Geoderma 322 (2018) 48–55

Contents lists available at ScienceDirect

Geoderma

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

Comparisons between USDA soil taxonomy and the Australian Soil Classification system II: Comparison of order, suborder and great group taxa

Philip Hughes[a,*], Alex B. McBratney[a], Budiman Minasny[a], Jingyi Huang[a,b], Erika Micheli[c], Jonathan Hempel[d], Edward Jones[a]

[a] Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Eveleigh, NSW 2015, Australia
[b] School of Biological, Earth and Environmental Sciences, UNSW Sydney, NSW 2052, Australia
[c] Department of Soil Science and Agricultural Chemistry, Szent István University, Gödöllő, Hungary
[d] National Soil Survey Centre, U.S. Department of Agriculture Natural Resources Conservation Service, Lincoln, NE 68508, USA

ARTICLE INFO

Handling Editor: M. Vepraskas

Keywords:
Principal component analysis
Centroids
Euclidean distance

ABSTRACT

Soil taxonomies over the world are incongruent- based on different tiers and different sets of properties. This second paper is concerned with understanding the relationships of each tier (Order, Suborder and Great Group) in both Soil Taxonomy (ST) and the Australian Soil Classification system (ASC) using mean nearest neighbour distances and convex hull areas in two principal component dimensions. It is determined that in most instances, convex hull comparisons using only two principal components, representing 30% of the variation in the data describe much of the variability between and within the orders, suborders and great groups of each classification system. These are useful for visual comparisons of taxa at various levels. Mean nearest-neighbour distances can include all 414 variables if necessary, which is more rigorous but complex. Both distance calculations and convex hulls highlight the same associations between taxa from ASC and ST. Both these methods demonstrate the robust soil classification capability of the ST and the ASC, but with convex hull sizes and nearest-neighbour distances that are smaller, the ASC proves to be slightly more coherent. The two systems occupy somewhat different areas in PC space, and ST covers a larger overall area, demonstrating that ASC is a purpose-built classification for Australian conditions while ST Is a more general system that can cover a wider variety of soils and management issues. We also show that great groups in ST are at about the same level of taxonomic generalization as great groups of the ASC. Combining the best elements and taxa of both these systems would be a positive step in the creation of a comprehensive system.

1. Introduction

Part I of this two-part paper is a roadmap which leads the way to data harmonization between two disparate classification systems, which are typically defined by a range of different factors. The significant achievement from this is twofold: the creation of the Total REference taXonomic database (TREX) - a combined database consisting of data from both the ASC and ST- and the calculation of taxonomic precision and accuracy - a measure of how well each taxonomic system defines its individual sub-units. This has never been attempted before, partially because data collection is not necessarily consistent between nations, partially because differing data sets from differing classification systems emphasize different aspects of soil- typically pertaining to important local issues, and partially because each different classification scheme is constructed differently, making a harmonious definition between soil profile descriptions (SPDs) and taxa difficult.

The solution which allows the initiation of TREX involves an ST centroid database as a starting point. This database has mean values of soil properties of ST taxa, described to the great group level, arranged according to depth along single vectors. This creates a dataset that is 299 rows (one row for each taxon) and 414 columns (one column for each of 23 properties at 18 standard depth increments). This data can be compared with data from other taxonomies, so long as the properties and depths are consistent. Data which could be used for each of the 23 properties at a variety of non-standard depths was taken from Australian SPDs in the Terrestrial Ecosystem Research Network (TERN) database. With this, each of the 23 required properties at the 18 standard depth increments could be estimated and splined. The mean values for this data at each tier of the ASC were calculated. TREX is the amalgamation of this data and the original ST centroid set. With TREX, side by side comparisons of ASC and ST taxa at the Order, Suborder or

* Corresponding author.
E-mail address: Philip.hughes@sydney.edu.au (P. Hughes).

https://doi.org/10.1016/j.geoderma.2018.02.022
Received 31 July 2017; Received in revised form 1 February 2018; Accepted 11 February 2018

Great group level can be made via a metric such as Euclidean or Mahalanobis distances. When converted, the 414 soil properties at depth could be represented by a lower number of components. The first 2 components represent 30% of the variability, 36 components represent 95% of the total variability. Accurate Euclidean distances can be rapidly achieved on 36 components as opposed to the original 414 columns as the overall size of the data is reduced by 90% without a corresponding loss to data quality. With these distances, the classification accuracy of each taxonomy could be obtained (Hughes et al., 2017). The way in which this was done was by the creation of a distance matrix of ST and ASC, pairing the taxa with the shortest Euclidean distances and obtaining the number of Great Groups which had Great Groups of the same Order as their nearest neighbour. This was expressed as a ratio of the total number of Great Groups within the Order-taxonomic accuracy. It was determined that ASC taxa have a higher taxonomic accuracy (Hughes et al., 2017).

1.1. Intra-distance analysis

Distance calculations are recognized as a method of understanding taxonomic features in the soil. Distances are a key component of cladistics analysis (Miltenyi et al., 2015) but can also be used for a variety of other analysis. In Hughes et al. (2017), the distances between objects have already been used to calculate taxonomic precision and accuracy. Other aspects of distance can be used to ascertain details on the construction and integrity of taxonomic systems.

The nearest-neighbour distance between principal components of properties of soil taxa can be used to determine the closest taxon to another. Useful for comparisons on a case-by-case basis, nearest neighbour distances lose meaning as the number of individual comparisons increase. The mean value, however, can be used as a measure of the coherence of a given group. It can be inferred that a group from a given taxonomy fits into a defined space if the nearest neighbour distance between all of its subgroups is small.

Both taxonomic systems studied, however, have further tiers to which SPDs can be classified. ST classifies to Order, Suborder, Great Group, Sub Group, Family and Series (Soil Survey Staff, 2010). ASC classifies to Order, Suborder, Great Group and Sub Group (Isbell, 2016). Beyond Sub Group of the ASC, the level of Family technically exists but it is rarely used or defined. For consistency, this study only evaluated the levels of Order, Suborder, and Great Group of the ST and ASC systems.

1.2. Inter-distance analysis

While intra-distance calculations can provide details on the construction of the taxonomy, inter-distance calculations can provide details on similarity. A distance matrix between means of groups such as ST Orders against ASC Orders can be calculated. The smallest Order to Order distance for a given Order is the taxonomically closest taxon. Comparing Orders alone (12 for ST and 14 for ASC) provides 168 distances. These can be ranked numerically from smallest to greatest distances between the orders of ST and ASC, creating a snapshot of which ST Orders have similarity to ASC Orders.

It should be noted that the ST uses soil climate to classify soils while the ASC does not use it at all. Therefore, these two systems can have a fundamental incompatibility. In this study, we only focus on the similarities of the soil properties measured at different depths between the soil taxa of the ST and the ASC systems.

1.3. Intra-hull analysis

With TREX operational, inter and intra taxonomic distances can be used to create a picture of the performance of each taxonomy on a global scale. Convex-hull analysis is used for a variety of applications from image correction to mapping (Barber et al., 1996). This simple technique can also be used in the assessment of taxonomic systems. The simplest method of visualizing convex hulls is by imagining what a particular group of data would look like if it were plotted on an x-y axis and a rubber band were stretched around the outside of the data points. The points that touch the rubber band and the shape which the band takes is the convex hull (Fadel et al., 2001).

The TREX database can provide metric axes in n-space on which the convex hull calculations can be carried out. All the soil properties of the different soil taxa are used to create a principal component space. If a soil classification system is hierarchical, then all soil taxa at the bottom tier pertaining to a particular taxonomy should occupy a large hull representing the greatest dissimilarity between the taxa. Hull taxa from the next tier up should be completely contained within the hull of the bottom tier, and the third tiers contained by the second. Graphically the setup should resemble Babushka dolls, a series of hulls each within the other.

The structure should reflect the priorities of the hierarchy; one hull should be within another and the size of the hulls should decrease with tier number. If a Suborder can exist independently of the Order, then the convex hulls would overlap two or more Orders. Such a scenario can be realized with simple Suborder definitions such as colour in the ASC and climate in ST. Colour, for example, can cross boundaries as it can be influenced by a variety of different mechanisms (Simonson and Boersma, 1972; White, 2013). A *Chromosol* is typically higher in clay than a *Podosol* (**Spodosol**-ST). The colour present in *Podosols* is typically from mineral accumulation down a profile, while the colour in the *Chromosol* is more likely from weathering of the in-situ minerals.

1.4. Inter hull analysis

TREX is a database which places taxa from two separate systems into the same property space. Convex hulls can be used to determine the similarity of groups and tiers between taxonomies. The rationale for such an analysis is simplicity and visuality. A distance matrix of all taxa is immense. Simplification by using only the Orders (12 in ST and 14 in ASC) creates 168 potential comparisons. The overall distance can be interrogated for pertinent properties and depths, but determining interactions of groups of Suborders between Orders and Suborders becomes time-consuming and complex.

Convex hulls provide a more simple comparison. Overlap area can be determined numerically, and large overlaps can be visually assessed. Conspicuous overlaps can be attributed to the proportion of hull 1 or the proportion of hull 2, and hull plots are useful even if no overlap exists. The proximity of hulls can sometimes indicate if a group has any similarities with another, and such commonalities can be picked up visually. The method is, however, constrained by dimensionality, such visualizations being difficult in anything greater than three dimensions, which would capture less of the total variability in the PC space.

1.5. Aims

The aims of this paper are to:

- Calculate empirical cumulative distribution functions of the nearest neighbour distances for Order, Suborder, and Great Groups in the ST and the ASC;
- Plot convex hulls for Orders, Suborders and Great Groups for taxa in ST and ASC;
- Estimate overlapping areas of convex hulls of Orders between ST and ASC, and determine any inferences from this.

2. Materials and methods

2.1. Soil profile description (SPD) centroids

The centroids in TREX were calculated using TERN data, which was