Geoderma 393 (2021) 115012

Contents lists available at ScienceDirect

Geoderma

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

A modelling framework for pedogenon mapping

Mercedes Román Dobarco [a,*], Alex McBratney [a], Budiman Minasny [a], Brendan Malone [b]

[a] Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Eveleigh, NSW 2015, Australia
[b] CSIRO Agriculture and Food, Black Mountain, ACT, Australia

ARTICLE INFO

Handling Editor: Kristin Piikki

Keywords:
Soil forming factors
Soil classification
Digital soil mapping
Genon
Unsupervised classification

ABSTRACT

Soil entities are generally defined based on soil properties, using morphological, genetic, or utilitarian criteria. Alternatively, soil entities could be characterized by groupings of homogeneous soil-forming factors under the assumption that the dominant soil-forming processes occurring over a time period within each group are similar, and therefore develop unique soil entities with similar soil properties. We define the pedogenon as a conceptual soil taxon defined from a set of quantitative state variables that represent the soil-forming factors for a given reference time. The objective of this study was to develop a methodology for mapping pedogenon classes at the time of the European settlement in New South Wales (Australia). This period was chosen as reference because from 1788 onwards the intensification of land use has accelerated the rate of change of soil properties. We implemented a two-step modelling approach with a set of environmental covariates representing the soil-forming factors, including the estimated natural vegetation at 1750. The *k*-means algorithm was applied to generate pedogenon classes suitable for local management. Then, hierarchical clustering was applied to identify the organization of pedogenons into families or "branches" of higher level taxa. We tested the ability of the pedogenon classes for explaining the variance of stable soil properties (particle size fractions) in the subsoil (30–60 cm depth) with redundancy analysis (RDA). The results indicated that between 800 and 1000 pedogenon classes provide the desired level of detail for both local and regional management across New South Wales. The influence of the pre-1750 vegetation types (e.g. *Acacia* open woodlands and shrublands, *Callitris* forests and woodlands) was apparent in the distribution of some pedogenon branches. Pedogenon classes differed in their characteristics (median area ≈750 km^2), but overall showed meaningful spatial patterns at local scale and formed regional assemblages. The RDA models indicated that pedogenon classes explained about 30% of the variance of silt and clay content. This flexible modelling framework allows the creation of pedogenon maps over large areas at high resolution (90 m) and is applicable at different scales. Potential applications of pedogenon maps include the quantitative assessment of soil change and designing soil monitoring surveys.

1. Introduction

The soils we observe today are the result of natural soil-forming processes that vary widely over the development time of a soil, and the direct and indirect effects of human activities on pedogenetic processes (Richter, 2007; Richter and Yaalon, 2012). The complexity of soils and their diverse response to human forcings require the identification of soil-class specific references for monitoring and assessing changes in soil multifunctionality due to management practices (Bunemann et al., 2018; McBratney et al., 2019). Soil condition and capability are components of the broader concept of soil security (McBratney et al., 2014), which was conceived as a holistic approach for evaluating soil functions and their connections with societal needs. Soil capacity refers to soil attributes that evolve at a slower rate (e.g., soil texture) and that are not readily changed by human activities whereas soil condition refers to more dynamic soil properties that are modified at a faster rate by soil management (e.g., particulate organic matter, pH). Soil capacity and condition determine the soil's capability, i.e. the potential functionality of the soil. While these concepts are relatively new and continue to evolve, they can be implemented in a spatially explicit framework for quantifying soil change. For this purpose, it is necessary to identify and map soil classes that result from natural multimillennial pedogenesis and historic anthropedogenesis (Richter, 2007) at a relevant taxonomic level for local and regional management (McBratney et al., 2019). Most soil classification systems group soils into classes characterized by similar soil properties following morphological criteria or diagnostic

* Corresponding author.
E-mail address: mercedes.romandobarco@sydney.edu.au (M. Román Dobarco).

https://doi.org/10.1016/j.geoderma.2021.115012
Received 8 September 2020; Received in revised form 9 February 2021; Accepted 10 February 2021
Available online 12 March 2021

features that indicate similar long-term pedogenetic processes (Krasilnikov et al., 2009). An alternative way of delimiting unique soil entities relies on identifying groups of homogeneous soil-forming factors, in accordance with earlier factor-based approaches (Dokuchaev, 1883; Jenny, 1941). These classes represent a soil system at quasi steady-state for a set of soil-forming factors at a given time and may evolve into different soil classes depending on the type and intensity of land use and management.

The concepts of soil genoform and phenoform proposed by Rossiter and Bouma (2018) can be applied for investigating soil change in the context of conventional soil classifications. Soil genoforms are "soil classes as identified by the soil classification system used as basis for detailed soil mapping in a given area" whereas soil phenoforms are defined as "persistent, non-cyclical variants of a soil genoform with sufficient physical or chemical differences to substantially affect soil functions"(Rossiter and Bouma, 2018). This framework requires that the soil classification system and the soil survey consider human-induced changes on soil polypedons when delineating soil mapping units (Smeck and Balduff, 2002). Otherwise, soils that share similar historic pedogenesis may be assigned to different classes if soil attributes used as diagnostic criteria have been severely modified by human activities. For example, accelerated erosion rates in croplands causing the loss of upper horizons and progressive mixing with subsoil horizons with tillage operations may lead to the separation of heavily eroded pedons and adjacent uneroded forest pedons into different soil series or soil orders according to Soil Taxonomy (Soil Survey Staff, 2010; Smeck and Balduff, 2002). In parts of the world where soil surveys are not available at the level of detail required for local or sub-regional soil management, digital soil mapping (DSM) approaches can be an alternative for identifying and mapping genoforms (Huang et al., 2018), but a different term may be necessary for designating soil classes with common historic pedogenesis. Boulaine (1969) indicated that for relatively large areas where the soil-forming processes are not markedly heterogeneous and act over the same time period on similar parent material, there would exist units of soil evolution or genons.

Boulaine (1969) defined a genon as "a soil volume comprising all the pedons that have the same structure, the same characteristics and result from the same pedogenesis". The genons were conceived as soil mapping units that vary in their composition (e.g., simple genons consist of pedons belonging to a single taxon, complex genons are associations of different taxons) and their spatial arrangement in the landscape (Campbell and Edmonds, 1984), and are not associated to any particular classification system. The genon follows the concept of polypedon defined by Johnson (1963): "a soil individual (polypedon) is also a real soil body; it is a parcel of contiguous pedons all of which have characteristics lying within the defined limits of a single soil series". Hence, the polypedon is a taxonomically homogeneous unit that can constitute a relatively pure map unit with a single dominant polypedon (consociation) or composite map units comprising several dominant polypedons (association, complex) (Zinck, 2016). On the other hand, Fridland (1972) suggested that the degree of homogeneity of the genon in terms of taxonomy remains somewhat obscure, since the classificational unit that defines its boundaries is not too specific. Other soil mapping units that are related to the concepts of polypedon and genon are the pedotop (Haase, 1968) and the soil combinations (Fridland, 1972). The latter are constituted by the spatial arrangement of elementary soil areals (i.e., soils belonging to a single taxa of the lowest rank) due to pedogenetic processes. The pedotop is a cartographic unit with homogeneous pedological attributes that results from uniform combinations of soil-forming factors, although it can present transition areas into neighbouring units and within unit variation (Campbell and Edmonds, 1984).

DSM approaches apply mostly supervised classification models (e.g., logistic regression, machine learning algorithms) for predicting the spatial distribution of soil properties or classes (Heung et al., 2016). Huang et al. (2018) developed a DSM approach for mapping soil classes prior to the European settlement in New South Wales (Australia), designated as genosoils and their derived phenosoils (soil classes derived from contemporary land use) with supervised classification. This approach requires high certainty when estimating the existing soil classes at the moment chosen as benchmark and detailed local knowledge of the soils present in the study area, making it challenging to implement at large extents. Unsupervised classification has been applied to remotely sensed spectral data and topographic variables for identifying patterns in soil-landscape relationships as a preliminary step for soil classification (Saunders and Boettinger, 2006), and has been proposed as a rapid, first order soil mapping method for extensive regions lacking soil data (Regmi and Rasmussen, 2018). Several studies have applied unsupervised classification using climatic, topographic and stable soil attributes as state variables for digital terron mapping at regional and national scale (Carré and McBratney, 2005; Malone et al., 2014; Roell et al., 2020).

The conceptual "changing model of soil" by Cline (1961) later revised by Richter and Yaalon (2012) sees soils as natural-human bodies, according to which soils and human societies and culture interact and co-evolve. The Aboriginal Australians settled in the continent over 50,000 years ago, being amongst the cultures with longest continuous presence in a geographic region (Tobler et al., 2017; Bird et al., 2018). Thus, historic anthropedogenesis comprises the land management practices carried by the Aboriginal societies over millennia. Controlled fire was a very extended land-management practice that shaped the structure and composition of the native vegetation in many Australian landscapes. Post-fire regeneration would create grasslands where grazing animals would feed while adjacent open forests were preserved as wildlife shelter, working as a shepherding method due to the lack of big predators (Gammage, 2011). Deep knowledge of the landscape was also used for agriculture (Pascoe, 2014). These dynamic land management practices maintained the provision of ecosystem services and influenced soil pedogenesis through the interactions of landscape, vegetation dynamics and fire regime. The intensification of land management after the European settlement in Australia during the second half of the 19th century has modified soil properties at an accelerated rate (e.g., loss of soil organic carbon stocks, erosion, soil sealing in urbanized areas). Hence, creating a map of hypothetical soil classes prior to the European settlement can be suitable for assessing soil change and guiding sustainable soil management (Huang et al., 2018) by identifying areas less affected by recent anthropogenic pressures that can be used as baseline within each soil class.

The objective of this study was to create soil classes characterized by homogeneous soil-forming factors with the assumption that the dominant soil-forming processes over a temporal period within each class are similar, and therefore comprise soils with similar soil properties. We define the pedogenon as a conceptual soil taxon created from a regionalised set of quantitative state variables representing the soil forming factors for a given reference time. This concept is closely related to soil genoform (Droogers and Bouma, 1997; Rossiter and Bouma, 2018) and genon (Boulaine, 1969). This study presents the methodology for mapping pedogenon classes, discusses their spatial characteristics and tests their ability to explain the variation of stable soil properties with New South Wales as case study. The proposed DSM framework was designed to have the following advantages: 1) it is easy to implement over large extents (regional, state and continental) at high resolution, 2) it defines soil map units that can be applied for local and regional soil management, and 3) the output maps can be compared with soil profile data and existing soil-landscape maps for evaluating the correspondence with pedogenetic processes (Regmi and Rasmussen, 2018).

2. Methods

2.1. Digital soil mapping framework

The digital soil mapping framework stems from the conceptual

model of Jenny (1941), in which soil properties or classes result from the interaction of the classical soil-forming factors climate (cl), organisms (o), relief (r), parent material (p), and time (t). Groups of homogeneous quantitative state variables representing the soil-forming factors at the time selected as benchmark are identified with unsupervised classification. Environmental covariates were used as proxies of soil-forming factors that have remained relatively constant (relief, parent material) or that we assume representative of the conditions at the time of the European settlement (climate, estimated native vegetation). In this study, a pedogenon is defined as follows:

$$pedogenon = f(cl_t, o_t, r_t, p_t) where\ \ t = reference\ \ time \quad (1)$$

In the context of pedogenetic time, during which the characteristics of the soil-forming factors and processes vary widely (Richter and Yaalon, 2012), this represents only a brief and static moment. Hence, the modelling framework can be generalized as:

$$pedogenon = f(s, cl, o, r, p, t) where\ \ t = period\ \ from\ \ the\ \ origin\ \ of\ \ soil\ \ formation\ \ up\ \ to\ \ the\ \ reference\ \ time \quad (2)$$

where the environmental covariates may represent the conditions of the soil forming-factors across the time of soil formation (e.g., paleoclimate data, past vegetation), and soil (s) or soil attributes that inform of pedogenetic pathways can be also included.

2.2. Study area

The study area is the state of New South Wales (NSW), situated on the eastern side of the Australian continent with an area of 801,137 km^2. NSW has high diversity of environmental and soil conditions. The climate ranges from hot arid in the western areas, warm temperate in the north, temperate in the south, and sub-alpine in the southeaster highlands (Gray et al., 2016). Mean annual temperature ranges between 4 and 21 °C and mean annual rainfall varies from less than 200 mm in the northwest of NSW to more than 1500 mm along the north-east coast (Hobley et al., 2015). The diversity of landscapes and ecosystems is reflected in the 17 bioregions present in NSW, that vary from sandy deserts, riverine plains, wooded grasslands, lush rainforests to rugged mountains (NSW National Parks and Wildlife Service, 2003). A mountain range, the Great Dividing Range, runs in direction north–south about 100–300 km inland, separating the eastern seaboard from the western interior (Hobley et al., 2015).

The surface geology of NSW is characterized by Paleozoic and Mesozoic siliceous, intermediate igneous and sedimentary rocks in eastern regions, and mainly Tertiary alluvial sands, silts, and clays in the western plains (Gray et al., 2016). The soils in NSW include 12 of the 14 soil orders of the Australian Soil Classification (Isbell et al., 1997), of which Vertosols occupy the greatest area, followed by Calcarosols, Chromosols, Kurosols, and Kandosols (Pino et al., 2019). The majority of the region is dedicated to grazing and cropping, with other land uses being nature conservation and forestry (Hobley et al., 2015).

2.3. Environmental covariates

The output of the models (Section 2.4) was highly sensitive to covariate selection. Hence, in preliminary analyses we trialled several combinations of covariates (results not shown) and decided to keep the covariates presented in Table 1 as proxies of soil-forming factors. The raster covariates were reprojected to WGS84 (EPSG:4326) projection and resampled to 3-arc second grid cell resolution when necessary. Continuous variables were resampled with the bilinear interpolation method and categorical variables with the nearest neighbour method.

2.3.1. Climate

Soil water and temperature are key drivers of pedogenesis. Soil temperature controls the rate of chemical reactions and biological processes (e.g., release of soil minerals into the soil solution, soil organic matter decomposition) and this influences the rates of soil formation (Buchan, 2011). Precipitation enables the transport of materials down the soil profile and soil moisture influences on-site transformations of minerals and organic matter (Kleber et al., 2015). We selected nine bioclimatic indices primarily developed for biodiversity and ecological models (Williams et al., 2012) that characterized the energy, temperature and precipitation gradients (Table 1). Gridded historical climate data are available for precipitation (1900–2020) and temperature (1910–2020) (Jones et al., 2009), but are scarce and with limited spatial coverage prior to 1900 (Ashcroft et al., 2014). Hence, although the bioclimatic indices are based on ANUCLIM 6.1 (Xu and Hutchinson, 2011) 30-year average climate surfaces (1975–2005), we follow the assumption that the general spatial patterns are representative of relatively stable climate conditions since the last ice age (Malone and Searle, 2020).

2.3.2. Parent material

Four variables based on airborne gamma-ray spectrometry imagery (Minty et al., 2009) were included as proxies of the parent material. The concentration of radioelements is related to the geochemistry and mineralogy of the bedrock and weathered materials (Wilford, 2012). We included the concentration of potassium (K), thorium (Th), and their ratio (Th/K). A weathering intensity index (WII) developed by Wilford (2012) estimates the degree of weathering of primary minerals into secondary minerals and oxides, and correlates well with regolith properties. The WII is complexly related to the soil-forming factor *t*, time or age, as it reflects the history of landscape processes, weathering and erosion.

2.3.3. Relief

Five covariates derived from the 3-second digital elevation model (DEM) produced from the Shuttle Radar Topographic Mission (SRTM) (Farr and Kobrick, 2000) data were included for describing relief (Table 1). Besides elevation and slope, two variables described geomorphic features (ridge top, valley bottom). The topographic wetness index informs on the likelihood of soil water saturation and the accumulation of transported materials.

2.3.4. Organisms

The historic natural vegetation was represented with a map of the estimated major vegetation groups (MVG) at the time of the European settlement (hereafter referred to as pre-1750 vegetation) developed by the National Vegetation Information System (NVIS Version 5.1, 2018) for the extent of Australia at 100 m resolution. MVG are defined based on vegetation structure, floristic composition and association information. Map units represent the dominant vegetation group. The scale of the source data is 1:1,000,000 or coarser for most of NSW. However, this map is the latest and spatially exhaustive summary of Australia's pre-clearing native vegetation. Several MVG classes were aggregated based on biotope descriptions and the comparison of the soil C:N:P stoichiometry under natural vegetation (Bui and Henderson 2013) leading to 15 classes of natural vegetation (Supplementary Material S1).

Table 1 (*continued*)

Covariate	Description	Clorpt factor	Variable type	Original raster resolution (m)	Reference
	(annual mean) (°C)				
TXX	Maximum temperature (monthly maximum) (°C)	cl	C	270	Williams et al. (2012)
TNI	Minimum temperature (monthly minimum) (°C)	cl	C	270	Williams et al. (2012)
TRX	Maximum monthly mean diurnal temperature range (°C). high variation in temperature conditions (inland or continental locations).	cl	C	270	Williams et al. (2012)
TRI	Minimum monthly mean diurnal temperature range (°C). Consistent temperature conditions (coastal locations).	cl	C	270	Williams et al. (2012)
RSM	Short-wave solar radiation - annual mean (MJ/m2/day)	cl	C	90	Wilson and Gallant (2000)
K	Radiometrics: filtered K element concentrations (%)	s, p	C	100	Minty (2019a); Geoscience Australia (2019)
Th	Radiometrics: filtered Th element concentrations (ppm)	s, p	C	100	Minty (2019b); Geoscience Australia (2019)
K/Th	Radiometrics: Ratio Th/K derived from the filtered Th and K grids	s, p	C	100	Minty (2019c); Geoscience Australia (2019)
WII	Weathering intensity index	p, t	C	100	Wilford (2012)
Elevation	SRTM-derived 3 Second Smoothed Digital Elevation Model	r	C	90	Gallant et al. (2009)
Slope	Slope (%)	r	C	90	Gallant et al. (2009)
TWI	Topographic wetness index	r	C	90	Quinn et al., (1991)
MRVBF	Multi-resolution valley bottom flatness index	r	C	90	Gallant and Dowling (2003)
MRRTF	Multi-resolution ridge top flatness index	r	C	90	Gallant and Dowling (2003)
MVG	Estimated pre-1750 major vegetation groups.	o	N	100	National Vegetation Information System V5.1 (2018)

Table 1
Covariates used to describe *clorpt* (Jenny, 1941) or *scorpan* (McBratney et al., 2003) factors and generate pedogenon classes. p: parent material; s: soil; t: time; r: relief; cl: climate; o: organisms; C: continuous variable; N: nominal or categorical variable.

Covariate	Description	Clorpt factor	Variable type	Original raster resolution (m)	Reference
PTA	Annual precipitation (mm)	cl	C	270	Williams et al. (2012)
PTS1MP	Precipitation: ratio of annual contrast in regional rainfall conditions between summer and winter solstice conditions.	cl	C	270	Williams et al. (2012)
PTS2MP	Precipitation: ratio of annual contrast in regional rainfall conditions between spring and autumn equinox conditions.	cl	C	270	Williams et al. (2012)
TNM	Minimum temperature	cl	C	270	Williams et al. (2012)

(*continued on next page*)

2.4. Clustering mixed data

Clustering was used to partition the dataset of environmental covariates into groups where the soil-forming factors within a group are homogeneous, and distinct from the elements of different groups. Most clustering algorithms are either specific for numerical or categorical data, and few can deal efficiently with large datasets of mixed data (Huang, 1998; Budiaji and Leisch, 2019). A common practice when dealing with mixed data is to calculate the Gower similarity coefficient (Gower, 1971) and apply a hierarchical or partitioning clustering method (e.g., partitioning around medoids). This approach is not feasible for very large datasets (e.g., $>10^5$–10^6 observations) due to the high computational cost and software memory limitations. Huang (1998) developed the *k*-prototypes algorithm, which combines numerical and categorical distances for finding the cluster centroids. *K*-prototypes requires careful weighing of both distance components for avoiding favouring either categorical or numerical variables. Recent partitioning algorithms increase the flexibility for calculating distance metrics for mixed datasets (Budiaji and Leisch, 2019). Alternatively, categorical variables can be transformed into numerical in a pre-processing step for applying numerical clustering algorithms.

The *k*-means algorithm is a popular non-hierarchical clustering method for numeric data (Hartigan and Wong, 1979) and efficient for very large datasets. The algorithm searches a partition of a numeric dataset X into k clusters that minimises the within-cluster sum of squared errors (WCSS), i.e., sum of squared distance to the cluster centroids. The centroids of the clusters are the means of the variables. The *k*-means algorithm operates with the following steps (Han et al., 2012):

- An initial k starting cluster centroids are selected (normally randomly).

- Each data point is assigned to the cluster to which is most similar based on the Euclidean distance between the point and the cluster mean. The cluster centroids are recalculated with the points assigned to the cluster.
- The data points are re-assigned to its nearest new cluster centroids.
- The iterations continue until the assignment is stable, that is, the clusters formed in the current round are the same as those formed in the previous round. In practice, the optimal solution is achieved when the squared distance of the centroids is smaller than a defined tolerance value (e.g., 0.0001).
- After a defined maximum number of iterations, if the assignment is not stable, the algorithm stops.

The 15 natural vegetation classes were transformed into binary variables (for each class, 1 = presence and 0 = absence) and a regular sample of 5,000,000 pixels subject to principal component analysis. The first 7 principal components explaining 50% of the variance were retained and the principal component scores predicted across NSW added to the dataset of 18 continuous variables. MVGs were well discriminated with 7 PCs: PC1, PC2 and PC7 characterized the *Eucalyptus* MVG, PC3 and PC4 described the MVG dominant in western NSW (e.g., *Acacia*, Mallee, *Chenopod* shrublands), and PC5 and PC6 discriminated diverse MVG that occupied smaller areas (e.g., rainforests and vine thickets, *Tussock* grasslands, *Casuarina*, *Callitris* forests and woodlands, heathlands).

The 25 continuous variables were centered and standardized. A regular sample of 259,000 observations were taken across NSW on a grid of 1.6 km × 1.6 km. The clustering dataset was rescaled by applying the inverse of the Cholesky transformation of the variance–covariance matrix:

$$\Sigma_X = LL^T$$

$$Y = XL^{-1}$$

where Σ_X is the variance–covariance matrix of the environmental covariates sample X, L is the Cholesky factor of Σ_X, a lower triangular matrix with positive diagonal values, and Y is the rescaled covariates dataset. The Euclidean distance calculated on the dataset Y is equivalent to the Mahalanobis distance calculated in X (Wicklin, 2012). Hence, the correlation among environmental covariates was accounted for in the clustering process.

The *k*-means algorithm is known for converging at local optima rather than at the global minimum. Hence, the initial assignment of cluster centroids is relevant for the outcome. The *k*-means++ is an initialization algorithm that chooses the first cluster centroid randomly and the remaining centroids are subsequently selected from the points with a probability proportional to the squared distance to its closest centroid (Arthur and Vassilvitskii, 2007). The clustering process was repeated 10 times with the *k*-means++ initialization and a maximum of 5000 iterations. The best initialization in terms of total WCSS was selected. The *k*-means algorithm was implemented with the *Kmeans_rccp* function of the ClusterR package (Mouselimis, 2020). The process was repeated with 18 continuous variables, excluding the pre-1750 estimated natural vegetation variables. This was done for the purpose of assessing the influence of the pre-1750 vegetation layer on the pedogenon maps. Pedogenon classes were mapped with 90 m grid cell resolution assigning each pixel to its closest cluster centroid after rescaling with the inverse Cholesky transformation.

2.5. Optimal number of soil classes and cluster evaluation

A requirement of distance-based clustering methods is that the number of clusters needs to be specified. We selected the number of pedogenon classes based on three criteria: 1) area-pedodiversity relationships (Guo et al., 2003), 2) the elbow method, and 3) visual assessment of the spatial patterns of different pedogenon maps.

Ibañez et al. (1998), McBratney et al. (2000) and Minasny et al. (2010) found a positive linear relationship between the area occupied in a studied region and pedodiversity (soil class richness and Shannon's entropy index) characterized with major soil groups (FAO major soil groups (FAO, 1993), World Reference Base (WRB) soil groups (IUSS, 2006)). We assumed that the pedodiversity-area relationship observed at continental and subregional scale (Guo et al., 2003; Minasny et al, 2010) would be inherent in the unsupervised classification of *pedogenon* classes. Currently, no soil classification system represents comprehensively the diversity of soils across the world (Ibañez et al., 1998) although some efforts have been made to unify various classification systems (Hughes et al., 2017). We chose Soil Taxonomy (Soil Survey Staff, 2010) as reference for selecting the number of classes because it was defined for a wide range of pedogenetic conditions and area-pedodiversity equations were available at various taxonomic levels (Guo et al., 2003), even though the purpose of Soil Taxonomy was not to characterize pedodiversity but rather detailed mapping. An objective of this study was to generate soil mapping units that can serve for local soil management. Hence, the desired final number of classes should be of a similar magnitude as for the taxonomic level of families from the Soil Taxonomy system (Soil Survey Staff, 2010). This level is one above the series but is considered a useful general group for management. We estimated the approximate number of pedogenon classes based on taxa richness-area equations published by Guo et al. (2003) for the USA based on Soil Taxonomy (Soil Survey Staff, 2010). The number of classes at an equivalent taxonomic level to soil family, for the area of mainland New South Wales (801,137 km^2) would be around 1040 classes.

The optimal number of classes was also evaluated in terms of total WCSS (i.e., sum of within-cluster squared errors per cluster) and the ratio between-CSS to total WCSS (i.e., (total SSE – total WCSS)/ total SSE) for the range between 100 and 1500 in intervals of 100 classes with the elbow method (Han et al., 2012). The optimal number of clusters is indicated by the turning point in the curve of total WCSS with respect to the number of clusters, after which the marginal effect on reducing the WCSS with increasing the number of classes is minor (Han et al., 2012). Finally, relative cluster evaluation was performed by inspecting visually the spatial patterns of pedogenon maps for different number of classes in smaller study areas where the soils have been well studied previously, which include Namoi-Edgeroi (Triantafilis and McBratney, 1993; Ward, 1999) and the Hunter Valley region (Malone et al., 2014). The evaluation did not compare pedogenon maps with soil maps using Soil Taxonomy, but relied in the resemblance between the spatial patterns of pedogenons and the variation previously observed in these soils, with maps of continuous soil properties and soil classes according to the Australian Soil Classification (Isbell and the National Committee on Soil and Terrain, 2016).

Summary statistics of the environmental variables sampled at the clustering dataset locations (259,000 cases) were calculated for each pedogenon class, as well as the Mahalanobis distance among all cluster centroids and to the closest centroid. The area occupied by each pedogenon class across NSW and the average distances between pixels of the same pedogenon class were calculated with the gridded predictions.

To illustrate the local and regional attributes of the pedogenon classes we examined their patterns across a smaller study area, Nowley farm. The E. J. Holtsbaum Agricultural Research Institute (Nowley farm) is a 2083 ha farm dedicated to dryland cropping and cattle located in the Liverpool plains area in north-western NSW. A detailed description of the topography, lithology and soils found at Nowley farm can be found in Stockmann et al. (2016). The entropy-based local indicator of spatial association (ELSA) (Naimi et al., 2019) was applied to explore the local patterns in clustering and their degree of association. ELSA is a local indicator of spatial association for categorical and continuous data that incorporates a measure of dissimilarity (attribute distance) between neighbouring sites and normalized entropy. The index varies between 0 and 1 from homogeneous to more heterogeneous areas (Naimi et al.,

2019). In addition, we calculated the degree of association between the pedogenon classes and the soil landscape units found at Nowley farm (Curlewis 1:100,000 Soil Landscape Mapsheet area (Banks, 1995)) with biased corrected Cramér's V (Cramér, 1946). Carmer's V ranges between 0 and 1, with a 1 indicting a perfect association between categorical variables.

2.6. Hierarchical clustering

An agglomerative hierarchical clustering was applied for assessing the similarities among clusters and their organization, treating the cluster centroids as individuals. The dissimilarity between cluster centroids was estimated with Ward's method, which minimizes the total within-cluster variance. Ward's method merges the pair of clusters with minimum between-cluster distance at each successive step. The resulting dendrogram was divided into branches and different colour ramp scales were assigned for enabling the interpretation of the spatial distribution of families or assemblages of similar *pedogenon* classes. The number of branches was selected after trialling several partition numbers and visualizing the dendrograms, so that the groups that we could identify were separated but the amount of isolated leaves or very small clusters was minimized.

2.7. Soil data

Legacy soil data was accessed with the Soil Data Federator (http://esoil.io/TERNLandscapes/SoilDataFederatoR/R/help/index.html), a web API that gathers soil data from different sources and is managed by the Terrestrial Ecosystem Research Network (TERN). All available data for clay (%), silt (%) and sand (%) measured with different particle size analysis methods (i.e., Coventry and Fett pipette (Coventry and Fett, 1979), hydrometer (Day, 1953) and plummet balance (Marshall, 1956)) were extracted. The rate of change of soil texture would be very slow, so it is considered a relatively stable soil property. The data quality was checked, followed by modifying or eliminating incomplete, incorrect, and duplicated records. Since much of the legacy data comes from agricultural fields only, data between 30 and 60 cm depth were selected for statistical analyses, since past tillage operations may have homogenised texture in upper horizon layers. After these processes, there were 1102 observations of particle size fractions at 836 locations for depths ranging between 30 and 60 cm.

2.8. Redundancy analysis

We applied an exploratory ordination analysis to visualize the variability of texture across and within classes and test the ability of pedogenon classes for explaining the variation of particle-size fractions. Redundancy discriminant analysis (RDA) is a linear canonical ordination method designed for identifying patterns of variation in a multivariate dataset that can be associated to potential explanatory variables (Legendre and Legendre, 2012; Borcard et al., 2018). RDA can be considered as a constrained version of principal component analysis, wherein the canonical axes built from linear combinations of response variables are also linear combinations of the explanatory variables. Here, particle-size fractions were the response variables, pedogenon class was the explanatory variable and horizon depth the covariable (Legendre and Legendre, 2012; Borcard et al., 2018). The canonical axes represent environmental gradients. The pedogenon classes can be interpreted as a nominal designation of environmental gradients defined a priori. Given the correlation among the three fractions, sand was removed from the set of dependent variables and only those classes with at least 10 silt and clay observations were used in the analysis. A partial RDA was performed to remove the effect of mean horizon depth before evaluating the effect of pedogenon class on clay and silt (Zuur et al., 2007). The RDA analyses were performed with the vegan package in R (Oksanen et al., 2019).

3. Results

3.1. Optimal number of classes

The scree plots of the total WCSS and the ratio between-CSS to total WCSS did not show any abrupt changes that indicated the minimum optimal number of clusters (Fig. 1). Smaller values of total WCSS indicated more compact clusters (feature space) as the number of clusters increases. The ratios between-CSS to total WCSS were above 0.80 from 500 clusters for both sets of variables, indicating a relatively good clustering pattern. The ratio between-CSS to total WCSS was slightly higher when pre-1750 vegetation was included as a covariate. Visual examination of the spatial patterns and number of classes present in smaller study areas with good available knowledge of soil properties suggested that the maps with 1000 pedogenon classes were most suitable for the objectives of this study (Supplementary Material S2). Hence, only the maps of 1000 classes were subjected to further analyses.

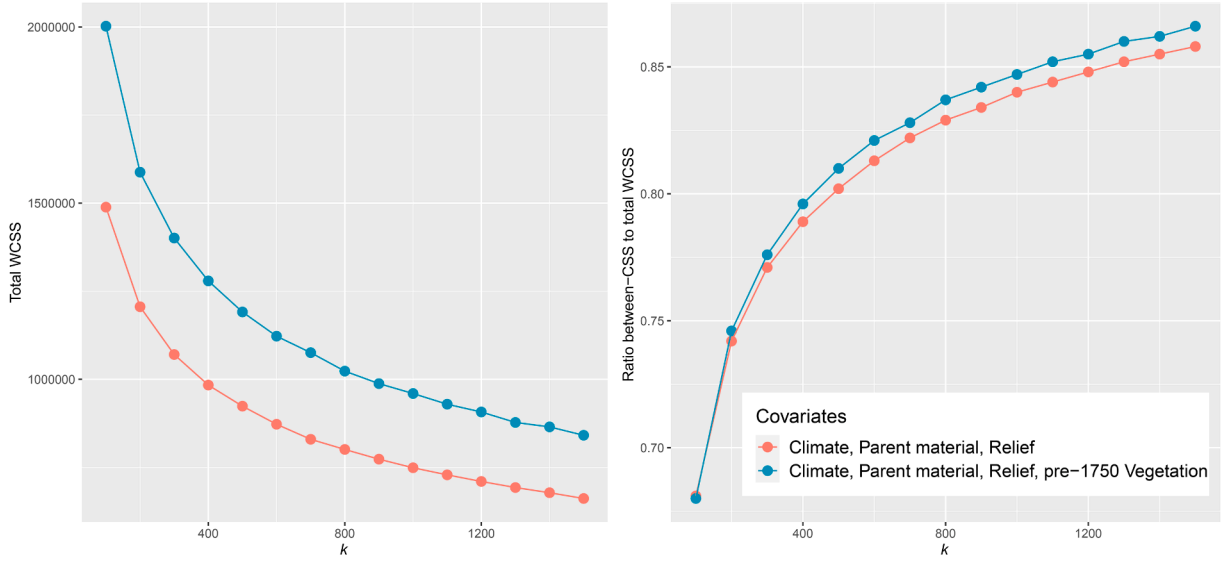

Fig. 1. a) Number of clusters (*k*) vs. total within cluster sum of squares (total WCSS) and b) Number of clusters (*k*) vs. ratio between-cluster sum of squares (between-CSS) to total within cluster sum of squares.

3.2. Pedogenon maps for NSW

The dendrogram of 1000 pedogenon classes created with pre-1750 vegetation and 18 continuous covariates was divided into small branches or families with the exception of a larger branch distributed along central NSW (purple-violet) that represented 26% of the total area (Fig. 2.a and Table 2 of Supplementary Material S1). The influence of the pre-1750 estimated vegetation was very apparent in the spatial patterns of some pedogenon classes, despite including only 7 principal components to keep a good balance between soil-forming factors. The distribution of some pedogenon classes reproduced the boundaries between MVG, but this also allowed to group together soils that likely present similar pedogenesis (e.g., Vertosols along the Darling Riverine Plains) (Fig. 3). The predominant major vegetation groups across all branches were *Eucalyptus* woodlands or open forests, although six pedogenon branches had other distinct dominant vegetation (e.g., *Acacia* woodlands and shrublands in the north-western corner, *Callitris* forests and woodlands, or Mallee woodlands and shrublands) (Table 2, Supplementary Material S1).

Towards the west of NSW, the pedogenon classes located in areas characterized by low precipitation, higher minimum temperatures and maximum diurnal temperature range, depositional areas (higher TWI), flat landscapes and relatively high weathering intensity index were organized into five branches (Table 2 and Fig. 1 of Supplementary Material S1). Towards the Great Dividing Range and the coast, the pedogenon branches were more scattered along the north–south and reflected differences in relief conditions and microclimate. The alpine, high elevation region in southeast NSW was grouped into a distinct branch (Table 2 and Fig. 1 of Supplementary Material S1).

The dendrogram of 1000 pedogenon classes generated with 18 covariates (climate, parent material, relief) was divided into 18 branches (Fig. 2.b). The broad spatial patterns agreed to some extent with bioregions of NSW (bioregions are areas characterized by broad landscape-scale geophysical and ecosystem features, designed for biodiversity conservation planning). There were clear divisions of NSW into areas with dominant branches, the largest one occupying 35% of the total area (Fig. 2 of Supplementary Material S1). For example, the north-western area (north-west of the Darling river) corresponds mostly to a branch of the dendrogram (blue-green) that also has some classes present in the centre of NSW (Cobar Peneplain bioregion). Similarly, the classes located in the Murray Darling Depression (southwestern NSW) belong mostly to the same branch (purple-pink). However, other bioregions like the Darling Riverine Plains, are divided into two branches. Towards the east, the families of pedogenon classes are distributed along a north-south gradient, as response to the patterns in relief and climate variables parallel to the coast (Table 3 and Fig. 3 of Supplementary Material S1). Similarly to the previous model, the alpine region was separated into a distinct branch. The areas attributed to a single branch are smaller than in central and western NSW.

The patterns of the pedogenon classes from both maps were similar in some regions, like in north-western and south NSW, despite following a different organization in their respective dendrograms. Perhaps the most apparent differences were in central NSW, where the pedogenon map with pre-1750 vegetation reflected the *Callitris* forests and woodlands and Mallee woodlands and shrublands. The degree of association between the categories of both maps measured with Cramér's V = 0.53 was strong considering the number of categories.

The summary statistics of the pedogenon classes' attributes were similar for both maps (Table 2). The median area occupied by a pedogenon class in NSW was 770 and 729 km^2, ranging between 4 and 2842 km^2. The summary statistics of the within-pedogenon geographic distances suggest that most of the pixels from the same pedogenon class are located between 30 km and 120 km apart. The mean maximum distance around 500 km indicated that some pixels are scattered far from the main area of occurrence of its pedogenon class. The difference between the maximum Mahalanobis distance to the closest pedogenon and the third quantile indicated that there were few pedogenon classes whose centroids are far from all other classes. The spatial patterns of pedogenon classes were highly variable. Some pedogenons were quite compact, whereas others were more disperse, occupying relatively small areas over an extensive geographical range. Overall, pedogenon classes were distributed over relatively extensive areas forming assemblages, but they can be considered mappable, relatively compact units.

3.3. Case study: Nowley farm

There were 33 and 31 pedogenon classes present at Nowley farm for the maps with and without pre-1750 vegetation respectively (Nowley farm is recognised as having a large pedodiversity). The first 7 pedogenon classes of the map created with pre-1750 vegetation represented 25%, 19%, 12%, 9%, 7%, 7% and 6% of the farm whereas secondary classes together occupied 15% of the area. Similarly, in the map with 18 continuous variables, the predominant pedogenons occupied 27%, 26%, 16%, 7% and 6% of the farm respectively, whereas the remaining 26 classes represented only 18% of the area (Fig. 4.b). The total extension of the 33 and 31 pedogenon classes across NSW was 25921 km^2 and 27902 km^2. Most of the pedogenons were located in proximity to the study area (approximately within a 150 km radius) forming compact units (Fig. 4.c) Few pedogenons spread towards the west of NSW with a more scattered pattern (Fig. 4.a).

The level of spatial association among the pedogenon classes at Nowley farm was calculated using the Mahalanobis distance among centroids as attribute distance. The values of ELSA for each cell were calculated within a local distance of 0.005 decimal degrees (approximately 475 m) indicated relatively high homogeneity across the study area, with a maximum value of 0.33. There was some variation in the degree of spatial association, with higher heterogeneity (higher ELSA values) in the eastern and western extremes of the farm (Fig. 5), which corresponded with different soil types in the farm (Stockmann et al., 2016).

Two soil landscape units were dominant across Nowley farm ('Noojee' and 'Trinkey Forest') and two soil landscape units had a smaller representation towards the eastern and western extremes of the farm ('Mount Milbulla' and 'Quirindi Creek') (Fig. 6.c) (Stockmann et al., 2016). The association between the 4 soil landscape units and the pedogenon classes was strong, with Cramér's V of 0.62 and 0.56 for the pedogenon maps with and without pre-1750 vegetation respectively. When we expanded the analysis area to the whole Curlewis Mapsheet (38 soil landscape classes, Fig. 6.a), the degree of association decreased, with Cramér's V equal to 0.33 and 0.28, respectively. Although these values still indicated a moderate association between the soil landscape units and the pedogenon maps. The association was higher with groups of main geomorphic processes (Fig. 6.b), with Cramér's V of 0.39 and 0.33 for the pedogenon maps with and without pre-1750 vegetation.

The soil landscape unit 'Mount Milbulla' overlapped with three pedogenon classes that were close to each other in the feature space (designated with the numeric codes 135, 139 and 222) (Fig. 3.b and 3.a dendrogram). These slopes have shallow and stony soils derived from weathered basalt and Vertosols in the benches (Stockmann et al., 2016). Similarly, the soil landscape unit 'Quirindi Creek' was mainly occupied by a distinct subgroup of pedogenon classes (code 272, 830, 86 and 318 in Fig. 4.b and dendrogram; also see Fig. 6.e for the map with pre-1750 vegetation). The soils in the floodplain of mixed origin vary, but towards the northwest of the farm are poorly drained grey-brown Vertosols (Stockmann et al., 2016). The soil landscape unit 'Trinkey Forest', located on undulating low hills with alluvial fan systems derived from sedimentary rocks, presents Sodosols in the southwest of the farm (Stockmann et al., 2016). This unit was not captured well by the pedogenon map without pre-1750 vegetation when we only consider the pedogenons present in the farm (Fig. 4.b), but the correspondence improved when all pedogenon classes (with pre-1750 vegetation) present in the mapsheet were represented (Fig. 6.e). The soil landscape unit

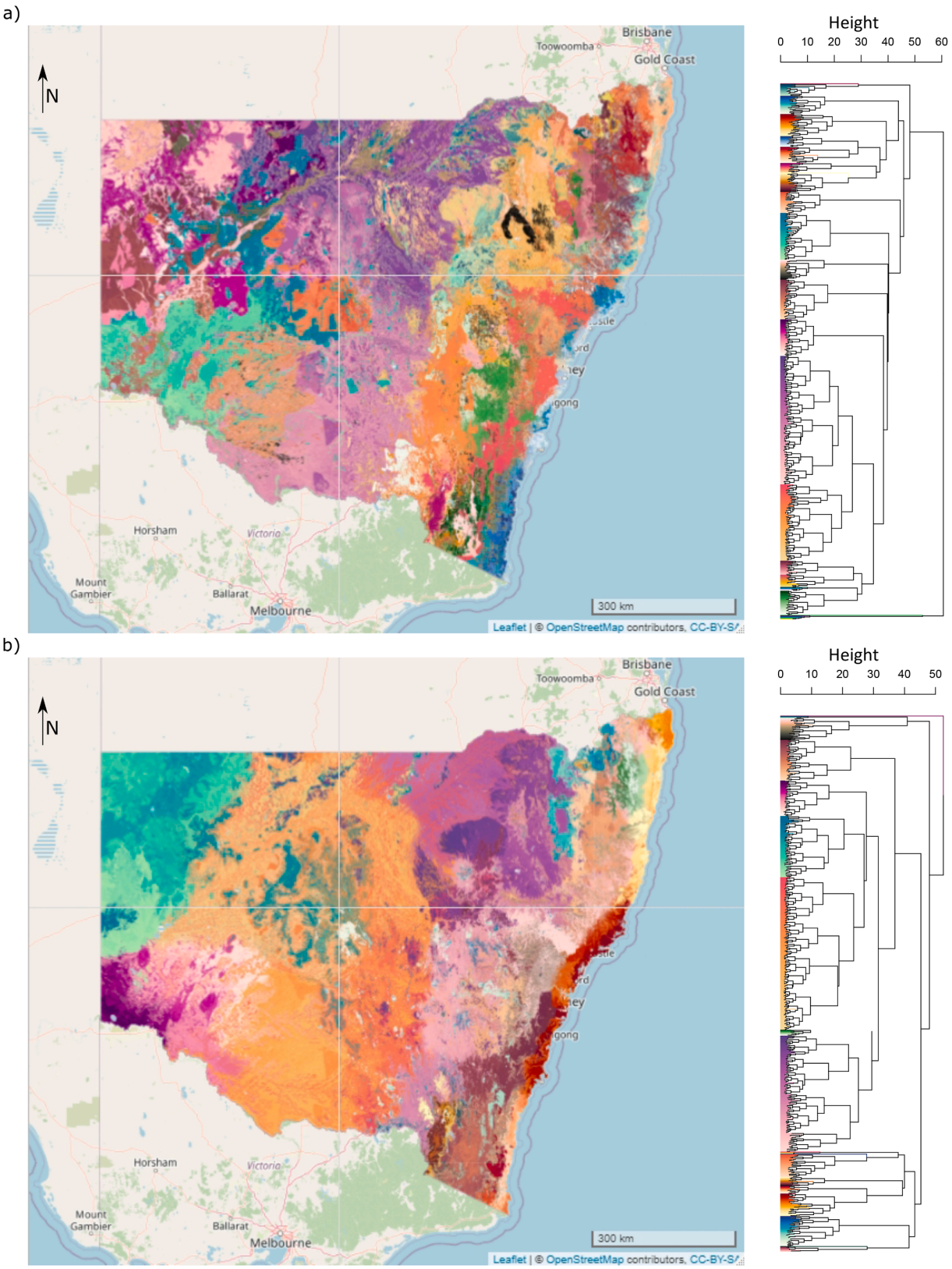

Fig. 2. Pedogenon classes for New South Wales (n = 1000) generated with a) 18 continuous environmental covariates and 7 principal components of the estimated pre-1750 vegetation, and b) 18 continuous variables. The dendrograms of pedogenon centroids were divided into 21 and 18 branches respectively, and a different colour branch assigned to each of them for identifying patterns in spatial distribution of more similar classes and showing a possible aggregation into higher level taxa. Background map from © OpenStreetMap contributors.

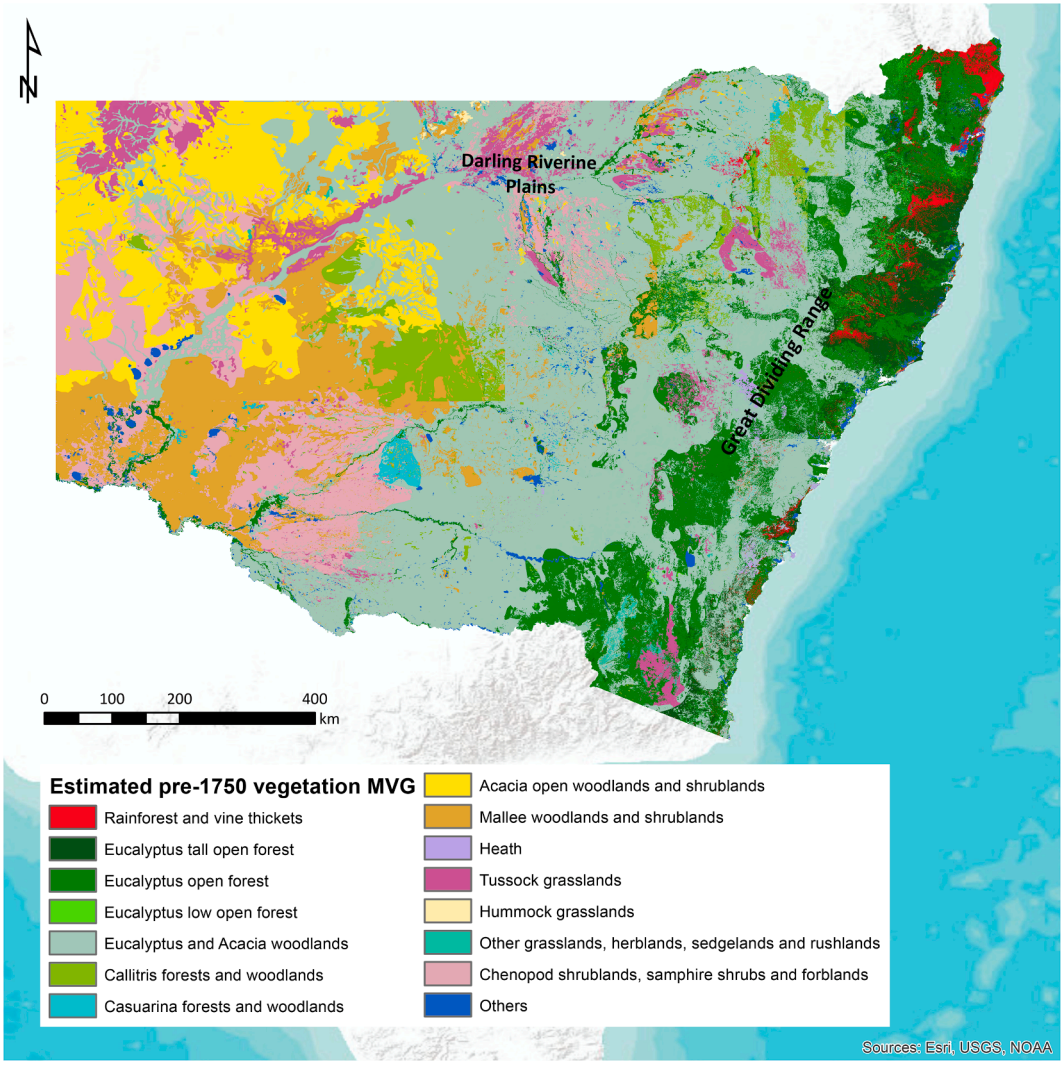

Fig. 3. Estimated pre-1750 vegetation (NVIS 5.1, 2018) reclassified into 15 major vegetation groups.

Table 2
Summary statistics of pedogenon classes for NSW. The distance* presents the mean ± standard deviation across all pedogenons of the summary statistics (1st quantile, median, 3rd quantile) of the distances between the pixels of the same pedogenon.

Covariates	Statistic	Pedogenon area (km^2)	Distance between pixels of the same pedogenon (km)*	Mahalanobis distance among all centroids	Mahalanobis distance to the closest centroid
Climate	Min	4	–	0.81	0.81
Relief	Q25	501	34 ± 19	4.4	1.37
Parent material	Median	770	65 ± 38	5.6	1.77
	Mean	798	–	6.17	2.12
	Q75	1062	107 ± 67	7.11	2.53
	Max	2415	473 ± 246	50.14	33.31
Climate	Min	4	–	0.92	0.92
Relief	Q25	513	37 ± 20	5.38	1.59
Parent material	Median	729	71 ± 41	6.56	2.04
Vegetation	Mean	799	–	7.02	2.38
(pre-1750)	Q75	1017	120 ± 74	7.96	2.81
	Max	2842	535 ± 253	49.86	34.77

'Noojee' is located in broad and long footslopes, with deep Vertosols and Chromosols developed over alluvium, basalts and dolerites (Stockmann et al., 2016). 'Noojee' had representation of several pedogenon classes that were not unique to the soil landscape unit (Fig. 6).

3.4. Redundancy analysis of stable soil properties and pedogenon classes

There were 19 pedogenon classes with at least 10 observations (23 ± 10 observations (mean ± standard deviation) for silt and clay in the pedogenon map created with pre-1750 vegetation for a total of 436 observations. The pedogenon map without pre-1750 vegetation had 20 pedogenon classes with at least 10 silt and clay observations (22 ± 11

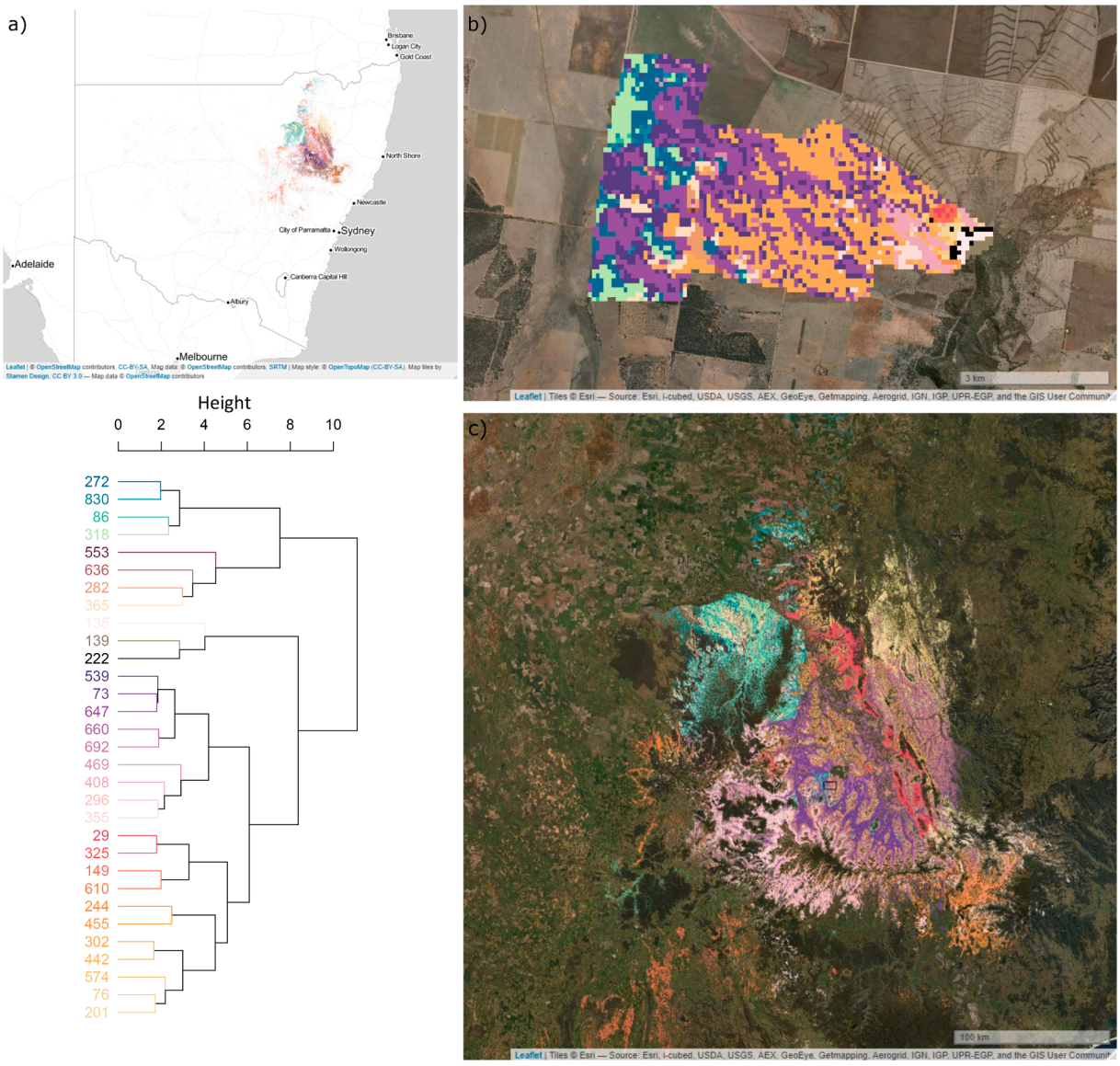

Fig. 4. Pedogenons present at Nowley Farm from the map created with 18 continuous variables (without pre-1750 vegetation). a) Distribution of the pedogenon classes across NSW and dendrogram of the pedogenons present at Nowley Farm, with a custom colour palette for this study area. The labels indicate the number designation of the pedogenon classes. b) Nowley Farm. c) Distribution of the pedogenon classes in the surroundings of the study area. A small black rectangle indicates the location of Nowley Farm.

observations) and a total of 431 observations. The results of the partial RDA indicated that 29% of the particle size fraction variance was explained by the pedogenon map with pre-1750 vegetation (constrained variance), 5% variance was accounted by the conditioning horizon depth, and the remaining 65% was unexplained variance. The first and second axes of the RDA explained respectively 54% and 46% of the constrained variance. A permutation test indicated that the global RDA model and both constrained axes were statistically significant ($p = 0.001$). The results of the partial RDA for the pedogenon map without pre-1750 vegetation indicated that the constrained variance was 35% and 61% of the variance remained unexplained. The effect of mean horizon depth accounted for 4% of the variance of silt and clay. The RDA axes explained 74% and 26% of the constrained variance. The RDA model and both canonical axes were statistically significant ($p = 0.001$). The pedogenon classes were not separated in the ordination plots for both RDA models but rather overlapped, with variability among observations of the same pedogenon class (Fig. 7).

4. Discussion

4.1. Optimal number of classes

We hypothesized that the diverse combinations of soil-forming factors identified by clustering would result in unique soil entities or classes. The number of combinations should have a positive relationship with the size of the study area as observed by Guo et al. (2003) for the conterminous USA. The results by Ibañez et al. (1998) and Minasny et al. (2010) suggested that pedodiversity (measured with Shannon's index) is smaller in Australia than in other continents at the level of major soil group. Hence, if differences in pedodiversity between continents were analogous at lower hierarchical level, the number of classes for NSW at an equivalent level of soil family should be smaller than that estimated from the equations by Guo et al. (2003) for the conterminous USA. However, when the mean taxonomic distance is used as pedodiversity index, this is not related to the area but rather to the level of detail of the soil surveys (Minasny et al., 2010). The pedodiversity-area relationship depends on the diversity index used.

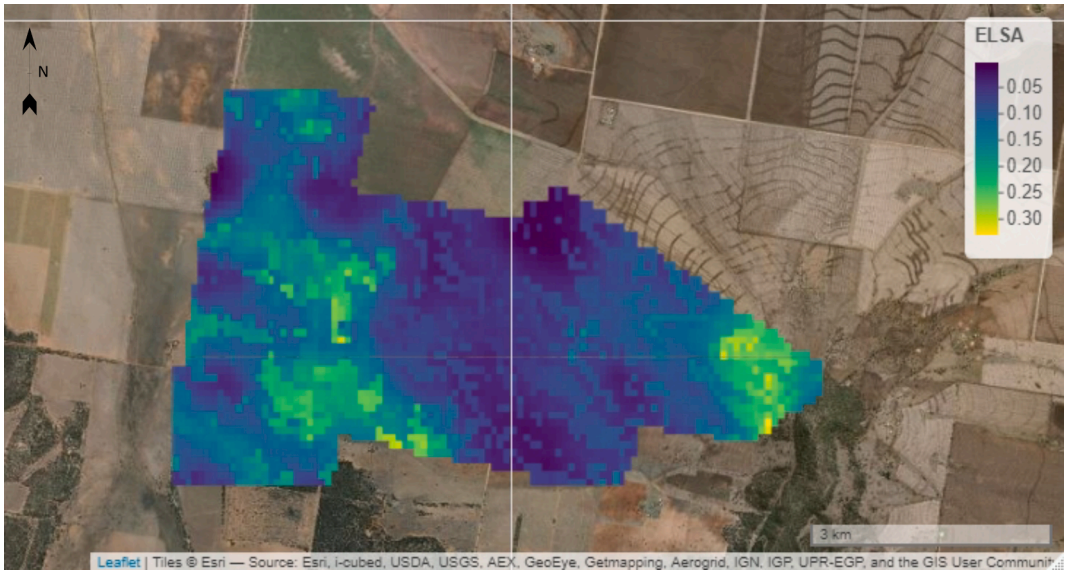

Fig. 5. Entropy-based local indicator of spatial association (ELSA) at Nowley farm.

The choice of the optimal number of classes is one of the critical points of the proposed methodology. Here, the elbow method did not provide a clear indication of the optimal number of classes, and we depended on rather subjective criteria. However, future studies should use intrinsic cluster evaluation metrics (e.g., silhouette index) (Han et al., 2012) for selecting the optimal number of classes. Alternatively, the comparison of pedogenon maps with traditional soil maps or survey data can also help for evaluating the number of classes (Supplementary Material S2). Overall, several pedogenon classes ranging between 800 and 1000 for New South Wales seem to provide the level of detail needed for both local and regional soil management and show moderate agreement with soil maps by the Australian Soil Classification (Isbell and the National Committee on Soil and Terrain, 2016) (Supplementary Material S2). The modelling framework is flexible and allows the end-users to aggregate pedogenons into higher level taxa with hierarchical clustering, treating the centroids as individual observations and optionally weighing the centroids by cluster size.

The maps of 1000 pedogenon classes had meaningful spatial patterns in smaller study areas (Fig. 6 and Supplementary Material S2), indicating that 1046 classes at the taxonomical level of families estimated with the equation by Guo et al. (2003) were a good first approximation for the optimal number of pedogenon classes. According to the taxa-area relationships for the Soil Taxonomy (Guo et al., 2003), for an area like New South Wales there should be around 300 pedogenon classes at the level of subgroups, 80 for great-groups, 26 for suborders and eight for soil orders. Considering the structure of the dendrograms (Fig. 2) and that in New South Wales there are 14 orders of the Australian Soil Classification (Isbell et al., 1997), the richness of pedogenon classes at higher levels may be higher than estimated based on analogies with the conterminous USA.

4.2. Representation of estimated vegetation at the reference state

The estimated pre-1750 vegetation was a central component of the modelling approach for mapping pedogenons. The classes created using pre-1750 vegetation presented spatial patterns that agreed with areas with distinct soil characteristics, e.g., Vertosols along the Darling Riverine Plains. However, since the dissimilarity between vegetation classes was assumed the same across all categories, this translated into higher heterogeneity in the hierarchical organization of the pedogenons (Fig. 2). In some cases, the imprint of the vegetation classes was very clear in the pedogenon map, like the regions of *Casuarina* forests and woodlands, *Callitris* forests and woodlands, or Hummock grasslands. Vegetation could be represented by the probability of occurrence of each MVG, which would describe the co-occurrence of different vegetation communities within an area instead of indicating the dominant vegetation class and would likely reduce the presence of crisp edges in the map.

Natural vegetation communities in Australia follow a climatic gradient while edaphic properties play a secondary role, although soil phosphorus and P:N are relevant for predicting the distribution of Eucalyptus communities (Bui and Henderson, 2013). Phosphorus cycling is controlled by biotic and geochemical processes. Thus, there are strong feedbacks between the spatial patterns of MVG and soil stoichiometry. Phosphorus is present in many rocks as apatite, but it is much less abundant in siliceous rocks (Binkley and Fisher, 2020). Instead of including the map of pre-1750 estimated vegetation, the potential distribution of natural vegetation communities can be indirectly included in the clustering dataset by including additional bioclimatic variables and information of soil parent material and mineralogy. In addition to gamma radiometrics that inform on surface geochemistry and mineralogy, the abundance of siliceous rocks may be estimated from remote sensed spectral band ratios (Cudahy et al., 2016). Vegetation communities adapted to singular edaphic conditions (e.g., saline soils, wetlands) are possibly captured by the combination of parent material and relief.

4.3. Identifying pedogenons in an ancient landscape

Climate has been considered the main driver of long-term pedogenesis, as it is reflected by some soil classification systems (Wilding, 1994; Bockheim et al., 2014). Soil-forming factors, particularly climate, are represented as relatively constant in many DSM models. The state variables used in this study identified the pedogenons with current climate and relatively recent vegetation, characterizing them somewhat as "zonal" soils (Marbut, 1935). However, most soils are formed by polygenesis and result from the combination of soil-forming processes that evolve over pedogenic time and the influence of palaeoclimates (Wilding, 1994; Richter and Yaalon, 2012). Soils in arid and semi-arid environments often present relict features developed by soil-forming processes occurring under past humid climates (Dergne, 1976). Former climatic conditions have a strong influence on the soil properties of highly weathered, ancient landscapes like those in western New South Wales. Thus, the application of a state factor model has its limitations for

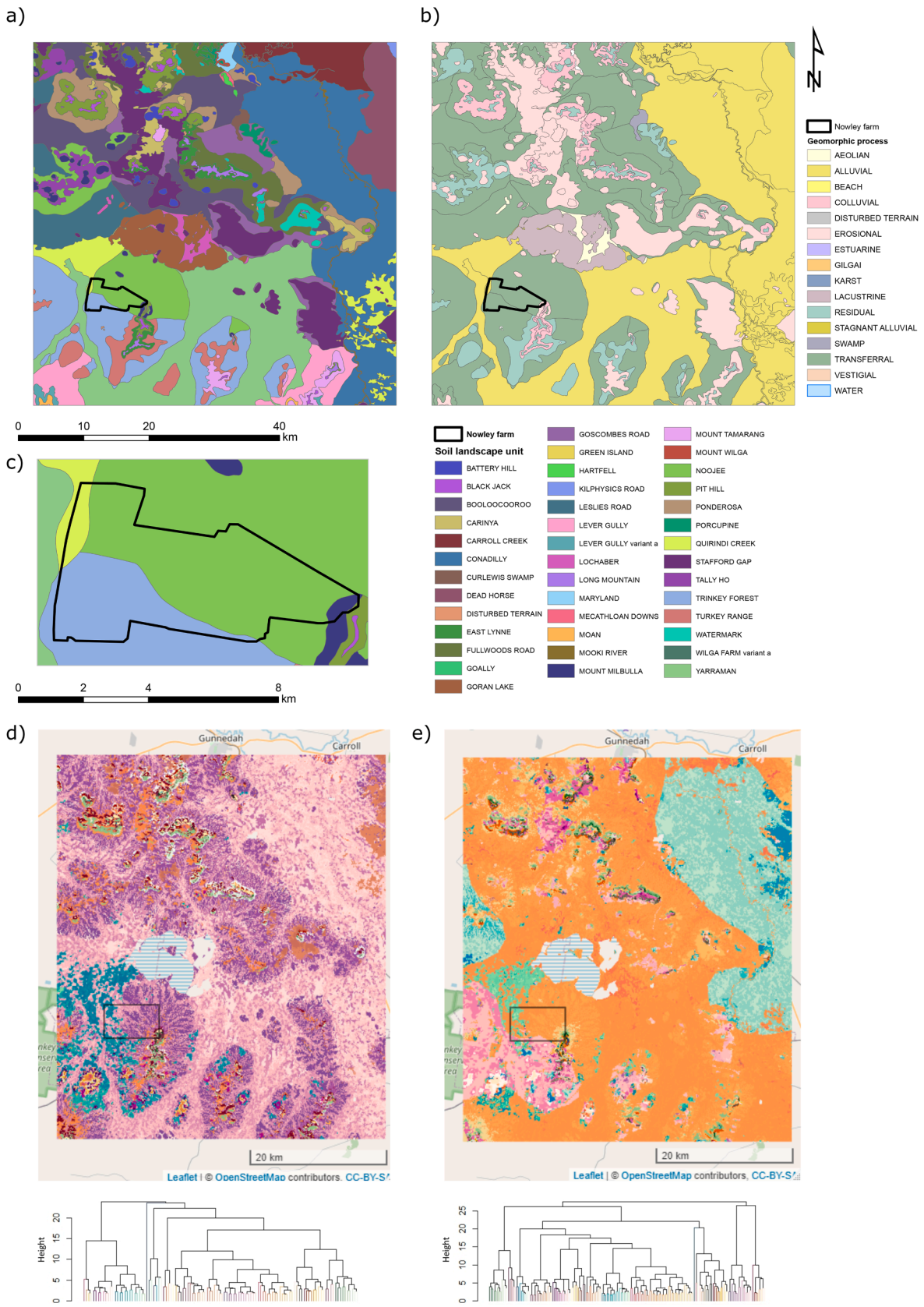

Fig. 6. a) Soil landscape units and), b) geomorphic process of the Curlewis 1: 100,000 sheet map (Banks, 1995), c) soil landscape units at Nowley farm, d) pedogenon classes present at the extent of the Curlewis 1: 100,000 sheet map (18 continuous environmental covariates) and e) pedogenon classes generated with 18 continuous environmental covariates and the estimated pre-1750 vegetation.

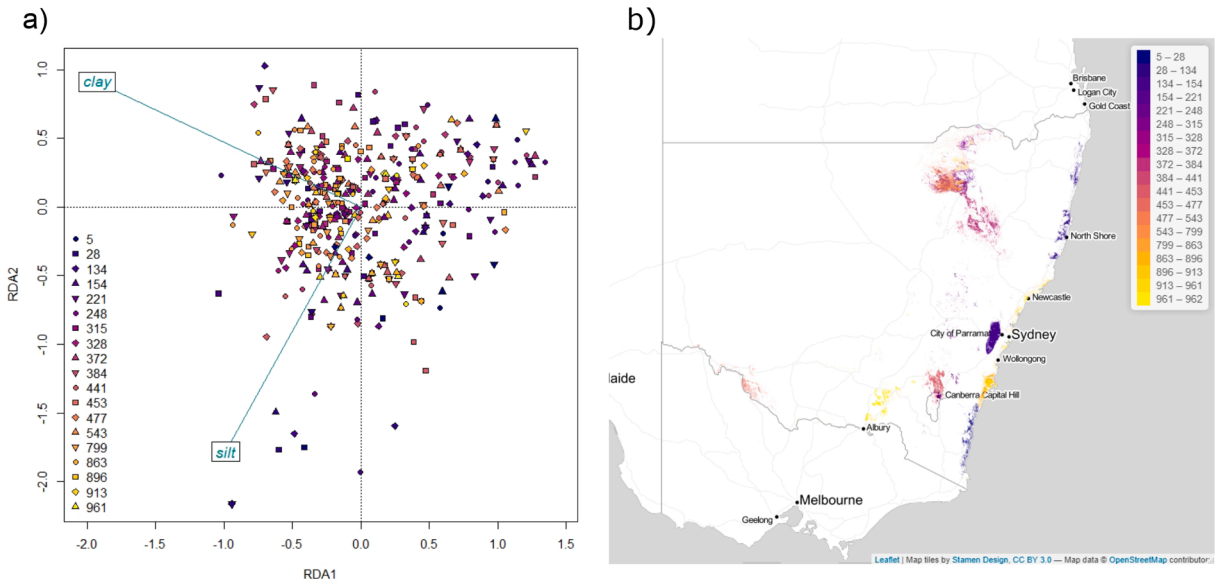

Fig. 7. Ordination plots (scaling = 1) of partial redundancy discriminant analysis (RDA) of silt and clay with pedogenon class as explanatory variable. a) RDA biplot for the map with pre-1750 vegetation, and b) pedogenons (map with pre-1750 vegetation) with soil texture data included in the RDA model. The legend indicates the number designation of the pedogenon classes.

old soils that have a higher probability of polygenesis (Wilding, 1994). The dynamic evolution of pedogenetic processes linked to climate is a challenge for the quantitative modelling of pedogenons. A partial solution for highlighting the relevance of former climates would be to integrate spatial information of paleoclimates (Brown et al., 2018) or the most detailed available information of geological stratigraphy or age of the geological substrate as a proxy for the soil-forming factor time. High-resolution paleoclimate surfaces available at global scale, such as WorldClim 1.4 (Hijmans et al., 2005) and PaleoClim (Brown et al., 2018), provide bioclimatic indices for several time periods of the Holocene, Pleistocene and Pliocene. However, the incorporation of paleoclimate data would raise the questions on how to weigh the data depending on the duration of the period they represent, and whether the cluster analysis should be done simultaneously on all covariates or designing a sequential workflow that grouped covariates by time periods.

4.4. Pedogenon modelling approach

The methodology presented here was implemented as a two-step clustering approach. In the first step we applied a non-hierarchical hard clustering algorithm (*k*-means) to identify mapping units suitable for local and regional studies. The second step consisted in a hierarchical clustering using Ward's criterion for quantifying the similarities among classes and describe its organization into higher-level taxa. Hierarchical clustering offers some flexibility to the end-user (e.g., environmental and agricultural agencies, land managers), who can select the desired number of classes for national or regional assessments and recalculate the summary statistics of environmental variables per class for higher-level taxa. We gave the same weight to each variable and tried to keep the balance between all represented soil-forming factors although climate had more representation than relief or parent material (nine climate variables, seven vegetation variables, four parent material variables and five relief variables). The output pedogenon maps are very sensitive to the representation of the soil-forming factors and covariates used as proxies. Here, we selected environmental covariates that have meaning for pedogenesis through trial and error, but the optimal selection of covariates may improve with automated algorithms. The covariate selection process could use soil data from traditional soil surveys (e.g., stable properties from subsoil horizons and information on pedogenetic pathways), preferentially with multivariate methods that allow for simultaneous selection of variables and consider multidimensional data. An alternative method would be to follow a successive clustering approach (Roell et al., 2020). Environmental variables, proxies of the dominant soil-forming factors at national scale (e.g., paleoclimate and parent material), would be used in a first clustering step. Afterwards these units would be divided into more classes using variables of soil-forming factors relevant at subregional scale (e.g., current climate, relief, parent material at higher resolution, pre-1750 vegetation). Since the choice of the variables affects greatly the resulting pedogenon classes (Fig. 2), this approach requires expert-knowledge on the hierarchy of soil-forming factors at national and subregional scale.

The pedogenon mapping approach is similar to digital terron mapping at regional (Carré and McBratney, 2005; Malone et al., 2014; Coggins et al., 2019) and national scale (Peng et al., 2020; Roell et al., 2020) in its use of unsupervised soil classification with clustering algorithms. However, there are some conceptual differences between the concepts of *terron* and *pedogenon*. The term terron was originally linked to the terroir, i.e., an area where the combination of agricultural management, history and culture with a particular soil, landscape and microclimate confers distinctive qualities to food products. A terron was defined as a soil-landscape entity described by soil and landscape attributes and their interactions (Carré and McBratney, 2005). Climate variables were later included for defining terron units (Coggins et al., 2019) in larger areas with heterogeneous microclimate (Peng et al., 2020; Roell et al., 2020). The definition of terron is oriented to crop production and environmental risk assessment, and the application determines the soil and environmental variables selected for the analysis. The objective of digital terron mapping is to create soil-landscape entities characterized by homogeneous soil, landscape, and climate attributes as a tool for supporting agricultural management. On the other hand, the objective of pedogenon mapping is to delineate soil entities with homogeneous soil-forming factors for a given reference time and among other applications, may serve as basis for soil change assessment. Besides differences in definition and application, there were differences with the modelling framework for terron mapping at national scale. Roell et al. (2020) followed a hierarchical clustering approach that started from the definition of national terrons that were further divided into regional terron classes. However, both terron and pedogenon

mapping are flexible and reproducible methods, applicable locally or over large areas, and face the same caveats of variable selection and optimization of the number of classes (Peng et al., 2020; Roell et al., 2020). Including some soil gridded information that could be used as proxy for long-term pedogenetic processes (e.g., subsoil mineralogy) would incorporate some elements of terron mapping but could be equally applied to generate soil mapping units for soil change assessment.

4.5. The concept of pedogenon and its potential applications for soil management

The definition of pedogenon as expressed in this study is broader than that of genon by Boulaine (1969) since the classes were generated exclusively with information on soil-forming factors and not on soil properties. Nevertheless, the methodology can be applied at a larger scale for identifying local *genons* of size closer to that conceptualized by Boulaine. At regional scale we missed some unique combinations of soil-forming factors due to the current sampling density (259,000 pixels represent less than 1% of the New South Wales gridded information at 90 m resolution), but higher sampling density requires longer processing time. Pedogenon maps at local scale can use a higher sampling density of the environmental covariates that would improve the identification of local patterns in soil variation. For example, in the southwest of Nowley Farm there is an area with sandy texture that was not discriminated when mapping at regional scale (Stockmann et al., 2016).

The spatial characteristics of the pedogenon classes mapped at the scale of New South Wales allows assessment of the soil condition and capability of a local study area within its regional context, meaning that knowledge on the effects of management practices on soil properties may be transferable between farms with similar pedogenon classes. Pedogenon classes with marginal presence in an area of interest can also be merged with their closest pedogenon class using the Mahalanobis distance between centroids as a criterion.

Pedogenon maps provide spatially explicit information for investigating the effect of anthropogenic activities on soil properties with space-for-time substitutions and transfer knowledge on the effect of different management practices between study areas of homogeneous environmental conditions. Conceptual frameworks of soil change (Yaalon and Yaron, 1966; Richter, 2007) can be implemented with pedogenon maps, if the classes are interpreted as genoforms for a reference time. Then, information on human activities since the reference time (e.g., land use change, agricultural management) can be overlayed to divide these soil classes into subclasses, and dissimilarity metrics between soil profiles can be used for identifying persistent soil variations (phenoforms). These variations are persistent enough that significant management intervention or a long period under a different land management would be necessary to revert them to the genoform (Rossiter and Bouma, 2018). Off-site alterations ranging from local to global change (e.g., atmospheric deposition, climate change) can be incorporated as explanatory variables in statistical models used for soil change assessment.

Another potential application is to provide the basis for designing soil monitoring surveys. A stratified random sampling scheme based on pedogenon information could be implemented for estimating baseline values of soil properties by pedogenon families (higher level taxa). For example, the strata may be defined by the combination of land use/cover with pedogenon families, resembling existing soil sampling schemes at national and continental scale (e.g., LUCAS Soil, Orgiazzi et al. (2018)).

4.6. Evaluation of pedogenon classes

The current study does not include an explicit evaluation of the pedogenon classes besides testing their explanatory power for subsoil texture. Pedogenon class explained only 29% and 35% of the variation of subsoil clay and silt, and the RDA ordination plot indicated high within-class variability (Fig. 7). Given the spatial extent of the pedogenon classes, it is expected that the soils found within a pedogenon have some degree of variability. Nevertheless, the RDA results suggest that the methodology for pedogenon mapping requires further improvement and these first pedogenon maps should be considered with prudence. However, the pedogenons' ability to explain particle size distribution may be enhanced by including more variables for parent material at a finer resolution, such as remote sensing spectral band ratios for unvegetated areas (Cudahy et al., 2016; Roberts et al., 2019), although the latter may be limited to local pedogenon maps in semi-arid and arid areas (Regmi and Rasmussen, 2018). Only two stable soil properties were included as response variables, but including other stable soil attributes like soil crystalline minerals, secondary minerals and Fe and Al oxides, presence of duripan (silcrete) or plinthite (Richter, 2007) could improve the evaluation of the pedogenon classes with legacy soil data. We did not include these variables in our analysis because the number of observations was not as numerous as particle size distribution. Hence it would limit the number of pedogenon classes included in the RDA. The evaluation of the pedogenon maps with new field data is unfeasible at the scale of New South Wales in terms of time and human resources. However, the measurement of soil properties and description of soil profiles by pedogenon class is feasible at local scale and will be carried in future studies.

The comparison of soil profile properties between pedogenon classes can use several metrics and sources of information, e.g., taxonomic distance based on soil horizon and profile attributes calculated with stable soil properties (Carré and Jacobson, 2009) or soil spectral information (Viscarra Rossel et al., 2011). Soil spectra in the mid and vis-near infrared range have intrinsic information on the biological, chemical and physical soil properties (Soriano-Disla et al., 2014). Hence, by selecting spectral bands correlated with the mineral soil fraction, it may be possible to quantify the similarities in soil properties among pedogenon classes and even compare the degree of diversity within and between different pedogenon classes using spectral-based diversity indices (Fajardo et al., 2017). The use of spectral information for characterizing pedogenon classes has to be done with caution, since it is sensitive to land use changes (Tivet et al., 2013; James et al., 2019). Hence, it may be more suitable to assess the effects of anthropogenic activities on soil spectral properties in interaction with pedogenon class.

5. Conclusion

This study introduced a methodology for mapping pedogenons, units defined by homogeneous quantitative state variables representing the soil-forming factors for a reference time. The main purpose of pedogenon mapping is to delineate units with similar multimillennial pedogenesis and historic anthropedogenesis that will serve as basis for soil change assessment, as well as for designing soil sampling and field surveys that can afterwards serve for soil mapping. The two-step clustering approach combined *k*-means (but could be replaced by fuzzy clustering or other partitioning algorithms) and hierarchical clustering. This framework allows us to generate pedogenon classes over large areas at fine resolution, is reproducible and easily applicable at different scales. This can be especially useful for large areas lacking detailed soil surveys (Regmi and Rasmussen, 2018) or where soil mapping units do not reflect common long-term pedogenesis. The application of the methodology at the state scale produced a detailed classification that had meaningful spatial patterns. The flexibility of the model enables us to merge pedogenon classes developed in a taxonomic level equivalent to soil family into higher level taxa depending on the management applications. Pedogenon mapping faces the challenges of optimizing the number of classes and variable selection, but more relevant for ancient landscapes is representing different polygenetic pathways. Indeed, the choice of covariates becomes an even more critical step in polygenetic soils (pedogenic paleosols *sensu* Richter and Yaalon, 2012). The

selection of the optimal number of classes and covariates needs to be improved by implementing objective cluster evaluation metrics and multivariate selection methods and soil information respectively. Future work will demonstrate how to apply pedogenon maps for assessing the effects of recent anthropedogenesis on soil condition and capability, and also focus on developing and evaluating pedogenon maps at local scale or support the design of soil monitoring surveys.

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.

Acknowledgments

We would like to thank Ross Searle of CSIRO for his help with the Soil Data Federator and for his suggestions on the covariate selection. The authors acknowledge the Terrestrial Ecosystem Research Network (TERN), an Australian Government NCRIS-enabled research infrastructure project, for facilitating and supporting this research.

Code availability

The scripts for implementing the modelling framework and visualize pedogenon maps and dendrograms are available at https://github.com/MercedesRD/Pedogenons and include a reproducible example.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2021.115012.

References

Zinck, J.A., 2016. The Pedologic Landscape: Organization of the Soil Material. In: Zinck, J.A., Metternicht, G., Bocco, G., Valle, DelH.F. (Eds.), Geopedology. Springer, Cham. https://doi.org/10.1007/978-3-319-19159-1_5.

Arthur, D., Vassilvitskii, S. 2007. K-means++: the advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA '07). Society for Industrial and Applied Mathematics, USA, 1027–1035. doi: 10.1145/1283383.1283494.

Ashcroft, L., Gergis, J., Karoly, D.J., 2014. A historical climate dataset for southeastern Australia, 1788–1859. Geosci. Data J. 1, 158–178. https://doi.org/10.1002/gdj3.19.

Banks, R.G., 1995. Soil Landscapes of the Curlewis. 1:100 000 Sheet. Department of Conservation and Land Management, Sydney.

Binkley, D., Fisher, R.F., 2020. Ecology and Management of Forest Soils. Wiley Online Library.

Bird, M.I., Beaman, R.J., Condie, S.A., Cooper, A., Ulm, S., Veth, P., 2018. Palaeogeography and voyage modeling indicates early human colonization of Australia was likely from Timor-Roti. Quaternary Sci. Rev. 191, 431–439. https://doi.org/10.1016/j.quascirev.2018.04.027.

Borcard, D., Gillet, F., Legendre, P., 2018. Canonical Ordination,in: Numerical ecology with R. Use R! (2nd ed.). Springer, Cham. 10.1007/978-3-319-71404-2_6.

Bockheim, J.G., Gennadiyev, A.N., Hartemin, A.E., Brevik, E.C., 2014. Soil-forming factors and Soil Taxonomy. Geoderma 226, 231–237. https://doi.org/10.1016/j.geoderma.2014.02.016.

Boulaine, J., 1969. Sol Pedon et Genon Concepts et definitions. Bull. AFES 5, 7–10.

Buchan, G.D., 2011. Temperature effects in soil. In: Gliński, J., Horabik, J., Lipiec, J. (Eds.), Encyclopedia of Agrophysics. Springer, Netherlands, Dordrecht, pp. 891–895.

Brown, J.L., Hill, D.J., Dolan, A.M., Carnaval, A.C., Haywood, A.M., 2018. PaleoClim, high spatial resolution paleoclimate surfaces for global land areas. Sci. Data. 5, 180254 https://doi.org/10.1038/sdata.2018.254.

Budiaji, W., Leisch, F., 2019. Simple K-medoids partitioning algorithm for mixed variable data. Algorithms 12. https://doi.org/10.3390/a12090177.

Bui, E.N., Henderson, B.L., 2013. C:N: P stoichiometry in Australian soils with respect to vegetation and environmental factors. Plant Soil 373, 553–568. https://doi.org/10.1007/s11104-013-1823-9.

Bunemann, E.K., Bongiorno, G., Bai, Z.G., Creamer, R.E., De Deyn, G., de Goede, R., Fleskens, L., Geissen, V., Kuyper, T.W., Mader, P., Pulleman, M., Sukkel, W., van Groenigen, J.W., Brussaard, L., 2018. Soil quality – a critical review. Soil Biol. Biochem. 120, 105–125. https://doi.org/10.1016/j.soilbio.2018.01.030.

Campbell, J.B., Edmonds, W.J., 1984. The missing geographic dimension to soil taxonomy. Ann. Assoc. Am. Geog. 74, 83–97. https://doi.org/10.1111/j.1467-8306.1984.tb01436.x.

Carré, F., Jacobson, M., 2009. Numerical classification of soil profile data using distance metrics. Geoderma 148, 336–345. https://doi.org/10.1016/j.geoderma.2008.11.008.

Carré, F., McBratney, A.B., 2005. Digital terron mapping. Geoderma 128, 340–353. https://doi.org/10.1016/j.geoderma.2005.04.012.

Coggins, S., Malone, B.P., Stockmann, U., Possell, M., McBratney, A.B., 2019. Towards meaningful geographical indications: validating terroirs on a 200 km^2 scale in Australia's lower Hunter Valley. Geoderma Reg. 16 https://doi.org/10.1016/j.geodrs.2019.e00209.

Coventry, R., Fett, D., 1979. A pipette and sieve method of particle size analysis and some observations on its efficacy. Commonwealth Scientific and Industrial Research Organization, 10.4225/08/5867f2ef8bfba.

Cramér, H., 1946. The Two-dimensional Case, Mathematical Methods of Statistics (43). Princeton University Press, Princeton.

Cudahy, T., Caccetta, M., Thomas, M., Hewson, R., Abrams, M., Kato, M., Kashimura, O., Ninomiya, Y., Yamaguchi, Y., Collings, S., Laukamp, C., Ong, C., Lau, I., Rodger, A., Chia, J., Warren, P., Woodcock, R., Fraser, R., Rankine, T., Vote, J., de Caritat, P., English, P., Meyer, D., Doescher, C., Fu, B., Shi, P., Mitchell, R., 2016. Satellite-derived mineral mapping and monitoring of weathering, deposition and erosion. Sci. Rep. 6, 23702. https://doi.org/10.1038/srep23702.

Day, P.R., 1953. Experimental confirmation of hydrometer theory. Soil Sci. 75, 181–186.

Dokuchaev, V.V., 1883. Russian Chernozem. SelectedWorks of V.V. Dokuchaev. Volume I (Translated in 1967). Israel Program for Scientific Translations. U.S. Department of Agriculture, Washington DC.

Droogers, P., Bouma, J., 1997. Soil survey input in exploratory modeling of sustainable soil management practices. Soil Sci. Soc. Am. J. 61, 1704–1710. https://doi.org/10.2136/sssaj1997.03615995006100060023x.

FAO, 1993. World Soil Resources. An Explanatory Note on the FAO World Soil Resources Map at 1 : 25,000,000 Scale. FAO World Soil Resources Reports 66, Rev. 1, Rome, 64 pp.

Fajardo, M.P., McBratney, A., Minasny, B., 2017. Measuring functional pedodiversity using spectroscopic information. Catena 152, 103–114. https://doi.org/10.1016/j.catena.2017.01.012.

Farr, T.G., Kobrick, M., 2000. Shuttle radar topography mission produces a wealth of data. Eos Trans. AGU 81 (48), 583–585. https://doi.org/10.1029/EO081i048p00583.

Fridland, V.M. 1972. Pattern of the soil cover. Moscow: Geographical Institute, Academy of Sciences of the USSR. (Israel Program for Scientific Translations, Jerusalem, 1976.).

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39, 1347. https://doi.org/10.1029/2002WR001426, 12.

Gallant, J., Wilson, N., Tickle, P.K., Dowling, T., Read, A., 2009. 3 Second SRTM Derived Digital Elevation Model (DEM) Version 1.0. Record 1.0. Geoscience Australia, Canberra.

Gammage, B., 2011. The Biggest Estate on Earth: How Aborigines made Australia. Allen & Unwin, Crows Nest, Australia.

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics 857–871. https://doi.org/10.2307/2528823.

Gray, J.M., Bishop, T.F.A., Wilford, J.R., 2016. Lithology and soil relationships for soil modelling and mapping. Catena 147, 429–440. https://doi.org/10.1016/j.catena.2016.07.045.

Guo, Y.Y., Gong, P., Amundson, R., 2003. Pedodiversity in the United States of America. Geoderma 117, 99–115. https://doi.org/10.1016/S0016-7061(03)00137-X.

Han, J., Kamber, M., Pei, J., 2012. 10-cluster analysis: Basic concepts and methods. In: Han, J., Kamber, M., Pei, J. (Eds.), Data mining, 3rd ed. Morgan Kaufmann, pp. 443–495. https://doi.org/10.1016/B978-0-12-381479-1.00010-1.

Haase, G., 1968. Pedon und Pedotop. In: Barthel, H. (Ed.), Landschaftsforschung. Hermann Haack, Leipzig, pp. 57–76.

Hartigan, J.A., Wong, M.A., 1979. A *K*-Means Clustering Algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics) 28, 100–108. https://doi.org/10.2307/2346830.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978. https://doi.org/10.1002/joc.1276.

Hobley, E., Wilson, B., Wilkie, A., Gray, J., Koen, T., 2015. Drivers of soil organic carbon storage and vertical distribution in Eastern Australia. Plant Soil 390, 111–127. https://doi.org/10.1007/s11104-015-2380-1.

Huang, J.Y., McBratney, A.B., Malone, B.P., Field, D.J., 2018. Mapping the transition from pre-European settlement to contemporary soil conditions in the Lower Hunter Valley, Australia. Geoderma 329, 27–42. https://doi.org/10.1016/j.geoderma.2018.05.016.

Huang, Z.X., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Disc. 2, 283–304. https://doi.org/10.1023/A:1009769707641.

Hughes, P., McBratney, A., 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 regional 11, 123–140. https://doi.org/10.1016/j.geodrs.2017.10.004.

Ibañez, J.J., Saldana, A., De-Alba, S., Lobo, A., Zucarello, V., 1998. Pedodiversity and global soil patterns at coarse scales. Geoderma 83, 171–192. https://doi.org/10.1016/S0016-7061(97)00147-X.

Isbell, R.F., McDonald, W.S., Ashton, L.J., 1997. Concepts and Rationale of the Australian Soil Classification. CSIRO Land and Water.
Isbell, R., the National Committee on Soil and Terrain, 2016. The Australian soil classification. CSIRO publishing, Melbourne.
IUSS Working Group WRB, 2006. World Reference Base for soil resources 2006, 2nd ed. World Soil Resources Reports No.103 FAO, Rome.
James, J.N., Gross, C.D., Dwivedi, P., Myers, T., Santos, F., Bernardi, R., de Faria, M.F., Guerrini, I.A., Harrison, R., Butman, D., 2019. Land use change alters the radiocarbon age and composition of soil and water-soluble organic matter in the Brazilian Cerrado. Geoderma 345, 38–50. https://doi.org/10.1016/j.geoderma.2019.03.019.
Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. Dover Publications, New York.
Johnson, W.M., 1963. The pedon and the polypedon. Soil Sci. Soc. Am. J. 27 (2), 212–215. https://doi.org/10.2136/sssaj1963.03615995002700020034x.
Jones, D.A., Wang, W., Fawcett, R., 2009. High-quality spatial climate data-sets for Australia. Aust. Meteorol. Oceanogr. J. 58 (4), 233.
Kleber, M., Eusterhues, K., Keiluweit, M., Mikutta, C., Mikutta, R., Nico, P.S., 2015. Mineral-organic associations: formation, properties, and relevance in soil environments. Adv. Agron. 130, 1–140. https://doi.org/10.1016/bs.agron.2014.10.005.
Krasilnikov, P.V., Ibañez, J.J., Arnold, R.W., Shoba, S., (Eds.), 2009. Soil Terminology, Correlation and Classification. Earthscan, London. 10.4324/9781849774352.
Legendre, P., Legendre, L.F., 2012. Canonical analysis. Numerical Ecology, 3rd ed. Elsevier, Amsterdam, The Netherlands.
Malone, B.P., Hughes, P., McBratney, A.B., Minasny, B., 2014. A model for the identification of terrons in the Lower Hunter Valley. Australia. Geoderma Reg. 1, 31–47. https://doi.org/10.1016/j.geodrs.2014.08.001.
Malone, B., Searle, R., 2020. Improvements to the Australian national soil thickness map using an integrated data mining approach. Geoderma 377, 114579. https://doi.org/10.1016/j.geoderma.2020.114579.
Marshall, T., 1956. A plummet balance for measuring the size distribution of soil particles. Aus. J. Appl. Sci. 7, 142–147.
McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.
McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. Geoderma 213, 203–213.
McBratney, A.B., Field, D., Morgan, C.L.S., Huang, J.Y., 2019. On soil capability, capacity, and condition. Sustainability 11, 3350. https://doi.org/10.3390/su11123350.
McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97, 293–327. https://doi.org/10.1016/S0016-7061(00)00043-4.
Minasny, B., McBratney, A.B., Hartemink, A.E., 2010. Global pedodiversity, taxonomic distance, and the World Reference Base. Geoderma 155, 132–139. https://doi.org/10.1016/j.geoderma.2009.04.024.
Minty, B., Franklin, R., Milligan, P., Richardson, L.M., Wilford, J.R., 2009. Radiometric Map of Australia. Geoscience Australia, Canberra http://pid.geoscience.gov.au/dataset/ga/68851.
Minty, B.R.S., 2019a. Radiometric Grid of Australia (Radmap) v4 2019 filtered pct potassium grid. Geoscience Australia, Canberra https://doi.org/10.26186/5dd48d628f4f6.
Minty, B.R.S., 2019b. Radiometric Grid of Australia (Radmap) v4 2019 filtered ppm thorium. Geoscience Australia, Canberra https://doi.org/10.26186/5dd48e3eb6367.
Minty, B.R.S., 2019c. Radiometric Grid of Australia (Radmap) v4 2019 ratio uranium over thorium. Geoscience Australia, Canberra https://doi.org/10.26186/5dd4a63603704.
Mouselimis, L., 2020. ClusterR: Gaussian Mixture Models, K-Means, Mini-Batch-Kmeans, K-Medoids and Affinity Propagation Clustering. R package version 1.2.2. https://CRAN.R-project.org/package=ClusterR.
Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., Toxopeus, A.G., Alibakhshi, S., 2019. ELSA: Entropy-based local indicator of spatial association. Spat. Stat.-Neth. 29, 66–88. https://doi.org/10.1016/j.spasta.2018.10.001.
National Vegetation Information System V5.1 ©, 2018. Australia - Pre-1750 Major Vegetation Groups – NVIS Version 5.1 (Albers 100m analysis product). Australian Government Department of Agriculture, Water and the Environment.
Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., 2019. vegan: Community Ecology Package.
NSW National Parks and Wildlife Service, 2003. The Bioregions of New South Wales: their biodiversity, conservation and history. NSW National Parks and Wildlife Service, Hurstville.
Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS Soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69, 140–153. https://doi.org/10.1111/ejss.12499.
Pascoe, B., 2014. Dark Emu: Aboriginal Australia and the Birth of Agriculture. Magabala Books, Broome, Australia.
Peng, Y., Roell, Y.E., Moller, A.B., Adhikari, K., Beucher, A., Greve, M.B., Greve, M.H., 2020. Identifying and mapping terrons in Denmark. Geoderma 363. https://doi.org/10.1016/j.geoderma.2020.114174.
Pino, V., McBratney, A., Fajardo, M., Wilson, N., Deaker, R., 2019. Understanding soil biodiversity using two orthogonal 1000km transects across New South Wales, Australia. Geoderma 354. https://doi.org/10.1016/j.geoderma.2019.07.018.
Quinn, P., Beven, K., Chevallier, P., Planchon, O., 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrol. Process. 5, 59–79. https://doi.org/10.1002/hyp.3360050106.
Regmi, N.R., Rasmussen, C., 2018. Predictive mapping of soil-landscape relationships in the arid Southwest United States. Catena 165, 473–486. https://doi.org/10.1016/j.catena.2018.02.031.
Richter, D.D., 2007. Humanity's transformation of Earth's soil: Pedology's new frontier. Soil Sci. 172, 957–967. https://doi.org/10.1097/ss.0b013e3181586bb7.
Richter, D.D., Yaalon, D.H., 2012. The Changing Model of Soil, Revisited. Soil Sci. Soc. Am. J. 76, 766-778. doi.org/10.2136/sssaj2011.0407.
Roell, Y.E., Peng, Y., Beucher, A., Greve, M.B., Greve, M.H., 2020. Development of hierarchical terron workflow based on gridded data – a case study in Denmark. Comput Geosci-Uk 138. https://doi.org/10.1016/j.cageo.2020.104454.
Rossiter, D.G., Bouma, J., 2018. A new look at soil phenoforms – definition, identification, mapping. Geoderma 314, 113–121. https://doi.org/10.1016/j.geoderma.2017.11.002.
Saunders, A.M., Boettinger, J.L., 2006. Chapter 28 Incorporating Classification Trees into a Pedogenic Understanding Raster Classification Methodology, Green River Basin, Wyoming, USA, in: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Developments in Soil Science. Elsevier, pp. 389-620. 10.1016/S0166-2481(06)31028-8.
Smeck, N.E., Balduff, D., 2002. Contrasting approaches for the classification of eroded soils in the USA. Paper no. 616 in Transactions of the 17th World Congress of Soil Science: Confronting New Realities in the 21st Century. Bangkok, Thailand.
Soil Survey Staff, 2010. Keys to Soil Taxonomy. United States Department of Agriculture, Soil Conservation Service, Washington, DC.
Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M. J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49 (2), 139–186. https://doi.org/10.1080/05704928.2013.811081.
Stockmann, U., Cattle, S.R., Minasny, B., McBratney, A.B., 2016. Utilizing portable X-ray fluorescence spectrometry for in-field investigation of pedogenesis. Catena 139, 220–231. https://doi.org/10.1016/j.catena.2016.01.007.
Tivet, F., de Moraes Sá, J.C., Lal, R., Milori, D.M.B.P., Briedis, C., Letourmy, P., Pinheiro, L.A., Borszowskei, P.R., da Cruz Hartman, D., 2013. Assessing humification and organic C compounds by laser-induced fluorescence and FTIR spectroscopies under conventional and no-till management in Brazilian Oxisols. Geoderma 207 (208), 71–81. https://doi.org/10.1016/j.geoderma.2013.05.001.
Tobler, R., Rohrlach, A., Soubrier, J., Bover, P., Llamas, B., Tuke, J., Bean, N., Abdullah-Highfold, A., Agius, S., O'Donoghue, A., O'Loughlin, I., Sutton, P., Zilio, F., Walshe, K., Williams, A.N., Turney, C.S.M., Williams, M., Richards, S.M., Mitchell, R. J., Kowal, E., Stephen, J.R., Williams, L., Haak, W., Cooper, A., 2017. Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. Nature 544, 180–184. https://doi.org/10.1038/nature21416.
Triantafilis, J.; McBratney, A.B., 1993. Application of continuous methods of soil classification and land suitability assessment in the Lower Namoi Valley. Canberra, ACT: CSIRO Division of Soils. 10.25919/5c9522aec8bad.
Viscarra Rossel, R.A., Chappell, A., De Caritat, P., McKenzie, N.J., 2011. On the soil information content of visible–near infrared reflectance spectra. Eur. J. Soil Sci. 62, 442–453. https://doi.org/10.1111/j.1365-2389.2011.01372.x.
Ward, W.T., 1999. Soils and landscapes near Narrabri and Edgeroi, NSW, with data analysis and using fuzzy k-means. CSIRO Land and Water Technical Report No.:22/99. http://hdl.handle.net/102.100.100/213385?index=1.
Wicklin, R., 2012. What is Mahalanobis Distance? https://blogs.sas.com/content/iml/2012/02/15/what-is-mahalanobis-distance.html (accesed 4 March 2020).
Wilford, J., 2012. A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis. Geoderma 183, 124–142. https://doi.org/10.1016/j.geoderma.2010.12.022.
Williams, K.J., Belbin, L., Austin, M.P., Stein, J.L., Ferrier, S., 2012. Which environmental variables should I use in my biodiversity model? Int. J. Geogr. Inf. Sci. 26, 2009–2047. https://doi.org/10.1080/13658816.2012.698015.
Wilson, J.P., Gallant, J.C., 2000. Secondary topographic attributes. In: Wilson, J.P., Gallant, J.C. (Eds.), Terrain Analysis: Principles and Applications. John Wiley & Sons, New York, pp. 87–131.
Wilding, L.P., 1994. Factors of soil formation: contributions to pedology. Factors of soil formation: a fiftieth anniversary retrospective, SSSA Special Publication No. 33, SSSA, Madison, WI. pp. 15-30. https://doi.org/10.2136/sssaspecpub33.c2.
Xu, T., Hutchinson, M.F., 2011. ANUCLIM Version 6.1 User Guide. Fenner School of Environment and Society. The Australian National University.
Yaalon, D.H., Yaron, B., 1966. Framework for Man-Made Soil Changes - an Outline of Metapedogenesis. Soil Sci 102 (4), 272–277.
Zuur, A.F., Ieno, E.N., Smith, G.M., 2007. Principal component analysis and redundancy analysis, in: Analysing Ecological Data. Statistics for Biology and Health. Springer, New York, NY. 10.1007/978-0-387-45972-1_12.