Computers and Electronics in Agriculture 176 (2020) 105631

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

A new approach to clustering soil profile data using the modified distance matrix

Vakhtang Shelia[a,b,*], Gerrit Hoogenboom[a,b]

[a] Agricultural and Biological Engineering Department, University of Florida, Gainesville, FL 32611, USA
[b] Institute for Sustainable Food Systems, University of Florida, Gainesville, FL 32611, USA

ARTICLE INFO

Keywords:
Soil horizon
Numerical classification
Hierarchical algorithm
Ward method
WISE

ABSTRACT

The application of different data mining methods for large soil profile datasets can be very useful for many agricultural and natural resource management applications, ranging from crop modeling to soil taxonomy. Distance or dissimilarity measures are key features of these methods. The proximity measure or the distance between vectors is calculated when they have the same dimensions. In the case of the soil profile data, the corresponding matrices representing different soils and their horizon properties usually have different dimensions. The objectives of this study were to explore a new approach for creating a semi-metric based on adjustment of the soil profile horizons, implement it in the computer application and to apply the modified distance matrix calculation to maximize the use of soil horizon properties for soil data mining. We assume that each soil horizon is homogenous while a vertical heterogeneity of soil profile is expressed through different soil horizons. Therefore, any sublayers of a horizon are characterized with the same values for its attributes as the horizon itself. In our approach, we developed matrices with the same dimension for soil profiles and calculated the proximity measure. The algorithm was implemented as an easy to use Fortran application that can calculate the modified distance matrix (MDM) for low- and high-dimensional soil profiles data. The proposed approach was shown to be effective when using existing reliable datasets, such as WISE Version 3.1, a global soil profile database developed by ISRIC. Hierarchical clustering was performed using the MDM based on the original algorithm of soil profile horizons adjustment with further integration into R. The principal finding shows that a proposed modified distance matrix can be used with different clustering methods for soil profile data clustering on a horizon-by-horizon basis. This study established a new methodology for using the modified distance matrix calculation and applying it with different clustering algorithms to large sets of soil profile data obtained from detailed soil surveys.

1. Introduction

Cluster Analysis (CA) is a data mining or data exploration method for unsupervised classification of patterns represented by observations, data items, or feature vectors for identifying natural groups within a set of entities. It is used to detect objects that share some selected property (e.g., time series, spatial patterns, soil profiles, etc.) and to create homogeneous groups of objects based on the observed similarities between the characteristics of these objects as described in the dataset (Everitt et al., 2001; Hair et al., 2009). CA is applicable to any type of data, and only a proximity matrix for a dataset is needed. This type of analysis is easy to interpret and has intuitive mathematical principles (Bandyopadhyay and Saha, 2013; King, 2015).

Various clustering and classification algorithms have been used for data mining for different applications across a wide range of disciplines, including economics, biology, meteorology, agriculture, soil science, and many others. Numerical soil classification began soon after the invention of the computer. For the first time, similarities between soil profiles were calculated and then were converted into distances, which subsequently enabled the reduction of multi-dimensional space, a process known as ordination (Hole and Hironaka, 1960; Rayner, 1966; Fitzpatrick, 1967; Moore and Russell, 1967). Since its inception, significant research has been conducted on numerical soil classification. Some classifications were created hierarchically (Webster and

Abbreviations: ISRIC, International Soil Reference and Information Centre; WISE, World Inventory of Soil Emission Potentials; FAO, Food and Agriculture Organization of the United Nations; UNESCO, United Nations Educational, Scientific and Cultural Organization; DSSAT, The Decision Support System for Agrotechnology Transfer

* Corresponding author at: 233 Frazier Rogers Hall, PO Box 110570, Gainesville, FL 32611-0570, USA.
E-mail address: vakhtang.shelia@ufl.edu (V. Shelia).

https://doi.org/10.1016/j.compag.2020.105631
Received 23 May 2020; Received in revised form 8 July 2020; Accepted 8 July 2020

Burrough, 1972) and others non-hierarchically (Crommelin and de Gruijter, 1973; Little and Ross, 1985) using various clustering algorithms. Later, by recognizing the intergrading nature of the soil populations, algorithms were applied using concepts of fuzzy sets and fuzzy-k means (Burrough et al., 1992; Lagacherie et al., 1997), and fuzzy-k means with extra grades (McBratney and de Gruijter, 1992; Mazaheri et al., 1995; Carré and Jacobson, 2009; Hughes et al., 2014). A hierarchical cluster analysis was used to group cone index layer profiles and to form groups that had similar penetrations, resulting in pedotransfer functions for cone penetrometers (Grunwald et al., 2001a; 2001b).

Three soil profile distances built from pedological, utilitarian and joint points of view were implemented in the Java Web Application, called OSACA (an acronym for "Outil Statistique d'Aide à la Cartogénèse Automatique") in order to allocate profiles to existing classes or to create a new classification of the profiles (Carré and Jacobson, 2009). Beaudette et al. (2013) provided a comparison of various existing soil profiles clustering approaches and developed a package called "aqp" (algorithms for quantitative pedology) for R that supports the analysis of soil databases. This package was designed to support data-driven approaches to common soils-related tasks such as visualization, aggregation, and classification of soil profile collections. In addition, Beaudette et al. (2013) sought to advance the study of numerical soil classification by building on previously published methods within an extensible and open source framework.

Clustering approaches as nonparametric methods have been successfully used in other applications as well. These techniques do not use any predefined mathematical functions and are based on similarities instead of fitting equations to data. Nemes et al. (2006, 2008) introduced a nonparametric lazy learning algorithm based on the k-Nearest Neighbor algorithm (k-NN) and they used this algorithm to estimate soil hydraulic properties. The technique showed little sensitivity in terms of how many nearest soils were selected and how those were weighed while formulating the output of the algorithm, as long as extremes were avoided. They found that the k-NN technique is a competitive alternative to other techniques for developing pedotransfer functions (PTFs).

Following the study by Gijsman et al. (2002) that compared eight methods for estimating water-retention parameters, Jagtap et al. (2004) developed a new approach based on the k-NN to determine values for missing soil properties. This procedure used a database of field-measured soil–water-retention data in relation to the soil physical properties of the sample. The soil most similar in texture and organic carbon concentration was considered the 'nearest neighbor' among all the soils in the database. These soil–water-retention values were assumed to be similar to those that needed to be estimated. To avoid estimating soil–water-retention values only based on one soil in the database, the six 'nearest neighbors' were used and weighted according to their degree of similarity. Gijsman et al. (2007) used this approach to convert 1,125 soil profiles from the international global soil database into a format that could be used as input data for commonly used biophysical computer models, such as the crop simulation models within the Decision Support System for Agrotechnology Transfer (DSSAT) (Jones et al., 2003; Hoogenboom et al., 2019).

The approach applied by Jagtap et al. (2004) averages the values of soil properties by depths, thereby implying that all soil information is contained in a single horizon. Therefore, when comparing soils, the approach does not consider soil as a profile that consists of a set of different horizons with various soil property values. In addition, the approach does not take into account the thickness of the soil profile horizons and the overall depth of the profile. The varying number of soil horizons makes it difficult to conduct a direct comparison between profiles. To resolve this issue, Jones and Thornton (2015) selected the most representative profile for each soil type of the WISE Database and used only three horizons from each soil profile, namely, the top horizon, the bottom horizon, and the horizon closest to the center. For soil profiles that consist of only one horizon, the single horizon was triplicated; for soil profiles that have only two horizons, the second horizon was duplicated. This repetition ensured consistency of the distance measure across profiles and simplified programming to only three horizons. All variables were normalized by dividing by the range; the average of each profile set for each soil type was calculated, and then the Euclidean distance from the average was calculated for each profile. The profile with the minimum distance from the mean was chosen as the most representative profile. However, this approach limited soil profiles to only three horizons and by duplicating horizons and creating additional ones, it goes beyond the existing soil profile depth.

Applying different data mining methods including CA, to large soil profile datasets can be very useful for agricultural and natural resource management applications, ranging from crop modeling to soil taxonomy. Many of these applications include modeling of water and solute transport processes, and they require detailed information about soil properties by horizon, which is not always readily available. Many of these properties can be estimated indirectly from soil texture and other properties included in soil databases that are available for many regions across the globe. Soil properties can also be estimated based on the homogeneous group of soil profiles in the absence of laboratory-measured soil physical and chemical properties (Minasny and McBratney, 2007). Therefore, soil profile data clustering could be very useful for these types of applications.

A definition of the distance or similarity between pairs of objects is necessary for CA. Given $\boldsymbol{p}$ objects, each composed of m × n elements organized in a data matrix X of dimension m × n, one commonly used measure of the proximity between two objects is simply the Euclidian Distance, which assumes that these two objects have the same dimension. For example, the World Inventory of Soil Emission Potentials (WISE) Harmonized Global Soil Profile Dataset (WISE3, ver. 3.1) of the International Soil Reference and Information Centre (ISRIC) (Batjes, 2008, 2009) contains data for 47,833 soil horizons for a total of 10,253 soil profiles. The number of soil horizons per profile ranges from 1 to 14. Of all the profiles, 2,604 have 5 horizons; 2,488 profiles have 4 horizons; 1,781 profiles have 3 horizons; and 1,742 profiles have 6 horizons. If we develop a matrix for each soil profile where the columns represent the attribute values for all horizons and rows are all attribute values by horizon, then the matrices will have a different number of rows and columns for the different soil profiles and, thus, a different dimension. As an example, let us consider two soil profiles: the first soil profile with 5 horizons (first dimension) and 6 attributes (second dimension) (a matrix of dimension 5 × 6) and the second soil profile with 6 horizons and 6 attributes (a matrix of dimension 6 × 6). Generally, we might have the same number of attributes as the second dimension, but the first dimension most probably will be different due to the different number of soil horizons. Therefore, the calculation of the distance, for example of the Euclidian Distance, is not possible.

In addition, the rows in the matrix might represent soil horizons that have different lower depths. In WISE3 the thickness of the soil profile horizons ranges from 0.01 to 4.30 m. Among all soil horizons, the majority (5,234) have a thickness of around 0.20 m, 4,169 horizons have a thickness of around 0.16 m, and 3,264 horizons have a thickness of around 0.30 cm, while the average thickness of a soil horizon is 0.28 m. Finally, the matrices might represent soil profiles that have a different soil depth. For example, in WISE3, the soil depth of the profiles ranges from 0.01 to 8.50 m.

The concepts of distance and similarity are crucial for many scientific applications. Distance and similarity measure are mostly needed to compute the distances or similarities between different objects, and are essential requirements in almost all pattern recognition applications, including classification, clustering, feature selection, outlier detection, regression, and search. A large number of distnace and similarity measures exist in the literature; thus, selecting one most appropriate for a particular task is a crucial issue (Bandyopadhyay and Saha, 2013; King, 2015). In the case of the soil profile data when comparing soil

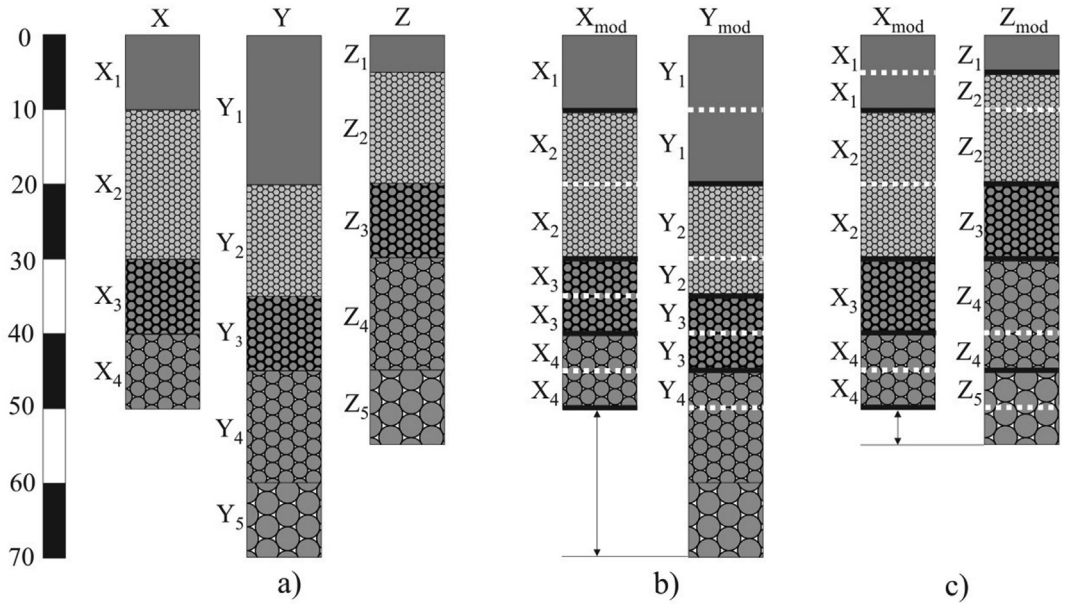

Fig. 1. Schematic representation of soil texture profiles X, Y, Z with different lower depth (cm) of horizons (a); adjusted soil profiles X_{mod}, Y_{mod} when calculating the modified distance between soil profiles X and Y (b), and adjusted soil profiles X_{mod}, Z_{mod} when calculating the modified distance between soil profiles X and Z (c). X_1 is the vector of attributes of the first horizon of X and X_{mod} soil profiles, X_2 - for the next horizon, etc. Y_1 is the vector of attributes of the first horizon of Y and Y_{mod} soil profiles, Y_2 – for the next horizon, etc. Z_1 is the vector of attributes of the first horizon of Z and Z_{mod} soil profiles, Z_2 – for the next horizon, etc.

profiles and estimating either the distance or similarity, one should take into account the properties of the individual horizons, the different number of horizons, the different depths for each horizon, and the soil depth of each profile. Therefore the objectives of this study were (i) to develop a new approach for defining the distance or dissimilarity measure between soil profiles based on horizon-by-horizon properties; (ii) to implement this approach in an algorithm and as a computer application; and (iii) to demonstrate its effectiveness for soil profile clustering using a comprehensive soil profile database.

2. Materials and methods

2.1. The Euclidian distance and soil horizons

The success of any pattern recognition technique largely depends on the choice of the proximity measure, which varies depending on the type of data used (Bandyopadhyay and Saha, 2013). The Euclidian Distance (ED) is the most commonly used dissimilarity measure for quantitative variables when vectors of the properties that characterize the objects have the same dimension. The ED between $\boldsymbol{p}$ dimensional vectors $X = (x_1, x_2, \cdots, x_p)$ and $Y = (y_1, y_2, \cdots, y_p)$ is defined as follows:

$$d(X, Y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \quad (1)$$

If measurement data for the soil profile attributes are provided for the individual soil horizons, then we might have X and Y vectors of dimension p, where $\boldsymbol{p} = \boldsymbol{m} \times \boldsymbol{n}$ and:

$$X = (x_{11}, x_{12}, \cdots, x_{1n}, x_{21}, x_{22}, \cdots, x_{2n}, \cdots, x_{m1}, x_{m2}, \cdots, x_{mn})$$

$$Y = (y_{11}, y_{12}, \cdots, y_{1n}, y_{21}, y_{22}, \cdots, y_{2n}, \cdots, y_{m1}, y_{m2}, \cdots, y_{mn}) \quad (2)$$

Here, $\boldsymbol{n}$ is the number of soil attributes; $\boldsymbol{m}$ is the number of soil horizons; $x_{11}, x_{12}, \cdots, x_{1n}$ are the first soil horizon attributes values of the first soil profile, etc., $y_{11}, y_{12}, \cdots, y_{1n}$ are the first soil horizon attributes values of the second soil profile, etc.

In this case, the ED can be expressed as follows:

$$d(X, Y) = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - y_{ij})^2} \quad (3)$$

Soil properties vary significantly horizontally across the landscape as well vertically through the soil profile. Therefore, using discrete intervals or genetic horizons to measure soil profile properties has been the de facto mode of operation in soil science for two main reasons. First, soils often have clear horizons that appear uniform in their texture, color, or structure. Second, field sampling and laboratory analysis are very costly. Thus, horizon-based sampling offers a reasonable compromise between the analysis cost and the information obtained (Myers et al., 2011). Soil datasets often consist of samples collected from many different locations and at different depth intervals. Studies may consider fixed intervals (e.g., 0–10 cm, 10–20 cm, and 20–30 cm), or they may use sampling intervals that are defined according to soil horizons. In this study, we use the term soil horizon in a generic way to refer to the fixed intervals in the depth of the soil profile; we also use the term in certain situations to refer to genetic horizons.

A common problem with soil profile data is that the vectors that represent measurement data for the soil profile attributes might have different numbers of elements and, thus, dimensions, and that the data from the horizons of the two soil profiles could refer to a different horizon thickness or soil depth. For the sake of convenience in using this methodology, we reconstruct the presented X and Y vectors (2) to matrices of $\boldsymbol{m} \times \boldsymbol{n}$ dimension as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix} \quad (4)$$

Here, each row represents attributes values of one horizon of a soil, and each column represents one soil attribute value for all horizons of a soil profile.

If the two matrices of Eq. (4) have the same dimension m × n, we can calculate the ED between them using the Eq. (3), where $\boldsymbol{i}$ and $\boldsymbol{j}$ are indices for horizons and the attributes, respectively. Here, m is the number of horizons; n is the number of attributes; and $\boldsymbol{x_{ij}}$ and $\boldsymbol{y_{ij}}$ are the standardized values of $\boldsymbol{j}^{th}$ attribute of $\boldsymbol{i}^{th}$ horizon point. It should be

noted that each attribute has to be standardized to minimize the scale difference. Thus, if we represent data for different soil profiles as matrices, then they generally have different dimensions because of the different number of the soil horizons for each profile and different lower depths of the individual soil horizons. To visualize this, Fig. 1a depicts the schematic representation for three different soil profiles X, Y, and Z. Profile X has 4 horizons with a lower depth of 10, 30, 40, and 50 cm, respectively, and X_1, X_2, X_3, and X_4 vectors of attributes by horizons. Here, $X_1 = (x_{11}, x_{12}, \cdots, x_{1n})$ is a vector of attributes for the first horizon; n is the total number of attributes, etc. Soil profile Y has 5 horizons with a lower depth of 20, 35, 45, 60, and 70 cm, respectively, and Y_1, Y_2, Y_3, Y_4, and Y_5 vectors of attributes. Here, $Y_1 = (y_{11}, y_{12}, \cdots, y_{1n})$ is a vector of attributes for its first horizon, etc. Soil profile Z has 5 horizons with a lower depth of 5, 20, 30, 35, and 45 cm, respectively, and accordingly Z_1, Z_2, Z_3, Z_4, and Z_5 vectors of attributes. Here, $Z_1 = (z_{11}, z_{12}, \cdots, z_{1n})$ are the attribute values for the first soil horizon, etc. In this case we have the following three matrices: X of dimension 4 × n, Y of dimension 5 × n, and Z of dimension 5 × n. Although the matrices Y and Z have the same dimension, we cannot compare a horizon that has Y_1 vector of attributes and a lower depth of 20 cm with a horizon that has Z_1 vector of attributes and 5 cm lower depth. Therefore, we cannot calculate the ED because of differences in lower depth of the soil horizons or because of a difference in horizon thickness. Accordingly, we developed a new approach to calculate the distance measure or dissimilarity in order to overcome the problem related to different dimensions of the matrices representing soil profiles and different horizon thickness.

2.2. Pair-wise soil profiles adjustment

In this study, each soil profile and the associated attributes for each horizon are identified with a matrix of dimension $\boldsymbol{m} \times \boldsymbol{n}$ as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad (5)$$

Here, $\boldsymbol{m}$ is the number of the soil horizons and $\boldsymbol{n}$ is the number of the soil profile attributes.

In general, normalization and standardization strategies are applied to not only scale data, but also to remove certain systematic biases that are inherent in the data. The biases are caused by the dependencies between attributes, which might not have a normal distribution. In normalization, where each attribute is treated independently, current methodologies include min–max normalization, z-score standardization, and normalization by decimal scaling (Dua and Chowriappa, 2013). It is advantageous to use z-score standardization when the dataset includes extensive outliers (Han and Kamber, 2006). Thus, for the first step we use z-score standardization for all attributes according to the following equation:

$$x_{ij,std} = \frac{x_{ij} - x_{j,aver}}{s_{j,std}} \quad (6)$$

In Eq. (6), $\boldsymbol{x_{ij,std}}$ is the standardized value of the $\boldsymbol{j}^{\text{th}}$ attribute of the $\boldsymbol{i}^{\text{th}}$ soil horizon; $\boldsymbol{x_{ij}}$ - a value of the $\boldsymbol{j}^{\text{th}}$ attribute of the $\boldsymbol{i}^{\text{th}}$ soil horizon; $\boldsymbol{x_{j,aver}}$ - an average value of the $\boldsymbol{j}^{\text{th}}$ attribute for all soil horizons; and $\boldsymbol{s_{j,std}}$ - standard deviation of the $\boldsymbol{j}^{\text{th}}$ attribute values for all soil horizons.

Soils often have clear horizons that appear uniform in their texture, color, or structure (Myers et al., 2011). Therefore, we assume that while a soil profile's vertical heterogeneity is expressed through different soil horizons, but for the given soil profile each of its horizons is homogenous. Hence, if any $\boldsymbol{i}^{\text{th}}$ horizon of the soil profile has a vector of attribute values $(x_{i1}, x_{i2}, \cdots, x_{in})$, then any layer whose lower depth is less than the current horizon and at the same time higher than an upper horizon is characterized with the same vector of attribute values. Based on this assumption we define any additional layer(s) after comparing lower depths of two horizons of the soil profiles. For example, the first horizons of soil profiles X and Y have a lower depth of 10 cm and 20 cm, respectively (Fig. 1a). Therefore, we can define an additional layer up to a depth of 10 cm to profile Y by constructing the new Y_{mod} soil profile as it is shown for Y_{mod} (Fig. 1b). The new layer constructed for soil profile Y_{mod} will have the same known Y_1 vector of attribute values of the initial horizon with a lower depth of 20 cm. Thus, vector Y_1 is shown twice (Fig. 1b). We continue this process of defining layers for each horizon of the soil profile considering their lower depths in the pair of soil profiles X_{mod} and Y_{mod} (Fig. 1b), as well as X_{mod} and Z_{mod} (Fig. 1c).

To calculate the distance between any two matrices that correspond to two different soil profiles, we create two new soil profiles X_{mod} and Y_{mod} according to the procedures described. An example is illustrated in Fig. 1, in which thick lines indicate the lower depths of existing soil profile horizons and dashed lines indicate the newly created additional layers with attribute values from soil profiles X and Y. After completing this process, we have seven layers in both the X_{mod} and Y_{mod} profiles that replace the four horizons for soil profile X and five horizons for soil profile Y with known attribute values. All soil layers in the profiles X_{mod} and in Y_{mod} have the same thickness except for the deeper section of the profile. At this point, we do not consider the section of the soil profile Y that is below a depth of 50 cm as it does not have any corresponding horizons in soil profile X.

Similar to the process for soil profiles X and Y (Fig. 1b), we apply the same procedures for soil profiles X and Z to create soil profiles X_{mod} and Z_{mod} (Fig. 1c). We have seven layers for the profiles X_{mod} and Z_{mod} that replace the initial four horizons for soil profile X and five horizons for soil profile Z with the same vectors of attribute values as for soil profiles X and Z (Fig. 1a). Again, at this point, we ignore the section of soil profile X below a depth of 45 cm as it does not have any corresponding horizons in soil profile Z. As a result, the modified soil profiles have known attribute values and the same number of soil layers with the same thickness. Following the adjustment of each pair of soil profiles, we now have matrices with the same dimension and we can calculate the ED as defined in equation (3) or introduce a proxy measure. In Eq. (3), $\boldsymbol{n}$ is the number of the soil profile attributes, which is constant for all soil profiles; $\boldsymbol{m}$ is the number of the soil profile layers, which will vary depending on the pair of the soil profiles being adjusted.

2.3. Modified distance matrix calculation

Up to this point, we have not considered the differences in soil depth between soil profiles and uncertainty inherent in those differences (Fig. 1). We now consider a case in which there are three soil profiles. Two profiles have exactly the same number of horizons, thickness of horizons and soil depth. The third profile has one section that is the same as the second profile and another section with additional horizons, and thus a greater soil depth. If we calculate, the distances between the first two soil profiles and then between the first and the third soil profile, the distances will be the same. However, the third profile differs from the second one and has additional horizons below the soil depth of the second profile. For such a case, we introduce a correction coefficient of soil profile closeness by depth between any two soil profiles X and Y using the following equation:

$$k_{corr}(X, Y) = 1 + \frac{|h_X - h_Y|}{\max(h_X, h_Y)} \quad (7)$$

Here $\boldsymbol{h_X}$ and $\boldsymbol{h_Y}$ are the soil depths for each of the two soil profiles X and Y.

Using the correction coefficient, we can also take into account the differences in soil depth of two soil profiles and the uncertainty inherent in those differences. This allows us to calculate the modified distance (d_m) between any two X and Y soil profiles using both the correction coefficient (k_{corr}(X,Y)) and the distance value between

matrices of adjusted soil profiles as follows:

$$d_m(X, Y) = k_{corr}(X, Y)d(X_{mod}, Y_{mod}) \quad (8)$$

Here, $d(X_{mod}, Y_{mod})$ is the ED between adjusted soil profiles X_{mod} and Y_{mod}. The correction coefficient introduces the degree of uncertainty related to the unknown part in one of them if the soil depths of the two soil profiles are different. It increases the distance or dissimilarity between two soil profiles when the difference between soil depths of the two soil profiles increases.

Based on equation (7), the correction coefficient k_{corr} for two soil profiles that have the same soil depth ($h_X = h_Y$) equals 1 and the distance for those two profiles will be same as for the modified soil profiles. The larger the difference between soil depths for two soil profiles the larger $|h_X - h_Y|$ and k_{corr}, and, thus, the value for the distance (Eq. (8)). For example, for the soil profiles X and Z shown in Fig. 1a and for the soil profiles X_{mod} and Z_{mod} shown in Fig. 1c, with soil depths differences in 5 cm k_{corr} equals 1.1. For the soil profiles X and Y from Fig. 1a, and for the soil profiles X_{mod} and Y_{mod} (Fig. 1b) with soil depth differences of 20 cm, k_{corr} value equals 1.3. Therefore, the values of k_{corr} (1.1 and 1.3) will affect the final distance values based on the distances between the modified soil profiles.

Distances that are not straight lines follow four rules (Fielding, 2007; Bandyopadhyay and Saha, 2013). Let $d_m(X, Y)$ be the modified distance between two soil profiles X and Y, then:

1. $d_m(X, Y) \geq 0$ (if profiles are identical, $d_m(X, Y) = 0$; if they are different, $d_m(X, Y) > 0$);
2. $d_m(X, Y) = d_m(Y, X)$ (the distance from X to Y is the same as that from Y to X);
3. $d_m(X, X) = 0$ (a soil profile is identical to itself);
4. $d_m(X, Z) \leq d_m(X, Y) + d_m(Y, Z)$

A distance measure is called a metric if it satisfies all four of the above conditions. Hence, not all distances are metrics, but all metrics are distances (Fielding, 2007; Bandyopadhyay and Saha, 2013). In the case of semi-metrics, the distance measures obey the first three conditions but may not obey the fourth 'triangle' condition (Fielding, 2007).

The first three conditions are simple and the modified distance for the soil profiles(d_m) satisfies them. The fourth condition is more complex and for the modified distances of the soil profiles, this condition is generally not true. This means that when considering three soil profiles X, Y, Z, the distance between two of them might exceed the sum of the distances between the other two. In other words, the modified distances of the soil profiles cannot be constructed as a triangle; thus, this proxy measure is a semi-metric. Semi-metrics have been applied successfully in many clustering applications (Gowda and Diday, 1992; Jain et al., 1999), thereby confirming the viewpoint that the distances or dissimilarities do not necessarily need to be metric (Gowda and Diday, 1992; Jain et al., 1999).

2.4. Algorithm for modified distance matrix calculation

Based on the previously described approach, we developed an algorithm and implemented it in a Fortran program to calculate the modified distance (d_m) for every pair of soil profiles from a given file with soil profile data and to construct a symmetric matrix consisting of distances between two soil profiles, hereafter referred to as the *MDM* (Modified Distance Matrix). The calculation of the MDM was implemented using the following steps (Fig. 2):

1. The program reads soil profile data from a *.csv (comma separated values) file. The example of the file structure is given in Table 1 and is as follows: the first line contains the column headers with soil profile attribute names. The first column should be the identifiers for each horizon; WISE3_ID is the unique soil profile number and is the same of all horizons for the same profile. The second column represents the lower depth for each horizon (BOTDEP, cm). These two columns must be present. The remaining columns are other attribute values of the horizons and can be arbitrarily configured;
2. The number of soil attributes are estimated;
3. The preliminary calculations are conducted according to Eq. (6), including the average and standard deviation for each attribute for all horizons, and the standardized values of attributes;
4. The number of soil profiles and the number of horizons in each soil profile are calculated;

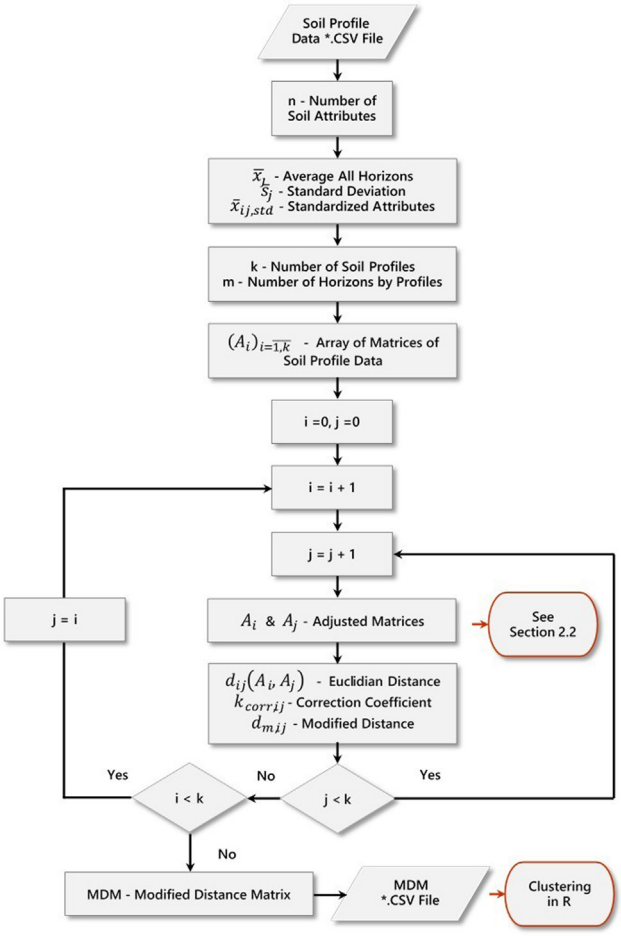

Fig. 2. Flowchart of the Modified Distance Matrix Calculation Algorithm.

Table 1
An example for a section of the input file for a distance matrix calculation with headers and attribute values, including soil profile names (IDs), the lower depth for each soil horizon, soil texture, and other relevant information.

WISE3_ID, BOTDEP, SAND, SILT, CLAY, BULKDENS, ORGC, PHH2O, CECSOIL, TOTN
US0008, 23, 19, 51, 30, 1.44, 25.6, 6.8, 27.2, 2.4
US0008, 33, 10, 59, 30, 1.44, 11.8, 6.8, 26.2, 1.7
US0008, 48, 12, 60, 27, 1.54, 10.9, 7.2, 23.7, 0.9
US0008, 81, 36, 37, 27, 1.64, 3.3, 8.5, 12.8, 0.4
US0009, 19, 11, 65, 25, 1.26, 21.1, 5.9, 20, 2
US0009, 33, 24, 45, 31, 1.27, 12.7, 6.4, 23.5, 1.2
US0009, 44, 22, 47, 32, 1.29, 9.7, 6.8, 23, 1
US0009, 67, 24, 51, 26, 1.32, 3.7, 7,22.7, 0.5
US0009, 92, 22, 58, 20, 1.35, 2.4, 7.3, 21, 0.4
US0009, 142, 28, 57, 15, 1.34, 1.1, 7.8, 25.8, 0.3

Note: WISE3_ID - the unique soil profile number, BOTDEP - the lower depth of a horizon (cm), SAND - sand content (w/w%), SILT - silt content (w/w%), CLAY - clay content (w/w%), BULKDENS - a bulk density (g cm^{-3}), ORGC - organic carbon (g kg^{-1}), PHH2O - pH, measured in water, CECSOIL - cation exchange capacity ($cmol_c$ kg^{-1}), TOTN - total nitrogen (g kg^{-1}).

5. The soil profile data by horizons and attributes are arranged in an array of matrices, where each matrix represents one soil profile data;
6. From the array of soil profile matrices, a pair of matrices is selected and adjusted layers and their lower depths are created by comparison of the lower depths of the horizons;
7. Two adjusted matrices with the same dimension are constructed, and the attribute values for the newly created layers are set equal to those of the container horizons;
8. Based on Eq. (3), the ED is calculated for the two modified soil profile matrices;
9. Based on Eq. (7), the correction coefficient is calculated using soil depths of two soil profiles;
10. According to Eq. (8) the modified distance is calculated using the correction coefficient from step 9;
11. The modified distance value is saved as an element of the MDM;
12. Steps 6 through 11 are repeated for the next soil profile matrix until all pairs of matrices have been completed;
13. The MDM is created as a symmetric matrix filled with values of modified distances for each pair of matrices and with zero values in the diagonal;
14. Finally, the MDM is written as a table in an output file in *.csv format. The output file will have the soil profile IDs as row and column names.

The Fortran program that was based on the above algorithm creates a *.csv format output file. The file with the symmetric matrix of $p \times p$ dimension, where p is the number of soil profiles, will contain the modified distances between soil profiles and can be used for further cluster analysis.

2.5. Example application with the global soil profile dataset

The WISE3 database was used to evaluate the performance of the proposed approach for clustering of soil profiles. The database was developed by ISRIC in Wageningen, the Netherlands. It is one of the most comprehensive soil databases, encompassing selected attribute data for 10,253 soil profiles and 47,800 horizons from 149 countries collected by many international soil professionals and distributed across the world. All profiles have been harmonized with respect to the original Legend (1974) and Revised Legend (1988) of FAO-UNESCO (Batjes, 2008, 2009). The individual profiles were sampled, described, and analyzed according to methods and standards used in the originating countries. There is no uniform set of properties for which all profiles have analytical data, because only selected measurements were planned during the original surveys. Because of compilation of legacy soil data derived from traditional soil surveys WISE3 contains gaps that are taxonomic, geographic, and soil analytical in nature. As a result, the amount of data available for modeling is sometimes much less than expected (Batjes, 2008, 2009).

WISE3 is a relational database compiled using MS Access and can be downloaded from the ISRIC web-portal (International Soil Reference and Information Centre (ISRIC), 2015). The structure of its attributes comprises several tables. The table WISE3_Site contains the precise sampling location for each soil profile, including longitude, latitude, elevation, profile location description and slope, and the soil classification according to FAO-UNESCO (1974), FAO (1990), and/or USDA (1999), or other appropriate entities. The table WISE3_Horizon provides information about soil horizon distribution and classification, soil color, organic-carbon and total nitrogen content, pH in water and in KCl, CEC, sand, silt, clay, coarse fraction, and bulk density. Both tables are organized alphabetically and share a unique profile reference number (WISE3_ID). The reference number contains a two-character abbreviation ISO code for the country of origin. The sampling depth (BOTDEP – lower depth of horizon) varies widely, sometimes to a depth of 8.50 m. For soil horizons, the WISE3 database defines the horizon boundaries, which are the upper depth of the horizon (TOPDEP), and the bottom depth of the horizon (BOTDEP). In this study, we used BOTDEP.

Table 2
Summary of the soil profile attributes from WISE Version 3.1 that are included in the soil profile datasets.

n	Attribute Abbreviation	Description of Soil Attribute	Unit
1	WISE3_ID	Unique soil profile number	
2	BOTDEP	Lower depth of horizon	cm
	Dataset1		
3	SAND	Sand content	w/w%
4	SILT	Silt content	w/w%
5	CLAY	Clay content	w/w%
	Dataset2		
6	BULKDENS	Bulk density	g cm $^{-3}$
	Dataset3		
6	ORGC	Organic carbon	g kg^{-1}
	Dataset4		
6	BULKDENS	Bulk density	g cm $^{-3}$
7	ORGC	Organic carbon	g kg^{-1}

There are many gaps for all 30 attributes of the soil profile horizons in the WISE3 dataset. Therefore, from the 30 attributes, several sets of attribute combinations were selected based on only three criteria: texture (i.e., sand, silt, and clay), bulk density, and soil fertility (represented by organic carbon) to define different datasets. Table 2 shows a summary of the soil profile attributes from the WISE3 database that are included in these combinations for creating different datasets. Queries for the Access database tables resulted in four datasets ranging from three to five dimensions not including WISE3_ID and for which all horizons had complete data. Five soil profile attributes, including WISE3_ID (unique soil profile number), BOTDEP (lower depth of horizon, cm), SAND (sand content, w/w%), SILT (silt content, w/w%), and CLAY (clay content, w/w%) are common for all datasets.

The first dataset (Dataset1) is three-dimensional and includes three soil texture attributes from the global soil dataset WISE3 with the largest number of complete horizon data for a total of 43,919 horizons. The second dataset (Dataset2) is four-dimensional and has all the attributes as Dataset1 with the addition of bulk density, for a total of 14,886 horizon data. The third data set (Dataset3) is four-dimensional and includes the same attributes as Dataset1, with the addition of organic carbon (ORGC, g kg^{-1}) and a total of 39,493 horizon data. Finally, the fourth dataset (Dataset4) is five-dimensional and includes the same attributes as Dataset1, with the addition of bulk density (BULKDENS, g cm $^{-3}$) and ORGC, with a total of 13,691 horizon data. We can see how the number of horizons in the datasets is reduced when the number of attributes increases because of missing values for certain soil attributes and soil horizons in WISE3.

Data in WISE3 were subjected to a rigorous, computerized data-checking scheme as well as numerous visual checks. Nonetheless, this large dataset is still likely to have inconsistencies or even errors and users should be aware of these potential limitations when analyzing or applying the data (Batjes, 2008, 2009). To remove outliers from the attribute values we applied a 5-σ criterion in accordance with Chebyshev's Theorem, which implies that at least 96% of the statistically acceptable data are within these boundaries without restriction related to the distribution of the data (Gnedenko, 1988; Amidan et al., 2005; Schmidt et al., 2011). Thus, if a value of the attribute deviated from the average value of the attribute by more than five standard deviations it was considered as an outlier and we excluded it from further analysis.

2.6. Cluster analysis

There are many clustering methods, and each results in a different grouping of a dataset (Gelbard et al., 2007). The two that are most widely used are the hierarchical and k-means clustering methods. In the

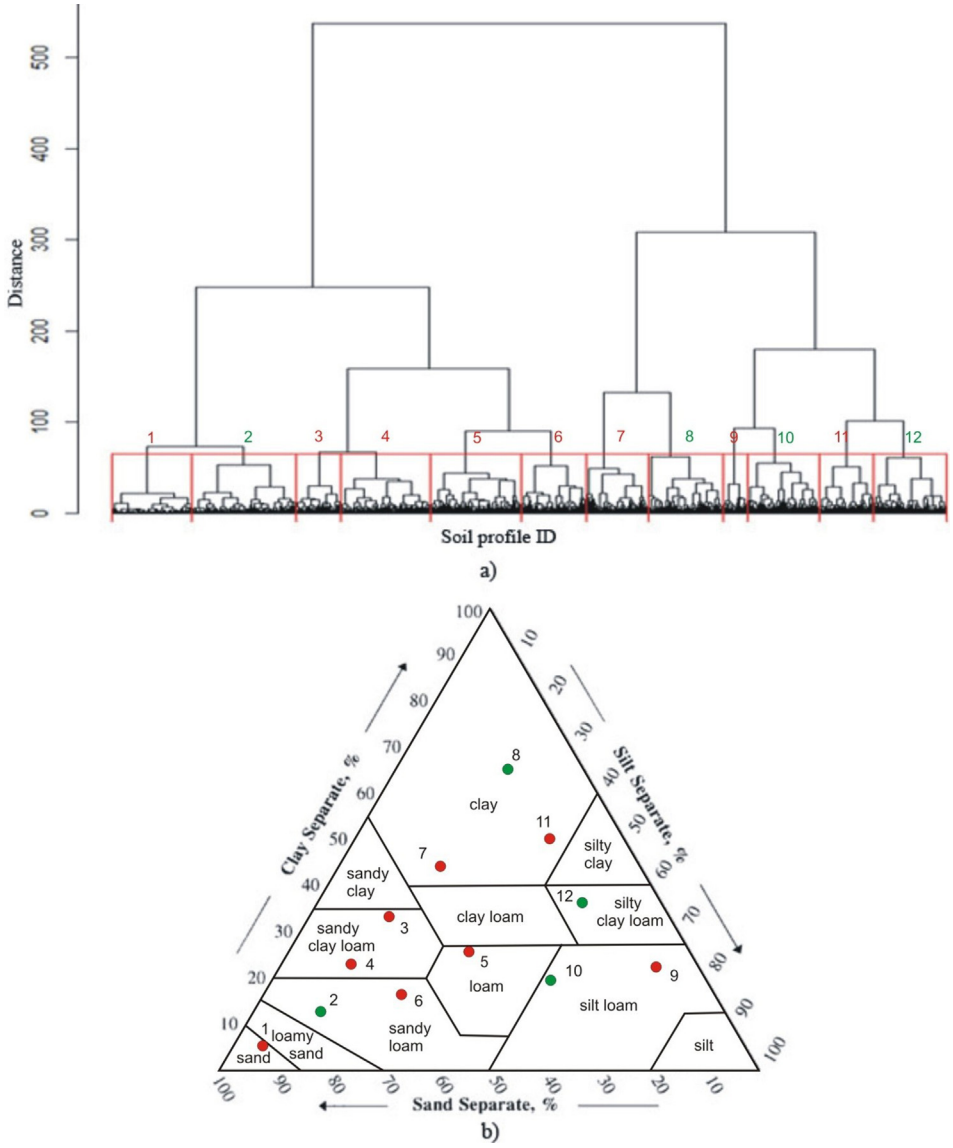

Fig. 3. Dendrogram of soil profile clustering (Ward's method) contained in Dataset1 (a) and average values of texture variables of all horizons for clusters of Dataset1 in red and green dots as shown in a standard soil-texture triangle (Brady and Weil, 1999) when the number of clusters k = 12 (green dots will be further split when k = 16) (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

hierarchical clustering procedure, a tree-like structure is built to determine the relationship among entities. In the k-means clustering procedure, one position of the measurement is taken as a central place and the distance from this central point is determined. An important step for either of these procedures is to select a distance measure that determines how the proximity of two objects is calculated. Unfortunately, none of the distance measures in hierarchical clustering or k-means is directly suitable for soil profile data clustering because of the different number of soil horizons in the profiles. In our study, we used hierarchical agglomerative methods because we were able to input the predefined distance matrix, the MDM in the case of the soil profile data.

Among the many agglomerative methods of hierarchical clustering, Ward's method (Ward, 1963) has been successfully used for a variety of analyses (Gong & Richman, 1995; Wilks, 2011; Calmanti et al., 2015). We adopted the Ward's method in our study to identify soil profile clusters and applied it to the MDM after calculating the distance matrix for the standardized soil profile data. The *.csv output file containing the MDM was imported into R (R Core Team, 2015) for clustering using Ward's hierarchical agglomerative method. After importing the MDM into R as a data frame, it was converted into an R distance object and was used as a distance or dissimilarity matrix by the "hclust" function of the R "stats" package. For this function, a clustering algorithm was represented by its argument value as "ward.D2" (Murtagh & Legendre, 2014; R Core Team, 2015).

The dendrogram is a useful graphical representation of the clustering structure of a set of objects. It is a branching diagram that represents the relationships of similarity among a group of entities and consists of many U-shaped lines connecting objects in a hierarchical tree. The height of each U represents the distance between two objects that are connected and each 'leaf' in the dendrogram represents one data object. Dendrograms are computed by using standardized variables and, therefore, the distance between clusters has no unit.

Selecting the significant number of clusters into which the initial dataset should be split is a critical aspect in practical applications of cluster analysis. Although some degree of subjectivity is often unavoidable, a certain level of clustering may be more appropriate than others. In this study, we adopted the method that uses the concept of 'plateau' and monitoring the distance between the clusters as they were merged at each step. A plateau is a situation where the distance

Table 3
The number of soil profiles (n), their percentage (%) in each cluster, and their types by soil texture for the different datasets (as presented in Table 2) from WISE3.

	Dataset1(9926 profiles & 43,919 horizons)			Dataset2(3644 profiles & 14,886 horizons)			Dataset3(9672 profiles & 39,493 horizons)		
#	n	%	Type by Texture	n	%	Type by Texture	n	%	Type by Texture
1	954	9.6	Sand	170	4.7	Sandy Loam	853	8.8	Sandy Clay Loam
2	1240	12.5	Sandy Loam	366	10.0	Loamy sand	426	4.4	Sandy Loam
3	537	5.4	Sandy Clay Loam	301	8.3	Sandy Clay Loam	608	6.3	Clay Loam
4	1067	10.7	Sandy Clay Loam	218	6.0	Sandy Clay Loam	1209	12.5	Loamy sand
5	1081	10.9	Loam	175	4.8	Clay	242	2.5	Sandy Loam
6	769	7.7	Sandy Loam	209	5.7	Clay	654	6.8	Sandy Clay Loam
7	734	7.4	Clay	183	5.0	Clay	971	10.0	Sandy Clay Loam
8	885	8.9	Clay	90	2.5	Sandy Clay	783	8.1	Clay
9	296	3.0	Silty Loam	49	1.3	Clay Loam	70	0.7	Clay
10	853	8.6	Silt Loam	118	3.2	Silt Loam	344	3.6	Clay
11	654	6.6	Clay	299	8.2	Silty Clay Loam	323	3.3	Clay
12	856	8.6	Silty Clay Loam	101	2.8	Silty Clay Loam	356	3.7	Clay
13				62	1.7	Sandy Loam	410	4.2	Clay
14				148	4.1	Loam	568	5.9	Silty Clay Loam
15				226	6.2	Sandy Clay Loam	515	5.3	Silt Loam
16				292	8.0	Clay	238	2.5	Loam
17				259	7.1	Sandy Loam	752	7.8	Loam
18				124	3.4	Loam	85	0.9	Silt Loam
19				254	7.0	Clay Loam	149	1.5	Sandy Loam
20							27	0.3	Clay Loam
21							89	0.9	Clay Loam

between merged clusters remains relatively constant. In order to define the number of clusters, a cutoff level should be set to the level that corresponds to the transition from a plateau to a sudden increase in the distance between merged clusters (Calmanti et al., 2015). In this study, we refer to the cutoff level as the number of clusters identified by setting the cutoff level at a certain distance depending on the transition from a plateau.

3. Results and discussion

3.1. Clustering a three-dimensional soil profile dataset

Several cluster analyses were performed based on the MDM using four datasets from WISE3 with different soil attributes (Table 2) to illustrate and assess the quality and impact of the proposed semi-metric proximity measure. For Dataset1 with three soil texture attributes, there were a total of 9,926 soil profiles, and the MDM had a dimension of 9,926 × 9,926. After using Ward's method of the hierarchical agglomerative clustering and graphically representing results in a dendrogram, we set the cutoff level at a distance of 65, which was close to the plateau, resulting in 12 clusters (Fig. 3a). We then calculated the number of soil profiles (n) and their percentage (%) by clusters. In cluster 1, there were 954 (9.6%) soil profiles from the total number of soil profiles of Dataset1, and in cluster 2, there were 1,240 (12.5%) soil profiles and so on (Table 3). Cluster 2 had the maximum number (1,240) of soil profiles, while cluster 9 had the minimum number of soil profiles (296).

When analyzing the profile numbers by clusters for Dataset1, all soil profiles in Dataset1 as part of the global soil profile dataset WISE3 were quite uniformly distributed among the clusters (Table 3). These clusters were then matched to the soil types found in the traditional soil-texture triangle (Brady and Weil, 1999) in which one can determine the texture class for a single point based on sand, silt, and clay content. Therefore, we calculated the average values for the texture variables for all soil horizons for each cluster and compared them with the texture-triangle using the web-tool Soil Texture Calculator (NRCS, 2020) (Fig. 3b). The average values for the soil texture attributes for all soil profile horizons in each cluster (Fig. 3b) matched the different soil types found in the soil-texture triangle very well. The green dots correspond to the clusters that will be split further when the number of clusters increases from 12 to 16.

We then reduced the cutoff level to a distance of 52 (Fig. 4a), which was also close to a plateau and resulted in 16 unique clusters. Again, we calculated the number of soil profiles (n) and their percentage (%) by clusters. We found that all soil profiles from Dataset1 as part of the dataset WISE3 are quite uniformly distributed among the clusters in this case as well. We calculated the average values for texture variables for all soil profile horizons for each cluster and again placed them in the texture-triangle using the web-tool Soil Texture Calculator (Fig. 4b). Clusters 2, 8, 10, and 12 (green dots) from the first case of clustering (k = 12; Fig. 3a,b) were split further, resulting in 8 new clusters represented as green dots with a new location, and 8 old clusters that did not change their location. When the new values of the average texture variables of the clusters where allocated in the soil-texture triangle, one new soil type was clay loam (Fig. 4b). The clusters that were identified based on the MDM calculation for soil profiles in WISE3 represented the range of soil texture types quite well.

3.2. Clustering four- and five-dimensional soil profile datasets

We selected two datasets from WISE3 to explore the impact of adding additional attributes to the clustering analysis. The first, Dataset2, included the three soil texture variables and soil bulk density (BULKDENS, g cm^{-3}), and the second, Dataset3, included three texture variables and soil organic carbon (ORGC, g kg^{-1}) (Table 2). Dataset2 contained 3,644 soil profiles and 14,886 horizons, and Dataset3 contained 9,672 soil profiles and 39,493 horizons. The MDM was calculated for both datasets, resulting in a dimension of 3,644 × 3,644 matrix for Dataset2 and a dimension of 9,672 × 9,672 matrix for Dataset3, followed by a cluster analysis using Ward's method of the hierarchical clustering. The final results are depicted in a dendrogram (Fig. 5a and Fig. 6a). Cutoff levels were selected at 40 for Dataset2 and 55 for Dataset3, which were close to the respective plateaus, resulting in 19 and 21 clusters, respectively. We then determined the number of soil profiles (n) and their percentage (%) in each cluster for Dataset2 and Dataset3 (Table 3). We calculated the average values for the texture variables for all soil horizons for each cluster and added them to the texture-triangle using the web tool Soil Texture Calculator (Fig. 5b and Fig. 6b).

In the case of Dataset2 and the corresponding 19 clusters, 9 soil texture types were presented from 12 soil texture types of the texture-triangle, versus 21 clusters and 8 soil texture types for Dataset3. Three soil types, i.e., sand, silt, and silty clay, were not present in either case of clustering; and a fourth type, sandy clay, was not present in the

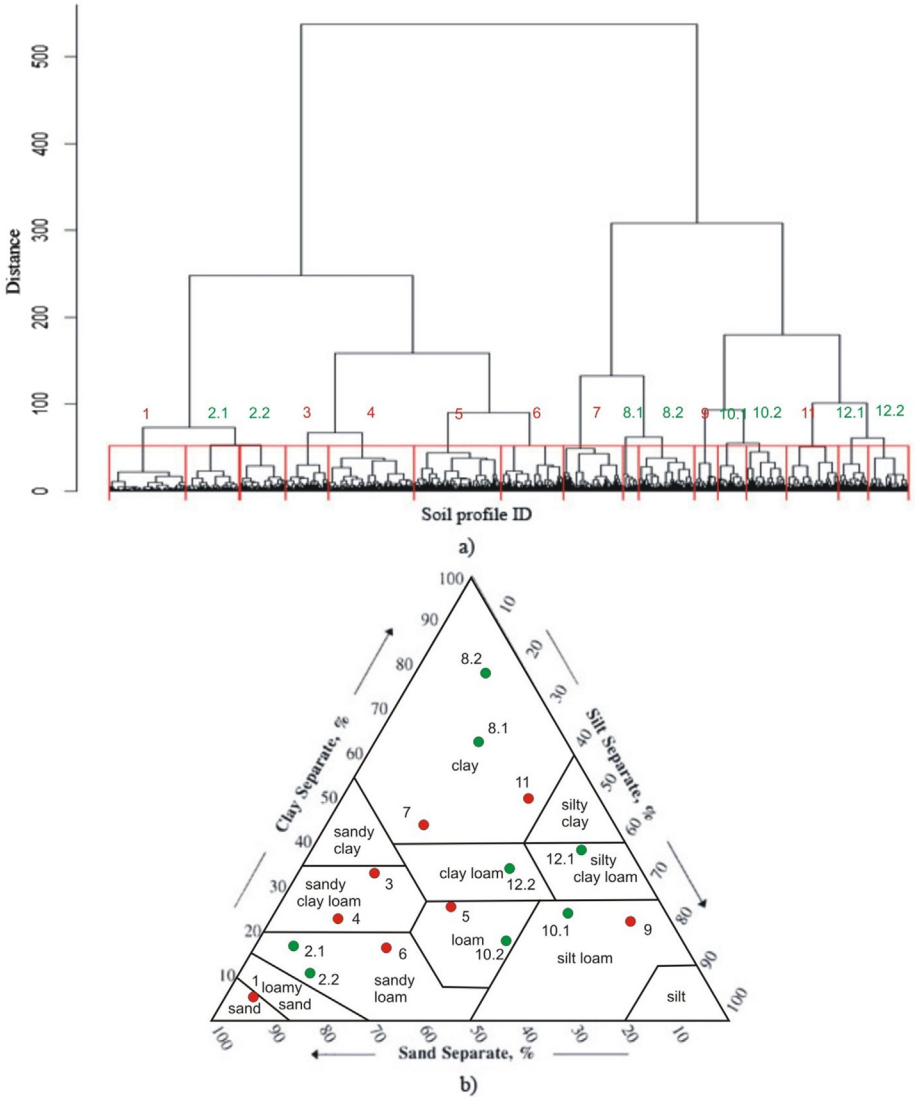

Fig. 4. Dendrogram of soil profile clustering (Ward's method) contained in Dataset1 (a) and average values of texture variables of all horizons for clusters of Dataset1 in red and green dots as shown in a standard soil-texture triangle (Brady and Weil, 1999) when the number of clusters k = 16 (green dots were split from k = 12 case) (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

second case only. As Dataset2 and Dataset3 initially contained a different number of profiles and clusters, we compared the percentage of soil profiles by soil texture types for two clustering cases. In both datasets for the two soil types, the high percentage of soil profiles were for the clay type with the same value of 23.6% of the profiles and 4 and 6 clusters, respectively, compared to the sandy clay loam type that had close values of 3 clusters and 20.4% of the profiles versus 3 clusters and 25.6% of the profiles, respectively (Table 3).

For the remaining soil types, the differences were significant in clustering. For the silty clay loam, two clusters were identified in Dataset2 with 11% of the soil profiles versus one cluster in Dataset3 with 5.9% of the soil profiles (Table 3). For the silty loam, one cluster was identified in Dataset2 versus two clusters in Dataset3 containing 6.2% of the soil profiles. For the sandy loam, three clusters were identified in Dataset2 with 13.5% of the soil profiles versus three clusters in Dataset3 with 8.4% the soil profiles. For the clay loam, two clusters were identified in Dataset2 with 8.3% of the soil profiles versus three clusters in Dataset3 with 7.5% of the soil profiles. For the loam, two clusters were identified in Dataset2 with 7.5% of the soil profiles versus two clusters in Dataset3 with 10.2% of the soil profiles. For the loamy sand, one cluster was identified in Dataset2 with 10% versus one cluster in Dataset3 with 12.5% of the soil profiles. Finally, for the sandy clay, one cluster was identified in Dataset2 with 2.5% of the soil profiles versus none for Dataset3. The results from this analysis showed that adding additional attributes changed the profile grouping in the clusters, the number of profiles for each cluster, and their distribution in the soil-texture triangle. However, the tendency of representing the different soil texture types remained unchanged.

In order to further explore the impact of adding a new attribute to soil clustering, we created Dataset4. It contained complete records for texture, bulk density, and organic carbon (Table 2), for a total of 13,691 horizons. Again, the MDM was calculated resulting in a matrix of dimension 3,571 × 3,571, followed by a cluster analysis using Ward's method of hierarchical clustering. The cutoff level was selected at 175, which was close to the plateau, and resulted in 29 clusters. Most of the clusters (18) or 64.1% of the profiles were represented as a clay loam soil; 1 cluster (18.5% of profiles) was a sandy clay loam; 2 clusters (14.6% of profiles) were a sandy loam; 7 clusters (2.8% of profiles) were a loam; and 1 cluster (0.1% of profiles) was a silty clay, for a total of only five soil types. The reason for this small number could be that the organic carbon varied widely in the soil with a coefficient of variance of about 200%, combined with bulk density and texture variables. Our findings showed that the semi-metric we introduced as a measure of dissimilarity was sensitive to the combination of the attributes that

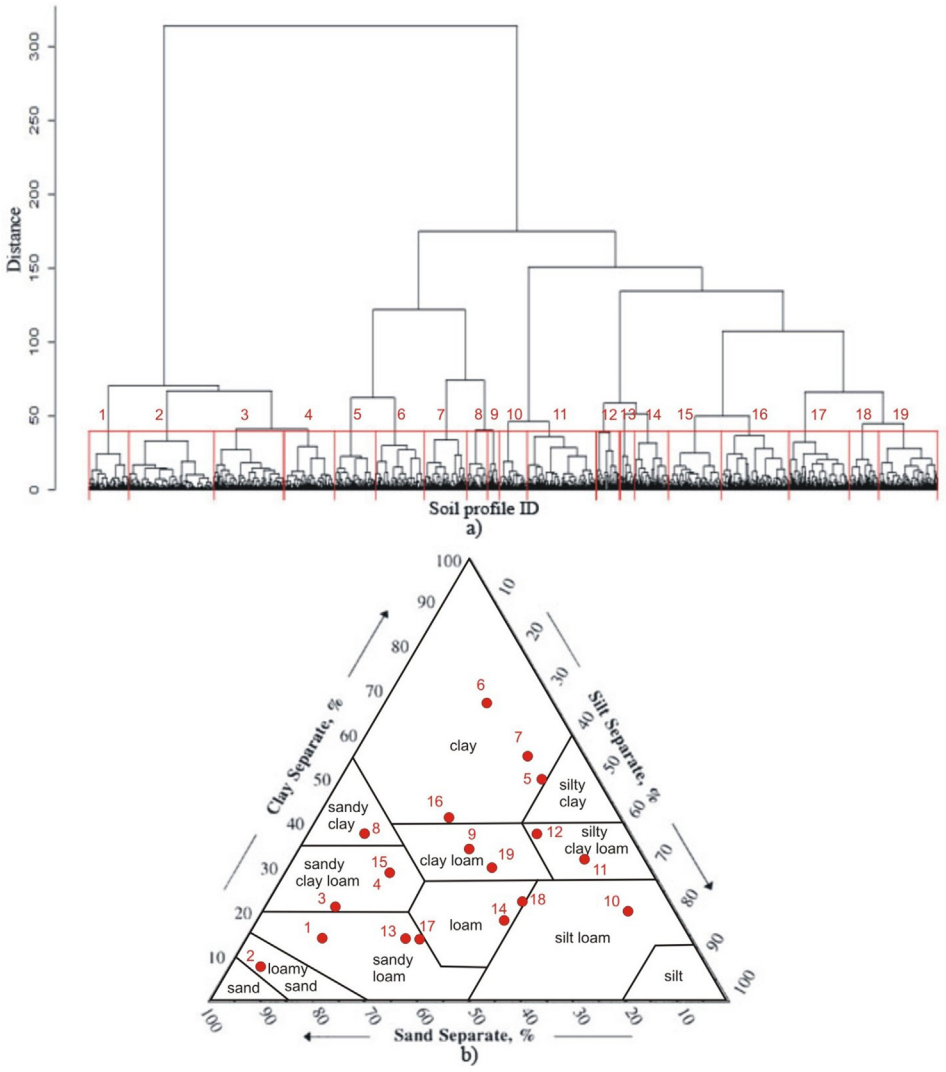

Fig. 5. Dendrogram of soil profile clustering (Ward's method) contained in Dataset2 (a) and average values of texture variables of all horizons for clusters of Dataset2 in red dots as shown in a standard soil-texture triangle (Brady and Weil, 1999) when the number of clusters k = 19 (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

were used for clustering and that the proper selection of attributes and the variability of their values should be considered carefully.

In this study, the global soil profile database WISE3 was used to evaluate the effectiveness of the proposed semi-metric and the corresponding MDM for clustering. Several clustering experiments that were conducted using four sub-datasets obtained from WISE3 by selecting different combination of the soil attributes demonstrated that the MDM was effective in identifying main soil groups. Although a database such as WISE3 with contributions from many soil scientists across the globe may contain some errors because of the differences in technical skills and access to instrumentation, our results showed that the proposed algorithm was able to predict the correct grouping in clusters despite the possible errors associated with the dataset.

The application MDM (Modified Distance Matrix) we developed in Fortran and is relatively easy to use. It accepts a CSV format file with soil profile data as an input that has a simple structure. The file can be created based on queries on the soil profile database. After invoking the application, it reads an input file and produces the modified distance matrix. The program output can then be used with a range of clustering packages that are available in R. The limitation of the application is that the implemented algorithm for a semi-metric currently handles only combinations of continuous variables and not binary and categorical data.

Designing new measures of distances or similarity to combine information on different soil profile horizons is not an easy task because of the different numbers of horizons in soil profiles, varying horizon thickness, and different soil profile depths. The approach we developed for the dissimilarity measure between soil profiles fully considers all horizons of a soil profile, their properties, lower depths of horizons and soil profile depths compared to other approaches that consider average values of soil attributes for the entire profile (Jagtap et al.,2004) or a fixed number of horizons per soil profile (Jones and Thornton, 2015). Therefore, our approach is more advanced and general; it does not change existing profile and related data and it provides reliable results, as we demonstrated in this study. This technique and metric could easily be extended to other proximity measures and can serve as a data mining technique for soil profile survey results.

Numerical approaches such as numerical soil classification have become attractive methodologies because of the availability of large national and international databases of soil profile descriptions and associated laboratory measured data. The proposed approach of soil profile horizons adjustment and modified distance matrix calculation can be used to estimate missing values for the attributes in WISE3 or other low- and high-dimensional soil profile datasets based on the known soil profile data from the same clusters after carefully selecting the attributes that have to be estimated. This study demonstrated how

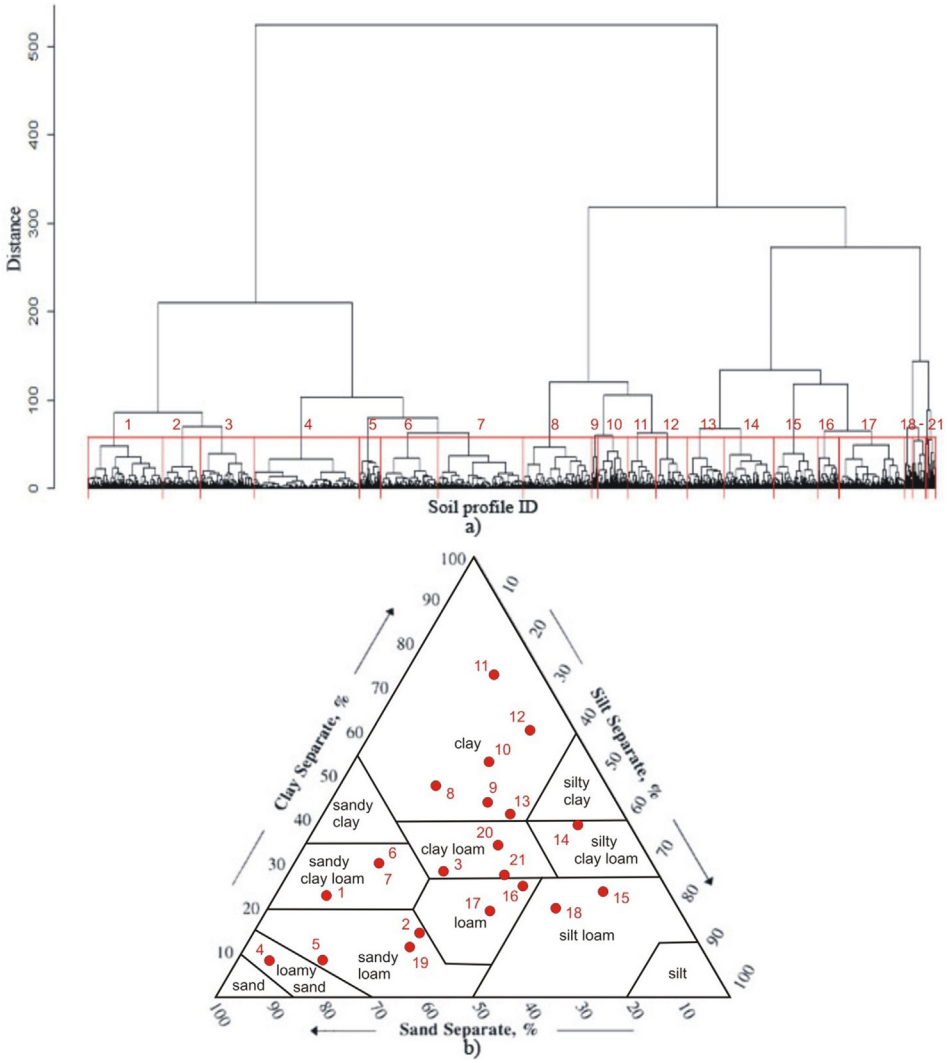

Fig. 6. Dendrogram of soil profile clustering (Ward's method) contained in Dataset3 (a) and average values of texture variables of all horizons for clusters of Dataset3 in red dots as shown in a standard soil-texture triangle (Brady and Weil, 1999) when the number of clusters k = 21 (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

improvements in clustering were obtained using the MDM. It can also be used to discover outliers or extreme profiles in the large soil profile datasets which as an unusual observations might be the most interesting in some data mining situations. Another use of clustering is to enable the selection of a smaller number of representatives from the entire data set. It has potential with respect to extracting knowledge from big soil databases and for machine learning methods for soil data. The MDM can be used in other ways as well, but this study was not intended to address all those strategies. The algorithm implemented in MDM provides a new tool for modern challenges facing soil surveys and scientists working with digital soil profile databases. In addition, data on the soil profile properties produced by the clustering using the MDM application can improve small- to large scale modeling applications, including crop, hydrologic, land surface and environmental models.

4. Conclusions

In this study, we developed a semi-metric and an algorithm implemented in a Fortran application. The application can facilitate grouping soil profiles with fairly high accuracy based on the MDM. The results show that this methodology is sufficient to detect the patterns of soil profiles across a large soil dataset that has an extensive spatial variability and represents many environmental conditions. Based on the success of this study, future research should evaluate if the use of different distance measures, other than Euclidean, for calculating the MDM or different data normalization and standardizing methods combined with various clustering algorithms could generate improved and more accurate results.

CRediT authorship contribution statement

Vakhtang Shelia: Methodology, Conceptualization, Data curation, Investigation, Formal analysis, Writing - original draft, Writing - review & editing. **Gerrit Hoogenboom:** Supervision, Methodology, Conceptualization, Formal analysis, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was partially supported by research funds allocated to the AgWeatherNet Program at Washington State University and the

CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS).

References

Amidan, B.G., Ferryman, T.A., Cooley, S.K., 2005. Data outlier detection using the Chebyshev theorem. IEEE Aerospace Conference. 3814–3819.

Bandyopadhyay, S., Saha, S., 2013. Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications. Springer.

Batjes, N.H., 2008. ISRIC-WISE Harmonized Global Soil Profile Dataset (Ver. 3.1). Report 2008/02, ISRIC –. World Soil Information, Wageningen, the Netherlands (with dataset).

Batjes, N.H., 2009. Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. Soil Use Manag. 25, 124–127.

Beaudette, D.E., Roudier, P., O'Geen, A.T., 2013. Algorithms for quantitative pedology: A toolkit for soil scientists. Comput. Geosci. 52, 258–268.

Brady, N.C., Weil, R.R., 1999. The Nature and Properties of Soils. Prentice Hall, Upper Saddle River, N.J.

Burrough, P.A., MacMillan, R.A., van Deursen, W., 1992. Fuzzy classification methods for determining land suitability from soil profile observations and topography. J. Soil Sci. 43, 193–210.

Calmanti, S., Dell'Aquila, A., Maimone, F., Pelino, V., 2015. Evaluation of climate patterns in a regional climate model over Italy using long-term records from SYNOP weather stations and cluster analysis. Climate Research. 62, 173–188.

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

Crommelin, R.D., de Gruijter, J.J., 1973. Cluster analysis applied to mineralogical data from the coversand formation in the Netherlands. Soil Survey Paper No. 7, Soil Survey Institute, Wageningen.

Dua, S., Chowriappa, P., 2013. Data Mining for Bioinformatics. Auerbach Publications.

Everitt, B.S., Landau, S., Morven, L., 2001. Cluster analysis. Oxford University Press, New York.

Fielding, A.H., 2007. Cluster and Classification Techniques for the Biosciences. Cambridge University Press, Cambridge, New York, USA.

Fitzpatrick, E.A., 1967. Soil nomenclature and classification. Geoderma 1, 91–105.

Gelbard, R., Goldman, O., Spiegler, I., 2007. Investigating diversity of clustering methods: An empirical comparison. Data Knowledge Engineering. 63 (1), 155–166.

Gijsman, A.J., Jagtap, S.S., Jones, J.W., 2002. Wading through a swamp of complete confusion: how to choose a method for estimating soil water retention parameters for crop models. Eur. J. Agron. 18 (1–2), 77–106.

Gijsman, A.J., Thornton, P.K., Hoogenboom, G., 2007. Using the WISE database to parameterize soil inputs for crop simulation models. Comput. Electron. Agric. 56, 85–100.

Gnedenko, B.V., 1988. Theory of Probability, 6th ed. Mir Publishers, Moscow, Russia.

Gong, X., Richman, M.B., 1995. On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. J. Clim. 8, 897–931.

Gowda, K.C., Diday, E., 1992. Symbolic clustering using a new dissimilarity measure. IEEE Trans. Syst. Man Cybern. 22, 368–378.

Grunwald, S., Rooney, D.J., McSweeney, K., Lowery, B., 2001a. Development of pedotransfer functions for a profile cone penetrometer. Geoderma 100, 25–47.

Grunwald, S., Lowery, B., Rooney, D.J., McSweeney, K., 2001b. Profile cone penetrometer data used to distinguish between soil materials. Soil Tillage Res. 62, 27–40.

J.F. Hair W.C. Black B.J. Babin R.E. Anderson Multivariate Data Analysis seventh ed. 2009 Prentice Hall Upper Saddle River, NJ, USA.

J. Han M. Kamber Data mining: Concepts and techniques Morgan Kaufmann Series in Data Management Systems 2nd ed., 2006 Morgan Kaufmann San Francisco, CA.

Hole, F.D., Hironaka, M., 1960. An experiment in ordination of some profiles. Soil Sci. Soc. Am. Proc. 24, 309–312.

Hoogenboom, G., Porter, C.H., Boote, K.J., Shelia, V., Wilkens, P.W., Singh, U., White, J. W., Asseng, S., Lizaso, J.I., Moreno, L.P., Pavan, W., Ogoshi, R., Hunt, L.A., Tsuji, G. Y., and Jones, J.W., 2019. The DSSAT crop modeling ecosystem. In: [K.J. Boote, editor] Advances in crop modeling for a sustainable agriculture. Burleigh Dodds Science Publishing, Cambridge, United Kingdom. 173-216.

P.A. Hughes A.B. McBratney B. Minasny S. Campbell End members, end points and extragrades in numerical soil classification Geoderma. 226–227 2014 2014 365 375.

International Soil Reference and Information Centre (ISRIC), 2015. ISRIC-WISE International Soil Profile Data Set [Online]. Available at https://data.isric.org/geonetwork/srv/eng/catalog.search#/metadata/a351682c-330a-4995-a5a1-57ad160e621c (Verified May 12, 2020).

Jagtap, S.S., Lall, U., Jones, J.W., Gijsman, A.J., Ritchie, J.T., 2004. Dynamic nearest neighbor method for estimating soil water parameters. Journal of the American Society of Agricultural and Biological Engineers. 47 (5), 1437–1444.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data Clustering: A Review. ACM Comput. Surv. 31 (3), 264–323.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. DSSAT Cropping System Model. Eur. J. Agron. 18, 235–265.

Jones, P., Thornton, P., 2015. Representative soil profiles for the Harmonized World Soil Database at different spatial resolutions for agricultural modelling applications. Agric. Syst. 139, 93–99.

King, R.S., 2015. Cluster Analysis and Data Mining - An Introduction. Dulles, Virginia; Boston, Massachusetts; New Delhi, Mercury Learning and Information.

Lagacherie, P., Cazemier, D.R., van Gaans, P.F.M., Burrough, P.A., 1997. Fuzzy k-means clustering of fields in an elementary catchment and extrapolation to a larger area. Geoderma 77, 197–216.

Little, I., Ross, D., 1985. The Levenshtein metric, a new means for soil classification tested by data from a sand-podzol chronosequence and evaluated by discriminant function analysis. Aust. J. Soil Res. 23, 115–130.

Mazaheri, S.A., Koppi, A.J., McBratney, A.B., 1995. A fuzzy allocation scheme for the Australian Great Groups Classification system. Eur. J. Soil Sci. 46, 601–612.

McBratney, A.B., de Gruijter, J.J., 1992. A continuum approach to soil classification by modified fuzzy k-means with extragrades. J. Soil Sci. 43, 159–175.

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

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

Murtagh, F., Legendre, P., 2014. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J. Classif. 31, 274–295.

Myers, D.B., Kitchen, N.R., Sudduth, K.A., Miles, R.J., Sadler, E.J., Grunwald, S., 2011. Peak functions for modeling high-resolution soil profile data. Geoderma 166, 74–83.

Natural Resources Conservation Service (NRCS), USDA, 2020. Soil Texture Calculator [Online] Available at http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_054167 (Verified May 12, 2020).

Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. 70, 327–336.

Nemes, A., Roberts, R.T., Rawls, W.J., Pachepsky, Y.A., van Genuchten, M.T., 2008. Software to estimate −33 and −1500 kPa soil water retention using the non-parametric k-Nearest Neighbor technique. Environmental Modeling and Software. 23 (2), 254–255.

Rayner, J.H., 1966. Classification of soils by numerical methods. J. Soil Sci. 17 (1), 79–92.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Schmidt, A., Hanson, C., Kathilankal, J., Law, B., 2011. Classification and assessment of turbulent fluxes above ecosystems in North-America with self-organizing feature map networks. Agric. For. Meteorol. 151, 508–520.

Ward Jr., J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.

Webster, R., Burrough, P.A., 1972. Computer-based soil mapping of small areas from sample data. I. Multivariate classification and ordination. J. Soil Sci. 23, 210–221.

Wilks, D.S., 2011. Statistical methods in the atmospheric sciences, 3rd ed. Academic Press, Waltham, MA.