FedUni ResearchOnline

<https://researchonline.federation.edu.au>

Copyright Notice

This is the submitted version of the following article:

Benke, K., Norng, S., Robinson, N. ., Chia, K., Rees, D. ., & Hopley, J. (2020). Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content. *Geoderma*, 366

Which has been published in final form at:

<https://doi.org/10.1016/j.geoderma.2020.114210>

Copyright © Elsevier Inc. All rights reserved.

Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content

K.K. Benke^{1,2}, S. Norng¹, N.J. Robinson³, K. Chia¹, D.B. Rees¹ and J. Hopley¹

¹ Centre for AgriBioscience, AgriBio, Ring Road, Bundoora, Victoria, 3083, Australia.

² School of Engineering, University of Melbourne, Parkville, Victoria, 3010, Australia.

³ School of Science, Information Technology and Engineering, Federation University Australia, University Drive, Mt Helen, VIC, 3350.

Corresponding author: Email: n.robinson@federation.edu.au

Abstract

The pedotransfer function is a mathematical model used to convert direct soil measurements into known and unknown soil properties. It provides information for modelling and simulation in soil research, hydrology, environmental science and climate change impacts, including investigating the carbon cycle and the exchange of carbon between soils and the atmosphere to support carbon farming. In particular, the pedotransfer function can provide input parameters for landscape design, soil quality assessment and economic optimisation. The objective of the study was to investigate the feasibility of using a generalised pedotransfer function derived with a machine learning method to predict soil electrical conductivity (EC) and soil organic carbon content (OC) for different regional locations in the state of Victoria, Australia. This strategy supports a unified approach to the interpolation and population of a single regional soils database, in contrast to a range of pedotransfer functions derived from local databases

with measurement sets that may have limited transferability. The pedotransfer function generation was based on a machine learning algorithm incorporating the Generalized Linear Mixed Model with interactions and nested terms, with Residual Maximum Likelihood estimation, and a predictor-frequency ranking system with step-wise reduction of predictors to evaluate the predictive errors in reduced models. The source of the data was the Victorian Soil Information System (VSIS), which is a database administered for soil information and mapping purposes. The database contains soil measurements and information from locations across Victoria and is a repository of historical data, including monitoring studies. In total, data from 93 projects were available for inputs to modelling and analysis, with 5158 samples used to derive predictors for EC and 1954 samples used to derive predictors for OC. Over 500 models were tested by systematically reducing the number of predictors from the full model. Five-fold cross-validation was used for estimation of model mean-squared prediction error (MSPE) and mean-absolute percentage error (MAPE). The results were statistically significant with only a gradual reduction in error for the top-ranked 50 models. The prediction errors (MSPE and MAPE) of the top ranked model for EC are 0.686 and 0.635, and 0.413 and 0.474 for OC respectively. The four most frequently occurring predictors both for EC and OC prediction across the full set of models were found to be soil depth, pH, particle size distribution and geomorphological mapping unit. The possible advantages and disadvantages of this approach were discussed with respect to other machine learning approaches.

Keywords: Artificial Intelligence, Big Data, Biostatistics, Carbon Farming, Climate Change, Predictive Analytics, Regional Analysis, Simulation

1. Introduction

1.1 Soil health

Soil data is used for the assessment of land condition, modelling soil function, and to support climate change research by analysis of carbon storage capacity and stability. Primary threats to soil function include soil salinization and decreasing levels of soil organic matter content (ITPS, 2015). Soil salinity is a major physico-chemical constraint in Mediterranean-type environments typical of the south-west Pacific region, with an impact on yields at the farm scale and the regional scale (Dang and Moody, 2016; Orton et al., 2018). The Millennium Drought was associated with a reduced impact of dryland salinity to both agricultural and natural ecosystems, although the future prognosis is unclear in response to climate change, variability and adoption of conservation land management practices (van Dijk et al., 2013). The decline in soil organic matter content (and carbon) in Australia has been considerable and is due to land clearance, removal of native vegetation and implementation of conventional cropping practices (Bell et al., 2007; Luo et al., 2010). The impacts, however, of current farming systems and management practices on organic matter distribution are not as clearly distinguished as climatic and soil differences (Allen et al., 2013; Robertson et al., 2016). While there have been recent efforts to understand the nature of key factors influencing the distribution of organic carbon content across Australia (Baldock et al., 2013), there are still significant gaps in contemporary soil information on the decline in soil organic matter. Likewise, accurate assessments of soils affected by salinization are problematic owing to this dearth of soil data and information (ITPS, 2015).

In many countries, soil data is used for applications such as land condition monitoring, predicting soil functions, understanding carbon storage capacity and stability, land use capability and agronomic modelling for edaphic needs. Soil organic carbon content and soil electrical conductivity are properties fundamental to these types of assessments. There has been considerable effort in the last decade to conserve and make available soil profile data for future applications. The GlobalSoilMap (GSM) project was a key driver in this significant conservation effort (Arrouays et al., 2017). With this collation of soil profile data, countries such as Australia achieved considerable gains in harmonising legacy soil

data collected using different standards and methods prior to adoption of agreed national standards in the 1980s. The National Soil Site Collation (NSSC) harmonised over 281,000 soil profiles for 2.4M laboratory measurements (Searle, 2014). As anticipated, the costlier or more time-consuming the measurement (e.g. mineralogy, available phosphorus or nitrogen, bulk density), the fewer the number of measurements available in this collation. For example, soil organic carbon (OC) and electrical conductivity (EC) measurements are available for 9.2% and 13.8 % of these profiles, and clay % (5.7%) respectively. Morphological observations such as coarse fragments (85% of profiles) and field texture (76% of horizons) are far more prolific in this collection.

Numerous gaps in regional soil property data are a continuing problem, as discussed at length by Biggs and Searle (2017), with two potential remedial options being available. First, one may seek access to private soil property data (Arrouays et al., 2017) using FAIR, i.e. *Findable, Accessible, Interoperable* and *Reusable* principles to fill spatial and temporal data gaps (Robinson et al., 2018). A key principle of making soil data FAIR is to encourage data sharing practices amongst data owners and users to enable maximum use and reuse of data inter-operably between machines and systems. Secondly, one may develop and apply predictive functions to populate gaps in soil site databases. A predictive model based on experimental measurements of soil properties is known as a pedotransfer function and is often developed using data from field surveys (Bouma, 1989; Minasny et al., 1999, McBratney et al., 2002), or data from smaller specialized databases (Van Looy et al., 2017).

1.2 Relevance of electrical conductivity and organic carbon content

Studies to date have shown that broad environmental influences such as climate, landform and parent material over variable time frames are key factors in explaining the distribution and occurrence of soil properties, including electrical conductivity (EC) and organic carbon content (OC). In the case of OC, Robertson et al. (2016) determined that OC was highly correlated with rainfall, temperature and to a lesser extent parent material. For example, in the past, OC was always measured below 1.5% in north-western Victoria (part of the North Western Dunefields and Plains Geomorphological Mapping Unit. The potential for OC accumulation in such an area requires conservative land management practices

for crop/pasture rotations to positively accrue OC in these soils (Robertson and Nash, 2013) with benefits to agricultural production as well as a factor in soil and landscape stability. The examples in Robertson et al. (2016) clearly illustrate that higher OC values (5-10%) are associated with higher rainfall and cooler temperatures, such as the Otway Ranges in southern Victoria. Similarly, the Strzelecki Ranges with parent material of feldspathic sandstone that has weathered to clays is more favourable for OC retention in comparison to sandier soils nearby that have also been influenced by past agricultural management (Baldock and Skjemstad, 1999). At regional to local scales, factors such as topographic position, slope and water accumulation/dispersion were found to be relevant to OC accumulation in the Southern Uplands and Western Uplands landscapes (Robertson et al., 2016).

Dryland salinity in Australia remains a significant land degradation threat to agricultural production with 5.7M hectares affected (NLWRA, 2001). In Victoria, around 2% of dryland agricultural land (247,000 hectares) is impacted by salinity. For soil salinity, similar environmental factors to OC (rainfall, elevation, GMU) were influential in explaining the distribution and quantity of salinisation in soils of Victoria. Accumulation of salts in soils of south-eastern Australia has been driven by vast quantities of salt sourced from the Southern Ocean during Quaternary arid weathering events, either as wind borne sediments (e.g. parna) and/or rain. Salts have also been sourced from the weathering of parent materials (e.g. Palaeozoic marine sediments, igneous incursions and volcanic rocks) and from marine incursions of the Cenozoic era. Combined with low rainfall conditions and subsoil constraints such as dense subsoils, transient salinity and sodicity are common for many texture contrast soils of western Victoria (MacEwan et al., 2010). While insufficient leaching has been associated with accumulation of salts in soil, there remain landscapes with high rainfall conditions where salts are still in high quantities in the root zone of crops and pastures (Ford et al., 1993). Other groundwater and dryland salinity processes (e.g. perched water tables, geological faulting) – which often result in lateral flow, have been linked to these waterlogging and salinity occurrences (Dahlhaus and MacEwan, 1997).

It is recognised that saline soils are strongly associated with broad geomorphic regions of Victoria which are spatially defined as coarse mapping units (Tier 1 divisions) in the Victorian Geomorphological Framework (Rees et al., 2010). These Geomorphological Mapping Units (GMU) are nested in a 3-tier

hierarchy based principally on differences in physiography, soils, vegetation and climate and align with landscape processes such as salinization. For example, salts mobilised south of the Great Dividing Range (GDR: watershed) have a vastly quicker path to the Tasman sea (an order of magnitude) than those north of the GDR where drainage gradients are much lower, as is rainfall and higher evaporation. The effect of salt accumulation (predominantly sodium chloride) is mainly observed along the footslopes of the Great Dividing Range which feed into rivers systems of the Murray Darling Basin.

1.3 Pedotransfer functions

The pedotransfer function (PTF) is used to convert *direct* soil measurements into *inferred* soil properties, or soil characteristics, and provides information for simulation and quality assessment in soils research, hydrology, environmental science and landscape design and optimisation. The formulation of a PTF should be driven by clear objectives to predict properties and parameters that are needed and that lead to improvement of existing soil databases and reliant soil information systems (Nemes, 2015). Populating missing property measurements including OC and EC in the world soil databases is a key step to progress knowledge on soil processes and biophysical interactions.

Soil properties used as inputs to the PTF include pH, soil depth, particle size distribution, texture class and soil structure. The PTF is often used to predict soil functions such as soil hydraulic properties, carbon density and storage potential and phosphorus sorption and availability (see Minasny et al., 2010; Van Looy et al., 2017; McNeill et al., 2018).

Many measured soil properties are inter-dependent and the use of a PTF is widely recognised as a means for modelling relationships in soil research (McBratney et al., 2002). An example is provided by Grunwald (2016), who notes that drainage is very dependent on texture, structure and bulk density which can be linked by a predictive model and defined mathematically subject to calibration on a training set of data. The model inputs may be statistical or categorical in nature. The PTF is commonly used to predict soil attributes at remote locations in a rapid and cost-effective manner in the absence of measurement data and is used for interpolation and extrapolation. There are also additional variables associated with soil forming factors and processes, such as terrain and climate (which are now digitized,

and widely accessible and available) and which have been included in PTF development (Jenny 1941; Van Looy et al., 2017).

A possible limitation with the classical approach of using a diverse range of PTF models to predict soil properties of interest at different locations is that the validity and reliability of independent derivations varied across the landscape. Many databases used in development of PTF models are based only on locally relevant and specialised datasets (Van Looy et al., 2017) rather than extensive regional soil profile databases that may also contain data gaps due to methodology changes over time or where resource constraints precluded more extensive laboratory analyses at the time of the survey. The multiplicity of proposed models adds complexity and challenges to modellers and analysts in other disciplines when they were confronted with PTF models that are training data dependent, therefore might not be applicable elsewhere. Examples include different soil methods used for classification and texture analysis in different countries (Minasny and McBratney, 2001). In Victoria, PTFs have been developed and used in the past to predict soil hydraulic parameters (Rab et al., 2011, 2016; Selle et al., 2011) and variation in SOC stock (Robertson et al., 2015) using comprehensively analysed datasets.

A recent detailed and informative study by McNeill et al. (2018) compared three statistical methods for producing empirical *generalised* pedotransfer functions for soil water content and uncertainty using a soil information system from New Zealand. They concluded that a logit model was best in this application because it was simple and produced the lowest statistical root-mean-square-error (RMSE) estimate, which is a criterion for uncertainty with respect to statistical variability. This study was quite novel in many respects, for example using the Akaike Information Criterion (AIC) value for backwards selection of explanatory variables aimed at finding the best model and using a combination of explanatory variables that included soil order and continuous variables.

The use of conventional multivariate statistical methods has been augmented by the emergence of related *machine learning* approaches that offer the prospect of a generalised approach that could be calibrated, if necessary, in a variety of jurisdictions. The proposed new approaches provide computational flexibility with multiple inputs and a statistical methodology that is optimised for

prediction accuracy. As a result, a variety of methods are now under investigation and there is a continuing need to study the advantages and disadvantages of different approaches.

Machine learning has been described as a process of goal-directed learning by iterative adjustment of model coefficients in order to maximise a performance index (Nilsson, 2014; Russell and Norvig, 2016). Note that classical multiple linear regression is not a learning algorithm because the coefficients are obtained by matrix inversion of the state equations in a single step. Examples of machine learning approaches include artificial neural networks, which are used for classification and function approximation and have been used for soil water retention models, such as the Van Genuchten equation (Minasny and McBratney, 2002; Minasny et al., 2004; Baker and Ellison, 2008). Another common machine learning approach is the so-called Random Forest algorithm (Chen et al., 2018a). In a comparative study, Hengl et al. (2018) investigated multinomial logistic regression and classical machine learning approaches, such as the Artificial Neural Network (ANN), Random Forest (RF), Generalised Boosted Regression and k -Nearest Neighbour approaches. They used five-fold cross-validation and mean-squared-error criteria and found that the RF algorithm was more effective than the ANN in the specialized application of global mapping of vegetation. Many of these approaches, however, have a common theme in that they are predictive models rather than explanatory models. It is noted that the use of partial dependence plots can enhance the explanation of the machine learning approaches. The relationship between the covariates in the model structure, however, is generally not known, or whether there are interactions between inputs, and this has resulted in criticisms that, for example, the ANN is a ‘black box’ prediction approach without an explanatory basis (Castelevecchi, 2016).

The ANN is the most famous and widely used example of a machine learning model and is often used as an exemplar of machine learning (Nilsson, 2014). It is a prototypical approach illustrating the common conceptual basis of many machine learning approaches, i.e. inputs connected to an array of weights which are connected to outputs (such as occurs also with RF algorithms). For example, to link it with the current approach, the ANN back-propagation model employs a nonlinear transfer function

(not to be confused with the PTF) in each layer for each output node, which is often characterised by an exponential model of the form :

$$y = \frac{1}{1 - e^{-f(X)}} \quad (1)$$

where y is an output node and $f(X)$ is the weighted sum of inputs from the previous layer. This transfer function is conceptually similar to a *logistic* model being applied to each node in the network, where the exponent $f(X)$ is a multivariate linear model (Nilsson, 2014; Russell and Norvig, 2016).

An alternative to the ANN as a predictive model is the well-known logistic model, which can be implemented in a linearized form by logarithmic transformation of Eqn (1), as shown in Eqn (2). A more complex form of $f(X)$ can now be introduced to include linear and nonlinear additive terms, and interactions and nested terms. In this paper, the Generalised Linear Mixed Effects Model (GLMM) model structure comprises of a so-called *log-link* function and the parameter estimation is carried out by the Residual Maximum Likelihood (REML) method, i.e.

$$\log Y = f(X) \quad (2)$$

where $Y = f(y)$ and $f(X)$ is a function of several predictors and their interactions (Galwey, 2014), where $X_i \in \{X_1, \dots, X_n\}$. For example, proposed models for electrical conductivity or soil organic carbon content could include the predictors $X_1 = \text{pH}$, $X_2 = \text{rainfall}$, $X_3 = \text{soil depth}$ and $X_4 = \text{elevation data}$, as recorded in the VSIS database. A GLMM may start with the linearized form shown in Eqn (2), as the log transform has the advantages of making the analysis more tractable and supports variance stabilisation with respect to the residuals (Galwey, 2014). The general form of the proposed model can be explored further by initially plotting pairwise relationships between variables, with possible identification of non-linear relationships. Potential for variable interactions and nesting may also be probed within these preliminary trials.

1.4 Aims of the study

The aim of the study was to develop and evaluate a generalised PTF approach, using GLMM and REML, to predict EC and OC using soil properties from the VSIS database (that may be incomplete or have a reduced set of covariates). The intention was to provide a more explanatory approach to machine learning that reveals more information on relationships between covariates, with interactions, together with identification of significant predictors. The investigation involved ranking models that were generated with progressively reduced sets of variables to satisfy different requirements for interpolation which may be constrained by cost, predictor availability, computational efficiency and uncertainty specification. The database used contains soil measurements and descriptions collected using different standards and is representative of a diverse set of landscapes, climates and soil types for the state of Victoria in south-eastern Australia.

2. Materials and Methods

2.1 Soil Data and Software

The source of the dataset was the VSIS, an initiative of the State Government of Victoria (Hunter et al., 2010). This database contains soil measurements and information collected across Victoria. Soil data from 93 projects in the database (between 1948 and 2014) were used in this study and, within each project, the number of sites available varied between 1 and 1368. At each site, samples were recorded at depth intervals representing the nature of the observed profile (e.g. pedological characterisation, pre-defined intervals for agronomic assessment or monitoring purposes). In total, there were $n = 27,669$ samples, representing 8746 sites: see Fig. 1), with each sample consisting of multiple soil measurements and properties (see variable list below).

Not all samples contained measured values for all soil properties. Samples were excluded from analysis if there were missing values in potential soil explanatory attributes such as Texture group, Control section (a contiguous model classification of soil horizon and depth) and Soil order. Both texture and

horizon (classified into control section for this study) are considered required minimum fields in soil site assessment at ‘level B’ for land resource survey (see Table 17.7 in McKenzie et al., 2008). Soil electrical conductivity is measured nearly twice as frequently as soil organic carbon content (Searle, 2014), although still considerably less than morphological properties. Other available environmental variables included were GMU, mean annual rainfall and elevation. This reduced the dataset available by 9.2% to $n = 25,119$.

Fig. 1. Study location (Victoria, Australia) with soil sites and Tier 1 Geomorphological Mapping Units.

In the EC dataset, only samples with an EC observation were included in the analysis. Where $EC > 10$ or equal to 0, these were also excluded from this analysis using quality-coding rules (Biggs and Searle, 2017). This resulted in 12,068 samples (48%) remaining in the dataset. Since particle size distribution (PSD: %clay, %silt, and %sand) would be used as a predictor, we further restricted analysis to rows where PSD data is present, this left 5158 samples (20.5%).

There were 2591 samples in the dataset with OC measurements. Of these, only 1954 samples (7.8%) also have records for the PSD, which was required to formulate the pedotransfer function.

265 Analysis was carried out in the *R* language for statistical calculation, data manipulation and graphical
266 display (Pinheiro et al., 2013; *R* Core Team, 2014).

267 2.2 Explanatory variables

268 The following predictants (EC and OC) and explanatory variable include both continuous and
269 categorical data (which are listed below):

- 270 • EC – soil electrical conductivity.
- 271 • OC – soil organic carbon content (Walkley-Black corrected).
- 272 • Control section – horizon and soil depth.
- 273 • Elevation (mASL) – elevation above sea level (assigned using spatial query).
- 274 • GMU – Geomorphological Mapping Unit.
- 275 • Particle size distribution (PSD), %clay, %silt, and %sand content.
- 276 • pH – laboratory pH measurement.
- 277 • Project – the program of field sampling that a site was collected for.
- 278 • Rainfall – mean annual rainfall 1960-1989.
- 279 • Soil Order – Australian Soil Classification (Order): Isbell and National Committee on Soil
280 and Terrain (2016).
- 281 • Texture Group – soil texture group.

282 The EC of a 1:5 soil/water suspension is based on Method 3A1 in Rayment and Lyons (2011). This
283 method is widely used across Australasia with most currently operating conductivity meters
284 accounting for solution temperature differences. Legacy measurements may be prone to lower
285 measurement precision than more recent measurements.

286 The OC content was determined using the wet oxidation method of Walkley and Black (Method 6A1
287 in Rayment and Lyons, 2011) as this method was applied consistently between projects in the dataset.
288 OC values vary from almost zero in silica rich sand to over 60% in some peaty soils.

289 Soil horizons were classified using a 5-layer contiguous model where soil layers were discriminated
290 according to major functions (e.g. water and gas movement, nutrient supply, physical behaviour and
291 plant growth). The layers (otherwise known as control sections) were assigned to manage the diversity
292 and variability of soil profile sequences for modelling and mapping purposes. These ‘control sections’
293 were developed based upon concepts of the ASRIS technical specifications of McKenzie et al. (2005).
294 Generally, Layer 1 = A1 and O horizon, Layer 2 = A2 and A3 horizon; Layer 3 = B1 and B21
295 horizon; Layer 4 = B22, B23 and B3 horizon, and Layer 5 = C and D horizon.

296 Terrain relief is represented by Elevation (mASL) sourced from the Vicmap elevation DTM20 dataset
297 (<https://www2.delwp.vic.gov.au/>).

298 Landforms and landscapes of Victoria have been mapped using a hierarchy known as the Victorian
299 Geomorphology Framework (Rees et al., 2010). The broadest geomorphological grouping of Victoria
300 (Tier 1) signify the coarse aggregations of similar landscapes and grouping of similar soil forming
301 processes. Tier 1 consists of 8 Divisions including the Eastern Uplands, Western Uplands, Southern
302 Uplands, Riverine Plains, North Western Dunefields and Plains, Western Plains, Eastern Plains and
303 Coast (Fig. 1).

304 Particle size was determined either using the Piper (1944) pipette or the Mikhail and Briner (1978)
305 plummet balance method. The relative composition of sand, silt and clay of the mineral soil which are
306 less than 2mm in diameter is represented in the ternary plot by control section (Fig. 2). The size
307 fractions used are consistent with those of the International soil texture classification (see Marshall,
308 1947).

309 Soil pH measurements in a 1:5 soil/water solution (Method 4A1; Rayment and Lyons, 2011) were
310 included as an explanatory soil property.

311 The Project for which the site and sample was collected (between 1948 and 2014) were included.

312 The mean annual rainfall (mm) between 1960 and 1989 was used as an environmental input variable.

There are 14 Soil Orders (Table 1) in the Australian Soil Classification (ASC) and all are represented in the dataset.

Six classes of soil texture (where diameter < 2mm) as described by Northcote (1979), i.e. Sands, Sandy Loams, Loams, Clay Loams, Light Clays and Medium to Heavy Clays.

Table 1. Soil Orders of the Australian Soil Classification and key features: see Isbell and National Committee on Soil and Terrain (2016) for further details.

Soil Order	Abbreviation	Features
Anthroposols	AN	man-made
Vertosols	VE	clay soils
Hydrosols	HY	saturated for prolonged periods
Tenosols	TE	soils with weak to negligible pedological organisation
Rudosols	RU	
Organosols	OR	organic soil material
Podosols	PO	organic matter, aluminium and/or iron accumulation
Kurosols	KU	texture contrast soils with strongly acidic subsoils
Chromosols	CH	texture contrast soils, not strongly acidic subsoil
Sodosols	SO	texture contrast soils, not strongly acidic, but sodic subsoil
Ferrosols	FE	non-texture contrast soils, high free iron
Dermosols	DE	non-texture contrast soils, well structured
Kandosols	KA	non-texture contrast soils, weakly or non-structured
Calcarosols	CA	usually calcareous throughout

2.3 Theoretical framework

In this study, the machine learning approach used is the GLMM with parameter estimates by the method REML, which is described by Galwey (2014). The approach taken produces unbiased estimators of mean and variance in a process that requires integration over random effects (Cnaan et al., 1997; Stroup, 2012; Galwey, 2014). The term ‘mixed’ refers to both fixed and random effects in the model. The REML algorithm is used for estimates of variance for the random-effect terms, and then for the fixed-effect terms, and requires iterative application of a non-linear algorithm, the so-called AI (Average Information) method, until convergence is reached.

An initial set of potential predictors (independent variables) was selected from the database from prior knowledge and field experience, including consultations with subject matter experts. Exploratory data

analysis was carried out to identify useful predictors by objective methods, such as statistical correlation and plotting database variables as pairs to identify possible relationships.

The GLMM approach includes interactions between predictors, noting that the random effects in this model were used to account for the nested effect of sites within project sample subsets. The general form of the model is described below in matrix notation,

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\epsilon} \quad (3)$$

where $\mathbf{y}$ is a vector of observations; $\boldsymbol{\beta}$ is an unknown vector of fixed effects; $\mathbf{u}$ is a vector of random effects; X and Z are matrices relating $\mathbf{y}$ to $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively; $\boldsymbol{\epsilon}$ is a vector of random errors with mean zero. In order to remove the unwanted variation due to different sites within a project, a random intercept model was first fitted with the random intercept accounting for the nested effect of sites within projects (Galwey, 2014). The residual from this model was plotted against various potential predictors for exploratory analysis. In the next step, all sub-models with fewer predictors were nested within the initial full model with the full set of predictors. The rule imposed when forming the nested models was that any interaction between predictors can only be included in the model if their main effects are in the model.

2.3.1 Model prediction error

Prediction error of a model can be assessed using the two standard error metrics, the mean-squared-prediction-error (MSPE) and (ii) mean-absolute-percentage-error (MAPE), described in Eqns (4) and (5), and a *k-fold cross-validation* scheme (where $k = 5$). In this method, the dataset is partitioned into *five* subsets, and *four* sets of data are used for training and one set for testing (i.e. 20% holdout for testing). This method is repeated *five* times with all *five* subsets used in the validation of the model. All the data were used in the initial exploratory experiments to identify potential pair-wise relationships between covariates, to aid in model development, and therefore there exists a degree of epistemic uncertainty due to sampling (Robinson et al., 2015; Benke et al., 2018).

Fig. 2. Soil texture triangle plots of the five control sections.

The prediction error was calculated using the standard metrics, i.e. MSPE and MAPE, as follows:

$$MSPE = \sum_{i=1}^n \left[\frac{(g(x_i) - \hat{g}(x_i))^2}{n} \right] \quad (4)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{g(x_i) - \hat{g}(x_i)}{g(x_i)} \right| \quad (5)$$

where $\hat{g}$ is the predicted value and g is the actual value. Note that Eqn (4) is also known as the mathematical expectation value of the squared residuals (differences between predictions and observations). The "P" in MSPE refers explicitly to prediction error in a k -fold cross-validation experiment. These metrics are fundamental as, for example, in the case of MSPE, a value of zero implies zero scatter and optimum fit between predictions and observations. It is also a measure of statistical

uncertainty associated with the model. Note that Eqn (4) is sometimes referred to as the mean-squared-error (MSE) of predictions and its square root is the *root-mean-squared-error* (RMSE), and both can be sourced or computed from MSPE results for each ranked model, k . Note that Eqn (5) is also referred to as the mean-average-percentage error (MAPE).

2.3.2 Predictor importance

Assessment of the importance of each predictor in the model is less straightforward than model validation. A metric of importance should favour predictors that appear more frequently in the reduced models. One way of ranking predictors is to inspect the best 100 models and see how often each predictor appears as a percentage. Alternatively, instead of inspection of an arbitrary number of models, one may observe the frequency that a predictor appears in the best X models, where X runs from 1 to N , and N is the number of models considered. When $X = 1$, the best model is considered, thus it is likely (but not necessarily) that all predictors will be present in the model.

2.3.3 Model for electrical conductivity (EC)

Prior experience from subject matter experts suggested that potential predictors for *EC* are particle size distribution (*PSD*), as described in Fig. 2, together with *GMU*, *Texture group*, *Soil order*, *Rainfall*, and *Control section*, as depicted in Eqn (6). Initial exploration involving plotting the random intercept model against various predictors suggests that one should also include in the model *pH* and *elevation*. In the first panel of Fig. 3 (left), as *pH* increases, so does the $\log(\text{EC})$, but the rate of increase differs between the five control sections (representing soil depth). The second (middle) panel shows that, as $\log(\text{Rainfall})$ increases, the $\log(\text{EC})$ decreases and the rate of decrease differs between control sections. The third (right) panel is a plot of the $\log(\text{EC})$ versus $\log(\text{Elevation})$ – with similar trends to $\log(\text{Rainfall})$. Overall, these plots show that control section is an important variable when modelling *EC*, because of its interaction with other predictors such as *pH*, rainfall and elevation. There was uncertainty about the

interaction between *Control section* and *pH*, *Elevation* and *Rainfall*, but they were added for completeness.

The initial additive linear mixed-effects model for EC was based on the following relationship, which contains continuous variables and categorical groups with interactions:

$$\begin{aligned} \log(EC) = & \text{alr}(PSD) + GMU + \text{Texture group} + \text{Soil order} \\ & + (pH + \text{Rainfall} + \text{Elevation}) \times \text{Control section} + (\text{Project/site}) \end{aligned} \quad (6)$$

In the above expression, ' + ' indicates addition of predictors, ' $\times$ ' indicates both the main and first order interaction of the associated predictors, *alr* indicates the additive-log-transformation, and *log* indicates natural-logarithmic transformation. The term *Project/site* is the random effect of site ID nested within Project. The expression *alr* is an R function developed by Van den Boogaart and Tolosana-Delgado (2013), based on Aitchison (1986), to perform additive log ratio transformation on the PSD compositional data, i.e. on the proportion of sand, silt and clay in the soil sample analysed (Fig. 2). They are equal to the *alr*(PSD)1 and *alr*(PSD)2 representing the 1st and 2nd component of the transformed PSD. They are equal to $\log\left(\frac{\%clay}{\%sand}\right)$ and $\log\left(\frac{\%silt}{\%sand}\right)$ respectively. The coefficients for this conceptual model can be calculated using records from the VSIS database and are reported later in the Results section.

2.2.4 Model for Organic Carbon Content (OC)

The model development for OC follows that described in the last section for EC. Subject matter experts and recent findings from Robertson et al. (2016) suggested that *PSD*, *GMU*, *Texture group*, *Soil order*, *Rainfall*, *Elevation*, *pH* and *Control section* were all useful for predicting *OC* (see Section 3.2 for further details). The mixed-effects model for OC is represented as follows.

$$\begin{aligned} \log(OC) = & \text{alr}(PSD) + GMU + \text{Texture group} + \text{Soil order} \\ & + (pH + \text{Rainfall} + \text{elevation}) \times \text{Control section} + (\text{Project/site}) \end{aligned} \quad (7)$$

The variables and model structure definitions follow those described in section 2.2.3.

Fig. 3. The plot of $\log(\text{EC})$ versus pH, $\log(\text{Rainfall})$, and $\log(\text{Elevation})$. The different lines (1-5) represent data from the five control sections.

3. Results

The results are documented in this section for electrical conductivity and soil organic carbon content, including the summary statistics for the predictant and explanatory variables (see Table 2).

Table 2. Summary Statistics for dataset used.

	Samples (n)	Min.	1 st Quartile	Median	Mean	3 rd Quartile	Max.
EC (dS/m ⁻¹)	12818	0.000	0.060	0.110	0.309	0.270	18.300
OC (%)	2708	0.000	0.570	1.300	1.956	2.700	17.000
pH _w	13132	3.800	5.500	6.419	6.754	8.081	11.180
Clay (%)	5970	0.05	15.50	28.50	32.21	47.00	93.50
Silt (%)	5969	0.05	7.00	14.00	15.55	22.00	60.00
Sand (%)	5971	1.90	31.00	47.00	48.19	64.00	100.00
Elevation (m)	27668	0.00	94.32	146.50	208.30	252.20	1781.00
Rainfall (mm)	27668	268.2	471.8	630.3	656.4	793.7	2336.0

GMU (tier 1)	1	2	3	4	5	6	7	8
#. Obs	3180	5783	1052	2344	5828	5795	1107	30

GMU (tier 1) units: 1. Eastern Uplands; 2. Western Uplands; 3. Southern Uplands; 4. Northern Riverine Plains; 5. North Western Dunefields and Plains; 6. Western Plains; 7. Eastern Plains; 8. Coast.

Texture group	Sands	Sandy loams	Loams	Clay loam	Light clays	Medium-heavy clays
#. Obs	1275	3450	3058	4198	5175	7963

Soil Order [#]	AN	CA	CH	DE	FE	HY	KA
	6	1273	6005	3726	385	165	1027
	KU	OR	PO	RU	SO	TE	VE
	1025	10	504	126	7070	1016	2778

[#] Soil order abbreviations from Table 1.

3.1 Electrical conductivity (EC)

The natural logarithmic transformation was applied to the EC value (Table 3). For Eqn (6), the coefficient estimates and standard errors for explanatory variables (including interaction terms of control section with pH, rainfall and elevation) are shown in Table 4. After considering all possible (main and nested) models and interactions of explanatory variables, there were 560 models fitted. This large suite of models (based on possible combinations of explanatory variables and interactions to be included in the sub-models) allows prediction of EC with the best model for the suite of available explanatory variables. For example, the first ranked model, which was derived on 5158 samples, uses all 11 explanatory variables with a MSPE = 0.686. In contrast, the 35th ranked model with a MSPE = 0.710 excludes PSD as an explanatory variable and can be applied to a significantly larger set of samples (n=26,914). This would enable all combinations of explanatory variables (including those without soil variables) to predict EC with a known prediction error and is symptomatic of many soil site databases across the globe that are populated with soil measurements to varying degrees.

Table 3. Summary statistics for samples used in the log EC and log OC PTF models.

	Min.	1 st Quartile	Median	Mean	3 rd Quartile	Max.
Log(EC)	-12.300	-2.813	-2.207	-2.165	-1.309	2.286
Log(OC)	-7.601	-4.962	-4.186	-4.287	-3.478	-1.778

In the soil texture triangles (Fig. 2), we depict ratios/proportions of sand, silt and clay (PSD) for each control section. The distribution of sand, silt and clay differs between all five control sections with sand dominated samples in the surface horizons (control sections 1 and 2) and samples with increasing clay content at lower depths (control sections 3 to 5). EC (and OC) demonstrate trends (EC increasing and OC decreasing) with control section (data not presented) that suggest observed differences in PSD with depth (represented as control section) is likely to be a useful explanatory factor in the prediction of EC, OC and other soil properties.

447 **Table 4.** Coefficient estimates and standard error of log(EC) prediction model (n=5158).

	Estimate	Standard Error
(Intercept)	-2.26	0.21
alr(psd)1	0.20	0.01
alr(psd)2	0.03	0.01
GMU(tier1)=2	0.24	0.07
GMU(tier1)=3	0.40	0.10
GMU(tier1)=4	0.24	0.08
GMU(tier1)=5	0.48	0.10
GMU(tier1)=6	0.57	0.08
GMU(tier1)=7	0.54	0.11
Texture Group=Light clays	-0.04	0.03
Texture Group=Loams	-0.04	0.03
Texture Group=Medium-heavy clays	0.04	0.04
Texture Group=Sands	-0.32	0.06
Texture Group=Sandy loams	-0.14	0.04
Soil Order=CH	-0.28	0.09
Soil Order=DE	-0.19	0.09
Soil Order=FE	-0.33	0.13
Soil Order=HY	0.05	0.16
Soil Order=KA	-0.22	0.11
Soil Order=KU	-0.27	0.10
Soil Order=PO	-0.16	0.13
Soil Order=RU	-0.15	0.19
Soil Order=SO	0.08	0.08
Soil Order=TE	-0.19	0.13
Soil Order=VE	0.15	0.09
pH	0.02	0.02
rain	0.01	0.15
elevation	-0.30	0.17
control2	-0.98	0.25
control3	-0.85	0.21
control4	-0.77	0.25
control5	-0.89	0.56
pH:control2	0.06	0.03
pH:control3	0.15	0.02
pH:control4	0.22	0.02
pH:control5	0.25	0.05
rain:control2	0.07	0.15
rain:control3	-0.57	0.13
rain:control4	-0.88	0.15
rain:control5	-1.03	0.35
elevation:control2	0.23	0.15
elevation:control3	0.18	0.13
elevation:control4	-0.08	0.18
elevation:control5	0.92	0.44

448

449 In the case of *prediction accuracy*, Table 5 shows the 5-fold pseudo-cross-validated results for the best
 450 and worst 50 models using the metrics MSPE and MAPE and ranked by the results for MSPE. The best

model ranked has a MSPE = 0.686 and the worst ranked model has a MSPE = 1.272. In the case of MAPE, the best and worst results were 0.635 and 0.911, respectively, showing less spread or resolution and therefore the MSPE was the preferred metric in this experiment. The plot of observed vs predicted values of the full/initial/best model is shown in Fig. 4. This is the optimal model if all the predictors are available. The black line is the 1:1 line and it is evident that the trend is linear.

In the case of *predictor importance*, the running frequency of each predictor appearing in the models is shown in Fig. 5, where the running frequency is a measure of how often a particular term appears in the fitted model; the higher the running frequency the more important the variable is as a predictor of the relevant response variable - in this case it is log(EC). One can see that control section appears to be the most important variable followed by pH, alr (PSD) component 1 and 2, GMU and Soil Order.

3.2 Organic carbon content (OC)

The variables included in the OC model are selected by analogy with the case presented for the EC model. Fig. 6 shows the data from five different control sections of log(OC) versus pH, log(Rainfall), and log(Elevation) respectively. The different lines represent data from five different levels of control. In the first panel (left), as pH increases, log(OC) decreases, but the rate of decrease differs between the five control sections. The second (middle) panel shows that, as log(Rainfall) increases, log(OC) increases and the rate of increase differs between control sections. The third (right) panel is a plot of log(OC) versus log(Elevation), showing similar trends to log(Rainfall). Overall, these plots show that control section is an important variable when modelling OC, because of its interaction with other predictors like pH, rainfall and elevation (control section relates to soil layer and therefore soil depth, while OC generally declines with soil depth).

The natural logarithm of OC was used in modelling due to the large spread in values. The summary statistics corresponding to $n = 1954$ samples are presented in Table 3. Estimates for the required coefficients and standard error for each variable (including interaction terms of control section with pH, rainfall and elevation) in the relationship are shown in Table 6. After considering all possible nesting

models, there were 560 models generated to reflect this combination of explanatory variables and interactions.

In the case of *prediction accuracy*, Table 7 shows the 5-fold pseudo-cross-validated results for the best 50 and worst 50 models using the metrics MSPE and MAPE and ranked by the results for MSPE. Once again, MSPE is the best metric for ranking and shows a greater range between best and worst model in the ranking (0.413 and 0.904, ratio of 1:2). The observed *vs* predicted plot of the full/initial/best model is also presented (Fig. 7) with a linear trend evident.

In the case of *predictor importance*, the running frequency of each predictor appearing in the models is shown in Fig. 8. The four most important running frequencies are labelled. Control section is by far the most important predictor, which is followed by GMU, the $\ln(\text{PSD})$ and pH.

Fig. 9 and Fig. 10 show data extracted from tables 5 and 7 for EC and OC, respectively. The model rankings are plotted from 1, ..., n, according to the error criteria MSPE and MAPE (showing the best 50 models and worst 50 models for illustration purposes). The small change in model prediction error metrics for the best 50 models indicates that step-wise reduction in covariates is possible without significant decrease in performance.

In summary, the approach for estimating EC and OC is described by Equations 6 and 7, augmented by ranking the predictor importance by frequency of occurrence. Model coefficients and their uncertainties (standard errors) are depicted in Tables 4 and 6, respectively. The model prediction accuracies (MSPE and MAPE) are also model uncertainties, because they are estimates of error, and are depicted in Tables 5 and 7. The latter two tables also list the predictors included or excluded in the ranked models.

497 **Table 5.** Cross-validated MSPE and MAPE of the best 50 and worst 50 models using the GLMM and
498 REML for predicting log(EC), ranked by order of MSPE. ‘YES’ indicates that variable coefficients are
499 in the model, whereas ‘—’ indicates otherwise. The leftmost column relates to the model ID.

Ran k	alr(psd)	GMU _(tier1)	Texture Group	Soil Order	pH	Rain fall	Elevation	Control section	pH:Control section	Rainfall:Co ntrol section	Elevation:Control section	MSPE	MAPE
1	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	0.686	0.635
2	YES	YES	--	YES	YES	YES	YES	YES	YES	YES	YES	0.688	0.635
3	YES	YES	YES	YES	YES	YES	--	YES	YES	YES	--	0.688	0.635
4	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	--	0.688	0.635
5	YES	YES	YES	YES	YES	--	YES	YES	YES	--	YES	0.688	0.638
6	YES	YES	YES	YES	YES	YES	YES	YES	YES	--	YES	0.689	0.636
7	YES	YES	--	YES	YES	--	YES	YES	YES	--	YES	0.689	0.638
8	YES	YES	--	YES	YES	YES	--	YES	YES	YES	--	0.690	0.635
9	YES	YES	--	YES	YES	YES	YES	YES	YES	--	YES	0.690	0.637
10	YES	YES	--	YES	YES	YES	YES	YES	YES	YES	--	0.690	0.636
11	YES	YES	YES	YES	YES	--	--	YES	YES	--	--	0.691	0.639
12	YES	YES	YES	YES	YES	--	YES	YES	YES	--	--	0.691	0.639
13	YES	YES	YES	YES	YES	YES	--	YES	YES	--	--	0.692	0.638
14	YES	YES	YES	YES	YES	YES	YES	YES	YES	--	--	0.692	0.638
15	YES	YES	YES	YES	YES	YES	YES	YES	--	YES	YES	0.692	0.636
16	YES	YES	--	YES	YES	--	--	YES	YES	--	--	0.692	0.640
17	YES	YES	--	YES	YES	--	YES	YES	YES	--	--	0.692	0.639
18	YES	YES	--	YES	YES	YES	--	YES	YES	--	--	0.693	0.638
19	YES	YES	--	YES	YES	YES	YES	YES	YES	--	--	0.693	0.638
20	YES	YES	YES	YES	YES	YES	YES	YES	--	YES	--	0.693	0.637
21	YES	YES	YES	YES	YES	YES	--	YES	--	YES	--	0.693	0.636
22	YES	YES	--	YES	YES	YES	YES	YES	--	YES	YES	0.694	0.637
23	YES	YES	--	YES	YES	YES	--	YES	--	YES	--	0.696	0.638
24	YES	YES	--	YES	YES	YES	YES	YES	--	YES	--	0.696	0.638
25	YES	YES	YES	YES	YES	--	YES	YES	--	--	YES	0.703	0.645
26	YES	YES	YES	YES	YES	YES	YES	YES	--	--	YES	0.704	0.644
27	YES	YES	--	YES	YES	--	YES	YES	--	--	YES	0.705	0.646
28	YES	YES	YES	YES	--	YES	YES	YES	--	YES	YES	0.706	0.651
29	YES	YES	--	YES	YES	YES	YES	YES	--	--	YES	0.706	0.645
30	YES	YES	YES	YES	--	YES	--	YES	--	YES	--	0.707	0.651
31	YES	YES	YES	YES	--	YES	YES	YES	--	YES	--	0.707	0.651
32	YES	YES	YES	--	YES	YES	YES	YES	YES	YES	YES	0.709	0.647
33	YES	YES	--	YES	--	YES	YES	YES	--	YES	YES	0.709	0.653
34	YES	YES	YES	--	YES	--	YES	YES	YES	--	YES	0.710	0.651
35	--	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	0.710	0.645
36	YES	YES	YES	--	YES	YES	YES	YES	YES	--	YES	0.710	0.648
37	YES	YES	YES	--	YES	YES	--	YES	YES	YES	--	0.711	0.647
38	YES	YES	YES	--	YES	YES	YES	YES	YES	YES	--	0.711	0.648
39	YES	YES	--	YES	--	YES	--	YES	--	YES	--	0.711	0.653
40	YES	YES	--	YES	--	YES	YES	YES	--	YES	--	0.711	0.654
41	--	YES	YES	--	YES	--	YES	YES	YES	--	YES	0.712	0.647
42	YES	YES	--	--	YES	YES	YES	YES	YES	YES	YES	0.712	0.648
43	--	YES	YES	YES	YES	YES	--	YES	YES	YES	--	0.712	0.646
44	--	YES	YES	YES	YES	YES	YES	YES	YES	YES	--	0.712	0.646
45	YES	YES	YES	YES	YES	--	--	YES	--	--	--	0.712	0.650
46	YES	YES	--	--	YES	--	YES	YES	YES	--	YES	0.712	0.652
47	YES	YES	--	--	YES	YES	YES	YES	YES	--	YES	0.713	0.649
48	YES	YES	YES	YES	YES	--	YES	YES	--	--	--	0.713	0.650
49	--	YES	YES	YES	YES	YES	YES	YES	YES	--	YES	0.713	0.646
50	YES	YES	YES	--	YES	YES	--	YES	YES	--	--	0.713	0.650
511	--	--	--	--	--	YES	YES	YES	--	YES	--	0.896	0.740
512	--	YES	--	--	YES	YES	YES	--	--	--	--	0.896	0.720
513	YES	--	YES	--	--	YES	--	--	--	--	--	0.901	0.753
514	YES	--	YES	YES	--	--	--	--	--	--	--	0.902	0.759
515	--	--	--	--	--	YES	--	YES	--	YES	--	0.906	0.746
516	--	YES	--	--	--	--	YES	YES	--	--	--	0.906	0.749
517	YES	--	--	--	--	--	YES	YES	--	--	--	0.910	0.765
518	--	--	YES	--	--	--	YES	YES	--	--	--	0.911	0.762
519	--	YES	--	--	--	--	--	YES	--	--	--	0.911	0.752
520	--	--	--	--	--	YES	YES	YES	--	--	YES	0.913	0.749
521	YES	--	--	--	--	YES	YES	--	--	--	--	0.913	0.757
522	YES	--	--	YES	--	--	--	--	--	--	--	0.916	0.765
523	YES	--	--	--	--	YES	--	--	--	--	--	0.919	0.761
524	--	--	--	--	YES	--	YES	--	--	--	--	0.929	0.736
525	--	--	YES	--	--	YES	YES	--	--	--	--	0.930	0.762
526	--	--	YES	YES	--	--	--	--	--	--	--	0.930	0.769
527	--	--	--	--	YES	YES	YES	--	--	--	--	0.934	0.737
528	--	YES	--	YES	--	YES	--	--	--	--	--	0.936	0.757
529	--	YES	--	YES	--	YES	YES	--	--	--	--	0.936	0.757
530	--	--	YES	--	--	YES	--	--	--	--	--	0.936	0.766
531	--	--	--	--	YES	--	--	--	--	--	--	0.939	0.743
532	--	--	--	--	YES	YES	--	--	--	--	--	0.941	0.742
533	--	YES	--	YES	--	--	YES	--	--	--	--	0.943	0.762
534	--	--	--	--	--	YES	YES	YES	--	--	--	0.943	0.761
535	--	YES	--	YES	--	--	--	--	--	--	--	0.946	0.763
536	--	--	--	--	--	YES	--	YES	--	--	--	0.948	0.764
537	YES	--	YES	--	--	--	YES	--	--	--	--	0.955	0.786
538	YES	--	YES	--	--	--	--	YES	--	--	--	0.956	0.790
539	--	--	--	--	--	--	YES	YES	--	--	YES	0.965	0.786
540	--	--	YES	--	--	--	--	YES	--	--	--	0.970	0.794
541	YES	--	--	--	--	--	--	YES	--	--	--	0.977	0.800
542	YES	--	--	--	--	--	YES	--	--	--	--	0.979	0.796

543	--	--	YES	--	--	--	YES	--	--	--	--	0.984	0.796
544	--	--	--	YES	--	YES	YES	--	--	--	--	0.986	0.781
545	--	--	--	YES	--	YES	--	--	--	--	--	0.993	0.788
546	--	--	--	--	--	--	YES	YES	--	--	--	0.998	0.798
547	--	--	--	YES	--	--	YES	--	--	--	--	1.006	0.793
548	YES	--	YES	--	--	--	--	--	--	--	--	1.029	0.822
549	--	YES	--	--	--	YES	--	--	--	--	--	1.044	0.799
550	--	YES	--	--	--	YES	YES	--	--	--	--	1.044	0.799
551	--	--	--	YES	--	--	--	--	--	--	--	1.047	0.815
552	--	--	YES	--	--	--	--	--	--	--	--	1.055	0.830
553	YES	--	--	--	--	--	--	--	--	--	--	1.058	0.833
554	--	--	--	--	--	--	--	YES	--	--	--	1.060	0.830
555	--	YES	--	--	--	--	YES	--	--	--	--	1.065	0.812
556	--	YES	--	--	--	--	--	--	--	--	--	1.074	0.816
557	--	--	--	--	--	YES	YES	--	--	--	--	1.107	0.831
558	--	--	--	--	--	YES	--	--	--	--	--	1.114	0.836
559	--	--	--	--	--	--	YES	--	--	--	--	1.185	0.871
560	--	--	--	--	--	--	--	--	--	--	--	1.272	0.911
Rank	alr(psd)	GMU _(tier1)	Texture Group	Soil Order	pH	Rain fall	Elevation	Control section	pH:Control section	Rainfall:Control section	Elevation:Control section	MSPE	MAPE

Fig. 4. Observed log(EC) vs Predicted log(EC) of the full initial model with all predictors (where EC = dS/m).

Fig. 5. Running frequency of the predictor in the model for $\log(\text{EC})$.

Fig. 6. The plot of $\log(\text{OC})$ versus pH , $\log(\text{Rainfall})$ and $\log(\text{Elevation})$. The different lines (1-5) represent data from the five control sections.

508 **Table 6.** Coefficient estimates and standard errors of the log(OC) model (n=1954).

	Estimate	Standard Error
(Intercept)	-2.32	0.24
alr(psd)1	0.11	0.02
alr(psd)2	0.11	0.02
GMU (tier 1)=2	0.05	0.08
GMU (tier 1)=3	0.29	0.12
GMU (tier 1)=4	-0.28	0.08
GMU (tier 1)=5	-0.22	0.10
GMU (tier 1)=6	0.33	0.10
GMU (tier 1)=7	0.40	0.11
Texture Group=Light Clays	-0.18	0.06
Texture Group=Loams	0.16	0.04
Texture Group=Medium-Heavy Clays	-0.21	0.07
Texture Group=Sands	-0.27	0.12
Texture Group=Sandy Loams	0.05	0.05
Soil Order=CH	-0.40	0.09
Soil Order=DE	-0.28	0.09
Soil Order=FE	-0.19	0.13
Soil Order=HY	-0.26	0.17
Soil Order=KA	-0.44	0.11
Soil Order=KU	-0.33	0.11
Soil Order=PO	-0.33	0.16
Soil Order=RU	-0.19	0.18
Soil Order=SO	-0.32	0.08
Soil Order=TE	-0.38	0.13
Soil Order=VE	-0.08	0.09
pH	-0.21	0.03
rain	0.25	0.15
elevation	0.53	0.16
control=2	-1.30	0.38
control=3	-1.00	0.34
control=4	-0.85	0.45
control=5	3.97	2.53
pH:control=2	0.05	0.05
pH:control=3	0.06	0.04
pH:control=4	-0.04	0.04
pH:control=5	-0.41	0.25
rain:control=2	-0.04	0.21
rain:control=3	-0.58	0.21
rain:control=4	-0.15	0.36
rain:control=5	-5.06	1.45
elevation:control=2	0.13	0.24
elevation:control=3	-0.08	0.22
elevation:control=4	-0.72	0.33
elevation:control=5	3.80	1.19

Table 7. Cross-validated MSPE and MAPE of the best 50 and worst 50 models using the GLMM and REML for predicting log(OC), ranked by MSPE. ‘YES’ indicates that variable coefficients are in the model, whereas ‘—’ indicates otherwise. The ranking corresponds also to the model ID.

Rank	alr(psd)	GMU _(tier1)	Texture Group	Soil Order	pH	Rain fall	Elevation	Control section	pH:Control section	Rainfall:Control section	Elevation:Control section	MSPE	MAPE
1	YES	YES	YES	YES	YES	YES	YES	YES	--	YES	YES	0.413	0.474
2	YES	YES	YES	YES	YES	YES	YES	YES	--	YES	--	0.414	0.474
3	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	0.416	0.475
4	YES	YES	YES	YES	YES	--	YES	YES	YES	--	--	0.416	0.477
5	YES	YES	YES	--	YES	YES	YES	YES	--	YES	YES	0.416	0.473
6	YES	YES	YES	YES	YES	YES	YES	YES	YES	--	--	0.416	0.477
7	YES	YES	YES	--	YES	YES	YES	YES	YES	YES	YES	0.416	0.473
8	YES	YES	YES	--	YES	YES	YES	YES	--	YES	--	0.417	0.473
9	YES	YES	YES	--	YES	--	YES	YES	YES	--	--	0.419	0.476
10	YES	YES	YES	--	YES	YES	YES	YES	YES	--	--	0.419	0.476
11	YES	YES	YES	YES	YES	YES	--	YES	--	YES	--	0.420	0.476
12	YES	YES	YES	YES	YES	YES	--	YES	YES	--	--	0.422	0.478
13	YES	YES	YES	YES	YES	--	YES	YES	--	--	--	0.423	0.483
14	YES	YES	YES	YES	YES	YES	YES	YES	--	--	--	0.423	0.482
15	YES	YES	--	YES	YES	YES	YES	YES	--	YES	YES	0.424	0.481
16	YES	YES	YES	--	YES	YES	--	YES	--	YES	--	0.424	0.475
17	YES	YES	--	YES	YES	--	YES	YES	YES	--	--	0.424	0.482
18	YES	YES	--	YES	YES	YES	YES	YES	YES	--	--	0.424	0.482
19	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	--	0.424	0.476
20	YES	YES	--	YES	YES	YES	YES	YES	--	YES	--	0.425	0.481
21	YES	YES	--	YES	YES	YES	YES	YES	YES	YES	YES	0.425	0.481
22	YES	YES	YES	--	YES	--	YES	YES	--	--	--	0.425	0.481
23	YES	YES	--	--	YES	YES	YES	YES	YES	YES	YES	0.425	0.480
24	YES	YES	--	--	YES	YES	YES	YES	--	YES	YES	0.425	0.480
25	YES	YES	YES	--	YES	YES	YES	YES	--	--	--	0.426	0.481
26	YES	YES	--	--	YES	YES	YES	YES	--	YES	--	0.426	0.480
27	YES	YES	YES	--	YES	YES	--	YES	YES	--	--	0.426	0.478
28	YES	YES	--	--	YES	--	YES	YES	YES	--	--	0.426	0.482
29	YES	YES	--	--	YES	YES	YES	YES	YES	--	--	0.426	0.481
30	YES	YES	YES	YES	YES	--	YES	YES	--	--	YES	0.427	0.481
31	YES	YES	YES	YES	YES	YES	YES	YES	--	--	YES	0.427	0.480
32	YES	YES	YES	YES	YES	YES	--	YES	YES	YES	--	0.429	0.477
33	YES	YES	YES	--	--	YES	YES	YES	--	YES	YES	0.429	0.482
34	YES	YES	YES	YES	YES	YES	--	YES	--	--	--	0.429	0.484
35	YES	YES	YES	YES	YES	--	YES	YES	YES	YES	YES	0.430	0.487
36	YES	YES	YES	--	--	YES	YES	YES	--	YES	--	0.430	0.482
37	YES	YES	--	YES	YES	YES	--	YES	YES	--	--	0.430	0.483
38	YES	YES	YES	--	YES	--	YES	YES	--	--	YES	0.430	0.479
39	YES	YES	YES	--	YES	YES	YES	YES	--	--	YES	0.431	0.478
40	YES	YES	--	YES	YES	YES	--	YES	--	YES	--	0.431	0.483
41	YES	YES	--	YES	YES	--	YES	YES	--	--	--	0.431	0.488
42	YES	YES	YES	--	YES	YES	YES	YES	YES	YES	--	0.432	0.475
43	YES	YES	YES	YES	--	YES	YES	YES	--	YES	--	0.432	0.487
44	YES	YES	--	YES	YES	YES	YES	YES	--	--	--	0.432	0.488
45	YES	YES	--	--	YES	--	YES	YES	--	--	--	0.433	0.487
46	YES	YES	YES	--	YES	YES	--	YES	--	--	--	0.433	0.483
47	YES	YES	--	--	YES	YES	YES	YES	--	--	--	0.433	0.487
48	YES	YES	YES	YES	YES	--	YES	YES	YES	--	YES	0.433	0.480
49	YES	YES	YES	YES	YES	YES	YES	YES	YES	--	YES	0.434	0.480
50	YES	YES	--	--	YES	YES	--	YES	YES	--	--	0.434	0.484
511	--	--	--	YES	YES	--	YES	--	--	--	--	0.780	0.694
512	--	YES	--	--	YES	YES	--	--	--	--	--	0.782	0.692
513	YES	--	YES	--	--	--	--	--	--	--	--	0.784	0.690
514	YES	--	--	--	YES	--	--	--	--	--	--	0.787	0.695
515	YES	--	YES	--	--	--	YES	--	--	--	--	0.792	0.694
516	YES	--	--	--	YES	YES	--	--	--	--	--	0.801	0.702
517	--	--	--	--	YES	--	--	--	--	--	--	0.806	0.706
518	--	--	YES	YES	--	--	YES	--	--	--	--	0.807	0.709
519	YES	--	--	--	YES	--	YES	--	--	--	--	0.809	0.707
520	YES	--	--	--	YES	YES	YES	--	--	--	--	0.810	0.707
521	YES	YES	--	YES	--	YES	--	--	--	--	--	0.813	0.699
522	YES	YES	--	YES	--	YES	YES	--	--	--	--	0.815	0.702
523	YES	YES	--	YES	--	--	YES	--	--	--	--	0.816	0.702
524	--	--	--	--	YES	YES	--	--	--	--	--	0.816	0.711
525	--	--	YES	YES	--	--	--	--	--	--	--	0.822	0.718
526	--	--	YES	--	--	YES	--	--	--	--	--	0.823	0.707
527	--	--	--	--	YES	YES	YES	--	--	--	--	0.825	0.718
528	--	--	--	--	YES	--	YES	--	--	--	--	0.826	0.718
529	YES	YES	--	--	--	YES	--	--	--	--	--	0.827	0.704
530	YES	YES	--	--	--	YES	YES	--	--	--	--	0.827	0.708
531	YES	YES	--	--	--	--	YES	--	--	--	--	0.827	0.708
532	YES	YES	--	YES	--	--	--	--	--	--	--	0.828	0.704
533	YES	YES	--	--	--	--	--	--	--	--	--	0.838	0.707
534	--	--	YES	--	--	YES	YES	--	--	--	--	0.840	0.717
535	YES	--	--	YES	--	YES	--	--	--	--	--	0.872	0.733
536	--	--	YES	--	--	--	YES	--	--	--	--	0.876	0.737
537	YES	--	--	--	--	YES	--	--	--	--	--	0.879	0.735
538	--	--	YES	--	--	--	--	--	--	--	--	0.890	0.745
539	--	YES	--	YES	--	YES	--	--	--	--	--	0.890	0.746
540	YES	--	--	YES	--	YES	YES	--	--	--	--	0.891	0.744
541	--	YES	--	YES	--	YES	YES	--	--	--	--	0.893	0.751
542	--	YES	--	--	--	YES	--	--	--	--	--	0.898	0.747

543	YES	--	--	--	--	YES	YES	--	--	--	--	0.899	0.746
544	--	YES	--	YES	--	--	YES	--	--	--	--	0.899	0.754
545	--	YES	--	--	--	YES	YES	--	--	--	--	0.900	0.752
546	--	YES	--	--	--	--	YES	--	--	--	--	0.904	0.754
547	YES	--	--	YES	--	--	--	--	--	--	--	0.921	0.758
548	--	YES	--	YES	--	--	--	--	--	--	--	0.922	0.761
549	YES	--	--	--	--	--	--	--	--	--	--	0.923	0.756
550	YES	--	--	YES	--	--	YES	--	--	--	--	0.925	0.762
551	YES	--	--	--	--	--	YES	--	--	--	--	0.929	0.762
552	--	YES	--	--	--	--	--	--	--	--	--	0.930	0.759
553	--	--	--	YES	--	YES	--	--	--	--	--	0.975	0.794
554	--	--	--	--	--	YES	--	--	--	--	--	0.977	0.790
555	--	--	--	YES	--	YES	YES	--	--	--	--	0.995	0.805
556	--	--	--	--	--	YES	YES	--	--	--	--	1.000	0.803
557	--	--	--	YES	--	--	YES	--	--	--	--	1.054	0.834
558	--	--	--	--	--	--	YES	--	--	--	--	1.060	0.832
559	--	--	--	YES	--	--	--	--	--	--	--	1.074	0.842
560	--	--	--	--	--	--	--	--	--	--	--	1.093	0.846
Rank	alr(psd)	GMU _(tier1)	Texture Group	Soil Order	pH	Rain fall	Elevation	Control section	pH:Control section	Rainfall:Control section	Elevation:Control section	MSPE	MAPE

Fig. 7. Observed log(OC) vs Predicted log(OC) of the full initial model with all predictors (where OC = %).

Fig. 8. Running frequency of each predictor appearing in the model for log(OC).

Fig. 9. Ranking of EC models from 1, ..., n, according to the error criteria.

Fig. 10. Ranking of OC models from 1, ..., n, according to the error criteria.

4. Discussion

The PTF generation was based on using a machine learning approach incorporating GLMM and REML together with a predictor-frequency ranking system using a set of predictors consisting of continuous and categorical variables with interactions and nested terms. The two key soil properties of interest were electrical conductivity and soil organic carbon content. By a process of step-wise reduction in the number of predictors used (including many combinations and interaction terms), over 500 models were generated and tested by 5-fold cross-validation. The models were ranked in order by standard error metrics so that reduced versions could be evaluated and chosen by researchers and modellers in case trade-offs are needed in the number of variables, computational efficiency or acceptable error margins. The study revealed that the four main predictors for electrical conductivity and soil organic carbon content were control section, pH, particle size distribution and geomorphological mapping unit (Tier 1).

The approach has theoretical similarity with machine learning methods that use variants of GLMM and log-link functions derived from logistic models (Benke, 2019; Hengl et al., 2018; Liao et al., 2015; McNeill et al., 2018; Stein et al., 2019). An advantage of GLMM with REML is that it is not a predictive

‘black box’ model, but rather provides added explanatory behaviour by exposing linear or non-linear relationships between covariates and possible interactions, and therefore more information on the nature of the actual physical processes driving predictions.

The PTF models produced can be used to augment the information provided by a database (in this case, VSIS) by supporting the mapping of soil properties in a rapid and cost-effective manner by interpolation and extrapolation. Reduced versions of the full model can be used in other applications by analysts who have limited resources but need to satisfy constraints on the number of predictors relating to cost, computational efficiency and uncertainty specification. This issue is important because the quantity of available soil property data varies between projects and sites, both over time and spatially in the regions.

The results were very encouraging given that there were potential limitations relating to data quality, including the period of time covering data acquisition (from 1948), different measurement and field techniques used and different operators. There were also epistemic uncertainties, including factors such as depth resolution of the control section, data veracity due to data entry errors and measurement error (Biggs and Searle, 2017; Benke et al., 2018). The signals (predictors) identified in the presence of noisy data implies a significant level of robustness in the information produced.

Some strengths and weaknesses of machine learning approaches in general have been discussed in the literature (see also Castelvechi, 2016; Liao et al., 2015; McNeill et al., 2018; Nilsson, 2014; Russell and Norvig, 2016; Stein et al, 2019). Physical models have advantages aside from being more explanatory than non-parametric machine learning algorithms in that they provide a theoretical basis for understanding soil processes, they support planning new research strategies, and are more robust, reliable and less sensitive to noise and artefacts in limited datasets. However, a physical model may struggle to match the high prediction accuracy of a machine learning algorithm with many adjustable coefficients (such as occurs with ANN and RF approaches). Machine learning algorithms generally require a large set of weights and a very large set of training data, which is not always possible in the real world where only small datasets may be available.

In comparison to other machine learning techniques, the GLMM-based approach with REML has potential advantages and disadvantages (e.g. Castelvechi, 2016; Nilsson, 2014; Russell and Norvig, 2016). For example, a conventional ANN, such as the common back-propagation feed-forward network, is a classical ‘black box’ model, essentially suitable for prediction only (Castelvechi, 2016). In contrast, the GLMM with REML has the advantage of providing coefficients for associated model variables that are explicit and identifiable, including possible linear or non-linear relationships and interactions, and a nested hierarchical structure with both fixed and random effects that are associated with a formal experimental design. It therefore has elements of an explanatory model rather than just a predictive model.

Although a GLMM can also have additive terms that are non-linear, it does not have the potential degrees of freedom of a large-scale non-parametric ANN or RF model. The ANN is capable of more accurately fitting the training data, with sufficient nodes and layers, but there is the risk of overfitting and training on noise and therefore the ANN may not always generalise as well as a parametric regression model. It is noted that the use of partial dependence plots can enhance the explanation of the machine learning approaches, including the ANN approach. The relationship between the covariates in the model structure, however, is generally not known, or whether there are interactions between inputs, and this has resulted in criticisms that, for example, the ANN is a ‘black box’ prediction approach without an explanatory basis. Unlike the GLMM, the ANN is more vulnerable to errors in training and prediction due to noise, artefacts in the data, and correlations or interactions between inputs or predictors (Castelvechi, 2016). The ANN cannot be applied readily to small datasets because it requires large training sets to avoid overfitting of parameters.

There is clearly a motivation for further research into machine learning approaches that can combine high prediction accuracy with the explanatory nature of physical models. The GLMM with REML is an approach that lies within the spectrum between these extremes and is consistent with regression models that are additive non-linear, iterative, and require a learning algorithm for parameter estimation.

As a generalised approach to PTF generation, the GLMM can be applied to small or large datasets for interpolation and extrapolation in regional locations without the need to generate hundreds of diverse local PTF models. The GLMM approach provides weights for the model variables that are identifiable soil properties, whilst also addressing the effect of interactions between predictors in a nested hierarchical structure with fixed and random effects that are associated with a formal experimental design. The rigorous formal definition distinguishes it from other more heuristic non-parametric machine learning methods. It therefore has many elements of an explanatory model rather than only a predictive model.

The identification of the four main predictors produces a reduced set of variables that would be valuable predictors in proposed new experiments for comparing the mean-square-prediction-error (MSPE) from a machine learning algorithm with the MSPE of a human expert. This would provide more information on the potential effectiveness of automated computer-based methods. The authors suggest that the soil research community agree on an open-source benchmark dataset of soils, with documented properties or attributes, to use as controls to compare machine learning algorithms with respect to prediction accuracy.

5. Conclusion

There is growing interest in the application of artificial intelligence to analyse large sets of data to support studies of soil health and condition, the carbon cycle, and climate change research. An emerging theme in contemporary modelling has been the advent of machine learning algorithms for predictions that are data-driven without any explicit model definition. There now appears to be a modelling spectrum where, at one extreme, there are traditional physical models with explicit structures revealing relationships between covariates and are explanatory in nature, while at the other extreme are more opaque data-driven statistical approaches such as machine learning algorithms that behave more like the ‘black box’ of artificial intelligence described by Castelvechi (2016), but produce high prediction accuracy. Regulatory authorities and stakeholders in business are aware of the black box

problem and are increasingly interested in transparent or ‘explainable’ machine learning that reveals the logical basis and physical attributes that support the predictions by the algorithm.

The strengths and weaknesses of machine learning were discussed as well as the advantages and disadvantages of the approach taken here which applied GLMM and REML to the prediction of electrical conductivity and soil organic carbon content. It was concluded that the GLMM with REML has potential as a compromise between the two competing paradigms of physical models and statistical models. The results, when applied to the VSIS soil database, with 5-fold cross-validation, revealed that the four main predictors for electrical conductivity and soil organic carbon content were soil depth, pH, particle size distribution and geomorphological mapping unit (Tier 1). The prediction accuracy was similar to algorithms reported in a recent study on machine learning, using similar error metrics and 5-fold cross-validation, although in a different application, which compared neural networks, random forest and *K*-nearest neighbour approaches (Hengl et al., 2018).

A primary consideration of this study was to investigate a predictive algorithm that is more explanatory in nature than a conventional non-parametric machine learning algorithm, and to identify the most important covariates and interactions, rather than maximising prediction accuracy. Because of this consideration, an extensive comparison between machine learning algorithms based on prediction accuracy alone was not appropriate but a subject for future research in controlled experiments.

631 **References**

- 632 Aitchison, J. (1986). The statistical analysis of compositional data. Chapman & Hall, London, UK.
- 633 Allen, D.E., Pringle, M.J., Bray, S., Hall, T.J., O'Reagain, P.O., Phelps, D., Cobon, D.H., Bloesch,
634 P.M., Dalal, R.C., 2013. What determines soil organic carbon stocks in the grazing lands of north-
635 eastern Australia? *Soil Research* 51(8), 695–706. <https://doi.org/10.1071/SR13041>.
- 636 Baldock, J., McDonald, L., Sanderman, J., 2013. Foreword to 'Soil Carbon in Australia's Agricultural
637 Lands'. *Soil Research*, 51, i–ii.
- 638 Baldock, J.A., Skjemstad, J.O. 1999. Soil organic carbon/soil organic matter. In: Peverill, K.A.,
639 Sparrow, L.A., Reuter, D.J. (Eds.), *Soil Analysis: An Interpretation Manual*. CSIRO Publishing,
640 Melbourne, pp. 159–170.
- 641 Arrouays, D., Leenaars, J., G., Richer-de-Forges, A., C., Adhikari, K., Ballabio, C., Greve, M.,
642 Grundy, M., Guerrero, E., Hempel, J., Hengl, T., Heuvelink, G., Batjes, N., Carvalho, E., Hartemink,
643 A., Hewitt, A., Hong, S., Krasilnikov, P., Lagacherie, P., Lelyk, G., Libohova, Z., Lilly, A.,
644 McBratney, A., McKenzie, N., Vasquez, G., M., Mulder, V., L., Minasny, B., Montanarella, L., Odeh,
645 I., Padarian, J., Poggio, L., Roudier, P., Saby, N., Savin, I., Searle, R., Solbovoy, V., Thompson, J.,
646 Smith, S., Sulaeman, Y., Vintila, R., Viscarra Rossel, R., Wilson, P., Zhang, G., Swerts, M., Oorts,
647 K., Karklins, A., Feng, L., Navarro, A., R., Levin, A., Laktionova, T., Dell'Acqua, M., Suvannang,
648 N., Ruam, W., Prasad, J., Patil, N., Husnjak, S., Pásztor, L., Okx, J., Hallett, S., Keay, C., Farewell,
649 T., Lilja, H., Juilleret, J., Marx, S., Takata, Y., Kazuyuki, Y., Mansuy, N., Panagos, P., Liedekerke,
650 M., V., Skalsky, R., Sobocka, J., Kobza, J., Eftekhari, K., Alavipanah, S., K., Moussadek, R.,
651 Badraoui, M., Da Silva, M., Paterson, G., da Conceição Gonçalves, M., Theocharopoulos, S.,
652 Yemefack, M., Tedou, S., Vrscaj, B., Grob, U., Kozák, J., Boruvka, L., Dobos, E., Taboada, M.,
653 Moretti, L., Rodriguez, D., 2017. Soil legacy data rescue via GlobalSoilMap and other international
654 and national initiatives. *GeoResJ* 14, 1–19.

655 Baker, L., Ellison, D., 2008. Optimisation of pedotransfer functions using an artificial neural network
656 ensemble method. *Geoderma*, 144(1-2), 212–224.

657 Bell, M., J., Sirling, G., R., Pankhurst, C., E., 2007. Management impacts on health of soils
658 supporting Australian grain and sugarcane industries. *Soil and Tillage Research* 97 (2), 256–271.

659 Benke, K. K., Norng, S., Robinson, N. J., Benke, L. R., Peterson, T. J., 2018. Error propagation in
660 computer models: analytic approaches, advantages, disadvantages and constraints. *Stochastic*
661 *Environmental Research and Risk Assessment*, 32(10), 2971-2985.

662 Benke, K. K. 2019. Data Analytics and Machine Learning for Disease Identification in Electronic
663 Health Records. *JAMA Ophthalmology*, 137(5), 497-498.

664 Biggs, A.J.W., Searle, R. 2017. Towards improved quality of soil morphology and analytical data in
665 Australia: starting the discussion. *Soil Research*, 55, 309–317.

666 Bouma, J., 1989. Using soil survey data for quantitative land evaluation. *Advances in Soil Science*, 9,
667 177–213.

668 Castelvechi, D., 2016. Can we open the black box of AI? *Nature News*, 538, 20–23.

669 Chen, G., Li, S., Knibbs, L.D., Hamm, N.A.S., Cao, W., Li, T., Guo, Y., 2018. A machine learning
670 method to estimate PM 2.5 concentrations across China with remote sensing, meteorological and land
671 use information. *Science of the Total Environment*, 636, 52–60.

672 Cnaan, A., Laird, N.M., Slasor, P., 1997. Tutorial in biostatistics: using the general linear mixed
673 model to analyse unbalanced repeated measures and longitudinal data. *Stat Med*, 16(2349), 80.

674 Dahlhaus, P.G., MacEwan, R.J., 1997. Dryland Salinity in South West Victoria – Challenging The
675 Myth. In: McNally, G. (ed), *Collected Case Studies in Engineering Geology, Environmental Geology,*
676 *Hydrogeology, Third Series*. Geological Society of Australia Inc, pp. 165-181.

677 Dang, Y.P., Moody, P.W., 2016. Quantifying the costs of soil constraints to Australian agriculture: A
 678 case study of wheat in north-eastern Australia. *Soil Research*, 54, 700–707.
 679 <https://doi.org/10.1071/SR15007>.

680 Ford, G.W., Martin, J.J., Rengasamy, P., Boucher, S.C., Ellington, A. 1993. Soil Sodicity in Victoria.
 681 *Australian Journal of Soil Research*, 31, 869–909.

682 Galwey, N.W., 2014. Introduction to mixed modelling: beyond regression and analysis of variance.
 683 John Wiley & Sons, Chichester, United Kingdom.

684 Grunwald, S. (Ed.). 2016. Environmental soil-landscape modeling: Geographic information
 685 technologies and pedometrics. CRC Press, Boca Raton, USA.

686 Hengl T., Walsh M.G., Sanderman, J., Wheeler, I., Harrison, S.P., Prentice, I.C., 2018. Global mapping
 687 of potential natural vegetation: an assessment of machine learning algorithms for estimating land
 688 potential. *PeerJ* 6:e5457 <https://doi.org/10.7717/peerj.5457>.

689 Hunter, D., Williams, S., Robinson, N., 2010. VSIS an new system for Victorian soil data. In
 690 *Proceedings of the 19th World Congress of Soil Science*, Brisbane, August (pp. 1-6).

691 Isbell, R.F., National Committee on Soil and Terrain, 2016. *Australian Soil Classification*, 2nd Edition.
 692 CSIRO Publishing, Clayton South, Australia.

693 ITPS. 2015. Status of the World's Soil Resources. Food and Agriculture Organization of the United
 694 Nations and Intergovernmental Technical Panel on Soils, Rome, Italy.

695 Jenny, H., 1941. In: *Factors of soil formation*. McGraw-Hill, New York.

696 Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., Karlson, E. W., Ananthakrishnan, A. N., Gainer,
 697 V.S., Shaw, S.Y., Xia, Z., Szolovits, P., Churchill, S., Kohane, I., 2015. Development of phenotype
 698 algorithms using electronic medical records and incorporating natural language processing. *British*
 699 *Medical Journal*. 350, h1885.

700 Marshall T.J., 1947. Mechanical composition of soil in relation to field description of texture. CSIR,
701 Aust. Bull. No. 224, Government printer, Melbourne, Victoria.

702 Luo, Z., Wang, E., Sun, O., J., 2010. Soil carbon change and its responses to agricultural practices in
703 Australian agro-ecosystems: A review and synthesis. *Geoderma* 155, 211–223.

704 MacEwan, R.J., Crawford, D.M., Newton, P.J., Clune, T.S., 2010. High clay contents, dense soils, and
705 spatial variability are the principal subsoil constraints to cropping the higher rainfall land in south-
706 eastern Australia. *Soil Research* 48, 150–166.

707 McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002. From pedotransfer functions to
708 soil inference systems. *Geoderma*, 109(1–2), 41–73.

709 McKenzie, N.J., Jacquier, D.W., Maschmedt, D.J., Griffin, E.A., Brough, D., 2005. The Australian
710 Soil Resource Information System, Technical Specifications. Version 1.5. Australian Collaborative
711 Land Evaluation Program, Canberra, ACT.

712 McKenzie, N.J., Grundy, M.J., Webster, R., Ringrose-Voase, A.J., 2008. Guidelines for Surveying
713 Soil and Land Resource. 2nd edition. CSIRO Publishing, Collingwood, Victoria.

714 McNeill, S.J., Lilburne, L.R., Carrick, S., Webb, T.H., & Cuthill, T., 2018. Pedotransfer functions for
715 the soil water characteristics of New Zealand soils using S-map information. *Geoderma*, 326, 96-110.

716 Mikhail, E.H., Briner, G.P., 1978. Routine particle size analysis of soils using sodium hypochlorite
717 and ultrasonic dispersion. *Australian Journal of Soil Research*, 16 (2), 241–244

718 Minasny, B., Hopmans, J.W., Harter, T., Eching, S.O., Tuli, A., Denton, M.A., 2004. Neural networks
719 prediction of soil hydraulic functions for alluvial soils using multistep outflow data. *Soil Science*
720 *Society of America Journal*, 68(2), 417–429.

721 Minasny B., McBratney, A., Malone, B., Sulaeman, Y., 2010. Digital mapping of soil carbon. 19th
 722 World Congress of Soil Science, Soil Solutions for a Changing World 1 – 6 August 2010, Brisbane,
 723 Australia. Published on DVD.

724 Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the
 725 development of pedotransfer functions for water-retention curves. *Geoderma*, 93(3–4), 225–253.

726 Minasny, B., McBratney, A.B., 2001. The Australian soil texture boomerang: a comparison of the
 727 Australian and USDA/ FAO soil particle-size classification systems. *Soil Research* 37, 1443–1451.

728 Minasny, B., McBratney, A., 2002. The neuro-m method for fitting neural network parametric
 729 pedotransfer functions. *Soil Science Society of America Journal*, 66(2), 352–361.

730 Nemes, A., 2015. Why do they keep rejecting my manuscript – do's and don't's and new horizons in
 731 pedotransfer studies. *Agrokémia és Talajtan*, 64(2), 361-371.

732 Nilsson, N.J., 2014. *Principles of Artificial Intelligence*. San Francisco, Morgan Kaufmann.

733 NLWRA (National Land & Water Resources Audit), 2001. Australian dryland salinity assessment 2000:
 734 extent, impacts, processes, monitoring and management options. NLWRA, Canberra.

735 Northcote, K.H., 1979. A Factual Key for the Recognition of Australian Soils. Rellim Technical
 736 Publications Pty Ltd, Adelaide, South Australia.

737 Orton, T.G., Mallawaarachchi, T., Pringle, M.J., Menzies, N.W., Dalal, R.C., Kopittke, P.M., Searle,
 738 R. Hochman, Z., Dang, Y.P., 2018. Quantifying the economic impact of soil constraints on Australian
 739 agriculture: A case-study of wheat. *Land Degradation & Development*, 29(11): 3866–3875.

740 Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. 2013. R package version 3.1–108. The R development
 741 core team nlme: linear and nonlinear mixed effect models.

742 Piper, C.S., 1944. *Soil and Plant Analysis*. Interscience Publishers, New York.

743 R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for
 744 Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.

745 Rab, M.A., Chandra, S., Fisher, P., Robinson, N.J., Kitching, M., Aumann, C.D., Imhof, M., 2011.
 746 Modelling and prediction of soil water contents at field capacity and permanent wilting point of
 747 dryland cropping soils. *Soil Research*, 49, 389-407.

748 Rab, M.A., Chandra, S., Imhof, M., 2016. Derivation and predictive performance of pedotransfer
 749 functions: an empirical investigation. *European Journal of Soil Science*, 67, 536-549.

750 Rayment, G.E., Lyons, D.J., 2011. *Soil Chemical Methods: Australasia* (Vol. 3). CSIRO Publishing,
 751 Collingwood, Victoria, Australia.

752 Rees, D., Robinson, N., Imhof, M., Joyce, E.B., MacEwan, R., 2010. Spatial frameworks to support
 753 digital soil mapping. 19th World Congress of Soil Science, Soil Solutions for a Changing World,
 754 Brisbane, Australia, pp. 50-53.

755 Robertson, F., Crawford, D., Partington, D., Oliver, I., Rees, D., Aumann, C., Armstrong, R., Perris,
 756 R., Davey, M., Moodie, M., Baldock, J., 2016. Soil organic carbon in cropping and pasture systems of
 757 Victoria Australia. *Soil Research*, 54, 64–77.

758 Robertson, F., Nash, D. 2013. Limited potential for soil carbon accumulation using current cropping
 759 practices in Victoria, Australia. *Agriculture, Ecosystems and Environment*, 165, 130–140.

760 Robinson, N.J., Benke, K.K., Norng, S., 2015. Identification and interpretation of sources of
 761 uncertainty in soils change in a global systems-based modelling process. *Soil Research*, 53(6), 592–
 762 604.

763 Robinson, N.J., Dahlhaus, P.G., Wong, M., Macleod, A., Jones, D., Nicholson, C., 2018. Testing the
 764 public-private soil data and information sharing model for sustainable soil management outcomes.
 765 *Soil Use and Management*. <https://doi.org/10.1111/sum.12472>.

766 Russell, S. J., Norvig, P., 2016. Artificial intelligence: a modern approach. 3rd Ed, Malaysia; Pearson
 767 Education Limited.

768 Searle, R., 2014. The Australian site data collation to support the GlobalSoilMap. In D. Arrouays, N.
 769 McKenzie, J. Hempel, A. C. R. deForges, & A. McBratney (Eds.), GlobalSoilMap: Basis of the global
 770 spatial soil information system. London: CRC Press.

771 Selle, B., Wang, Q.J., Mehta, B. 2011. Relationship between hydraulic and basic properties for
 772 irrigated soils in southeast Australia. *Journal of Plant Nutrition and Soil Science*, 174(1), 81-92.

773 Stein, J.D., Rahman, M., Andrews, C., Ehrlich, J.R., Kamat, S., Shah, M., Boese, E.A., Woodward,
 774 M.A., Cowall, J., Trager, E.H., Narayanaswamy, P., Hanauer, D.A. 2019. Evaluation of an algorithm
 775 for identifying ocular conditions in electronic health record data. *JAMA Ophthalmology*, 137(5), 491-
 776 497.

777 Stroup, W.W., 2012. Generalized linear mixed models: modern concepts, methods and applications.
 778 CRC press, Baton Rouge, USA.

779 Van den Boogaart, K.G., Tolosana-Delgado, R., 2013. Analyzing compositional data with R (Vol.
 780 122). Berlin, Springer, Germany.

781 Van Dijk, A.I.J.M., Beck, H.E., Crosbie, R.S., de Jeu, R.A.M., Liu, R.R., Podger, G.M., Timbal, B.,
 782 Viney, N.R., 2013. The Millennium Drought in southeast Australia (2001–2009): Natural and human
 783 causes and implications for water resources, ecosystems, economy, and society, *Water Resources*
 784 *Research*, 49, 1040–1057. doi:10.1002/wrcr.20123.

785 Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A.,
 786 Pachepsky, Y.A., Padarian, J., Schaap, M.G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg,
 787 M.J., Weihermüller, L., Zacharias, S., Zhang, Y., Vereecken, H., 2017. Pedotransfer Functions in
 788 *Earth System Science: Challenges and Perspectives. Reviews of Geophysics*, 2017, 1199–1256.
 789 <https://doi.org/10.1002/2017RG000581>.