A global numerical classification of the soil surface layer

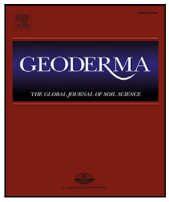
Alexandre M.J.-C Wadoux, Alex B Mcbratney

Geoderma 447 (2024) 116915

Contents lists available at ScienceDirect

Geoderma

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



A global numerical classification of the soil surface layer

Alexandre M.J.-C. Wadoux [a,b,*], Alex B. McBratney [b]

[a] LISAH, Univ Montpellier, AgroParisTech, INRAE, IRD, L'Institut Agro, Montpellier, France
[b] Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Australia

ARTICLE INFO

Handling Editor: Jingyi Huang

Keywords:
Clustering
Horizon classification
Identification
Allocation
Classes pattern
SoilGrids

ABSTRACT

The quest for a global soil classification system has been a long-standing challenge in soil science. There currently exist two, seemingly disjoint, global soil classification systems, the USDA Soil Taxonomy and the World Reference Base for Soil Resources, and many regional and national systems. While both systems are acknowledged as international, there remain various examples of their shortcoming in accounting of topsoil features, local applications and communication with established regional classification systems. This calls for a numerical soil classification that addresses these discrepancies and achieves harmonization with existing national systems. In this paper, we report on the development of a natural layer classification system — as opposed to the classification of soil profile entities, as a first step towards achieving a comprehensive global numerical soil classification not based on *a priori* defined classes. We implemented a modelling approach with a set of predicted key soil properties available globally for the soil surface layer with the same depth range of 0–5 cm. The set of properties was partitioned into a number of homogeneous and disjoint classes using the *k*-means clustering algorithm. Next, we investigated the pattern of variation of the clusters in association with the soil property map with principal component analysis. A three-component nomenclature system is derived in a transformed space of the class-specific centroids to account for the uneven distribution of the centroids in the principal component space. We show that it is possible to build a data-based objective numerical taxonomic classification of soil layers, and that existing sets of key soil properties, predicted separately, coalesce into identifiable clusters or classes and manifest discernible spatial and/or pedological patterns. This grouping of key soil properties to logical categories is a possible step to better define diagnostic horizon features and suggest new ones. The general-purpose map of soil surface layer classes of the world also has potential applications in assessing soil change and designing monitoring surveys.

1. Introduction

The quest for a global soil classification system has been a long-standing challenge in soil science. For more than a century, soil classifications have been derived for both theoretical and practical purposes (De Bakker, 1970; de Gruijter, 1977; Hallsworth, 1965). The intended purpose may vary but they have in common that they enable the coherent description of the soils through organization of multiple soil properties and their simplified representation into consistent classes. The introduction of computers in the 1960s allowed the development of numerical methods of soil classification involving the calculation of taxonomic distances — distances between two points in the multivariate soil character space. The numerical approach simplified the generation and tests of various solutions for the arrangement of soil individuals into classes and the allocation of new individuals into pre-existing classes (Rayner, 1966; McBratney, 1994). It also facilitated the identification and display of spatial and pedological patterns and the development of continuous classifications that do more justice to the intergrading nature of soil population (McBratney and De Gruijter, 1992; Burrough et al., 1997). The work, however, has not advanced much since the 1990s, mostly because of insufficient national and international databases of soil profiles to produce harmonized numerical national or international soil classification. There are, currently, few efforts devoted to providing an unified global soil classification which can serve as a basis to regional classification and synchronization between systems (McBratney, 2010; Hempel et al., 2013; Hartemink, 2015; Wadoux et al., 2021).

There currently exist two, seemingly disjoint, global soil classification systems, the USDA Soil Taxonomy (ST, Soil Science Division Staff, 2017) and the World Reference Base for Soil Resources (WRB, IUSS Working Group, 2015; IUSS Working Group WRB, 2022), and many regional and national systems (for an overview, see Krasilnikov et al., 2009). The WRB was created to harmonize classification by establishing consensus on key soil groupings. These groupings are defined

* Correspondence to: Laboratory of soil-agrosystem-hydrosystem interaction (LISAH), 2 place Pierre Viala, 34090 Montpellier, France.
E-mail address: alexandre.wadoux@inrae.fr (A.M.J.-C. Wadoux).

https://doi.org/10.1016/j.geoderma.2024.116915
Received 7 October 2023; Received in revised form 9 May 2024; Accepted 9 May 2024
Available online 25 May 2024

quantitatively based on measurable field properties and morphological characteristics, which reflect the processes of soil formation (pedogenesis). The WRB comprises two hierarchical levels: reference soil groups and uniquely defined qualifiers for specific soil characteristics. The ST is structured with six hierarchical levels (order, sub-order, great group, sub-group, family and series) and incorporates both soil properties and climate as expressions of pedogenesis. While both systems are acknowledged as international, there remain various examples of their shortcomings for local applications and communication with established regional classification systems. Charzyński (2006), for example, reclassified soil profiles from Poland according to the WRB taxonomy but found few correspondences between analytical procedures used in Poland and those suggested in WRB. Other examples of such shortcomings are described for Australia (e.g. Morand, 2013) or Benin (e.g. Azuka et al., 2015). There has been an attempt to coalesce a number of systems including ST and WRB by recognizing the centroids of existing classes to produce a universal soil classification system (Minasny et al., 2010) — that work is ongoing. This does not involve the creation of new classes. In order to do that, a numerical soil classification that addresses the discrepancies mentioned previously and achieves harmonization with existing national systems could be attempted.

The development of a soil horizon classification system — as opposed to the classification of soil profile entities, is sought to be a possible first step towards achieving a global numerical soil classification (see Hempel et al., 2013, Goal 4 in "Diagnostic and Soil Profile Information Harmonization"). The notion of a soil horizon as an individual was advocated by McBratney (1993) following the earlier proposal of FitzPatrick (1971) and FitzPatrick (1993) to use horizons as starting point from which to build a classification, and by Bouma (1989) in his "building block" approach.

In existing classification systems, however, topsoil layers are disregarded or incompletely characterized. This is yet necessary because topsoils vary greatly in space and are the support for agricultural production and drive most assessments of soil quality and health. The ST, for example, suggests that the placement of a soil in the taxonomy should not be influenced by topsoil management features, such as tillage, up to a depth of 8 to 25 cm (Soil Science Division Staff, 2017). In WRB, while numerous qualifiers can help in characterizing soils in the A horizon and there exist numerous reference soil groups in case of which the soil features of the A horizon are input for the classification of soils in the WRB, the WRB system does not explicitly describe dynamic topsoil horizon features such as organic matter and biological features (Broll et al., 2006). The need for a detailed description and classification of topsoils was recognized in Broll et al. (2006), but despite a few attempts in the literature (e.g. Buol et al., 1975; FAO, 1998), a detailed topsoil classification is still lacking. A notable exception is the Fertility Capability Classification system (Sanchez et al., 2003) which accounts for subsoil as well as topsoil features and biological properties, and group soils according to their potential fertility.

A systematic numerical approach to classification could be applied by grouping key soil properties to logical categories for horizons, the results of which can better define existing diagnostic horizons features or suggest new ones. In various countries and globally there now exist many common sets of predicted soil attributes with many predicted over the same depth ranges following the GlobalSoilMap specifications (Arrouays et al., 2014). The attributes are considered key soil properties. This raises the question of whether these sets of key soil properties, predicted separately, coalesce into identifiable clusters or classes and manifest discernible spatial or pedological patterns. If this is indeed the case this is the beginning of a solution to one of the remaining problems in soil classification, that is, the construction of a data-based objective numerical taxonomic classification of soil.

The objectives of this paper are to identify soil surface layer classes and to describe the application of a numerical soil classification to global soil data so as to produce a general purpose map of soil classes of the world surface layer.

2. Materials and methods

2.1. Purpose of the classification

The classification is natural and in support of two purposes: (i) organizing sets of soil properties predicted separately into identifiable classes so as to stimulate the understanding of whether these classes manifest discernible spatial or pedological patterns, and (ii) serving as a base to make an objective numerical taxonomic classification of soil, in this case soil layers. In this study we focus only on the surface material but in the longer run a set of horizon and profile classes.

2.2. Soil data

We collected a global dataset for ten surface (0–5 cm) soil properties. Owing to the difficulty to obtain surface measurements well spread throughout the globe, we used maps of soil properties obtained through SoilGrids 2.0 (Poggio et al., 2021). The study used the global predictions for bulk density, cation exchange capacity, texture, pH, and organic carbon content, which we used in combination with the total nitrogen content to create the CEC/clay, SOC/clay and SOC/N indices. Table 1 summarizes the soil properties along with their short name, description and unit.

2.3. Clustering

The maps of soil properties were partitioned into a number of homogeneous and disjoint classes using the k-means clustering algorithm (Lloyd, 1982; Hartigan and Wong, 1979). It is a popular and computationally efficient algorithm for large datasets. In k-means clustering, a number of k non-overlapping clusters are created by partitioning the matrix $\mathbf{X}$ containing the data to be clustered. The partitions are created so that a criterion is minimized. The criterion is the within-cluster sum of squared error, calculated as the squared distance to the cluster centroids (i.e. the means are the cluster-specific means of the variables). The criterion is minimized using an optimization algorithm that searches for the optimal value of the clustering criterion by rearranging existing partitions and keeping the ones that provide improvements (Landau et al., 2011). The steps in k-means clustering are as follows:

1. An initial set of k centroids are selected randomly.
2. All points in $\mathbf{X}$ are assigned to their nearest centroids in terms of Euclidean distance between the point and the centroids.
3. The cluster centroids are re-calculated with the points assigned to the cluster.
4. The process is repeated until the partitions are stable, that is, when the centroids of the previous rounds are close to the centroids of the current round using a user-defined tolerance value, or when the maximum number of iterations is reached.

Prior to clustering, the dataset was transformed by Cholesky decomposition of the variance–covariance matrix of $\mathbf{X}$, i.e. $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ where $\mathbf{C}$ is the variance–covariance matrix of $\mathbf{X}$ and $\mathbf{L}$ is a lower triangular matrix with positive diagonal values. The transformation was applied by matrix multiplication of $\mathbf{X}$ by $\mathbf{L}^{\mathrm{T}}$ so that $\mathbf{Y} = \mathbf{X}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{Y}$ is the transformed dataset. Clustering on $\mathbf{Y}$ using the Euclidean distance is equivalent to clustering using the Mahalanobis distance calculated on $\mathbf{X}$ (Anderson, 2003, p. 80).

We selected the optimal number of clusters with the elbow method for a range of clusters between 5 and 200 by steps of 5 and between 500 to 1000 in steps of 50. The evaluation criteria of the clustering quality were (i) the total within-cluster sum of square (WSCC), i.e. the sum of the within-cluster sum of squares averaged over all clusters, (ii) the variance explained, that is, the sum of the within-cluster sum of squares averaged over all clusters divided by the total sum of squares, and (iii) the Calinski–Harabasz statistic (Caliński and Harabasz, 1974),

Table 1
List of soil maps used as input to the numerical classification system along with their short names, description and unit.
Source: Adapted from Poggio et al. (2021).

Variable	Short name	Description	Unit
Bulk density	Bulk density	Bulk density of the fine earth fraction oven dry	cg/cm^3
Cation exchange capacity	CEC	Capacity of the fine earth to hold exchangeable cations	$cmol_c/kg$
Clay content	Clay	Gravimetric contents of clay in the fine earth fraction of the soil	percent
Silt content	Silt	Gravimetric contents of silt in the fine earth fraction of the soil	percent
Sand content	Sand	Gravimetric contents of sand in the fine earth fraction of the soil	percent
pH in water	pH	Negative common logarithm of the activity of hydronium ions in water	unitless
Organic carbon concentration	SOC	Gravimetric content of organic carbon in the fine earth fraction of the soil	percent
Ratio of cation exchange capacity to clay	CEC/clay	Index of clay mineralogy or of "CEC Activity"	mol_c/kg clay
Ratio of organic carbon to clay	SOC/clay	Indicator of soil structure quality (Prout et al., 2021)	unitless
Ratio of organic carbon to nitrogen	SOC/N	Indicator of the nitrogen immobilization or mineralization during organic matter decomposition by micro-organisms (Swift et al., 1979)	unitless

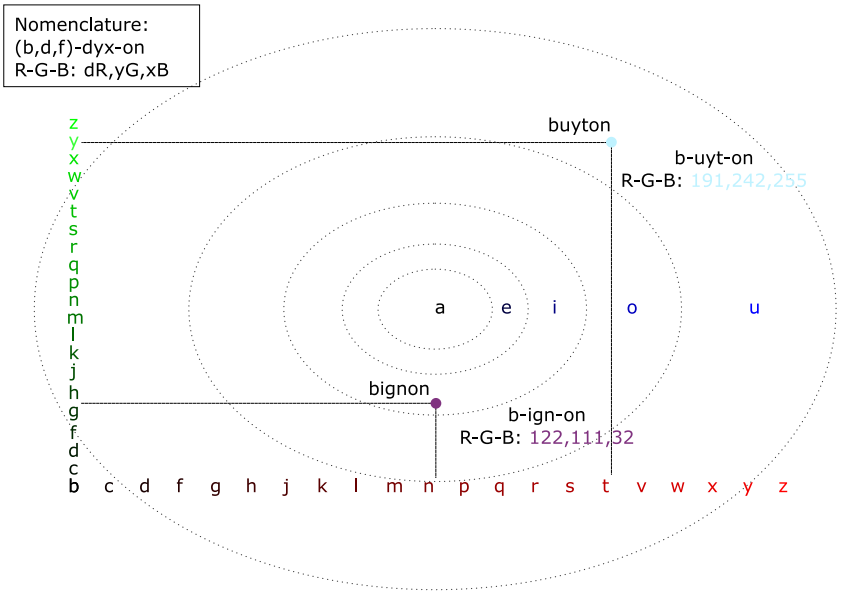


Fig. 1. Representation of the transformed PCA space with indications on the nomenclature system.

which is the ratio of the sum of between-cluster dispersion and the sum of within-cluster dispersion, for all clusters. With these three criteria, the optimal number of clusters was decided heuristically by plotting the number of clusters against the criterion value, and by selecting the elbow of the criterion curve as the optimal number.

2.4. Principal component analysis

A principal component analysis (PCA) was used to investigate the pattern of variation of the clusters in association with the soil property maps of Table 1. PCA is a dimensionality reduction method designed to transform a set of variables into a reduced number of uncorrelated variables called principal components, each of which is a linear combination of the original variables. In PCA, it is expected that few components account for most of the variation of the original set of variables. PCA was applied on the set of ten attributes (Table 1) after which the class-specific centroids were calculated for each component.

2.5. Space transformation and nomenclature

The locations of the individual class-specific centroids displayed in the PCA biplot are likely to yield an uneven distribution, making the development of a surface layer nomenclature challenging. Before naming the classes, we performed a space transformation using the spacebender algorithm of McBratney and Minasny (2013). The algorithm aims at equalizing the space between observations by projecting them onto a new t space. Transformation is made by computing a pairwise distance matrix between the transformed and original space using a fat-ruler algorithm. The algorithm aims to make the observation more equally distributed than the original observation (i.e. to equalize the variance). It makes use of a user-defined parameter w which is the distance between two points. Next, a principal coordinate analysis is applied on the distance matrix to convert it to a centred matrix.

The nomenclature is defined in the transformed space and has three components (Fig. 1), two of which are obtained from the location of the centroid in the transformed space of the PCA biplot. In the transformed space, consider a series of five circles around the origin at coordinate (0,0). The radius of the circles is obtained by taking the quintiles of the transformed distance, and the area of the circles is named by a vowel: -a for the smallest circle, followed by -e, -i, -o and -u. Both the x- and y-axes are defined by a consonant, from left to right (x-axis) or bottom to top (y-axis). The combination of letters (i.e. distance from the origin, x- and y- axes) gives information to name the centroids.

The three-component nomenclature is defined as follows: The first component is a consonant, either -b, -d or -f. The letter -b is used in all cases, and if two classes have the same name, the -b is replaced by -d and so on. For example, three classes called "buyton" become "buyton", "duyton" and "fuyton". The second component is a letter, one of the three (i.e. in the order d, y and x) corresponding to the centroid location on the graph: d is the distance from the origin associated to a vowel (i.e. *a*, *e*, *i*, *o*, or *u*); y corresponds to a consonant of the y-axis; x corresponds to a consonant in the x-axis. Finally, the qualifying suffix of *-on* is added onto the word.

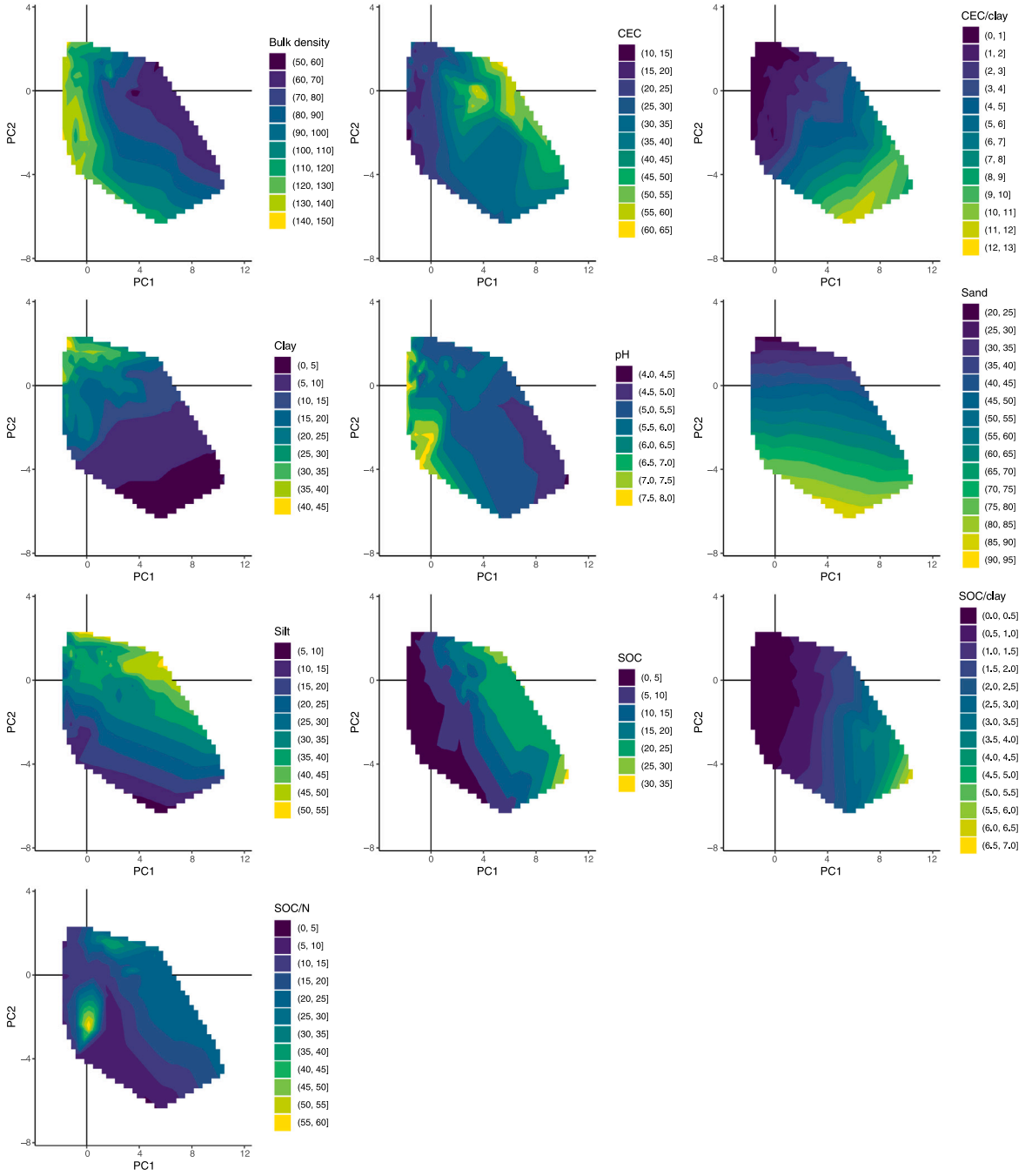


Fig. 2. Contour plot of the soil properties values in the first two principal components space.

2.6. Practical implementation and computational aspects

To speed up processing, clustering was made on a subset of 10^6 locations obtained from the set of soil properties maps and using a systematic random sampling. The clustering was performed with a *k*-means++ initialization (Arthur and Vassilvitskii, 2007) with 50 repetitions, a maximum number of iterations set to 1000 and a tolerance value of 0.0001. Clustering was made with the KMeans_rcpp function of the ClusterR package (Mouselimis, 2023). After clustering, the centroids were used to assign each pixel to its closest centroids after the transformation using the Cholesky matrix.

Principal component analysis was performed on the same subset of 10^6 locations described. The subset was centred at 0 and standardized to unit variance before calculation. The centroids were taken as the mean of the class-specific values of the property, assuming that the sample is large enough to obtain a realistic estimate of the centroids.

The spacebender algorithm was implemented in MatLab. We tested various window size from 2 to 50 by steps 2. The coefficient of variation was plotted against the window size to select a window of size of 2.

3. Results

We applied the *k*-means classification algorithm to the data matrix for a range of potential class sizes and calculated the three cluster number metrics for the range of potential class sizes. These showed a reasonably even monotonic change with class number for the three criteria. This suggests that there might not be a natural number of classes, and that we are potentially partitioning a very weakly clustered

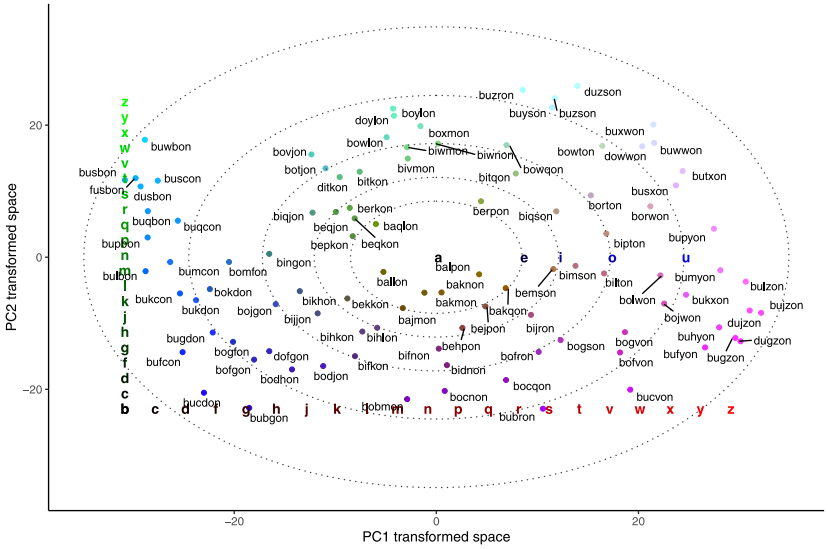


Fig. 3. Distribution of the 100 class centroids projected on to the transformed space of the first two principal components along with the class name. The colours refer to the nomenclature from Fig. 1; the x-axis represents a gradient of red, the y-axis a gradient of green and the distance from the origin at (0,0) a gradient a blue.

continuum. We choose a partitioning such as a further increase in number of classes resulted in relatively little gain for the three criteria. Hereafter the number of classes was set to 100.

The principal component analysis on the set of ten soil property maps (Table 1) showed that 6 components accounted for 97% of the variation. The first and second components accounted for 49% and 22% of the variance, respectively. Fig. 2 shows plots of the first two components with a contour plot of the soil property values. All properties have a clear pattern with detailed variation. Bulk density, CEC, pH, SOC and SOC/clay have a gradient of increasing (bulk density, pH) or decreasing (CEC, SOC, SOC/clay) values in the x-axis, whereas CEC/clay, clay, sand and silt have a variation in a y-axis.

Fig. 3 shows the distribution of the 100 class centroids projected on to the transformed space of the first two principal components. The name of each centroid is given by the nomenclature (see Section 2.5). The transformed space is occupied fairly well by the centroids of the 100 classes.

When the transformed space is collapsed on to the original space of the first two principal components (Fig. 4) the contribution of the soil properties to the location of the centroids appears more clearly. The thin grey lines indicate the density of observations for densities of 100,000, 10,000, 1000 and 10 values. The blue arrows are the vectors projected on to the component axes. All properties but texture contribute largely to the first component whereas clay, silt and sand contribute to the second. Dynamic properties seem represented in one dimension in the first component. The second component, uncorrelated to the first, reflects differences in the surface layer brought by stable properties. The distribution of the cluster centroids in the plane of first two components shows strong clustering around the origin. Two modes appear similarly in the first two components: most centroids have either slightly negative or positive in the component axes.

We searched for class representants searching for the closest individual to the centroid, for all centroids — these were called examplars by McBratney (1994). A map of examplars is shown in Fig. 5. There is great diversity in the density of such examplars globally. The northern hemisphere has more examplars than the southern one. Areas in north America, a large band spanning eastern Europe and Russia, and western Australia are covered with a high density of examplars, some being geographically close to each other. Africa, South America, and Southern Asia have, conversely, a relatively low density of such examplars. There are no examplars in the south of North America, in Antarctica, in Oceania excluding Australia.

Fig. 6 shows the spatial distribution of the 100 surface layer classes. There is a strong spatial patterning with a gradient of classes related to the latitude. For example, Australia, the continents of Africa and Asia have large bands of the same class spanning a longitudinal gradient. While all classes have a clear spatial pattern and seem geographically compact, some areas such as North America and Asia show a more patchy distribution of classes. The number of classes and the global scale make further discussion on the spatial distribution of the classes a challenging exercise.

4. Discussion

The total WSCC, variance explained and Calinkski–Harabasz values did not reveal a clear optimal number of classes. We clustered a weakly clustered continuum, which is mostly due to the nature of the data (i.e. maps of predicted properties presumably with some spatial and attribute smoothing) which we used as input. In the absence of a definite cut-off value, we used a number of classes that offered a compromise between within-class homogeneity and ease of use of the classification (i.e. tractability, see de Gruijter, 1977, Section 3.4.2.1). While it would have been preferable to obtain this number with a formal criterion, there are still advantages in partitioning a weakly-clustered continuum since we reveal classes while all input maps of soil attributes were originally predicted separately, and we do more justice to the way soils behave and vary in space.

The soil surface layer class map shows a strong spatial patterning without too much repetition from one place to another. Some classes, however, repeat in different continents or hemispheres. This is the case for the "bojwon" class, which occurs in eastern Africa, but also in India, or the "bodhon" class which occurs in a large latitude band covering Asia, North America and Africa (Fig. 6 and Supplementary Material, Fig. 1). To reveal groups of classes with similar characteristics, we further treated the class centroids as individuals and performed an agglomerative hierarchical clustering. The dissimilarity between centroids was assessed with the same distance as that described in Section 2.3 using Ward's method. The dendrogram (Supplementary Material, Fig. 2) divided into 10 branches revealed two main groups of 55 and 21 surface classes, respectively, and 5 groups with more than 2 soil layer classes, whereas the remaining 3 branches had only one centroid.

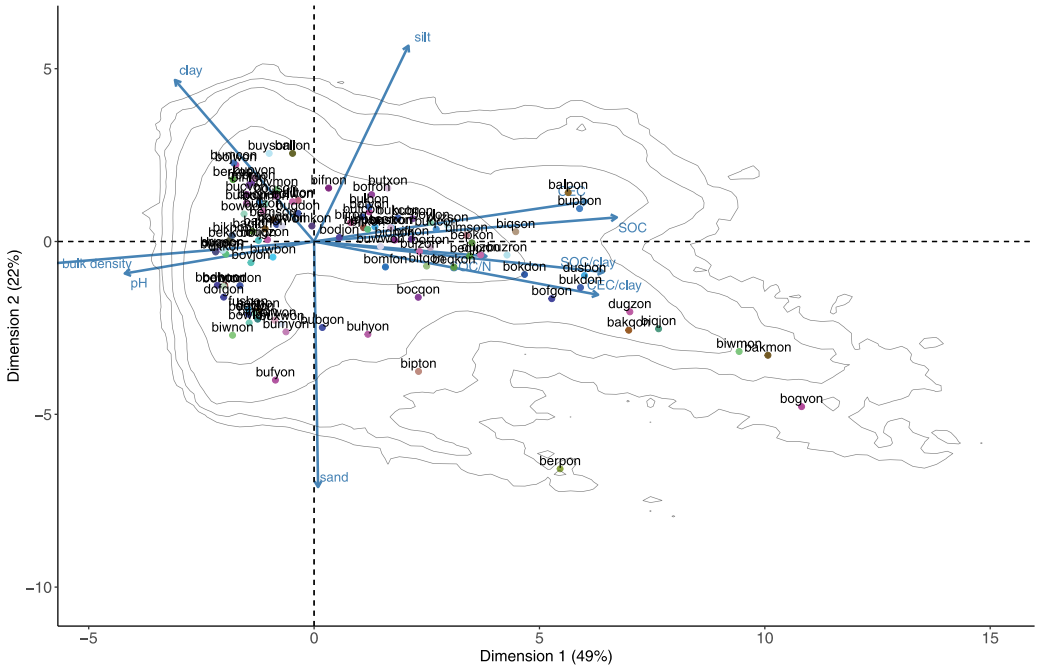


Fig. 4. Distribution of the centroids in the first two principal component axes along with vectors showing the contribution of the original properties (blue arrows) and density of observation (grey lines) for a densities of 100,000, 10,000, 1000 and 10 values.

This suggested that some layer classes have similar characteristics but that some classes are very different. The map of the group of similar classes (Supplementary Material, Fig. 3) showed that the two major groups of classes repeat in continents with a strong climate gradient. Comparison of Fig. 6 and Supplementary Material, Fig. 3 with grouping of soil regions given by Nachtergaele (1999) based on the FAO world reference base for soil resources (FAO/UNESCO, 2003) show strong similarities in pattern in all continents.

More important than the boundaries, although seldom reported, are the centroids and the examplars. They enable the transfer of information and knowledge between places and the allocation of new individuals to an existing classification based on the actual value of the centroid or examplar and taxonomic distance metrics. This approach, advocated in Minasny et al. (2010), was used to compare classifications systems (e.g. Láng et al., 2013; Van Huyssteen et al., 2014) and to build transfer functions between systems (e.g. Michéli et al., 2016; Hughes et al., 2017). Most classification systems, however, have a very coarse classification of surface horizons. The WRB and ST, for example, have 13 diagnostic surface horizons and 8 epipedons, respectively. The WRB has the Mollic, Chernic, Umbric, Vertic, Plaggic diagnostic horizons, and the Ochric qualifier. The FAO-Unesco Soil Map of the World, similarly, contains four A horizons and one H horizon (FAO-UNESCO-ISRIC, 1988; IUSS Working Group, 2015; Soil Science Division Staff, 2017; Broll et al., 2006; IUSS Working Group WRB, 2022). Our classification provides the first global-scale numerical classification of topsoil (surface layer), from which distance between centroids can be calculated with previous systems.

This work built on existing global maps of soil properties. This brings additional challenges as maps are predictions and not measured properties. This means that the maps usually contain more uncertainty than measured values and that they are a smooth representation of the soil properties. We were also unable to include some variables to better capture the set of properties needed in usual classification systems, such as rock content or colour. On these properties, colour is the most important one, being a principal qualifier for diagnostic horizons in most classification systems. The colour relates to the mineral and organic composition of the soil, as well as to a number of moisture-related characteristics. Despite a few attempts at regional (Poppiel et al., 2020) and national (Liu et al., 2020) scale, there exists currently, to our knowledge, no global map of soil colour. In a recent study, Rizzo et al. (2023) made a global map of bare soil areas — mostly cropland, including remote sensing imagery and ground-point spectroscopic data. Further work is needed, however, to interpolate the colour at various depth intervals and in vegetated areas. In future work, we would ideally use the exhaustive set of 23 properties suggested by Hughes et al. (2018), which next to colour and the set of properties used in this study also encompass water and ice content, carbonate content, exchangeable cations, acid saturation, exchangeable sodium percentage, electrical conductivity, and gypsum content. Since several of these properties are dynamic, that is, change over time in response to human activities, including them in addition to the ones presented in Table 1 would certainly increase the capacity of our natural classification to account for temporal alteration due to land management.

To many, especially those interested in soil quality, health or security (Karlen et al., 2003; Kibblewhite et al., 2008; Evangelista et al., 2023), the properties of the surface horizon are important and most subject to change over time in response to management. While current conventional soil classification systems recognize surface layer classes, of the order 10 or 20 (Broll et al., 2006) — it is too coarse for a realistic interpretation for soil condition and capacity assessment. A detailed classification of surface layers adds considerably to the potential understanding and monitoring of soil capacity and condition of surface soils. This could be achieved by building a technical classification for this surface layer natural classification. There are two main types of classification that are widely recognized in the literature. The first type are natural classifications and the second are technical classifications. Cline (1949) discusses the value of natural classification, from which many possible technical classifications can be derived. The WRB taxonomy, for example, attempts to be a natural classification that can be the basis for interpretation for research on land use capability or soil fertility. Our classification being natural, it could potentially be used for practical and technical purposes to build a technical classification that relate directly to soil health, quality and security assessment.

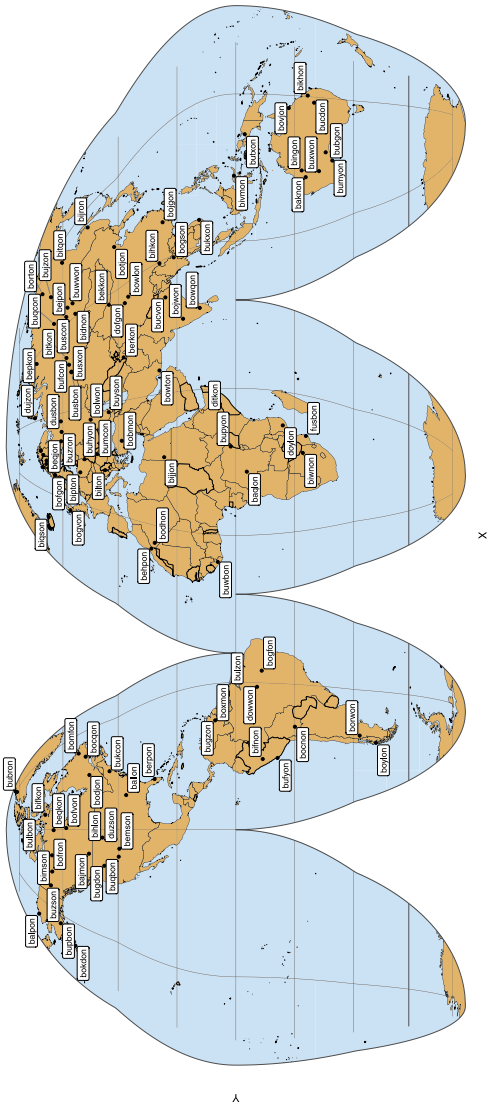


Fig. 5. Map of location of examplars (i.e. individual the closest to the centroid) for the 100 surface layer classes.

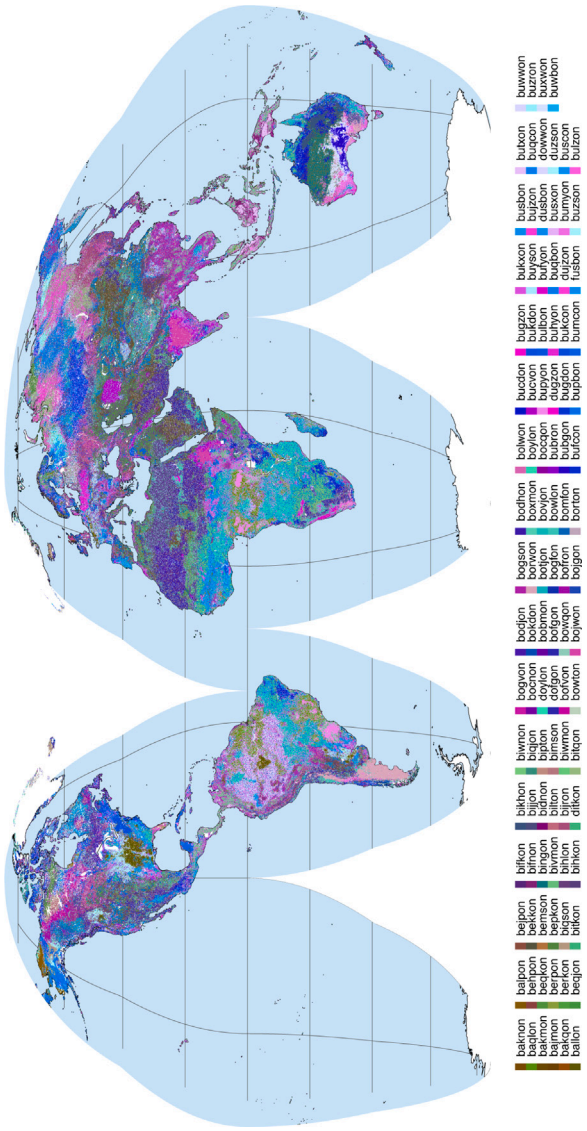


Fig. 6. Global map of the spatial distribution of the 100 surface layer classes.

The next step is a horizon or layer classification of soils for all layers followed by a profile classification based on the sequence of layers. A comprehensive classification for all layers would enable a comparison between systems. A first attempt was made with the dynamic Comprehensive Soil Classification System (CSCS, McBratney et al., 2022) by merging existing soil taxa from several systems by sequentially adding new centroids based on soil properties. While this approach has been developed and tested (see, for example, Shahbazi et al., 2018), much efforts remain to build a global representation of the soils. This was also highlighted as one of the most pressing challenges of our discipline (see Wadoux et al., 2021, Challenge 2).

5. Conclusion

We reported on the development and implementation of a numerical surface layer classification system. A total of 100 surface classes were obtained using maps of soil properties as input, and their centroids were named with a nomenclature that accounts for their location in the two-dimensional space of the principal components. A global map of the newly created soil classes is provided as well as the location of the 100 class-representants known as examplars, i.e. individual the closest to the centroid. From the results and discussion we draw the following conclusions:

- The 100 surface classes were obtained by clustering a weakly clustered continuum. This might be due to the nature of the data that we used as input (i.e. maps of soil properties).
- The pattern of the surface classes distribution followed well-defined soil geographies, in particular similar to that from the FAO world base resources map. Existing world soil maps, however, do not discriminate the surface layer in such detail.
- The nomenclature was based on the centroids. Centroids enable the transfer of information and knowledge between places and classification systems. The next step is a layer classification of soils for all layers. This would enable building a representative global soil classification system in which centroids from several systems could sequentially be added.
- In this study, we found that three out of 100 classes were significantly different from the rest of the classes, as shown by a grouping of the class centroids using a distance metric.
- Such a detailed classification of surface soil has merit for soil quality, health and security assessment.

Future work should focus on the validation of this new topsoil layer taxonomy. Focus should also be on the creation of a global soil taxonomy using a harmonized dataset of say 23 soil properties at various depth intervals. The methodology presented here could serve as basis for building such a global soil taxonomy. The use of fuzzy *k*-means and the "akromeson" algorithm (Hughes et al., 2014) could be investigated for this purpose.

CRediT authorship contribution statement

Alexandre M.J.-C. Wadoux: Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alex B. McBratney:** Writing – review & editing, Visualization, Validation, Resources, Investigation, Conceptualization.

Declaration of competing interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

Data availability

The authors do not have permission to share data.

Acknowledgements

We acknowledge the support of the Australian Research Council Laureate Fellowship (FL210100054) on Soil Security entitled "A calculable approach to securing Australia's soil". For the purpose of Open Access, a CC-BY public copyright licence has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.geoderma.2024.116915.

References

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, third ed. In: Wiley Series in Probability and Statistics.

Arrouays, D., McBratney, A.B., Minasny, B., Hempel, J.W., Heuvelink, G.B.M., MacMillan, R.A., Hartemink, A.E., Lagacherie, P., McKenzie, N.J., 2014. The GlobalSoilMap project specifications. In: Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.R., McBratney, A.B. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press, London.

Arthur, D., Vassilvitskii, S., 2007. k-means++: the advantages of careful seeding. In: SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, pp. 1027–1035.

Azuka, C.V., Igué, A.M., Diekkrüger, B., Igwe, C.A., 2015. Soil survey and soil classification of the Koupendri catchment in Benin, West Africa. Afr. J. Agric. Res. 10 (42), 3938–3951.

Bouma, J., 1989. In: Bouma, J., Bregt, A. (Eds.), Land Qualities in Space and Time. Pudoc, Wageningen.

Broll, G., Brauckmann, H.-J., Overesch, M., Junge, B., Erber, C., Milbert, G., Baize, D., Nachtergaele, F., 2006. Topsoil characterization—recommendations for revision and expansion of the FAO-Draft (1998) with emphasis on humus forms and biological features. J. Plant Nutr. Soil Sci. 169 (3), 453–461.

Buol, S.W., Sanchez, P.A., Cate, Jr., R.B., Granger, M.A., 1975. Soil fertility capability classification. In: Soil Management in Tropical America. North Carolina State Univ., USA, pp. 126–141.

Burrough, P.A., van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77 (2–4), 115–135.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Comm. Statist.-Theory Methods 3 (1), 1–27.

Charzyński, P., 2006. Testing WRB on Polish Soils. Association of Polish adult educators, Toruń Branch.

Cline, M.G., 1949. Basic principles of soil classification. Soil Sci. 67 (2), 81–92.

De Bakker, H., 1970. Purposes of soil classification. Geoderma 4 (3), 195–208.

Evangelista, S.J., Field, D.J., McBratney, A.B., Minasny, B., Ng, W., Padarian, J., Dobarco, M.R., Wadoux, A.M.J.-C., 2023. A proposal for the assessment of soil security: Soil functions, soil services and threats to soil. Soil Secur. 10, 100086.

FAO, 1998. Topsoil Characterization For Sustainable Land Management. Land and Water Development Division, Soil Resources, Management and Conservation Service, Rome.

FAO-UNESCO-ISRIC, 1988. Revised Legend, FAO-Unesco SoilMap of the World. World Soil Resources Reports 60, FAO, Rome.

FAO/UNESCO, 2003. FAO/UNESCO Soil Map of the World. FAO, Rome, https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/faounesco-soil-map-of-the-world/en/.

FitzPatrick, E.A., 1971. Pedology, a Systematic Approach to Soil Science. Oliver and Boyd, Edinburgh.

FitzPatrick, E.A., 1993. Principles of soil horizon definition and classification. CATENA 20 (4), 395–402.

de Gruijter, J.J., 1977. Numerical Classification of Soils and Its Application in Survey (Ph.D. thesis). Wageningen University and Research, the Netherlands.

Hallsworth, E.G., 1965. The relationship between experimental pedology and soil classification. In: Hallsworth, E., Crawford, D. (Eds.), Experimental Pedology. Butterworths, London, pp. 354–374.

Hartemink, A.E., 2015. The use of soil classification in journal papers between 1975 and 2014. Geoderma Reg. 5, 127–139.

Hartigan, J.A., Wong, M.A., 1979. A k-means clustering algorithm. Appl. Stat. 28, 126–130.

Hempel, J., Michéli, E., Owens, P., McBratney, A.B., 2013. Universal soil classification system report from the International Union of Soil Sciences Working Group. Soil Horiz. 54 (2), 1–6.

Hughes, P., McBratney, A.B., Huang, J., Minasny, B., Hempel, J., Palmer, D.J., Micheli, E., 2017. Creating a novel comprehensive soil classification system by sequentially adding taxa from existing systems. Geoderma Reg. 11, 123–140.

Hughes, P., McBratney, A.B., Huang, J., Minasny, B., Micheli, E., Hempel, J., 2018. A nomenclature algorithm for a potentially global soil taxonomy. Geoderma 322, 56–70.

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

IUSS Working Group, 2015. World Soil Resources Reports, World Reference Base for Soil Resources 2014: International Soil Classification System for Naming Soils and Creating Legends for Soil Maps, vol. 106, FAO, Rome, Italy, p. 192.

IUSS Working Group WRB, 2022. World Reference Base for Soil Resources. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps, World Soil Resources Reports, fourth ed. International Union of Soil Sciences (IUSS), Vienna, Austria.

Karlen, D.L., Ditzler, C.A., Andrews, S.S., 2003. Soil quality: why and how? Geoderma 114 (3–4), 145–156.

Kibblewhite, M.G., Ritz, K., Swift, M.J., 2008. Soil health in agricultural systems. Philos. Trans. R. Soc. B 363 (1492), 685–701.

Krasilnikov, P., Martí, J.-J.I., Arnold, R., Shoba, S., 2009. A Handbook of Soil Terminology, Correlation and Classification. Routledge, London, UK.

Landau, S., Leese, M., Stahl, D., Everitt, B.S., 2011. Cluster Analysis, fifth ed. John Wiley & Sons.

Láng, V., Fuchs, M., Waltner, I., Michéli, E., 2013. Soil taxonomic distance, a tool for correlation: As exemplified by the Hungarian Brown Forest Soils and related WRB Reference Soil Groups. Geoderma 192, 269–276.

Liu, F., Rossiter, D.G., Zhang, G.-L., Li, D.-C., 2020. A soil colour map of China. Geoderma 379, 114556.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inform. Theory 28 (2), 129–137.

McBratney, A.B., 1993. Some remarks on soil horizon classes. CATENA 20 (4), 427–430.

McBratney, A.B., 1994. Allocation of new individuals to continuous soil classes. Soil Res. 32 (4), 623–633.

McBratney, A.B., 2010. Numerical approaches to a Universal Soil Classification System. unpublished work, Soil Science Society of America 2010 International Annual Meeting Long Beach, California.

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 (1), 159–175.

McBratney, A.B., Minasny, B., 2013. Spacebender. Spat. Stat. 4, 57–67.

McBratney, A.B., Minasny, B., Huang, J., Arrouays, D., Richer-de Forges, A.C., Savin, I., Michéli, E., Cunha dos Anjos, L.H., Gelsleichter, Y., Jeon, S.H., 2022. CSCS2.0: A comprehensive soil classification system for quantitatively identifying soils across the world. Abstract - 22nd World Congress of Soil Science, 31 July - 5 August, Glasgow, UK.

Michéli, E., Láng, V., Owens, P.R., McBratney, A.B., Hempel, J., 2016. Testing the pedometric evaluation of taxonomic units on soil taxonomy—A step in advancing towards a universal soil classification system. Geoderma 264, 340–349.

Minasny, B., McBratney, A.B., Hartemink, A.E., 2010. Global pedodiversity, taxonomic distance, and the World Reference Base. Geoderma 155 (3–4), 132–139.

Morand, D.T., 2013. The World Reference Base for Soils (WRB) and Soil Taxonomy: an appraisal of their application to the soils of the Northern Rivers of New South Wales. Soil Res. 51 (3), 167–181.

Mouselimis, L., 2023. ClusterR: Gaussian mixture models, K-means, mini-batch-kmeans, K-medoids and affinity propagation clustering. URL: https://CRAN.R-project.org/package=ClusterR. R package version 1.3.1.

Nachtergaele, F.O., 1999. From the soil map of the world to the digital global soil and terrain database: 1960–2002. In: Sumner, M.E. (Ed.), Handbook of Soil Science. CRC Press, Boca Raton, pp. H5–17.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL 7 (1), 217–240.

Poppiel, R.R., Lacerda, M.P.C., Rizzo, R., Safanelli, J.L., Bonfatti, B.R., Silvero, N.E.Q., Demattê, J.A.M., 2020. Soil color and mineralogy mapping using proximal and remote sensing in midwest Brazil. Remote Sens. 12 (7), 1197.

Prout, J.M., Shepherd, K.D., McGrath, S.P., Kirk, G.J.D., Haefele, S.M., 2021. What is a good level of soil organic matter? An index based on organic carbon to clay ratio. Eur. J. Soil Sci. 72 (6), 2493–2503.

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

Rizzo, R., Wadoux, A.M.J.C., Demattê, J.A.M., Minasny, B., Barrón, V., Ben-Dor, E., Franco, N., Savin, I., Poppiel, R., Silvero, N.E.Q., Terra, F.S., Rosina, N.A., Rosas, J.T.F., Greschuk, L.T., Ballester, M.V.R., Gómez, A.M.R., Bellinaso, H., Safanneli, J.L., Chabrillat, S., Fiorio, P.R., Das, B.S., Malone, B.P., Zalidis, G., Tziolas, N., Tsakiridis, N., Karyotis, K., Samarinas, N., Kalopesa, E., Gholizadeh, A., Shepherd, K.D., Milewski, R., Vaudour, E., Wang, C., Mohamed, E.S., 2023. Remote sensing of the Earth's soil color in space and time. Remote Sens. Environ. 120 (7), 1197.

Sanchez, P.A., Palm, C.A., Buol, S.W., 2003. Fertility capability soil classification: a tool to help assess soil quality in the tropics. Geoderma 114 (3–4), 157–185.

Shahbazi, F., Huang, J., McBratney, A.B., Hughes, P., 2018. Allocating soil profile descriptions to a novel comprehensive soil classification system. Geoderma 329, 54–60.

Soil Science Division Staff, 2017. In: Ditzler, C., Scheffe, K., Monger, H.C. (Eds.), Soil Survey Manual. In: USDA Handbook, vol. 18, Government Printing Office, Washington, D.C., USA.

Swift, M.J., Heal, O.W., Anderson, J.M., Anderson, J.M., 1979. Decomposition in Terrestrial Ecosystems, vol. 5, University of California Press.

Van Huyssteen, C., Michéli, E., Fuchs, M., Waltner, I., 2014. Taxonomic distance between South African diagnostic horizons and the World Reference Base diagnostics. CATENA 113, 276–280.

Wadoux, A.M.J.-C., Heuvelink, G.B.M., Lark, R.M., Lagacherie, P., Bouma, J., Mulder, V.L., Libohova, Z., Yang, L., McBratney, A.B., 2021. Ten challenges for the future of pedometrics. Geoderma 401, 115155.