "point of origin" with respect to which a soil can be characterised and classified.

It makes sense that the observations themselves should consist of a small volume representation, which we shall call a *soil material*. The soil material would logically fit inside recognised soil horizons and as such is fairly congruent with Holmgren's (1988) sampling *locule*. In geostatistical terminology, we are saying that a soil sample must have a defined geometric support. A soil material may or may not also be the same as a representative elementary volume (REV), which Bouma (1985) considered is the smallest volume that can represent a given soil horizon and lead to a consistent population of data. In aggregated soils, REVs are peds and in sandy soils they are individual sand grains.

The substance of Holmgren's revised pedon seems to be more related to the practice in soil physics, chemistry and biology. It may be aligned with the geostatistical and REV concept above although the space-time geometric support of this operationally defined object is not clear. A set of nine cores of fixed radius sampled on a grid 20 m apart and centred on some location (x, y) on which a set of measurements are made and attributes recorded according to some schema could be considered an *observational pedon*. The set of data thus obtained could be called a *pedon description*.

8.1.2 Soil Material as a Collection of Soil Properties

A large number of soil properties, for example, physical, geochemical and biological, can be attributed to or measured on a particular soil material. We shall call a collection of soil properties describing the soil material a *soil material description* (Table 8.1). The infrared spectrum offers another multivariate description (Fidêncio et al. 2001; Leone and Sommer 2000; Valeriano et al. 1995). The soil material is a real entity, whereas a soil material description is a virtual one. If different sets of properties are used to describe a soil material, then the soil material descriptions are different entities.

Table 8.1 Descriptions of soil material from three profiles observed at Pokolbin in the Lower Hunter Valley, New South Wales, Australia

Australian Soil Classification suborder	Brown Dermosol	Red Chromosol	Brown Kandosol
Horizon designation	A1 horizon	B21 horizon	B21 horizon
Texture grade	Clay loam	Light-medium clay	Light clay
Moist colour hue	10YR	5YR	7.5YR
Moist colour value	4	3	3
Moist colour chroma	3	4	4
Dry colour hue	10YR	5YR	7.5YR
Dry colour value	5	3	5
Dry colour chroma	4	4	4
Structure	Moderately pedal	Strongly pedal	Apedal massive
pH (1:5 H_2O)	5.67	6.74	5.53
pH (1:5 $CaCl_2$)	4.67	5.43	4.25
Electrical conductivity ($\mu S\ cm^{-1}$)	145.6	54.9	87.3

8.2 Creation of Classes

Scientific classifications of soil have been made since at least the late nineteenth century. These were often induced from supposed genesis. Later, more objective classification schemes were devised. The most widely known is probably today's Soil Taxonomy and its antecedent, the so-called Seventh Approximation. Most soil classifications in use today are hierarchical systems in which the soil universe is segregated into progressively more detailed, mutually exclusive classes as one proceeds down the hierarchy. This is the same model that has traditionally been applied in biology, but researchers have occasionally questioned its application to the classification of soils, in part because of the genetic assumptions implicit in the hierarchical structure (e.g. Leeper 1956).

Hughes and Lindley (1955) were some of the first researchers to classify soils using a statistical procedure although research into numerical soil classification of soils really began in earnest in the 1960s (Bidwell and Hole 1964a, b; Campbell et al. 1970; Grigal and Arneman 1969; Rayner 1966) based on work done in the biological sciences (Sneath and Sokal 1962; Sokal 1963).

Cluster analysis is the application of numerical methods to the classification of multivariate data. The goal of cluster analysis is to partition a population of individuals into classes by finding groups of similar individuals in the multivariate attribute space. The aim is that individuals allocated to the same class should be similar to each other and dissimilar to individuals allocated to the other classes (Fisher and van Ness 1971).

An individual being classified is known as a pattern, $\mathbf{x}$, and is denoted as a vector of d features or attributes x_i:

$$\mathbf{x} = (x_1, x_2, x_3 \ldots, x_d)$$

The vector $\mathbf{x}$ is usually assumed to be a row vector. A *pattern set* is a set of individuals, denoted $\boldsymbol{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ \ldots, \mathbf{x}_n\}$. In our case the attributes x_i are soil attributes. They may be measured on a continuous scale or coded on an ordinal or nominal scale. In practice, most researchers avoid the inclusion of attributes measured on ordinal and nominal scales because they can complicate the computation of pairwise similarities between individuals.

It is important to note that the clustering process is agnostic about the true partition structure amongst the population. Cluster algorithms generally partition the population of individuals into a set of classes even if natural classes are not present (Jain 2010).

8.2.1 Similarity Measures

Numerical classification relies on being able to assess the degree of similarity or resemblance between a pair of individuals. Measures that can do so are generically known as coefficients of similarity. Sneath and Sokal (1973) identified four classes of similarity coefficients: *distance coefficients*, *association coefficients*, *correlation coefficients* and *probabilistic similarity coefficients*. Association coefficients typically measure similarity on the basis of agreement in the state of qualitative attributes and include the Jaccard coefficient (Jaccard 1908) and the simple matching coefficient (Sokal and Michener 1958). Correlation coefficients measure similarity on the basis of proportionality and independence between a pair of attribute vectors and include the Pearson product-moment coefficient. Probabilistic similarity coefficients involve information statistics which measure the homogeneity of a given partition of individuals; an example can be found in Estabrook (1967).

Distance coefficients measure the distance between individuals in various ways. According to Sneath and Sokal (1973), they have the greatest intellectual appeal to taxonomists compared to other kinds of similarity coefficients since they are the easiest to visualise. They are the similarity coefficients used most frequently by pedometricians over the last few decades. Strictly speaking distance coefficients are measures of *dissimilarity* since their values increase with decreasing similarity (Clifford and Williams 1976). In the remainder of this section, we briefly describe several popular distance coefficients.

8.2.1.1 Euclidean Distance

The Euclidean distance is one of the most familiar measures of distance and was introduced to numerical taxonomy by Sokal (1961). In the two-dimensional case, it is equivalent to Pythagoras' theorem but can easily be extended to compute dissimilarities in higher-dimensional spaces. The Euclidean distance D_E^2 between individuals i and j is computed as

$$D_E^2\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sum\nolimits_{p=1}^{d} \left(\mathbf{x}_{ip} - \mathbf{x}_{jp}\right)^2 = \left(\mathbf{x}_i - \mathbf{x}_j\right)^T \left(\mathbf{x}_i - \mathbf{x}_j\right) \tag{8.1}$$

Several pre- or post-treatments can be applied. For example, since D_E^2 grows larger as the number of attributes increases, it is a common practice to divide it by d (Clifford and Williams 1976; Webster and Oliver 1990). In addition, measurements of soil attributes tend to have different natural numerical ranges, irrespective of units of measurement. For example, soil particle-size fractions expressed gravimetrically have a range of 0–100 g kg^{-1}, whereas soil organic carbon measurements in mineral soils are frequently in the range of 0–5%. This means that attributes with larger ranges tend to have greater influence in the calculation of the Euclidean distance (Clifford and Williams 1976; Jain et al. 1999; Kantardzic 2011; Moore and Russell 1967), which may be undesirable. To avoid this we can standardise each attribute by range prior to computation of the distance (Arkley 1976):

$$x' = (x - x_{\min}) / (x_{\max} - x_{\min}) \tag{8.2}$$

or by variance:

$$x' = (x - \bar{x}) / SD_x \tag{8.3}$$

where x' is the standardised value and $\bar{x}$, x_{min} and x_{max} are the mean, minimum and maximum values of the observed range of x.

Finally, distributions of soil attributes are often skewed. Moore and Russell (1967) noted that the Euclidean distance appears to be sensitive to the shape of attribute distributions. Skew may be reduced either by transformation of the values, typically using a logarithmic or square root transformation, or by truncating extreme but rarely occurring values (Arkley 1976).

8.2.1.2 Mahalanobis Distance

Unlike the Euclidean distance, the Mahalanobis distance (Mahalanobis 1936) takes into account the differences in variance between the soil attributes and the correlations between them. As a result the attributes do not need to be standardised prior to calculation of the distance. The Mahalanobis distance is calculated as follows (McBratney and de Gruijter 1992):

$$D_M^2\left(\mathrm{x}_i, \mathrm{x}_j\right) = \left(\mathrm{x}_i - \mathrm{x}_j\right)^{\mathrm{T}} \mathbf{E}^{-1} \left(\mathrm{x}_i - \mathrm{x}_j\right) \tag{8.4}$$

where $\mathbf{E}$ is the sample covariance matrix of $\mathbf{X}$.

8.2.1.3 Gower Distance

Both the Euclidean distance and Mahalanobis distance require that the attributes are quantitative continuous variables like clay content or electrical conductivity;

Table 8.2 Values assign to s_{ijp} and δ_{ijp} for dichotomous nominal variables (Gower 1971)

	Values of attribute p			
i	+	+	−	−
J	+	−	+	−
s_{ijp}	1	0	0	0
δ_{ijp}	1	1	1	0

however, many soil attributes are measured on a qualitative scale. There are several types of qualitative variables (Williams 1976a). The first kind are called nominal variables and represent an attribute that may take many states, such as type of aggregate coatings (clay, manganese, iron, organic, etc.). A given individual can be in one state only. Although we may encode the states on an integer scale for convenience, the ordering of the states implies no special meaning. A special case of nominal variable is what Gower (1971) refers to as a *dichotomous variable*. Dichotomous variables record the presence or absence of a feature, such as mottling or stones. Two individuals have a higher degree of similarity to each other if both are in possession of the same feature than if it is absent in one individual. Absence of a feature in both individuals does not infer the same degree of similarity as its presence in both individuals since it may not be known whether the feature can occur in the populations to which the individuals belong.

Ordinal variables are similar to nominal variables in that they represent an attribute that may take many states. In this case the order of the states when encoded into an integer scale is important although the distance between states may be unknown (Williams 1976a). An example of an ordinal-valued attribute is stone size classes (e.g. 2–6 mm, 6–20 mm, 20–50 mm, >50 mm).

Gower (1971) described a general coefficient of similarity that is able to handle qualitative and nominal quantitative attributes. It is calculated as follows:

$$S_{ij} = \sum_{p=1}^{d} s_{ijp}\delta_{ijp} / \sum_{p=1}^{d} \delta_{ijp} \tag{8.5}$$

where s_{ijp} is the similarity score between individuals i and j for attribute p. The quantity δ_{ijp} represents the possibility of making a comparison between individuals i and j for attribute p; it takes a value of 1 when attribute p can be compared and 0 otherwise. When all d comparisons are possible, $\sum_{p=1}^{d} \delta_{ijp} = d$.

The method of computing the similarity score varies depending on the kind of attribute. In the case of dichotomous nominal variables, s_{ijp} and δ_{ijp} are assigned the appropriate values in Table 8.2.

In the case of multistate nominal variables, $s_{ijp} = 1$ if individuals i and j have the same state for attribute p and $s_{ijp} = 0$ otherwise.

For quantitative variables, s_{ijp} is computed as follows:

$$s_{ijp} = 1 - \left| x_i - x_j \right| / R_p \tag{8.6}$$

where R_p is the numerical range of attribute p either in the population or the sample.

Despite its flexibility, the Gower distance has not been widely used in pedometric studies although examples exist in spatial prediction research (e.g. Mallavan et al. 2010; McKenzie and Austin 1993; Zhang et al. 2013) and soil classification research (e.g. Beaudette et al. 2013; Oliver and Webster 1989; Roudier et al. 2016).

8.2.2 Preprocessing

Before we undertake a cluster analysis, it is frequently necessary to perform some preprocessing operations on the attributes of the individuals we wish to cluster. We may wish to do so for a number of reasons, including the following: we would like to choose an optimal subset of attributes or we may want to modify the influence that certain attributes have on the cluster analysis. Preprocessing objectives may be realised by undertaking several common tasks including examination of the correlations between attributes, standardisation to a common range of values, transformation to reduce skew and determination of appropriate weights (Arkley 1976). The steps we need to take are frequently determined by the assumptions underlying a particular clustering algorithm. In this section we briefly describe a couple of the more common tasks.

8.2.2.1 Optimal Subset of Attributes

We may be presented with a large number of attributes and wish to select an optimal subset. It is tempting to merely choose the several that we think are the most important, or the most interpretable or that we have the most experience with, but such a choice is likely to be suboptimal with respect to the information about the individuals' partition structure that a given set of attributes carry (Arkley 1976). In any dataset of soil material attributes, it is likely that many will be correlated with each other, and some highly so. For the sake of cluster analysis, an optimal set of attributes is the set in which the correlations between the attributes are minimised. On the other hand, for maximum pedological interpretability, they should also be as correlated as possible with other attributes not used in the cluster analysis (Norris 1971). For quantitative attributes this can be achieved by examining the Pearson correlation matrix arising from a pairwise comparison of all available attributes and selecting a subset according to some heuristic. An example of such a matrix for attributes of some undisturbed soils in Israel is presented in Table 8.3 (after Banin and Amiel 1970). For example, Sarkar et al. (1966), after assembling a dataset of 61 soil attributes, examined pairs having a correlation greater than or equal to 0.90 and eliminated the attribute that was most highly correlated with other attributes. By doing so they reduced the number of attributes to 51. Through trial and error, they were able to make bigger reductions by lowering the pairwise comparison threshold.

A similar but more rigorous outcome can be achieved by subjecting a dataset containing n soil attributes to the well-known method of principal components

Table 8.3 Correlation coefficients for attributes of some undisturbed soils in Israel, after Banin and Amiel (1970)

	S	CEC	$CaCO_3$	OM	Clay	Silt	Sand	W_H	W_{15}	$W_{1/3}$	W_S	W_A
S												
CEC	0.930											
$CaCO_3$	−0.195	−0.160										
OM	0.313	0.067	−0.063									
Clay	0.950	0.869	−0.085	0.393								
Silt	0.206	0.213	0.755	0.266	0.310							
Sand	−0.881	−0.828	−0.156	−0.423	−0.950	−0.589						
W_H	0.972	0.900	−0.187	0.419	0.962	0.282	−0.918					
W_{15}	0.922	0.897	0.033	0.374	0.900	0.379	−0.895	0.940				
$W_{1/3}$	0.886	0.885	0.183	0.290	0.881	0.540	−0.929	0.892	0.951			
W_S	0.865	0.882	−0.194	0.286	0.906	0.224	−0.847	0.899	0.848	0.834		
W_A	0.702	0.733	0.354	0.130	0.719	0.671	−0.829	0.710	0.738	0.911	0.686	

Cell values are correlation coefficients, *r*, between relevant pair of attributes
S specific surface area (measured in $m^2\ g^{-1}$), *CEC* cation exchange capacity (meq $100\ g^{-1}$), $CaCO_3$ calcium carbonate content (%), *OM* organic matter (%), *Clay* clay content(%), *Sand* sand content (%), W_H hygroscopic moisture content (%), W_{15} moisture content at 15 atm. (%), $W_{1/3}$ moisture content at 1/3 atm. (%), W_S moisture content of saturated paste (%), W_A available moisture (%)

analysis (Hotelling 1933). In simple terms, PCA involves the rotation of the n orthogonal axes of the feature space formed by the n attributes in such a way that the rotated first axis accounts for the most variance in the dataset, the rotated second axis accounts for the second largest component of variance in the dataset and so on. The rotated axes are called principal components, and the values of the individuals on each principal component axis are called scores. The principal component axes are orthogonal with each other, which guarantees that the scores measured on them are uncorrelated with each other. The full set of principal component axes are needed to describe the rotated feature space entirely, but it is probable that most of the variance is described by the first few components (Norris 1971). Thus, with respect to reducing the number of attributes required for cluster analysis, the original attributes can be replaced with the individuals' scores on the first few principal component axes with the certainty that the scores on each axis are independent of each other and that minimal information has been lost compared to that contained in the original dataset. The disadvantage is that the principal component scores are more difficult to interpret than the original attribute values and the original attributes cannot be recovered unless the scores of all the principal components are known. Kyuma and Kawaguchi (1976) used the approach when classifying Japanese paddy soils to reduce a set of 12 soil material attributes to the scores on two principal components prior to classification using a dendrogram.

8.2.2.2 Weighting of Attributes

We may also wish to weight certain attributes prior to the cluster analysis according to their perceived importance or by some other rule. Whether or not attributes are weighted often depends on whether the resulting classification is intended to reflect general-purpose or special-purpose use. If a general-purpose classification is desired, then it is generally accepted that all attributes should remain unweighted so that each has equal value and importance (Sneath and Sokal 1962). Arkley (1976) noted that this is in conflict with traditional general-purpose hierarchical soil classifications where attributes that are used as partitioning criteria at higher nodes of the hierarchy have greater effective weight in the classification than attributes appearing lower in the hierarchy.

Weighting specific attributes means that they can exert greater influence in the cluster analysis compared to the attributes that remain unweighted, and this can be reflected in the class definitions (Gibbons 1968). This has application in the construction of special-purpose classification systems if certain soil uses can be shown to depend on specific soil attributes.

8.2.3 Kinds of Clustering Algorithms

The field of cluster analysis is broad, and researchers have devised many clustering algorithms. Not all of them have been applied to soils, and, of those that have, some have proved more popular than others. The algorithms themselves can be classified in various ways. For example, Williams (1976b) made a classification with respect to application in agricultural science that reflects a time when fuzzy classification was still in its infancy. Clustering algorithms are often described in terms of (i) whether or not the classes are known a priori (supervised versus unsupervised algorithms), (ii) the trajectory of class formation (agglomerative versus partitional classification), (iii) the structure of the resulting classification (hierarchical versus non-hierarchical classification) and (iv) the exclusivity of their classes (exclusive versus nonexclusive classification). Criteria (i)–(iii) relate to the structure of the classification systems, whereas criterion (iv) relates to the nature of the classes. In this section we briefly explore these concepts.

8.2.3.1 Supervised Versus Unsupervised Classification

Supervised classification is carried out when the classes are known a priori. Supervised techniques were some of the first numerical classification techniques applied to soils. The ordination techniques that appeared in the late 1950s enabled researchers to group existing soil taxa on the basis of quantifiable similarity (Bidwell and Hole 1964b). Hole and Hironaka (1960) published one of the first studies to do so. They applied ordination to the grouping of soils in the Miami taxonomic family from Ohio and to representative profiles of great soil groups collected from around the world. Results of ordinations were often summarised using dendrograms (Bidwell and Hole 1964a, b; Rayner 1966).

Contemporary methods of classification tree analysis and artificial neural networks are also supervised classification techniques although they are more frequently used to calibrate environmental and landscape characteristics with soil observations for the purpose of spatial prediction. A recent exception was the study by Ribeiro et al. (2014) who used fuzzy classification trees to examine the relationships between soil properties and classes in the Brazilian soil classification system.

Unsupervised classification, on the other hand, attempts to discover natural classes amongst a group of individuals. It does not assume any pre-existing soil classification. For this reason it is often useful for exploring if natural groups are present in an unclassified collection of soil material samples. In the soil literature, popular unsupervised classification algorithms have included the *k*-means and, later, fuzzy *k*-means algorithms and their derivatives, described later in this chapter.

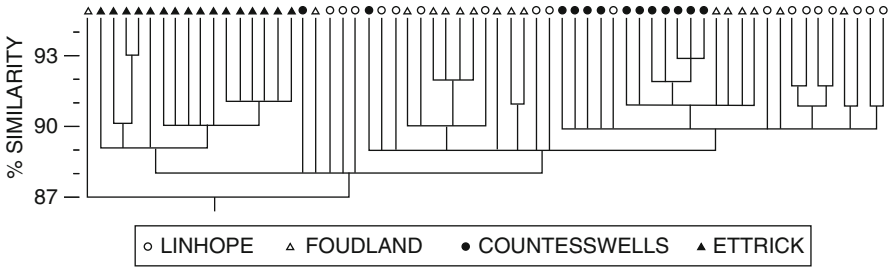

Fig. 8.1 Dendrogram of some Scottish soils produced via a single-linkage agglomerative hierarchical classification procedure (Muir et al. 1970, reproduced with permission)

8.2.3.2 Agglomerative Versus Partitional Classification

Agglomerative clustering works by grouping individuals together into larger and larger groups on the basis of the similarity between them. Given a population of n individuals, n initial clusters are formed that each contain one individual. Clusters are successively joined in n–1 steps until all have been joined together into one cluster (Fig. 8.1). At any stage, the two clusters that are chosen to be joined are those that are the closest according to some distance or similarity metric, typically in the Euclidean space. There are many versions of agglomerative clustering, and they differ primarily in how the intercluster distance is calculated. They are reviewed in detail by Anderson (1971) and Webster (1977).

The simplest agglomerative clustering method is known as nearest neighbour clustering. In this method, the intercluster distance is defined as the smallest of the distances between the members of each cluster. A more complex method known as centroid clustering (Gower 1967) defines the intercluster distance as the squared distance between cluster centroids.

Partitional, or *divisive*, clustering algorithms divide a population of individuals into a set of k classes. A key problem with this kind of clustering is choosing the number of clusters (Dubes 1987; Jain et al. 1999). Patterns corresponding to the initial class centroids may be chosen at random or by using expert knowledge. Using a similarity measure, individual patterns are allocated to the cluster whose centroid is the closest. After all patterns have been allocated, the cluster centroids are often recalculated and patterns reallocated. Iteration continues in this fashion until some measure of goodness is acceptable. One of the most well-known partitional clustering algorithms, k-means clustering, is frequently attributed to Lloyd (1982), who devised it in the 1950s. Hartigan published a more efficient version in the late 1970s (Hartigan 1975; Hartigan and Wong 1979).

8.2.3.3 Hierarchical vs. Non-hierarchical Classification

In a taxonomic hierarchy, individuals belong to groups at the lowest or most specific level of the hierarchy. Figure 8.1 depicts a hierarchical classification of some Scottish soils, although the lowest level of the hierarchy is at the top of the graph. These groups belong to more general groups at the next hierarchic level, and so on until all groups are united in a single, general group (Webster and Oliver 1990). Soil classes are frequently organised in this way due to their simplicity and relative ease of use for allocation. Hierarchies may be generated agglomeratively or divisively (Anderson 1971; Webster 1977).

Many contemporary soil classification systems employ hierarchical structures. Despite this, there are circumstances in which a hierarchical structure is neither applicable nor ideal (Dale et al. 1989; Webster and Oliver 1990). For example, hierarchies were commonly formed divisively by choosing one or a few so-called *diagnostic properties* as the subdividing criterion (or criteria) at each hierarchic level. The subdivision is typically mutually exclusive so, for instance, class A may have a topsoil organic carbon content of less than 4% and class B greater than or equal to 4%. The classes at the lowest level of the hierarchy are thus defined by a unique set of attribute values because their class limits do not overlap with those of other classes. Such classification has been called *monothetic* (Sokal and Sneath 1963). However, as Webster (1968) pointed out, soil is *polythetic* in the sense that a set of classes may possess many shared attributes that cannot be subdivided mutually exclusively.

8.2.3.4 Exclusive Versus Nonexclusive Algorithms

Clustering algorithms may also be classified according to the character of the classes that they produce. Once we allocate an individual to a class, we say that the individual has a degree of *membership*, m, in the class. This membership quantifies the degree to which the individual possesses the characteristics of the central concept of a class and can be expressed on a numerical scale from 0 to 1 (Burrough et al. 1997).

Exclusive algorithms produce classes with boundaries that are discontinuous, or hard, or crisp. Crisp classes are exclusive because an individual can belong to, or have *complete* membership in, one and only one crisp class. Numerically, in an exclusive system, an individual can have $m = 1$ to one and only one class and $m = 0$ to all others. A hard partition of n individuals into k classes can be represented by an $n \times k$ matrix of memberships $\mathbf{M} = (m_{ic})$. The following conditions apply in order to ensure that the classes are mutually exclusive, jointly exhaustive and non-empty (McBratney and de Gruijter 1992):

$$\sum_{c=1}^{k} m_{ic} = 1, \quad i = 1, \ldots, n \tag{8.7}$$

$$\sum_{i=1}^{n} m_{ic} > 0,\ c = 1, \ldots, k \tag{8.8}$$

$$m_{ic} \in \{0, 1\},\ i = 1, \ldots, n; c = 1, \ldots, k \tag{8.9}$$

The first condition ensures that, for a given individual, the memberships across all classes sum to unity; the second ensures that, for a given class, the sum of memberships of all individuals is greater than zero; the third ensures that membership can only take values of 0 or 1.

Class limits are often defined as a set of discriminating criteria using statements such as the following:

> members of class Alpha possess an A-horizon clay content of 10–20% clay.

In the preceding example, although some within-class variation in A-horizon clay is permitted, the limits themselves are crisp because an individual with an A-horizon clay content less than 10% or greater than 20% is excluded from membership in class Alpha. It is implicitly assumed that all change between classes occurs at the class boundaries and that the within-class variation is irrelevant at least for interpretive purposes (Burrough 1989). An advantage of crisp classes is that allocation is relatively simple and can often be achieved through the construction of a device such as a dichotomous key.

Crisp classes are not without drawbacks, however. Fundamentally, that their boundaries are hard means that they disregard the natural continuity in the soil attribute space (McBratney et al. 1992). As such they are incapable of representing vague concepts (Metternicht 2003). Furthermore, as variation in the soil attribute space is continuous, any placement of crisp boundaries is arbitrary. Other concerns are more practical. For example, although within-class variation in the diagnostic properties may be ignored in practice, it must be adequately known for the class limits to be established. A system of crisp classes is prone to misclassification in certain circumstances. For example, measurement error may lead to misclassification of an individual (Webster 1968) and a group of otherwise similar profiles may be allocated to different classes because they vary with respect to a single diagnostic property.

Nonexclusive algorithms produce classes with boundaries that are continuous, or fuzzy, or overlapping. Continuous classes are nonexclusive because an individual can belong to, or have *partial* membership in, more than one class simultaneously. Thus, they allow for vagueness in the class definitions that crisp classes cannot accommodate. In a nonexclusive system, the membership requirements are relaxed so that m can vary continuously between 0 and 1. In other words, the third assumption, above (Eq. 8.9), is replaced with the following:

$$m_{ic} \in [0, 1],\ i = 1, \ldots, n; c = 1, \ldots, k \tag{8.10}$$

Continuous classes are usually defined by a central concept or centroid which is essentially a soil material description consisting of the modal values for a suite

of soil attributes. Memberships are based on the degree of similarity between an individual and the class centroid. Such a definition is compatible with what Cline (1949) envisioned.

The vagueness in class allocation that is enabled by partial membership confers several advantages to continuous classes. For example, the placement of arbitrary class limits is avoided. Intergrade soils—soils with characteristics intermediate to those of two or more class definitions—are better accommodated. Finally, some of the risk of misallocation is avoided since an individual can still have partial membership in an alternate class.

8.2.4 k-Means Clustering and Its Derivatives

In this section we discuss the *k*-means clustering algorithm and several of its derivatives, which are amongst the most popular clustering algorithms used to classify soil material.

8.2.4.1 *k*-Means Algorithm

The *k*-means clustering algorithm is an unsupervised, partitional, non-hierarchical algorithm that partitions a population of individuals into crisp classes (Hartigan 1975).

The *k*-means clustering algorithm works as follows:

1. Choose a value of k with the restriction that $1 < k < n$.
2. Initialise k cluster centroids. Cluster centroid patterns may be initialised randomly or by using expert knowledge.
3. Compute distance between cluster centroids and patterns of individuals being classified.
4. Recompute cluster centroids once all individuals have been allocated to a class.
5. Repeat steps 3–4 until some convergence criterion is met (e.g. the allocation has not changed).

The similarity between individual patterns is typically computed using the Euclidean distance, although the Mahalanobis distance may also be used (Mao and Jain 1996). In step 3, an individual is allocated to the class of the cluster to which it is closest in Euclidean space. In step 4, recomputation of the cluster centroids is achieved by averaging the values of the attributes of the individuals assigned to it:

$$c_{j,p} = \frac{\sum_{\mathbf{x}_i \in \mathbf{X}} x_{ip}}{|C_j|}, 1 \leq p \leq d \tag{8.11}$$

where $|C_j|$ is the cardinality of the cluster C_j.

The algorithm reaches convergence when some criterion is met, such as when individuals cease to move from one cluster to another or when the value of an objective function ceases to decrease significantly from iteration to iteration (Jain et al. 1999). A common objective function, which computes the total Euclidean distance of the patterns to their cluster centres, is computed as follows:

$$J(\mathbf{M}, \mathbf{C}) = \sum_{i=1}^{n} \sum_{c=1}^{k} m_{ic} d^2 (\mathbf{x}_i, \mathbf{c}_c) \tag{8.12}$$

The *k*-means algorithm is easy to implement, which has made it a popular choice in cluster analyses in many fields. Despite this, users need to be aware of several factors. First, the algorithm is sensitive to the choice of the initial cluster centroids. A random selection of initial cluster centroids is likely to lead to a partition of individuals that is only *locally* optimal since an exhaustive search for the combination of individuals that yields the *global* minimum value of the objective function is computationally prohibitive. Indeed it may be impossible to prove that a given partition is globally optimal (Steinley 2006). Some researchers (Falkenauer and Marchand 2001; Hartigan 1975; Jain 2010) suggested selecting the best partition, in terms of the minimal objective value, from a pool of partitions created by running the *k*-means algorithm several times, but the best partition in this case is still unlikely to be the global optimum.

Second, because the Euclidean distance is typically used to quantify similarity between individuals, the algorithm produces convex, approximately spherical, clusters in the attribute space (Jain 2010). Since soil classes are rarely this shape (Odeh et al. 1992), this distance is likely to be inappropriate.

In soil science the *k*-means algorithm has been applied not only to soil classification (e.g. Bormann 2010; Minasny and McBratney 2006) but also digital soil mapping (Bui and Moran 2001) and sampling design (Brus et al. 2006).

8.2.4.2 Fuzzy *k*-Means

The fuzzy *k*-means algorithm is an unsupervised, partitional, non-hierarchical classification algorithm that produces continuous classes. The fuzzy *k*-means algorithm extends the notion of fuzzy logic (e.g. Zadeh 1965) to cluster analysis to allow for classes to overlap in the attribute space. Introduced in the early 1980s (Bezdek 1981), the algorithm began to be used for soil classification in the early 1990s (Odeh et al. 1990).

The fuzzy *k*-means algorithm functions in much the same way as the *k*-means algorithm except in the computation of membership functions. In the fuzzy *k*-means algorithm, the computation of cluster centroids is modified to account for the partial memberships of all the individuals associated with each cluster.

The degree of fuzziness in the fuzzy clustering can be modified and is controlled by the parameter φ, the so-called *fuzziness exponent*. The degree of fuzziness relates to the degree of overlap of the resulting classes in the attribute space. The minimum

value of φ is 1.0, which is equivalent to no fuzziness, and yields a hard partition into non-overlapping classes, much like the *k*-means algorithm. As φ increases, the degree of overlap of classes in the attribute space increases.

The membership *m* to class *c* of individual *i* is computed using the following:

$$m_{ic} = \frac{d_{ic}^{-2/(\varphi-1)}}{\sum_j^k d_{ij}^{-2/(\varphi-1)}}, i = 1, \ldots, n; c = 1, \ldots, k \tag{8.13}$$

The cluster centroids are computed using the following:

$$\mathbf{c}_c = \frac{\sum_{i=1}^n m_{ic}^{\varphi} \mathbf{x}_i}{\sum_{i=1}^n m_{ic}^{\varphi}}, c = 1, \ldots, k \tag{8.14}$$

Optimisation attempts to minimise the following objective function, which is a weighted sum of the distances between every pattern and every cluster centroid:

$$J_B(\mathbf{M}, \mathbf{C}) = \sum_{i=1}^n \sum_{c=1}^k m_{ic}^{\varphi} d_{ic}^2 \tag{8.15}$$

Compared to classes arising from the *k*-means algorithm, fuzzy classes are more robust because they have been shown to contain more information (Lagacherie et al. 1997) and be less sensitive to errors in the attribute data (Heuvelink and Burrough 1993).

Users need to be aware of some of the same factors relating to *k*-means applications. For example, continuous classes still tend to form spherical or hyperspherical classes in attribute space (Rousseeuw et al. 1996), which is inappropriate if we do not expect our classes to take such a shape. Second, the choice of *k* is still somewhat subjective although cluster validity measures can help to choose an appropriate number. Even so, some (e.g. McBratney and Moore 1985; Odeh et al. 1990) have cautioned that it may not be possible to know how many classes exist in our data because we frequently do not know how representative our soil observations are.

We may also ponder what is the appropriate value of φ. As φ determines the fuzziness of the fuzzy classification, a good value should reflect the fuzziness in the attribute space. This is usually not known in advance (Lagacherie et al. 1997). Odeh et al. (1992) suggested that the optimal φ should represent a balance between preserving natural partitional structures in the dataset and continuity of the classes. They reasoned that *k* should be established first by examining the partition entropies associated with the different *k*, and then φ could be set to reflect the appropriate level of fuzziness. The final choice remains somewhat arbitrary but should be guided by solid expert knowledge of the data.

Researchers typically use values of φ in the range of 1.1–1.5 (e.g. Burrough et al. 2000; Cockx et al. 2007; Dobermann et al. 2003; Triantafilis et al. 2001). Although researchers often do not describe the manner in which they determine φ in their fuzzy *k*-means cluster analyses, some have presented detailed studies in which they attempted to determine an appropriate φ empirically. For example, McBratney

and Moore (1985) used the derivative of J_B (**M**, **C**) with respect to φ to determine optimal values of φ and k. de Bruin and Stein (1998), in a fuzzy cluster analysis of landform components across a valley hillslope in the Netherlands, determined the optimal k and φ in terms of how well the fuzzy memberships could predict the variation in topsoil clay content across the hillslope. They found that the optimal fuzzy partition, with $k = 4$ and $\varphi = 2.1$, was able to account for about 70% variation in the soil property.

We illustrate the use of the fuzzy k-means algorithm by cluster analysis of 81 soil material samples from profiles observed at Pokolbin in the Lower Hunter Valley in New South Wales, Australia. The soil material samples were taken from within a pedogenetic horizon of their respective profile; while specific horizons were not favoured, B2 horizons were the most frequent sources of the soil material because profiles were sampled by auger and other horizons were often not thick enough to contain the volume of material required for laboratory analysis. The samples were attributed with a range of soil properties including clay content, pH and moist *Lab* colour (converted from Munsell colour notation) and effective cation exchange capacity (eCEC; Table 8.4). We ran the fuzzy k-means algorithm on the soil material samples several times in order to produce a set of partitions from $k = 2$ to $k = 25$ classes. A locally optimal partition was obtained when $k = 9$ and $\varphi = 1.2$ (see also Fig. 8.6 and associated discussion). The centroids of the resulting classes are presented in Table 8.4. Relationships between attributes appear to be pedologically sensible. For example, higher clay content is generally associated with lower sand content and vice versa, and darker-coloured soils (those with smaller moist L) tend to be associated with higher total carbon.

A biplot of the first two principal components of the observations' attributes is presented in Fig. 8.2. The 95% density ellipses were computed based on individuals with membership of 0.5 or higher in each class (after Triantafilis et al. 2001). The first two principal components account for about 74% of the variation in the data; the first four principal components account for nearly 95% of the variation. While some pairs of classes, such as E and G, are well separated, others, like A and C, transcend the distribution of several other classes in the two-dimensional representation. The loadings of the attributes enable an examination of the contribution of each attribute to the principal components (Fig. 8.3). Thus, clay content, sand content and electrical conductivity contribute most to the first principal component, whereas total carbon and $CaCO_3$ content contribute the most to the second principal component. Relationships between the soil attributes as expressed by the directions of their loadings vectors also affirm the relationships that are discernible in Table 8.4.

8.2.4.3 Fuzzy k-Means with Extragrades

Consider the synthetic dataset in Fig. 8.4a in which there are five natural classes and several outlying individuals. A weakness of the fuzzy k-means algorithm is that individuals with approximately equal memberships to all cluster centroids may

Table 8.4 Fuzzy *k*-means centroids of soil material classes derived from a collection of soil material samples from profiles observed at Pokolbin in the Lower Hunter Valley in New South Wales, Australia

Class	pH (1:5 H_2O)	Electrical conductivity (dS m^{-1})	Total C (%)	$CaCO_3$ (%)	Clay (%)	Sand (%)	Moist *Lab* colour			eCEC (meq 100 g^{-1})
							L	*a*	*b*	
A	5.6	0.348	0.27	0.41	40.2	47.7	63.5	2.0	15.5	14.98
B	4.8	0.865	0.86	0.41	8.2	15.3	21.1	4.5	12.1	28.27
C	6.8	0.399	0.39	0.40	39.0	44.2	38.9	4.5	18.6	16.89
D	5.5	0.267	0.51	0.41	50.0	39.5	39.6	20.9	29.1	16.45
E	5.2	0.254	0.63	0.40	62.3	23.0	40.5	5.9	19.2	22.62
F	5.3	0.877	0.47	0.52	52.0	32.3	41.3	4.7	17.6	21.79
G	5.9	0.130	0.81	0.41	23.7	67.5	42.1	4.9	17.9	7.83
H	6.2	0.245	0.42	0.86	37.4	52.1	47.7	10.4	42.9	16.23
I	7.1	0.131	1.61	0.61	53.6	29.2	29.0	6.7	17.0	30.97

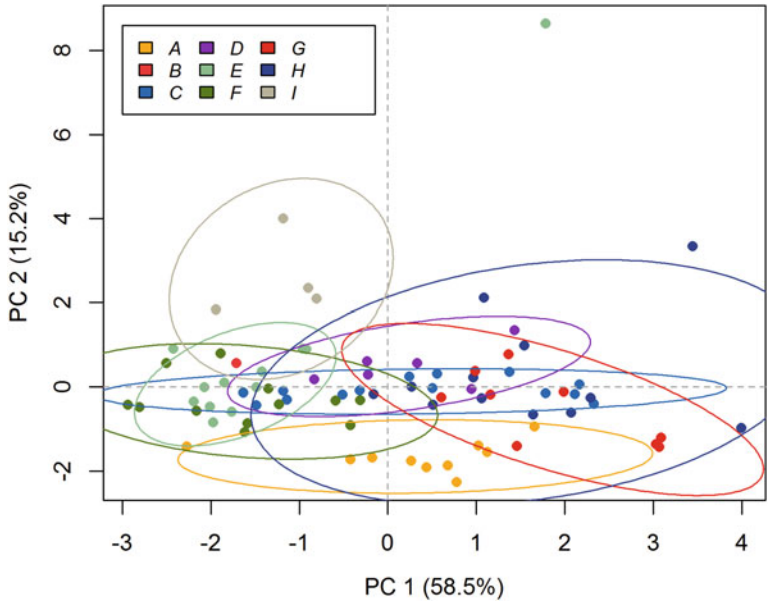

Fig. 8.2 Distribution of the soil material samples along the first two principal components of their attributes

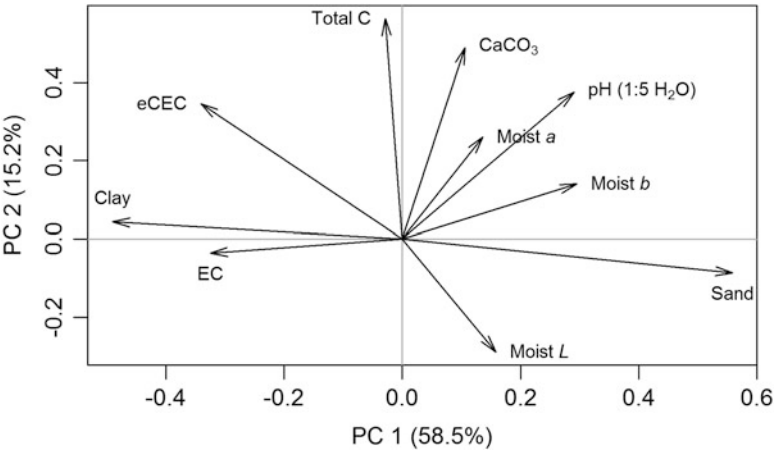

Fig. 8.3 Vectors of the 10 attributes in the space formed by the first two principal components

receive about equal memberships to their classes, whether or not the individuals lie in the centre or the outlying region of the attribute space (de Gruijter et al. 1997). This may lead to a distortion of the locations of the centroids in the attribute space (Fig. 8.4b).

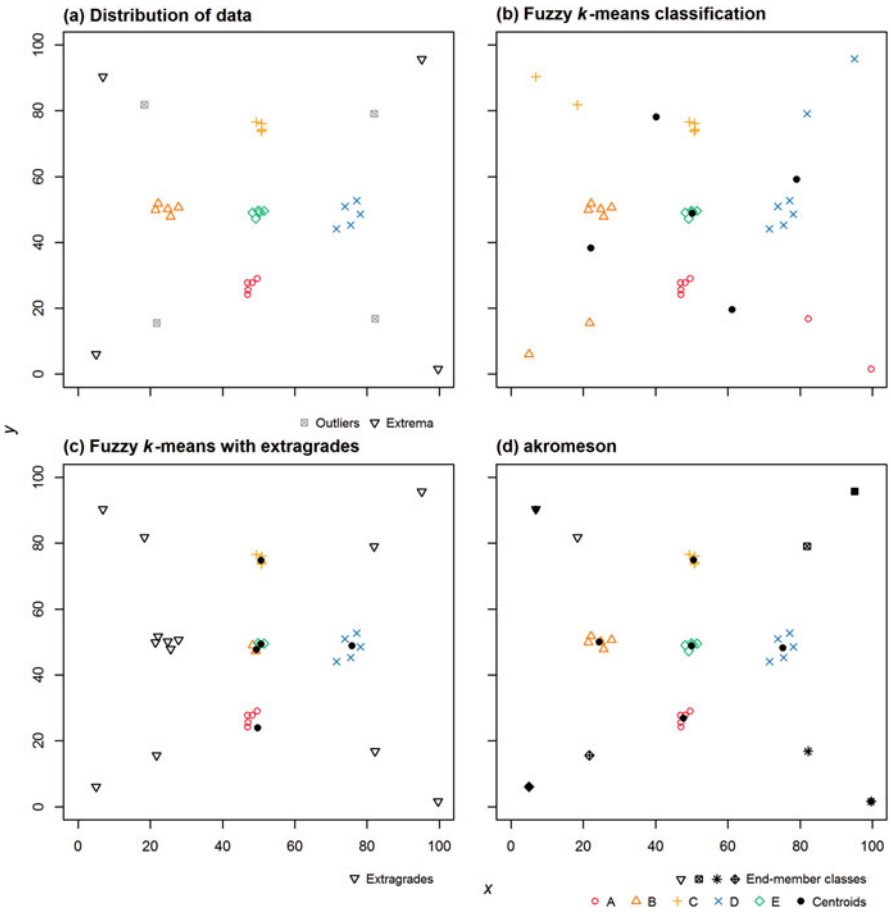

Fig. 8.4 Classification of a synthetic dataset (**a**) by several fuzzy variants of the *k*-means algorithm: (**b**) the fuzzy *k*-means algorithm, (**c**) fuzzy *k*-means with extragrades and (**d**) akromeson, after Hughes et al. (2014)

Fuzzy *k*-means with extragrades (de Gruijter and McBratney 1988; McBratney and de Gruijter 1992) is a modification to the standard fuzzy *k*-means algorithm that allows the modelling of an extragrade class of individuals. It attempts to overcome the weakness of the fuzzy *k*-means algorithm caused by outlying individuals. Although all individuals that do not possess a high degree of membership to any soil class may be termed intergrades, de Gruijter and McBratney (1988) distinguished intergrades located in the space *between* classes from those located in the outlying space. They are termed *intragrades* and *extragrades*, respectively. As failure to distinguish true outliers from mere intragrades may be misleading, de Gruijter and McBratney (1988) defined an extragrade class in which the memberships m_{i*} were made directly dependent on the distances to the class centroids. The membership and centroid-update equations are modified as follows (de Gruijter and McBratney 1988):

$$m_{ic} = \frac{d_{ic}^{-2/(\varphi-1)}}{\sum_{j=1}^{k} d_{ij}^{-2/(\varphi-1)} + \left(\frac{1-a}{a}\sum_{j=1}^{k} d_{ij}^{-2}\right)^{-1/(\varphi-1)}}, \; i = 1, \ldots, n; c = 1, \ldots, k \tag{8.16}$$

$$m_{i*} = \frac{\left(\frac{1-a}{a}\sum_{j=1}^{k} d_{ij}^{-2}\right)^{-1/(\varphi-1)}}{\sum_{j=1}^{k} d_{ij}^{-2/(\varphi-1)} + \left(\frac{1-a}{a}\sum_{j=1}^{k} d_{ij}^{-2}\right)^{-1/(\varphi-1)}}, \; i = 1, \ldots, n \tag{8.17}$$

$$\mathbf{c}_c = \frac{\sum_{i=1}^{n}\left\{m_{ic}^{\varphi} - \frac{1-a}{a} d_{ic}^{-4} m_{i*}^{\varphi}\right\}\mathbf{x}_i}{\sum_{i=1}^{n}\left\{m_{ic}^{\varphi} - \frac{1-a}{a} d_{ic}^{-4} m_{i*}^{\varphi}\right\}}, \; c = 1, \ldots, k. \tag{8.18}$$

The fuzzy objective function is modified accordingly:

$$J_{\mathrm{MG}}(\mathbf{M}, \mathbf{C}) = a\sum_{i=1}^{n}\sum_{c=1}^{k} m_{ic}^{\varphi} d_{ic}^{2} + (1-a)\sum_{i=1}^{n} m_{i*}^{\varphi}\sum_{c=1}^{k} d_{ic}^{-2} \tag{8.19}$$

The memberships to the extragrade class, m_{i*}, spread across regions at larger distances from the class centroids, unlike the memberships of the regular classes which tend to occupy fuzzy hyperspheres around the class centroids (McBratney and de Gruijter 1992).

The parameter α determines the mean extragrade membership; however, the function relating both quantities is generally unknown. Because of this, de Gruijter and McBratney (1988) estimated α empirically using a Regula-Falsi procedure. Lagacherie et al. (1997) estimated α by examination of a two-dimensional representation of the multivariate attribute data in conjunction with their expert knowledge of their study area.

8.2.4.4 Akromeson

The fuzzy *k*-means with extragrades algorithm enabled individuals lying in the outer parts of the attribute space to be recognised and placed into their own extragrade class. Doing so reduced the leverage the outlying points had on the formation of regular fuzzy classes in the more densely populated parts of the attribute space. Notwithstanding these advantages, estimation of the α parameter which determines the mean extragrade membership is not straightforward, and no clear procedure for doing so exists. Incorrect estimation of the extragrade class may lead to some extragrade individuals being treated as if they were individuals in the centre of the data and vice versa (Fig. 8.4c). In addition extragrades are placed into a single class regardless of their distribution in the attribute space. In reality there may be clusters of individuals in the outlying space, but the fuzzy *k*-means with extragrades algorithm is unable to resolve them if they exist.

Hughes et al. (2014) introduced the idea of end points, which are the individuals lying at the extremes of the attribute space (the points classified *extrema* in Fig. 8.4a). They can be detected as the vertices of a convex hull around the individuals in attribute space. Hughes et al. (2014) developed an algorithm they called *akromeson* that identifies end points and treats them as fixed centroids in a semi-supervised fuzzy *k*-means cluster analysis (Bensaid et al. 1996). A heuristic process is used to refine the number of end points, k_e, that are used as fixed centroids. The aim of the semi-supervised fuzzy *k*-means clustering algorithm, then, is to find additional clusters, k_d, in the attribute space so that $k = k_e + k_d$. The fuzziness exponent, φ, can be determined using the usual methods (Odeh et al. 1992).

The end result is a set of k classes that includes k_e end point classes (Fig. 8.4d). In Fig. 8.4d, $k_e = 4$ and $k_d = 5$ so $k = 4 + 5 = 9$. The end point classes supersede the single extragrade class completely and as such provide more information about the distribution of individuals in outlying parts of the attribute space. This extra information may be related to real environmental differences between akrograde classes that in turn may be important for managing land.

8.2.4.5 Fuzzy *c*-Numbers

The fuzzy *c*-numbers algorithm (Yang and Ko 1996) is an extension of fuzzy *k*-means clustering that allows the input attributes to be fuzzy numbers. As with the fuzzy *k*-means algorithm, the basic steps are similar to the standard *k*-means algorithm; however, unlike these two approaches, which cluster crisp attribute data, the fuzzy *c*-numbers algorithm clusters attribute data that are fuzzy.

Fuzzy attribute data are actually somewhat commonplace in soil material descriptions. For example, the pH value of a modal soil material representative might be described by a modal value and perceived lower and upper limits. Fuzzy attribute data can be represented using fuzzy numbers, which are fuzzy sets over the set of real numbers and are specified via membership functions. There are many different kinds of fuzzy numbers, but three of the most common are triangular, trapezoidal and Gaussian fuzzy numbers (Liu and Samal 2002a; Yang and Ko 1996), so-called due to the shape of their membership functions. Triangular fuzzy numbers may be symmetric or asymmetric, and we may estimate their parameters from a collection of soil profile measurements (Fig. 8.5). For example, say the distributions of pH (1:5 H_2O) in the 40–50 cm depth interval of a collection of Chromosols and Kurosols from the Lower Hunter Valley in New South Wales, Australia, take the values in Table 8.5. Kurosols are strongly acidic (pH (1:5 H_2O) less than 5.5) in the top 20 cm of the B_2 horizon, whereas Chromosols are not. A symmetric triangular fuzzy number $\tilde{x}_{ip}$ representing the pH may have a maximum membership at the mean pH—the so-called *apex*, x_{ip}, and minima of membership above and below this value at a distance defined by the *spread* S_{ip}, which we may choose to represent as two standard deviations from the mean pH value. Thus, a symmetric triangular fuzzy number $\tilde{x}_{ip}$ can be represented as (x_{ip}, S_{ip}).

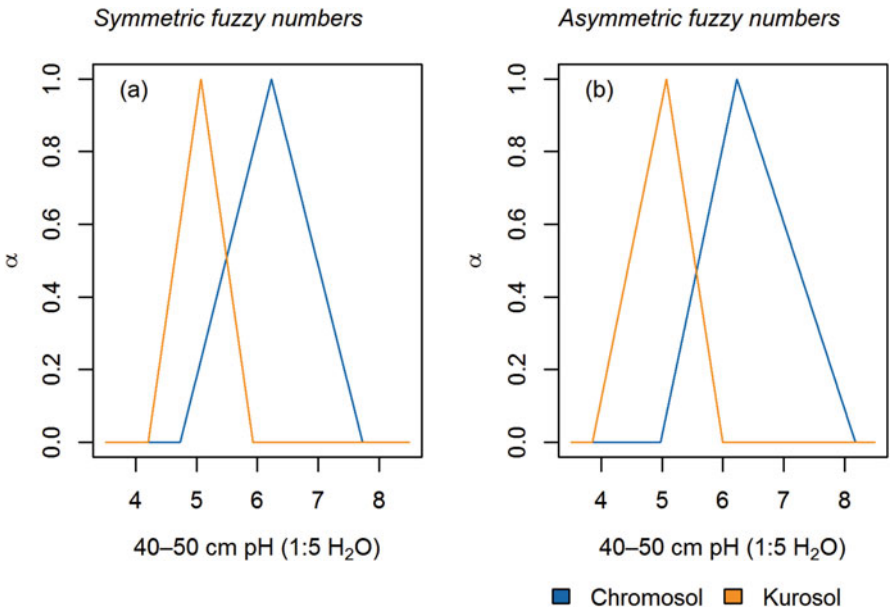

Fig. 8.5 Triangular numbers representing 40–50 cm pH (1:5 H_2O) for Chromosols and Kurosols in the Lower Hunter Valley, New South Wales, Australia

Table 8.5 Descriptive statistics of the distribution of pH (1:5 H_2O) for Chromosols and Kurosols observed at Pokolbin in the Lower Hunter Valley in New South Wales, Australia

	Chromosol	Kurosol
Mean	6.23	5.07
Standard deviation	0.75	0.43
Minimum	4.97	3.85
Maximum	8.18	6.00
n	57	28

On the other hand, the membership minima of an asymmetric triangular fuzzy number are not the same distance from the apex. We may choose to represent them as the observed maximum and minimum pH. An asymmetric triangular fuzzy number $\tilde{x}_{ip}$ can be represented as (x_{ip}, a_{ip}, b_{ip}) where a and b are the lower and upper minima of membership, respectively. Figure 8.5 demonstrates that the difference between symmetric and asymmetric triangular fuzzy numbers is relatively subtle if they are based on a distribution of values that is roughly normally distributed. For mathematical simplicity the symmetric fuzzy numbers are used hereforth.

A soil material description is usually represented as a vector of several attributes rather than a single attribute. A fuzzy vector of symmetric triangular fuzzy numbers can be defined as $\tilde{x}_i = (x_i, \mathbf{P}_{x_i})$ where x_i is the standard pattern vector $\mathbf{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$ and $\mathbf{P}_{xi}$ is the *panderance matrix* that contains the spread information (Celmiņš 1987):

$$\mathbf{P}_{x_i} = \begin{bmatrix} s_{i1}^2 & 0 & \cdots & 0 \\ 0 & s_{i2}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & s_{id}^2 \end{bmatrix} \tag{8.20}$$

Liu and Samal (2002a) computed the distance between two fuzzy vectors $\tilde{a} = (a, \mathbf{P}_a)$ and $\tilde{b} = (b, \mathbf{P}_b)$ as

$$d^2\left(\tilde{a}, \tilde{b}\right) = \|a - b\|^2 + tr\left((\mathbf{P}_a - \mathbf{P}_b)^T (\mathbf{P}_a - \mathbf{P}_b)\right) \tag{8.21}$$

where *tr* is the trace of the matrix product of $(\mathbf{P}_a - \mathbf{P}_b)^T(\mathbf{P}_a - \mathbf{P}_b)$.

The cluster centres are given by $\{\tilde{c}_1, \tilde{c}_2, \tilde{c}_3, \ldots, \tilde{c}_k\}$ where $\tilde{c}_c = (c_c, \mathbf{P}_c)$. The membership of a fuzzy pattern $\tilde{x}_i$ in class c is computed as

$$m_c\left(\tilde{x}_i\right) = \frac{1}{\sum_{j=1}^{k} d^2\left(\tilde{x}_i, \tilde{c}_c\right) / d^2(\tilde{x}_i, \tilde{c}_k)^{1/(\varphi-1)}} \tag{8.22}$$

Update of the centroid patterns $\tilde{c}_j$ is a two-step process because both the apex vector $\mathbf{c}_c$ and panderance matrix $\mathbf{P}_{cj}$ need to be recomputed for each centroid. The apex vector is calculated as

$$\mathbf{c}_c = \frac{\sum_{i=1}^{n} m_c(\tilde{x}_i)^{\varphi} \mathbf{x}_i}{\sum_{i=1}^{n} m_c(\tilde{x}_i)^{\varphi}} \tag{8.23}$$

and the panderance matrix is calculated as

$$\mathbf{P}_c = \frac{\sum_{i=1}^{n} m_c(\tilde{x}_i)^{\varphi} \mathbf{P}_i}{\sum_{i=1}^{n} m_c(\tilde{x}_i)^{\varphi}} \tag{8.24}$$

Liu and Samal (2002a, b) used the fuzzy *c*-numbers algorithm to identify agroecozones in the state of Nebraska in the United States. Although we could not find evidence of the application of the fuzzy *c*-numbers algorithm to soil material classification, and only very infrequent application of fuzzy numbers in the soil classification literature more generally (e.g. Bhattacharya and Solomatine 2006), fuzzy numbers have found application in other aspects of soil science and related fields, including soil sampling (Lark 2000), engineering (Dodagoudar and Venkatachalam 2000; Saboya Jr. et al. 2006), hydrology (Dou et al. 1999; Schulz and Huwe 1999; Verma et al. 2009) and forestry (Kaya and Kahraman 2011). Considering soil attributes are often recorded and presented as uncertain quantities, it appears that the fuzzy *c*-numbers algorithm and others capable of handling fuzzy attributes (e.g. d'Urso and Giordani 2006) could have natural application in soil classification.

8.3 Cluster Validation

Cluster validation usually involves the selection of the partition corresponding to the optimal k from amongst several alternatives. Many cluster validity metrics have been devised that assist in making this decision. Several authors have performed extensive comparisons between a range of these metrics (e.g. Arbelaitz et al. 2013; Milligan and Cooper 1985) although only few have been used in the numerical classification of soil material.

Cluster validity metrics should ideally be independent of the essential parameters of the cluster analysis, such as n, k and φ (Roubens 1982). They can be classified into two categories: *membership-based measures* and *geometry-based measures* (Liu and Samal 2002b). Membership-based validity measures attack the problem of cluster validity by examining the fuzziness of a partition in the membership space. On the other hand, geometry-based cluster validity measures attempt to solve the problem of cluster validity by examining the separation of a partition in the attribute space. This is achieved by quantifying the shape and distribution of clusters with respect to their compactness and their separation from each other, which can be measured via the intra-cluster distance and the intercluster distance, respectively.

8.3.1 Membership-Based Measures

A range of membership-based measures are commonly used. Two of the simplest are the partition coefficient, F, and the partition entropy, H (Bezdek 1981):

$$F = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} m_{ij}^2 \tag{8.25}$$

$$H = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} m_{ij}\log\left(m_{ij}\right) \tag{8.26}$$

The partition coefficient measures the fuzziness of the fuzzy classes and ranges from 0.5, corresponding to the most fuzzy partition, to 1.0, corresponding to the least fuzzy partition (Bezdek 1981). Thus, F is maximised when the partition is hard. The partition entropy is inversely proportional to the goodness of the fuzzy classes, and as such a better partition is indicated when H is minimised (Liu and Samal 2002b).

Roubens (1982) stated that the search of an optimal value of k is complicated by the fact that F and H tend to increase and decrease with k, respectively. He proposed that the optimal k could be found more easily using the fuzziness performance index, FPI, and the normalised classification entropy, NCE.

The FPI is computed as (Odeh et al. 1992)

$$\mathrm{FPI} = 1 - (kF - 1) / (k - 1) \tag{8.27}$$

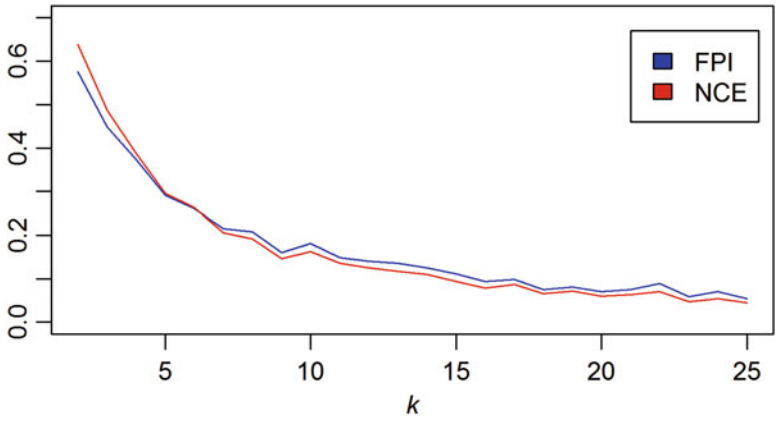

Fig. 8.6 Curves of fuzziness partition index and normalised classification entropy derived from runs of a fuzzy k-means clustering of soil material from the Lower Hunter Valley in New South Wales, Australia, using integer values of k from 2 to 25

The FPI varies between 0 and 1. Like the partition coefficient, it describes the fuzziness of a fuzzy partition. A low value of FPI implies that the continuous classes are relatively hard and that there is little sharing of membership between any pair of them. This suggests that there is a distinct natural partition structure amongst the individuals in the dataset (Odeh et al. 1992). The converse applies when FPI is high.

The *normalised classification entropy* (NCE) describes the uncertainty of the fuzzy partitioning of the individuals (Odeh et al. 1992):

$$\text{NCE} = \frac{H}{\log\ k} \tag{8.28}$$

where H is Bezdek's partition entropy (Bolliger and Mladenoff, 2005). The optimal value of k is usually found by determining a local minimum of both FPI and NCE (Odeh et al. 1992; Triantafilis et al. 2001). The set of classes at $k = 9$ satisfies this criterion in Fig. 8.6 for fuzzy clustering of soil material in Sect. 8.2.4.2.

Membership-based measures of cluster validity have been criticised on account of their lack of a direct connection to the geometry of the clusters (Xie and Beni 1991). Nevertheless, they remain relatively popular in pedometric research. While some soil researchers have used F and/or H (Burrough et al. 2001), most have used FPI and/or NCE (e.g. Bragato 2004; Cockx et al. 2007; Triantafilis et al. 2001; van Alphen and Stoorvogel 2000; Verheyen et al. 2001).

8.3.2 Geometry-Based Measures

Geometry-based measures can be classified as ratio-type and summation-type measures based on how the intra- and intercluster distances are combined (Kim and Ramakrishna 2005). Many of the more common geometry-based cluster validity metrics are of the ratio type, and several are based on the measure that Dunn originally devised (after Halkidi et al. 2001):

$$D = \min_{1\le i\le k}\left\{\min_{i+1\le j\le k-1}\left\{\frac{d\left(C_i, C_j\right)}{\max_{1\le c\le k}\left(\operatorname{diam}\left(C_c\right)\right)}\right\}\right\} \tag{8.29}$$

where $d(C_i, C_j)$ is the dissimilarity function between two clusters C_i and C_j, calculated as

$$d\left(C_i, C_j\right) = \min_{x\in C_i, y\in C_j}\left(d\left(\mathbf{x}, \mathbf{y}\right)\right) \tag{8.30}$$

and diam(Cc) is the diameter of cluster c, calculated as

$$\operatorname{diam}\left(C_c\right) = \max_{x,y\in C}\left(d\left(\mathbf{x}, \mathbf{y}\right)\right) \tag{8.31}$$

Dunn's index is primarily used to validate crisp classes. The numerator in the central term of Dunn's index pertains to the intercluster separation, whereas the denominator pertains to the intra-cluster dispersion. Large values of Dunn's index ought to indicate the presence of compact and well-separated clusters (Halkidi et al. 2001). The Dunn's index has only rarely been used in the pedometric literature (Ließ 2015).

The *compactness and separation validity function*, S (Xie and Beni 1991), is a ratio-type measure of cluster validity that has been used to validate fuzzy clusters. It has been used from time to time in the pedometric and related literature (e.g. Odgers et al. 2011; Sun et al. 2012; Vrindts et al. 2005). It is calculated as

$$S = \frac{\sum_{j=1}^{k}\sigma_j}{nd_{\min}^2} \tag{8.32}$$

where σ_j is the sum of squares of the fuzzy deviation, $m_{ij}\|\mathbf{x}_i - \mathbf{c}_j\|^2$, of individual i from centroid j:

$$\sigma_j = \sum_{i=1}^{n} m_{ij}\left\|\mathbf{x}_i - \mathbf{c}_j\right\|^2 \tag{8.33}$$

The term $\|\mathbf{x}_i - \mathbf{c}_j\|$ is simply the Euclidean distance between individual i and centroid j. σ_j is a measure of non-compactness: that is, the higher the value of σ_j, the further from the centroid are the members of class j (Liu and Samal 2002b).

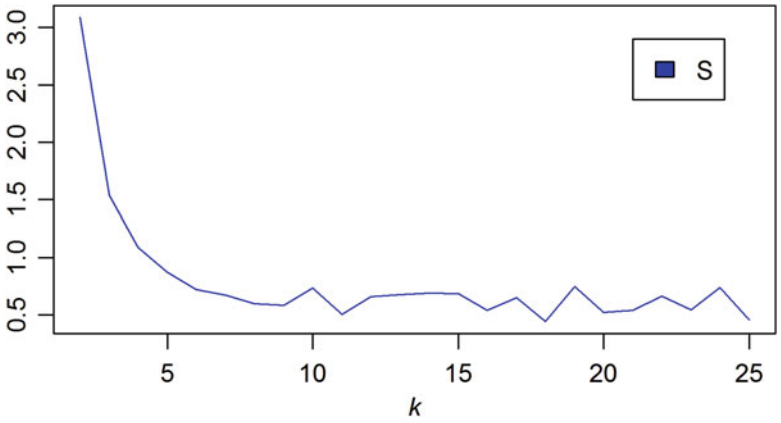

Fig. 8.7 Curve of the compactness and separation validity function, S, derived from runs of a fuzzy k-means clustering of soil material from the Lower Hunter Valley in New South Wales, Australia, using integer values of k between 2 and 25

Finally, $d^2_{\min}$ is the *separation* of the fuzzy partition and is calculated as

$$d^2_{\min} = \min \left\| \mathbf{c}_i - \mathbf{c}_j \right\|^2 \tag{8.34}$$

which is effectively the square of the minimum Euclidean distance between cluster centroids. A larger value of $d^2_{\min}$ indicates that all the clusters are well separated in the feature space (Xie and Beni 1991).

As with other cluster validity criteria, local minima in a curve of S across several values of k may indicate a locally optimal partition. Thus, Fig. 8.7 indicates that an optimal partition of the Lower Hunter Valley soil material in Sect. 8.2.4.2 may occur when $k = 11$.

Xie and Beni (1991) note that S is meaningless when c is very large and close to n because of its tendency to monotonically decrease in these circumstances. In practice this is not often a problem because the values of c that we are interested in are usually much lower than n.

8.4 Allocation to Pre-existing Classes

The placement of new individuals into a class of some classification system is known as *allocation*, *identification* or *diagnosis* (McBratney 1994) although the term *classification* has often been misappropriated to refer to the same process. The act of allocating an individual to a class implies that a classification system exists a priori. In soil science, considerably less has been written about allocation than classification although research has been performed for decades (Norris and

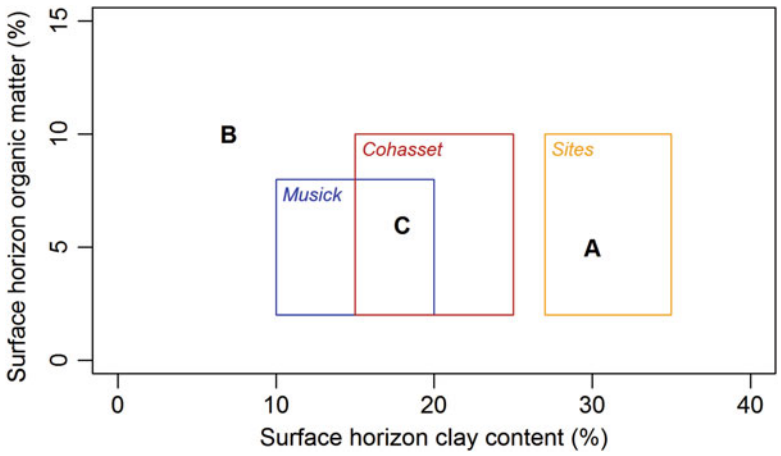

Fig. 8.8 Two-dimensional example of parallelepiped allocation. Parallelepipeds established using class limits for surface horizon clay content (%) and organic matter (%) of several soil series from El Dorado County in California, United States. Data retrieved from SSURGO database (Soil Survey Staff 2016)

Loveday 1971). Like classification itself, allocation can be crisp or fuzzy. Several methods are available, including parallelepiped methods, discriminant analysis and classification trees. Not all have been commonly applied in pedometric research.

8.4.1 Parallelepiped Method

Parallelepiped allocation is one of the most basic methods of allocation. It is sometimes also known as the *box decision rule* or the *level slice procedure* (Campbell 1996). For each of the k classes in a classification system, a parallelepiped is set up in the d-dimensional feature space using the class limits of the d attributes to set its boundaries. Alternatively the standard deviations of the attributes may also be used to set their boundaries. We illustrate with a simple two-dimensional example in Fig. 8.8 using class limits. It should be easy to see that classes are implicitly hypercubic or hypertrapezoidal in many dimensions. Allocation of an individual to class j is simply a matter of identifying which parallelepiped the individual is located inside of. Parallelepiped allocation is therefore crisp. For example, in Fig. 8.8, soil material A with surface horizon clay content of 30% and surface horizon organic matter content of 5% is clearly a member of the Sites series.

Mather and Koch (2011) identify difficulties that occur in two extreme cases. First, an individual may not lie inside any of the k parallelepipeds, making allocation impossible (B in Fig. 8.8). Second, an individual may lie inside the overlapping parallelepipeds of more than one class, in which case a protocol must be in place to determine which class the individual should be allocated to (C in Fig. 8.8). In

both these cases, more information (i.e. more attributes) is necessary to resolve the appropriate class, if any.

Parallelepiped classification has been commonly used to classify land cover from remotely sensed imagery (e.g. Goodenough and Shlien 1974; Jensen 1978; Robinove 1979) but appears not to have been used in pedometric studies.

8.4.2 Minimum Distance Method

The minimum distance method allocates an individual to the class of the centroid to which it has the shortest distance (i.e. is most similar to) in the d-dimensional attribute space. Like the parallelepiped method, allocation is a fairly simple process. Since the Euclidean or Mahalanobis distances are usually used to relate individuals to centroids, classes are implicitly hyperspherical or hyperellipsoidal in the attribute space, respectively. Minimum distance allocation may not be possible if classes do not have centroids.

8.4.3 Discriminant Analysis

According to the principle of discriminant analysis (Fisher 1936), a linear function of the attributes of a population of individuals belonging to two classes can be found that best discriminates between the two classes. Such a function can be represented as

$$X = a_1 x_1 + a_2 x_2 + a_3 x_3 + \cdots + a_d x_d \tag{8.35}$$

The coefficients of the linear function are chosen so as to maximise the distance of separation between the class means (i.e. the class centroids) relative to the within-class variability (Healy 1965). In matrix terms, they can be found by multiplying the vector of separation distances $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ by $\mathbf{S}^{-1}$, the inverse of the variance-covariance matrix of the attributes of the sample of individuals (Blackith and Reyment 1971):

$$\mathbf{a} = \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{8.36}$$

A linear function so found is known as a *discriminant function*. The discriminant function is ideally calibrated on representative individuals chosen randomly from members of each class (Webster 1977). A larger value of the function, when evaluated, implies a clearer separation of a pair of classes than does a smaller value. This brings us to an important point: as Blackith and Reyment (1971) point out, without including S^{-1} in the computation of the discriminant coefficients, continued weighted addition of attributes would cause X to increase indefinitely and thus give a false impression of the degree of separation between the pair of classes.

In other words, each attribute would contribute information about the separation of the classes without respect to the information provided by the other attributes. The incorporation of the correlations between the attributes in the computation of **a** constrains the size of X and ensures that each attribute only contributes the amount of information that is unique to it.

In discriminant analysis the attribute distributions of each class are assumed to be multivariate normally distributed and that the standard deviations are equal. Oftentimes these assumptions are unrealistic in practice, and the technique appears to be sufficiently robust against mild departures from them (Blackith and Reyment 1971).

Rao (1948) adapted the method for $k > 2$ classes, in which case there are always $k - 1$ discriminant functions. Healy (1965), Blackith and Reyment (1971) and Webster and Burrough (1974) are amongst those who have described discriminant analysis in geometric terms. Consider two bivariate classes whose distributions are plotted as ellipses in the *discriminant space* defined by their attribute axes. The classes are best separated by a line that passes through the intersection of the two ellipses. The axis drawn orthogonal to such a line has been called the *discriminant axis* and is the best axis for discriminating between the two classes (Webster and Burrough 1974). The concept can of course be extended into as many attribute dimensions d as are necessary, in which case the ellipses become hyperplanes (Webster and Oliver 1990).

In terms of allocation, distances in the discriminant space are Mahalanobis distances. Thus, new individuals may be allocated to the class to whose centroid it is closest in the Mahalanobis sense (Webster and Burrough 1974).

Discriminant analysis was first employed in soil science by Cox and Martin (1937), who used it to quantify the significance of several soil properties for predicting the presence of *Azotobacter*. It has been used subsequently by Hughes and Lindley (1955), Oertel (1961), Norris and Loveday (1971) and Webster and Burrough (1974), amongst others, for the purpose of allocation. Triantafilis et al. (2003) generalised the theory to a fuzzy linear discriminant analysis.

8.4.3.1 Classification Trees

Classification trees are a hierarchical non-parametric example of supervised classification as they require the set of classes to be known a priori. Because the classes are known a priori, they are readily useful for class allocation. Classification trees can be formulated algorithmically ab initio or by manual extraction of the classification rules in existing classification systems (see Fig. 8.9 for an example). Classification trees recursively subdivide a population of individuals into ever more specific subgroups. Points of subdivision are called nodes, and subdivision is conducted by performing a logical test on the threshold of some attribute. Attributes may be continuous or discrete valued. The logical test is formalised in a *decision rule* that also determines the subgroup implied by passing or failing the logical test. A terminal subgroup is called a leaf and contains a single class rather than a pointer

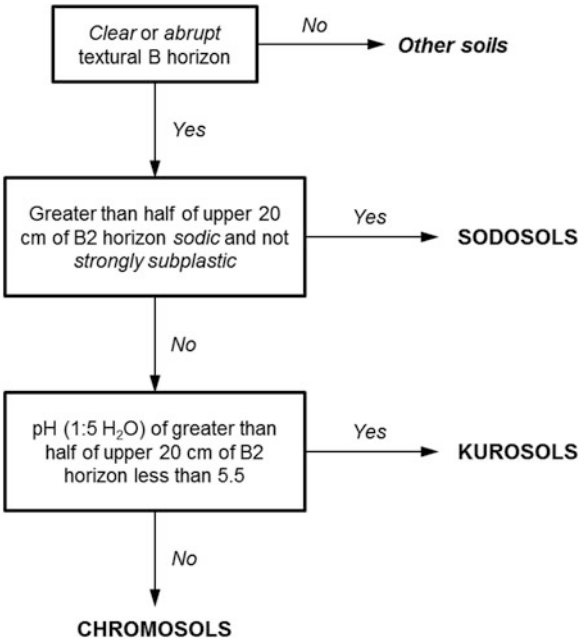

Fig. 8.9 Classification tree for orders of the Australian Soil Classification that possess a clear or abrupt textural B horizon

to a subsequent decision rule. Decision rules are analogous to the if-then rules that are ubiquitous in software programming and may be expressed accordingly. For example, a decision rule may be as follows:

If clay content $\geq$35%: subgroup A
Else subgroup B

When formulated algorithmically, decision rules are usually chosen to optimise some measure of goodness in the subgroups formed by the split, such as within-subgroup purity. According to Friedl and Brodley (1997), a range of metrics can be used to quantify how well this is done; for example, Lagacherie and Holmes (1997) used the Gini index. Pruning may be conducted in order to reduce the size of the tree and to avoid overfitting to the calibration data. Pruning involves merging pairs of leaves and may be accomplished by a cross-validation procedure (Scull et al. 2003).

Classification trees have several advantages over other approaches (Friedl and Brodley 1997). For example, because they are non-parametric, they are insensitive to the distributions of attribute values. They can handle nonlinear relationships between classes and attribute values, and they are able to handle both continuous and discrete-valued attributes. Finally, the tree structure is readily interpretable.

Lagacherie and Holmes (1997) were one of the first to apply classification trees in soil survey, and since then they have become a popular supervised classification tool. It appears, though, that their most frequent use has been in soil survey and mapping rather than in classification of soil material.

8.4.4 Fuzzy k-Means

McBratney (1994) demonstrated how individuals can be allocated to soil classes created using the fuzzy k-means with extragrades algorithm. Allocation is a matter of calculating the degree of membership of the unknown individual to each of the fuzzy classes using Eq. 8.16 and to the extragrade class using Eq. 8.17 described earlier. In order to do this, the parameters k, φ and α must be known, along with the matrix of class centroids and the variance-covariance matrix of the samples used to produce the classification. Furthermore, the Mahalanobis distances between the new individual and the class centroids must also be calculated.

Such an allocation scheme is not exclusively applicable to fuzzy soil classes produced by a numerical soil classification procedure. For example, Mazaheri et al. (1995) employed the technique in order to allocate new profiles to the classes of the Australian Great Soil Groups classification scheme (Stace et al. 1968) Although the Great Soil Groups classification scheme is not a numerical soil classification system, the classes are considered by some (Moore et al. 1983) to be fuzzy in the sense that they are described by a central concept, and, since no taxonomic key has been devised for the purpose of allocation, the class boundaries are somewhat vague.

Despite the fact that the classification scheme is likely to be obsolete now, as the universe of soils it caters for is too small for all Australian conditions (Moore et al. 1983), Mazaheri et al. (1995) reported positive results when they used the fuzzy allocation scheme to allocate six individuals to a class. They determined that the numerical allocation system was not only useful for allocating individuals to classes with partial membership but that it also enables a critical review of the existing classification system in at least two situations: (i) if profiles are allocated with more-or-less equal membership to several classes and (ii) if profiles possess a large extragrade membership.

8.4.5 Soil Horizon Classes

Soil scientists have long recognised soil horizons as the fundamental building blocks of the soil profile. For example, of the eight criteria Marbut (1920) proposed as grounds for the differentiation of soil profiles, seven related to various characteristics of soil horizons.

Attempts have been made to classify soil horizons since the late nineteenth century (FitzPatrick 1967). Dokuchaev was the first to use the A-B-C-horizon nomenclature in his description of Chernozem soils. By itself, though, the A-B-C notation does not convey a lot of information about the character of soil material belonging to a given horizon—or to put it more optimistically, the diversity of soil horizons is greater than can be readily captured by the A-B-C notation (Nikiforoff 1931). For example, a Vertosol B horizon has a very different character to a Podosol B horizon. Horizon subscripts were developed to provide more information in

this respect although the precise meaning of them can vary from jurisdiction to jurisdiction. Bridges (1990) describes a range of other difficulties with the A-B-C system.

The USDA devised a system of diagnostic horizons during its development of the Soil Taxonomy classification system (Soil Survey Staff 1999). Other soil classifications also use their own versions of these horizons (e.g. FAO 2014; Hewitt 2010). Diagnostic horizons are classes of horizons that possess specific soil material characteristics. Their descriptions are often lengthy and complex. They were developed as aids to soil profile classification because specific soil profile classes often require the presence of specific diagnostic horizons. Despite their utility, diagnostic horizons have not been without criticism. For example, FitzPatrick (1976) points out that uncertainty arises in situations where a soil profile contains more than one diagnostic horizon: because the profile is possibly eligible for allocation to more than one soil profile class, a judgement must be made on which diagnostic horizon is more important (or more diagnostic!).

Researchers have since developed more quantitative systems. The best-known and most comprehensive is that of FitzPatrick (e.g. 1993, 1988, 1976, 1967), who devised a system of about 81 classes of soil horizons, which he later called *segments*. Each segment was given a name ending in *-on* and a two-letter code, much like the chemical elements. The codes for a profile's horizons could be assembled to produce a code for the entire profile much like a chemical formula.

Researchers have also used fuzzy numerical classification to create soil layer classes. For example Powell et al. (1992) used the fuzzy *k*-means with extragrades algorithm (de Gruijter and McBratney 1988) to create a set of fuzzy soil layer classes for their study area in the Lockyer Valley in Queensland, Australia. Triantafilis et al. (2001) carried out a similar analysis for soils from the Edgeroi district in the Namoi Valley of New South Wales, Australia. In both cases the researchers found that the numerical soil layer classes were well able to explain pedological and landscape features in their respective study areas.

References

Anderson AJB (1971) Numeric examination of multivariate soil samples. Math Geol 3:1–14

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recogn 46:243–256. doi:10.1016/j.patcog.2012.07.021

Arkley RJ (1976) Statistical methods in soil classification research. Adv Agron 28:37–70. doi:10.1016/S0065-2113(08)60552-0

Banin A, Amiel A (1970) A correlative study of the chemical and physical properties of a group of natural soils of Israel. Geoderma 3:185–198. doi:10.1016/0016-7061(70)90018-2

Beaudette DE, Roudier P, O'Geen AT (2013) Algorithms for quantitative pedology: a toolkit for soil scientists. Comput Geosci 52:258–268. doi:10.1016/j.cageo.2012.10.020

Bensaid AM, Hall LO, Bezdek JC, Clarke LP (1996) Partially supervised clustering for image segmentation. Pattern Recogn 29:859–871. doi:10.1016/0031-3203(95)00120-4

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms, advanced applications in pattern recognition. Plenum Press, New York

Bhattacharya B, Solomatine DP (2006) Machine learning in soil classification. Neural Netw 19:186–195. doi:10.1016/j.neunet.2006.01.005
Bidwell OW, Hole FD (1964a) Numerical taxonomy and soil classification. Soil Sci 97:58–62
Bidwell OW, Hole FD (1964b) An experiment in the numerical classification of some Kansas soils. Soil Sci Soc Am Proc 28:263–268
Blackith RE, Reyment RA (1971) Multivariate morphometrics. Academic Press, London
Bolliger J, Mladenoff DJ (2005) Quantifying spatial classification uncertainties of the historical Wisconsin landscape (USA). Ecography 28:141–156
Bormann H (2010) Towards a hydrologically motivated soil texture classification. Geoderma 157:142–153. doi:10.1016/j.geoderma.2010.04.005
Bouma J (1985) Soil variability and soil survey. In: Nielsen DR, Bouma J (eds) Soil spatial variability: proceedings of a Workshop of the ISSS and SSSA, Las Vegas, USA, 30 November–1 December 1984. Pudoc, Wageningen, pp 130–149
Bragato G (2004) Fuzzy continuous classification and spatial interpolation in conventional soil survey for soil mapping of the lower Piave plain. Geoderma 118:1–16. doi:10.1016/S0016-7061(03)00166-6
Bridges EM (1990) Soil horizon designations (No. Technical Report 19). International Soil Reference and Information Centre, Wageningen
Brus, D.J., de Gruijter, J.J., van Groenigen, J.W., 2006. Designing spatial coverage samples using the k-means clustering algorithm, in: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital soil mapping—an introductory perspective, developments in soil science. Elsevier, pp. 183–192
Bui EN, Moran CJ (2001) Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103:79–94. doi:10.1016/S0016-7061(01)00070-2
Buol, S.W., 2003. Philosophies of soil classifications: from is to does, in: Eswaran, H., Rice, T., Ahrens, R., Stewart, B.A. (Eds.), Soil classification: a global desk reference. CRC Press LLC, pp. 3–10
Burrough PA (1989) Fuzzy mathematical methods for soil survey and land evaluation. J Soil Sci 40:477–492. doi:10.1111/j.1365-2389.1989.tb01290.x
Burrough PA, van Gaans PFM, Hootsmans R (1997) Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77:115–135. doi:10.1016/S0016-7061(97)00018-9
Burrough PA, van Gaans PFM, MacMillan RA (2000) High-resolution landform classification using fuzzy k-means. Fuzzy Sets Syst 113:37–52. doi:10.1016/S0165-0114(99)00011-1
Burrough PA, Wilson JP, van Gaans PFM, Hansen AJ (2001) Fuzzy k-means classification of topoclimatic data as an aid to forest mapping in the greater Yellowstone area, USA. Landsc Ecol 16:523–546. doi:10.1023/A:1013167712622
Campbell JB (1996) Introduction to remote sensing, 2nd edn. The Guilford Press, New York
Campbell NA, Mulcahy MJ, McArthur WM (1970) Numerical classification of soil profiles on the basis of field morphological properties. Aust J Soil Res 8:43–58. doi:10.1071/SR9700043
Celmiņš A (1987) Least squares model fitting to fuzzy vector data. Fuzzy Sets Syst 22:245–269. doi:10.1016/0165-0114(87)90070-4
Clifford HT, Williams WT (1976) Similarity measures. In: Williams WT (ed) Pattern analysis in agricultural science. CSIRO, East Melbourne, pp 37–46
Cline MG (1949) Basic principles of soil classification. Soil Sci 67:81–92
Cockx L, van Meirvenne M, de Vos B (2007) Using the EM38DD soil sensor to delineate clay lenses in a sandy forest soil. Soil Sci Soc Am J 71:1314–1322. doi:10.2136/sssaj2006.0323
Cox GM, Martin WP (1937) Use of a discriminant function for differentiating soils with different Azotobacter populations. Iowa State Coll J Sci 11:323–331
D'Urso P, Giordani P (2006) A weighted fuzzy c-means clustering model for fuzzy data. Computational Statistics & Data Analysis 50:1496–1523. doi:10.1016/j.csda.2004.12.002
Dale MB, McBratney AB, Russell JS (1989) On the role of expert systems and numerical taxonomy in soil classification. J Soil Sci 40:223–234. doi:10.1111/j.1365-2389.1989.tb01268.x

de Bruin S, Stein A (1998) Soil-landscape modelling using fuzzy c-means clustering of attribute data derived from a digital elevation model (DEM). Geoderma 83:17–33. doi:10.1016/S0016-7061(97)00143-2

de Gruijter JJ, McBratney AB (1988) A modified fuzzy k-means method for predictive classification. In: Bock HH (ed) Classification and related methods of data analysis. Elsevier Science Publishers B.V, Amsterdam, pp 97–104

de Gruijter JJ, Walvoort DJJ, van Gaans PFM (1997) Continuous soil maps—a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. Geoderma 77:169–195. doi:10.1016/S0016-7061(97)00021-9

Dobermann A, Witt C, Abdulrachman S, Gines HC, Nagarajan R, Son TT, Tan PS, Wang GH, Chien NV, Thoa VTK, Phung CV, Stalin P, Muthukrishnan P, Ravi V, Babu M, Simbahan GC, Adviento MAA (2003) Soil fertility and indigenous nutrient supply in irrigated rice domains of Asia. Agron J 95:913–923. doi:10.2134/agronj2003.9130

Dodagoudar GR, Venkatachalam G (2000) Reliability analysis of slopes using fuzzy sets theory. Comput Geotech 27:101–115. doi:10.1016/S0266-352X(00)00009-4

Dou C, Woldt W, Bogardi I (1999) Fuzzy rule-based approach to describe solute transport in the unsaturated zone. J Hydrol 220:74–85. doi:10.1016/S0022-1694(99)00065-7

Dubes RC (1987) How many clusters are best?—an experiment. Pattern Recogn 20:645–663. doi:10.1016/0031-3203(87)90034-3

Estabrook GF (1967) An information theory model for character analysis. Taxon 16:86–97. doi:10.2307/1216888

Falkenauer, E., Marchand, A., 2001. Using k-means? Consider ArrayMiner. Presented at the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, Las Vegas

FAO (2014) World reference base for soil resources (No. 106), World soil resources reports. Food and Agriculture Organisation of the United Nations, Rome

Fidêncio PH, Ruisánchez I, Poppi RJ (2001) Application of artificial neural networks to the classification of soils from São Paulo state using near-infrared spectroscopy. Analyst 126:2194–2200. doi:10.1039/B107533K

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugenics 7:179–188

Fisher L, van Ness JW (1971) Admissible clustering procedures. Biometrika 58:91–104. doi:10.1093/biomet/58.1.91

FitzPatrick EA (1967) Soil nomenclature and classification. Geoderma 1:91–105

FitzPatrick EA (1976) Soil horizons and homology. Classif Soc Bull 3:68–89

FitzPatrick EA (1988) Soil horizon designation and classification (Technical Paper No. 17), Technical paper. International Soil Reference and Information Centre, Wageningen

FitzPatrick EA (1993) Principles of soil horizon definition and classification. Catena 20:395–402. doi:10.1016/S0341-8162(05)80005-0

Friedl MA, Brodley CE (1997) Decision tree classification of land cover from remotely sensed data. Remote Sens Environ 61:399–409. doi:10.1016/S0034-4257(97)00049-7

Grigal DF, Arneman HF (1969) Numerical classification of some forested Minnesota soils. Soil Sci Soc Am Proc 33:433–438. doi:10.2136/sssaj1969.03615995003300030029x

Gibbons, F.R., 1968. Limitations to the usefulness of soil classification. Presented at the 9th International Congress of Soil Science, Adelaide, South Australia, pp. 159–167

Goodenough, D., Shlien, S., 1974. Results of cover-type classification by maximum likelihood and parallelepiped methods, in: Proceedings of the Second Canadian Symposium on Remote Sensing. Presented at the Second Canadian Symposium on Remote Sensing, Canadian Remote Sensing Society, University of Guelph, pp. 136–164

Gower JC (1967) A comparison of some methods of cluster analysis. Biometrics 23:623–637. doi:10.2307/2528417

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–874

Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. J Intell Inf Syst 17:107–145. doi:10.1023/A:1012801612483
Hartigan JA (1975) Clustering algorithms, probability and mathematical statistics. Wiley, New York
Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. J R Stat Soc Ser C (Appl Stat) 28:100–108
Healy, M.J.R., 1965. Descriptive uses of discriminant functions. In: Mathematics and computer science in biology and medicine: Proceedings of a Conference Held by Medical Research Council in Association with the Health Departments, Oxford, July 1964. Her Majesty's Stationery Office, London, 93–102
Heuvelink GBM, Burrough PA (1993) Error propagation in cartographic modelling using Boolean logic and continuous classification. Int J Geogr Inf Syst 7:231–246. doi:10.1080/02693799308901954
Hewitt AE (2010) New Zealand soil classification, Landcare Research Science Series, vol 1, 3rd edn. Manaaki Whenua Press, Lincoln. doi:10.7931/DL1-LRSS-1-2010
Hole FD, Hironaka M (1960) An experiment in ordination of some soil profiles. Soil Sci Soc Am Proc 24:309–312
Holmgren, G.G.S., 1986. The soil individual. Presented at the Proceedings of the 13th World Congress of Soil Science, Hamburg, pp. 1146–1147
Holmgren GGS (1988) The point representation of soil. Soil Sci Soc Am J 52:712–716
Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–441. doi:10.1037/h0071325
Hughes RE, Lindley DV (1955) Application of biometric methods to problems of classification in ecology. Nature 175:806–807. doi:10.1038/175806a0
Hughes PA, McBratney AB, Minasny B, Campbell S (2014) End members, end points and extragrades in numerical soil classification. Geoderma 226–227:365–375. doi:10.1016/j.geoderma.2014.03.010
Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bull Soc Vaud Sci Nat 44:223–270. doi:10.5169/seals-268384
Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recogn Lett 31:651–666. doi:10.1016/j.patrec.2009.09.011
Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31:264–323. doi:10.1145/331499.331504
Jensen JR (1978) Digital land cover mapping using layered classification logic and physical composition attributes. Am Cartographer 5:121–132. doi:10.1559/152304078784022917
Kantardzic, M., 2011. Preparing the data, data mining: concepts, models, methods, and algorithms. Wiley, Hoboken, 26–52
Kaya T, Kahraman C (2011) Fuzzy multiple criteria forestry decision making based on an integrated VIKOR and AHP approach. Expert Syst Appl 38:7326–7333. doi:10.1016/j.eswa.2010.12.003
Kim M, Ramakrishna RS (2005) New indices for cluster validity assessment. Pattern Recogn Lett 26:2353–2363. doi:10.1016/j.patrec.2005.04.007
Knox EG (1965) Soil individuals and soil classification. Soil Sci Soc Am Proc 29:79–84
Kyuma K, Kawaguchi K (1976) Soil material classification for paddy soils in Japan. Soil Sci Plant Nutr 22:111–124. doi:10.1080/00380768.1976.10432973
Lagacherie P, Holmes S (1997) Addressing geographical data errors in a classification tree for soil unit prediction. Int J Geogr Inf Sci 11:183–198. doi:10.1080/136588197242455
Lagacherie P, Cazemier DR, van Gaans PFM, Burrough PA (1997) Fuzzy k-means clustering of fields in an elementary catchment and extrapolation to a larger area. Geoderma 77:197–216. doi:10.1016/S0016-7061(97)00022-0
Lark RM (2000) Designing sampling grids from imprecise information on soil variability, an approach based on the fuzzy kriging variance. Geoderma 98:35–59. doi:10.1016/S0016-7061(00)00051-3

Leeper GW (1956) The classification of soils. J Soil Sci 7:59–64. doi:10.1111/j.1365-2389.1956.tb00861.x
Leone AP, Sommer S (2000) Multivariate analysis of laboratory spectra for the assessment of soil development and soil degradation in the southern Apennines (Italy). Remote Sens Environ 72:346–359. doi:10.1016/S0034-4257(99)00110-8
Ließ M (2015) Sampling for regression-based digital soil mapping: closing the gap between statistical desires and operational applicability. Spat Stat 13:106–122. doi:10.1016/j.spasta.2015.06.002
Liu M, Samal A (2002a) A fuzzy clustering approach to define agroecozones. Ecol Model 149:215–228. doi:10.1016/S0304-3800(01)00446-X
Liu M, Samal A (2002b) Cluster validation using legacy delineations. Image Vis Comput 20:459–467
Lloyd SP (1982) Least squares optimization in PCM. IEEE Trans Inf Theory 28:129–137. doi:10.1109/TIT.1982.1056489
MacVicar CN (1969) A basis for the classification of soil. J Soil Sci 20:141–152. doi:10.1111/j.1365-2389.1969.tb01563.x
Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci India 2:49–55
Mallavan, B.P., Minasny, B., McBratney, A.B., 2010. Homosoil, a methodology for quantitative extrapolation of soil information across the globe, In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S., McBratney, A.B. (Eds.), Digital soil mapping: bridging research, environmental application, and operation, progress in soil science. Springer Science+Business Media B.V., pp. 137–149. doi:10.1007/978-90-481-8863-5_12
Mao J, Jain AK (1996) A self-organizing network for hyperellipsoidal clustering (HEC). IEEE Trans Neural Netw 7:16–29. doi:10.1109/72.478389
Marbut, C.F., 1920. The contribution of soil surveys to soil science, In: Proceedings of the Forty-First Annual Meeting of the Society for the Promotion of Agricultural Science. Presented at the Forty-First Annual Meeting of the Society for the Promotion of Agricultural Science, Springfield, Massachusetts, pp. 116–142
Mather PM, Koch M (2011) Computer processing of remotely-sensed images: an introduction, 4th edn. Wiley, Chichester
Mazaheri SA, Koppi AJ, McBratney AB (1995) A fuzzy allocation scheme for the Australian great soil groups classification system. Eur J Soil Sci 46:601–612
McBratney AB (1994) Allocation of new individuals to continuous soil classes. Aust J Soil Res 32:623–633. doi:10.1071/SR9940623
McBratney AB, de Gruijter JJ (1992) A continuum approach to soil classification by modified fuzzy k-means with extragrades. J Soil Sci 43:159–175. doi:10.1111/j.1365-2389.1992.tb00127.x
McBratney AB, Moore AW (1985) Application of fuzzy sets to climatic classification. Agric For Meteorol 35:165–185. doi:10.1016/0168-1923(85)90082-6
McBratney AB, de Gruijter JJ, Brus DJ (1992) Spacial prediction and mapping of continuous soil classes. Geoderma 54:39–64. doi:10.1016/0016-7061(92)90097-Q
McKenzie NJ, Austin MP (1993) A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57:329–355. doi:10.1016/0016-7061(93)90049-Q
Metternicht GI (2003) Categorical fuzziness: a comparison between crisp and fuzzy class boundary modelling for mapping salt-affected soils using Landsat TM data and a classification based on anion ratios. Ecol Model 168:371–389. doi:10.1016/S0304-3800(03)00147-9
Milligan GW, Cooper MC (1985) An examination of the procedures for determining the number of clusters in a data set. Psychometrika 50:159–179. doi:10.1007/BF02294245
Minasny B, McBratney AB (2006) Mechanistic soil-landscape modelling as an approach to developing pedogenetic classifications. Geoderma 133:138–149. doi:10.1016/j.geoderma.2006.03.042
Moore AW, Russell JS (1967) Comparison of coefficients and grouping procedures in numerical analysis of soil trace element data. Geoderma 1:139–158. doi:10.1016/0016-7061(67)90006-7

Moore, A.W., Isbell, R.F., Northcote, K.H., 1983. Classification of Australian soils, In: Soils: an Australian viewpoint. CSIRO, Melbourne, pp. 253–266

Muir JW, Hardie HGM, Inkson RHE, Anderson AJB (1970) The classification of soil profiles by traditional and numerical methods. Geoderma 4:81–90. doi:10.1016/0016-7061(70)90035-2

Nikiforoff CC (1931) History of A, B, C. Bull Am Soil Surv Assoc 12:67–70

Norris JM (1971) The application of multivariate analysis to soil studies. I Grouping of soils using different properties. J Soil Sci 22:69–80. doi:10.1111/j.1365-2389.1971.tb01594.x

Norris JM, Loveday J (1971) The application of multivariate analysis to soil studies. II The allocation of soil profiles to established groups: a comparison of soil survey and computer method. J Soil Sci 22:395–400. doi:10.1111/j.1365-2389.1971.tb01625.x

Odeh IOA, McBratney AB, Chittleborough DJ (1990) Design of optimal sample spacings for mapping soil using fuzzy-k-means and regionalized variable theory. Geoderma 47:93–122. doi:10.1016/0016-7061(90)90049-F

Odeh IOA, McBratney AB, Chittleborough DJ (1992) Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. Soil Sci Soc Am J 56:505–516. doi:10.2136/sssaj1992.03615995005600020027x

Odgers NP, McBratney AB, Minasny B (2011) Bottom-up digital soil mapping. I Soil layer classes. Geoderma 163:38–44. doi:10.1016/j.geoderma.2011.03.014

Oertel AC (1961) Chemical discrimination of terra rossas and rendzinas. J Soil Sci 12:111–118. doi:10.1111/j.1365-2389.1961.tb00901.x

Oliver MA, Webster R (1989) A geostatistical basis for spatial weighting in multivariate classification. Math Geol 21:15–35. doi:10.1007/BF00897238

Powell B, McBratney AB, MacLeod DA (1992) Fuzzy classification of soil profiles and horizons from the Lockyer Valley, Queensland, Australia. Geoderma 52:173–197. doi:10.1016/0016-7061(92)90082-I

Rao CR (1948) The utilization of multiple measurements in problems of biological classification. J R Stat Soc Ser B (Methodol) 10:159–193

Rayner JH (1966) Classification of soils by numerical methods. J Soil Sci 17:79–92. doi:10.1111/j.1365-2389.1966.tb01454.x

Ribeiro MV, Cunha LMS, Camargo HA, Rodrigues LHA (2014) Applying a fuzzy decision tree approach to soil classification. In: Laurent A, Strauss O, Bouchon-Meunier B, Yager RR (eds) Information processing and management of uncertainty in knowledge-based systems, communications in computer and information science. Springer, Cham, pp 87–96

Robinove, C.J., 1979. Integrated terrain mapping with digital landsat images in Queensland, Australia (Geological Survey Professional Paper No. 1102). United States Geological Survey, Washington, D.C.

Roubens M (1982) Fuzzy clustering algorithms and their cluster validity. Eur J Oper Res 10:294–301. doi:10.1016/0377-2217(82)90228-4

Roudier, P., Manderson, A., Hedley, C., 2016. Advances towards quantitative assessments of soil profile properties, In: Hartemink, A.E., Minasny, B. (Eds.), Digital soil Morphometrics, progress in soil science. Springer International, pp. 113–132. doi:10.1007/978-3-319-28295-4_8

Rousseeuw PJ, Kaufman L, Trauwaert E (1996) Fuzzy clustering using scatter matrices. Comput Stat Data Anal 23:135–151. doi:10.1016/S0167-9473(96)00026-6

Saboya Jr F, da Glória Alves M, Dias Pinto W (2006) Assessment of failure susceptibility of soil slopes using fuzzy logic. Eng Geol 86:211–224. doi:10.1016/j.enggeo.2006.05.001

Sarkar PK, Bidwell OW, Marcus LF (1966) Selection of characteristics for numerical classification of soils. Soil Sci Soc Am Proc 30:269–272. doi:10.2136/sssaj1966.302269x

Schulz K, Huwe B (1999) Uncertainty and sensitivity analysis of water transport modelling in a layered soil profile using fuzzy set theory. J Hydroinf 1:127–138

Scull P, Franklin J, Chadwick OA, McArthur D (2003) Predictive soil mapping: a review. Prog Phys Geogr 27:171–197. doi:10.1191/0309133303pp366ra

Simonson, R.W., Gardiner, D.R., 1960. Concept and functions of the pedon. Presented at the 7th International Congress of Soil Science, pp. 127–131

Sneath PHA, Sokal RR (1962) Numerical taxonomy. Nature 193:855–860. doi:10.1038/193855a0
Sneath PHA, Sokal RR (1973) Numerical taxonomy: the principles and practice of numerical classification, A series of books in biology. W. H. Freeman and Company, San Francisco
Soil Survey Staff (1993) Soil Survey Manual, U. S. Department of Agriculture Handbook 18. United States Department of Agriculture Soil Conservation Service
Soil Survey Staff (1999) Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys, 2nd ed. United States Department of Agriculture Natural Resources Conservation Service
Soil Survey Staff (2016) Soil survey geographic (SSURGO) database. United States Department of Agriculture Natural Resources Conservation Service. https://sdmdataaccess.sc.egov.usda.gov. Accessed 9 Dec 2016
Sokal RR (1961) Distance as a measure of taxonomic similarity. Syst Zool 10:70–79
Sokal RR (1963) The principles and practice of numerical taxonomy. Taxon 12:190–199
Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38:1409–1438
Sokal RR, Sneath PHA (1963) Principles of numerical taxonomy, A series of books in biology. W. H. Freeman and Company, San Francisco
Stace HCT, Hubble GD, Brewer R, Northcote KH, Sleeman JR, Mulcahy MJ, Hallsworth EG (1968) A handbook of Australian soils. Rellim Technical Publications, Glenside
Steinley D (2006) K-means clustering: a half-century synthesis. Br J Math Stat Psychol 59:1–34. doi:10.1348/000711005X48266
Sun X-L, Zhao Y-G, Wang H-L, Yang L, Qin C-Z, Zhu A-X, Zhang G-L, Pei T, Li B-L (2012) Sensitivity of digital soil maps based on FCM to the fuzzy exponent and the number of clusters. Geoderma 171–172:24–34. doi:10.1016/j.geoderma.2011.03.016
Triantafilis J, Ward WT, Odeh IOA, McBratney AB (2001) Creation and interpolation of continuous soil layer classes in the lower Namoi valley. Soil Sci Soc Am J 65:403–413
Triantafilis J, Odeh IOA, Minasny B, McBratney AB (2003) Elucidation of physiographic and hydrogeological features of the lower Namoi valley using fuzzy k-means classification of EM34 data. Environ Model Softw 18:667–680. doi:10.1016/S1364-8152(03)00053-7
Valeriano MM, Epiphanio JCN, Formaggio AR, Oliveira JB (1995) Bi-directional reflectance factor of 14 soil classes from Brazil. Int J Remote Sens 16:113–128. doi:10.1080/01431169508954375
van Alphen BJ, Stoorvogel JJ (2000) A functional approach to soil characterization in support of precision agriculture. Soil Sci Soc Am J 64:1706–1713. doi:10.2136/sssaj2000.6451706x
Verheyen K, Adriaens D, Hermy M, Deckers S (2001) High-resolution continuous soil classification using morphological soil profile descriptions. Geoderma 101:31–48. doi:10.1016/S0016-7061(00)00088-4
Verma P, Singh P, George KV, Singh HV, Devotta S, Singh RN (2009) Uncertainty analysis of transport of water and pesticide in an unsaturated layered soil profile using fuzzy set theory. Appl Math Model 33:770–782. doi:10.1016/j.apm.2007.12.004
Vrindts E, Mouazen AM, Reyniers M, Maertens K, Maleki MR, Ramon H, de Baerdemaeker J (2005) Management zones based on correlation between soil compaction, yield and crop data. Biosyst Eng 92:419–428. doi:10.1016/j.biosystemseng.2005.08.010
Webster R (1968) Fundamental objections to the 7th approximation. J Soil Sci 19:354–366. doi:10.1111/j.1365-2389.1968.tb01546.x
Webster R (1977) Quantitative and numerical methods in soil classification and survey. Oxford University Press, Oxford
Webster R, Burrough PA (1974) Multiple discriminant analysis in soil survey. J Soil Sci 25:120–134. doi:10.1111/j.1365-2389.1974.tb01109.x
Webster R, Oliver MA (1990) Statistical methods in soil and land resource survey, spatial information systems. Oxford University Press, Oxford
Williams WT (1976a) Attributes. In: Williams WT (ed) Pattern analysis in agricultural science. CSIRO, Melbourne, pp 31–36

Williams WT (1976b) Types of classification. In: Williams WT (ed) Pattern analysis in agricultural science. CSIRO, East Melbourne, pp 76–83

Xie XL, Beni G (1991) A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell 13:841–847. doi:10.1109/34.85677

Yang M-S, Ko C-H (1996) On a class of fuzzy c-numbers clustering procedures for fuzzy data. Fuzzy Sets Syst 84:49–60. doi:10.1016/0165-0114(95)00308-8

Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353. doi:10.1016/S0019-9958(65)90241-X

Zhang S, Zhu A-X, Liu W, Liu J, Yang L (2013) Mapping detailed soil property using small scale soil type maps and sparse typical samples. Chin Geogr Sci 23:680–691. doi:10.1007/s11769-013-0632-7