OVERVIEW OF PEDOMETRICS

Alexandre M.J.-C. Wadoux

Affiliation: Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Australia
Corresponding author address: C81 - ATP - The Biomedical Building, The University of Sydney, 1 Central Avenue, Eveleigh NSW 2015, Sydney, Australia
Email: alexandre.wadoux@sydney.edu.au
Phone: +61 491 747 770

The late Inakwu O.A. Odeh

Affiliation: Formerly University of Sydney, Sydney, Australia

Alex B. McBratney

Affiliation: Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Australia
Email: alex.mcbratney@sydney.edu.au
Phone: +61 2 9351 1170

Abstract

Pedometrics is concerned with the application of mathematical and statistical methods to the study of the distribution and genesis of soils. Here, we describe the main areas that pedometric research addresses: distribution of the soil pattern in character space, spatial and spatio-temporal soil variation, quantitative evaluation of the utility and quality of soil, and quantitative pedogenesis. To these main areas akin to the problems of pedology, pedometrics considers and represents uncertainty. Pedometric research is undeniably in an expansion phase and has now many areas of application at the interface with many questions relevant to the sustainable growth of our societies.

Keywords: Pedometrics – Pedology – Digital soil mapping – Geostatistics – Numerical classification – Fuzzy logic – Soil functions – Soil landscape modelling – Space-time prediction – Land evaluation – Uncertainty

Key points

- Succinctly describes the four main areas of pedometrics.
- Pedometric research can contribute to many areas of soil science.

Introduction

The term ‘pedometrics’, first coined by McBratney in the late 1980s, ‘is a neologism derived from the Greek roots pedos or πεδοσ (soil) and metron or μετρον (measurement)’. It is used in a similar fashion to other words such as biometrics, psychometrics, econometrics, chemometrics, and, the oldest of all, geometry. Furthermore the term covers two main ideas – the ‘soil’ or ‘pedo’ part, which corresponds to that branch of soil science we call pedology, and the ‘metrics’ part, which is restricted to mathematical and statistical methods. Simply defined, pedometrics is the use of quantitative methods for the study of soil distribution and genesis, and as a sustainable resource. Another problem-oriented definition is ‘soil science under uncertainty.’ In this sense pedometrics deals with uncertainty in soil models that describe deterministic or stochastic variation, vagueness, and lack of knowledge of soil properties and processes (Table 1). Thus, mathematical, statistical and numerical methods could be applied to resolve the uncertainty and complexity inherent in a soil system model, including numerical approaches to classification, which deals with supposedly deterministic variation (Webster, 1994).

Table 1: Types of quantitative models used for soil studies with examples and sources of uncertainty (Courtesy AB McBratney and CJ Moran).

Causality and/or uncertainty	***Model***		
	Deterministic	***Stochastic***	***Nonstatistical***
Empirical	Generalized linear models, numerical taxonomy, Jenny functional relations and canonical ordination	Time series, spatial processes, temporal and spatial variation of soil properties, Markov chain models	Fuzzy systems, Machine learning
Mechanistic	Flow and diffusion of soil plasmic materials, profile and landscape development	Indeterminate	Indeterminate
Uncertainty	Imprecise measurements, model uncertainty	Random process, probability	Vagueness, ambiguity, and fuzzy geostatistics

Pedometrics is not new, as mathematical and statistical methods have been applied to soil studies since the 1900s (statistical research by R. Fisher at Rothamsted) and more intensively since the 1960s and 1970s. However, it is now a technical branch of soil science complementing traditional pedology. Over time the use of computers has increased in both fields, and the difference between the two has decreased and, in some cases, overlapped. Due to new demands for quantitative soil information required for global-scale models, regional environmental planning, and field-scale agricultural land management, traditional pedology has become more quantitative through the increased use of digital soil data in computerized soil information systems. Concurrently, pedometrics has emerged as a collection of quantitative tools, which are increasingly being used to account for conceptual pedological models of soil variation. As of 2021, there is a strong and growing overlap and synthesis between ‘traditional’ pedology and quantitative pedology or pedometrics (as shown in Figure 1).

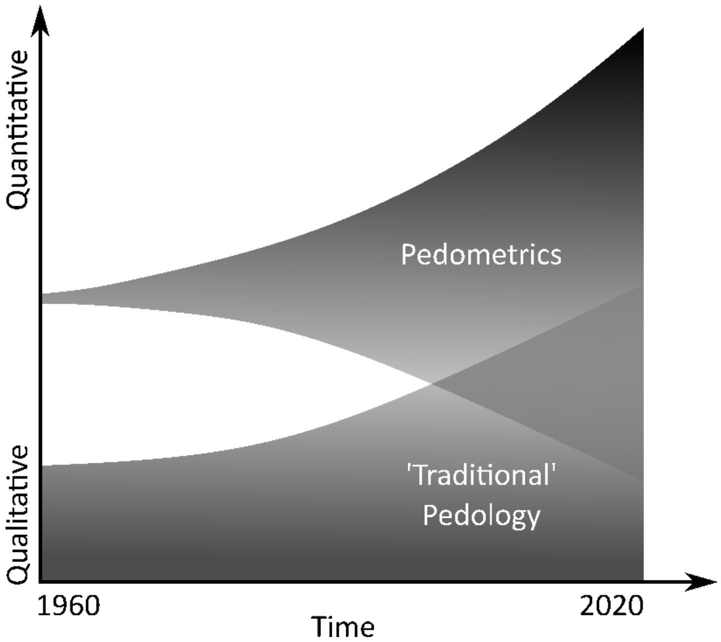

Figure 1. A timeline of the growth of pedology and pedometrics. (Modified from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. Geoderma 97: 293–327).

Methods and Problems

Pedometrics: Principles and Hypotheses

As the definition implies, pedometrics encompasses the whole quantitative approach to the study and description of soil. It also involves the use of mechanistic, stochastic, deterministic, and empirical models (Table 1). A more restricted form of pedometrics is one that involves applications of statistical and probabilistic modelling to soil variation. Pedometric principles are therefore embodied in (geo-) statistical modelling of pedological processes and the associated phenomena as they influence soil variation, and the associated model uncertainty.

Pedometrics is therefore aimed at resolving or testing a number of questions or hypotheses (McBratney and Lark, 2018):

1- What is the pattern of soil distribution in character space, i.e. soil taxonomy?
2- How does the soil vary spatially and temporally?
3- What are the utility and quality of soil?
4- What are the causal factors of soil formation, i.e. soil genesis?
5- What is the uncertainty of modelling the spatial pattern of soils, i.e. is a model an accurate representation of reality?

With the exception of (5), these hypotheses dovetail well with the requirements of conventional pedology. Although developments in pedometrics in the past were restricted to statistics and probability, increasingly the utility of soil is the subject of pedometric research and mechanistic models are being applied and developed to model soil processes.

Soil Measurements and Properties

To model soil quantitatively, we need to perceive it, i.e., sense it or measure some quality or quantity of it. In other words, how do we represent soil properties numerically to model their variation mathematically and/or statistically? Ever since the advent of modern environmental science, technological advances have improved immensely the way we characterize or measure the quality and/or quantity attributes of soil. Soil measurement techniques include direct measurements, visual observations, and remote sensing (including proximal, airborne, and space-based modes of sensing).

Field properties: hard and fuzzy descriptors?

Other than color, many soil morphological properties, required in routine soil surveys, are described in somewhat vague terms. For example, soil structure grade is described as 'structureless' or 'weak' or 'moderate' or 'strong' in many national soil survey handbooks. To make the grades more amenable to quantitative analysis (e.g., multivariate statistical analysis) these terms are usually coded as 0, 1, 2, and 3, respectively. But because of the vagueness in the linguistic characterization, fuzzy coding, which accounts for uncertainty or lack of clear boundary between the grades, has been used. Other morphological descriptors such as soil aggregate (structure) size classes, e.g., 'fine,' 'medium,' and 'coarse,' may be given linguistic variables which could be fuzzified to determine the degree of truth that a given soil layer is characterized by. Simple examples of fuzzy coding are presented in Table 2.

Table 2: Fuzzy coding of structural type

Type	*Horizontality*	*Verticality*	*Flatness*	*Accommodation*
Platy	1.0	0.1	1.0	1.0
Lenticular	1.0	0.3	0.3	1.0
Prismatic	0.2	1.0	1.0	1.0
Columnar	0.2	1.0	0.9	0.9
Angular blocky	1.0	1.0	1.0	1.0
Subangular blocky	0.7	0.7	0.5	0.5
Granular	0.2	0.2	0.0	0.1
Massive	0.0	0.0	0.0	1.0
Single grain	0.0	0.0	0.0	0.0

Reproduced with permission from Odeh IOA, McBratney AB, and Chittleborough DJ (1991) Elucidation of soil-landform interrelationships by canonical ordination analysis. *Geoderma* 49: 1–32.

Soil spectroscopy

Measurement of soil properties has gained tremendously from advances in technology, particularly in spectroscopy. Soil spectroscopy is loosely defined as the study of the spectral signature of a soil material. The spectral signature of a soil is primarily a function of its mineral and organic composition since the soil is formed from the transformation of rocks, organic matter from plants and organism activity. Thus, a soil type has a unique spectral reflectance curve. Spectroscopic measurements can be proximal or remote, when mounted on a satellite, plane or drone. Proximal measurements might be done *in situ* on the soil profile using portable field spectrometers or in the laboratory. Because of the large number of soil characteristics that a soil spectrum contains, it is said that spectroscopic measurements are cost-effective and fast, compared to conventional methods of soil analysis (Nocita et al., 2015).

The visible (400-700 nm) and infrared (400-25,000 nm) range of the electromagnetic spectrum is of particular interest to soil scientists. Infrared spectra have encoded information on both organic and inorganic soil material. The mid-infrared range (2,500-25,000 nm; 400-4,000 cm^{-1}), contains distinct absorption features for soil organic matter and mineralogy. These absorption features have overtones and combinations in the visible and near-infrared (vis-NIR) range (400-2,500 nm), which means that soil organic and mineral characteristic are also reflected in the vis-NIR range but with fewer, broader, and more complex (i.e. non-specific) absorption bands. Figure 3 shows three vis-NIR spectra for soil samples with different compositions.

Pedometricians have extensively investigated the predictive power of such spectra to replace and complement soil analyses. Mathematical transfer functions are used to correlate the spectral wavelengths (the independent variable) to laboratory-measured values of soil properties (the dependent variables). Relatively simple models can be used, for example linear regression on user-defined wavelengths, but more complex models exist, such as principal component analysis and partial least-squares regression. Both models consider the whole spectra for estimating their parameters and make use of reduction dimensionality (principal component analysis). More recently, pedometricians have also considered machine learning models, with some success. Once calibrated, the model is used to predict the soil properties using spectral information only. These soil property estimates are stored together with the soil spectral information into spectral libraries. Lately, efforts have been made to combine and standardize soil spectra from various instruments into large, regional and global soil spectral libraries. Research is also on-going to fuse multiple sensors and instruments to obtain more precise estimates of soil properties.

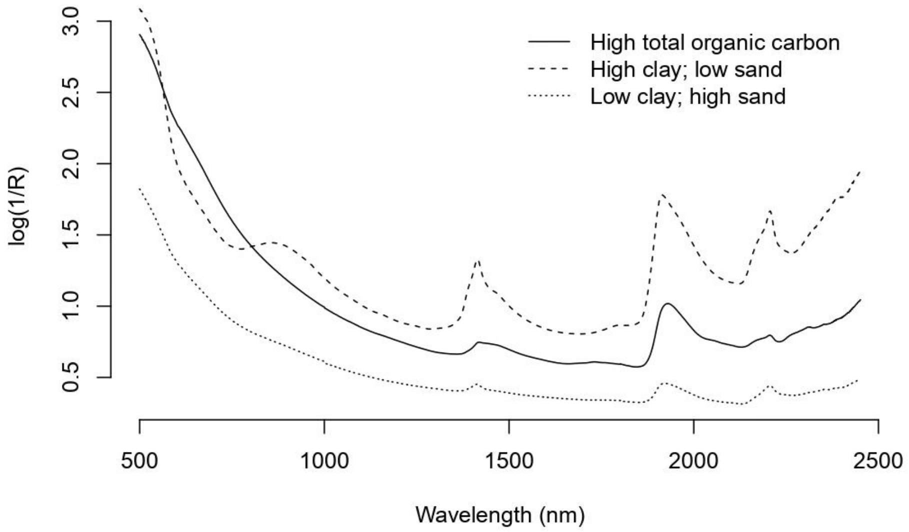

Figure 2: Vis-NIR spectra of three soil samples from wheatbelt of southern New South Wales, Australia: a soil with 12% total organic carbon and two soil samples with low total organic carbon content (<1%) of which one has 74% clay and 19% sand and the other 5% clay and 91% sand.

Pedotransfer functions

The term ‘pedotransfer function’ (PTF) is generally defined as translating the raw soil data into more useful information. It can also be defined as predictive functions of certain soil properties difficult to obtain from other easily, routinely, or cheaply measured properties (van Loy et al., 2017). The most readily available data come from soil survey, such as field morphology, texture, structure, and pH. The PTFs add value to the basic soil data by translating them into predictors of other more laborious and expensively determined soil properties. These functions fill the gap between the available soil data and other properties that are more useful or required for a particular model or quality assessment. A simple illustration of how to apply PTFs to practical problems is shown in Figure 4.

Soil spectral inference systems also rely on PTFs to populate soil databases. Spectral inference systems combine soil spectral data to infer a set of basic soil properties, and PTFs to estimate another set of soil properties more difficult to measure or without spectral response, such as permanent soil wilting point or field capacity. The PTFs are found in the literature and combined into a network structure to create an inference system. One of the main features of soil spectral inference systems is the quantification and propagation of uncertainty, using methods such as model ensemble by bootstrap and aggregation. (See chapter on this topic in Encyclopedia).

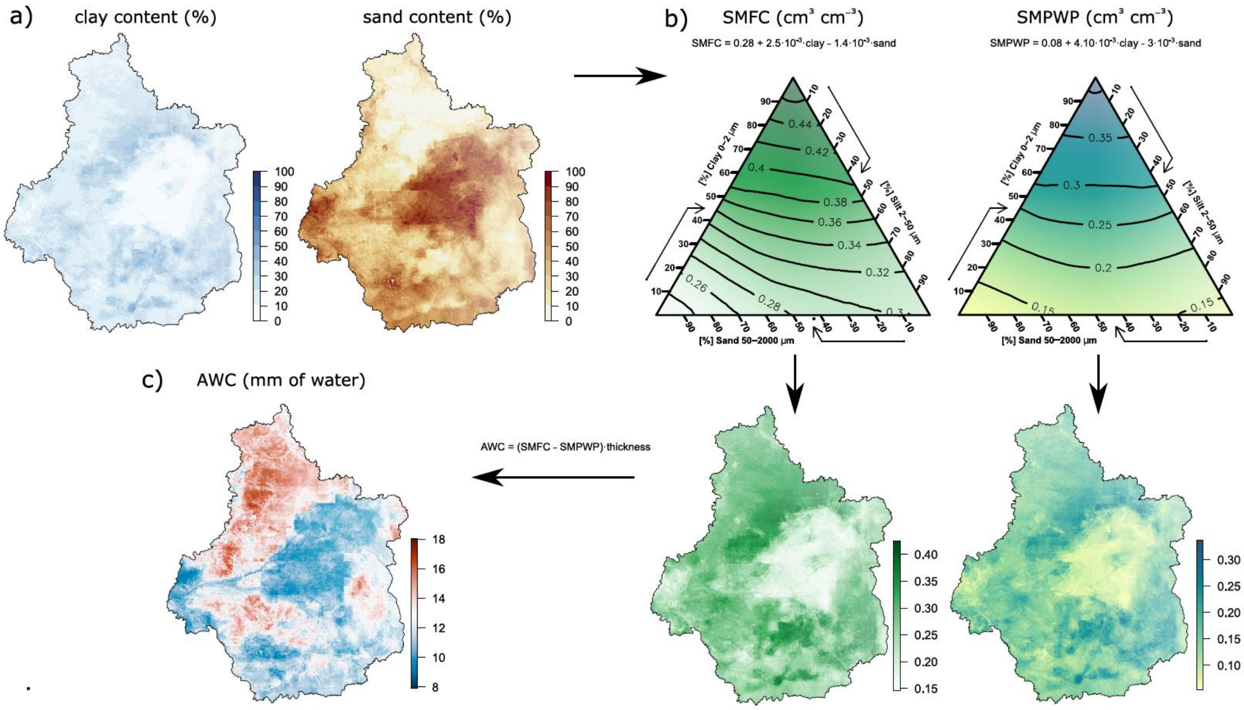

Figure 3: A simple example of pedotransfer functions (PTFs) for the Centre-Val de Loire region (France): a) particle-size fractions were interpolated on to a fine grid from a set of point measurements; b) horizon (5-15 cm) soil water at field capacity (SMFC) and soil moisture content at permanent wilting point (SMPWP) were in turn estimated through a PTF expressed as a function of particles sizes; c) available water capacity is estimated by multiplying the SMFC subtracted from the SMPWP by the soil layer thickness (in mm), assuming all soil is fine earth (Modified from Román Dobarco, MR, Bourennane, H, Arrouays, D, Saby, NP, Cousin, I, & Martin, MP (2019). Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. Geoderma, 344, 14-30).

Quantitative Soil Geography

Soil Classification and Soil Map Units

The pattern of soil in the landscape is as a result of effects of soil-forming factors, namely: climate, organisms, parent material, topography, and time (usually referred to as 'clorpt,' but recently extended to 'scorpan' – see Eqn [2], below). These factors, particularly climate and parent material, produce broad patterns of soil in geographic space. The complex combination of these factors is what causes repetitive patterns of soil in the landscape, which form the basis for soil classification, identification, and mapping. In traditional soil classification, both the conceptual and real classes are qualitative and have no clear-cut boundary. Pedologists have, in the past, attempted to avoid these by applying numerical classification to complex soil data. Soil classification, in both geographical and taxonomic space, is a simple representation of complex, and sometimes repetitive patterns of soil in the landscape. Class identification, on the other hand, is aimed at matching classes that are conceptually described (usually in a classification system) with reality – represented by data obtained by soil morphological description and measurements.

A soil map unit is a collection of areas that are relatively homogeneous in soil constitution or miscellaneous areas that constitute different soil components or both. In a given survey region, a map unit differs in some respect from all others and can be uniquely identified on a soil map. A soil map delineation, on the other hand, is an individual area on the map, bounded by other delineations or the map boundary. Usually a delineation contains the dominant components in the map unit name, but it may not always contain a representative portion of each kind of inclusion. Soil boundaries can seldom be shown with complete accuracy on soil maps; hence parts of adjacent polypedons are inadvertently included or excluded from delineations. This problem has been the stimulus for paradigm shifts in soil classification in the latter part of the 1990s; hence the applications of numerical hard and fuzzy classification systems in soil science.

From the 2010s onwards, there was a renewed interest in numerical soil classification, after nearly two decades of meagre activity and progress. Two classification systems, the World Reference Base (WRB) and Soil Taxonomy, are considered global soil classification systems. Both systems, however, also have several shortcomings for local applications. A better communication between regional soil classification systems was attempted, for example by standardization using taxonomic distance. All soil classifications can be evaluated in the multi-dimensional space of the soil data with dimensionality reduction, and by computing centroids for each class of any system. Numerical methods applied to soil classifications have other advantages. For example, they enable the quantification of similarity between soil profile descriptions or between genetic horizons (Hughes et al. 2018).

Fuzzy classes and the breakdown of the classification paradigm

Traditionally most classification systems are composed of mutually exclusive classes in order to conform to discontinuous soil variation embedded in the traditional concept of soil map units. But soil variation is more continuous than discrete. The pioneer work in pedometrics, involving computer-based numerical classification, was designed to address this limitation, among others. While the applications of numerical soil classification to soil studies are, to some extent, based on continuous representation of soil in space, their results are still interpreted in terms of ‘hard’ classes. There is also lack of any spatial coherence for the classes to be mapped, despite some attempts with, for example, spatially weighted classification. Recent advances are based on fuzzy sets to optimize the quality of prediction of the resulting classification, and which take cognizance of the continuous nature of soil variation and nonlinearity in the inter-attribute relationships.

The first application of fuzzy set theory in soil survey was principally for classification. Two different but complementary approaches to grouping individuals into fuzzy classes in soil science are: (1) fuzzy *c*-means (FCM, also known as ‘fuzzy *k*-means’), and (2) the semantic import model (SI).

The FCM algorithm for improved predictive classification by providing for membership to an extragrade class is based on the objective function, defining the within-class sum-of-square errors, J_E, expressed as (McBratney and Odeh, 1997):

$$J_E(M,c) = \alpha \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij}^{\varphi} d_{ij}^{2} + (1-\alpha) \sum_{j=1}^{c} m_{i*} \sum_{j=1}^{c} d_{ij}^{-2} , \qquad [1]$$

where c is the number of classes, n is the number of individuals or pedons; m_{ij} is the membership of an individual i in class j; φ is the fuzziness exponent $(1 < \varphi < \infty)$; d_{ij} is the character space between the feature value of an individual, i, and the feature centroidal value for class j; α is the parameter that determines the mean value of m_{i*}, which is the membership value of an individual, i, in the extragrade class.

The SI model was developed primarily for land evaluation. The need to use fuzzy set theory in land evaluation, such as that defined in the Food and Agriculture Organization of the UN (FAO) framework, arises because basic soil information used for land evaluation is mostly described by seemingly vague terms such as ‘poorly drained,’ ‘slightly susceptible to soil erosion,’ and ‘moderate nutrient availability.’ Not even when these terms are defined precisely is the qualitative ambiguity removed. Usually, the land evaluator's aim is to produce a set of clearly defined classes of land qualities based on specified land use requirements. These subsequently provide the means of transferring information about the soil and its use. As land qualities are complex attributes that are derived from land characteristics such as topography, soil, water, or biological and human activity, subsequent Boolean logical operations in the process of land evaluation tend to throw away much useful information.

Pedodiversity

Measuring pedodiversity within an area is a way of measuring soil variation, usually using pre-classified soil entities such as taxa or properties. The diversity of pedological entities such as pedotaxa, pedogenetic horizon and soil properties, as well as their spatial and temporal patterns influences several ecosystem functions that the soil provides. A more diverse soil is also more resilient against disturbance and stress caused by unsound soil management practices.

The concept of diversity contains two fundamental components which apply to pedodiversity (Ibáñez and Bockheim, 2013): i) the richness, i.e. the number of soil types (soil classes) in the area of interest or within a level of the taxonomic system and ii) the evenness, i.e. the relative number of soil types or classes. Soil scientists have contributed to developing indices of pedodiversity, most of them being similar to the well-known indices of ecological and biological diversity developed by ecologists and biologists. Such indices encompass species richness and abundance models, few of them incorporate the two into a single index, the most popular of which is the Shannon diversity index, defined by:

$$\text{Shannon diversity index} = -\sum_{i=1}^{n} p_i \times \ln p_i, \qquad [2]$$

where p_i is the proportion of the ith soil type relative to the total number of soil types n in the area of interest or level of the taxonomic system. When the area is composed of a single soil type, it equals zero and is unbounded otherwise as the number of soil types increases. When applied to existing soil classification systems, several authors have found clear relationships between the number of soil types (i.e. the pedodiversity) and the size of the area surveyed. Such indices were applied at the global scale, where it was found that some continents are more diverse than others.

Pedometricians contended that taxonomic distance between soil types should be included in the computation of pedodiversity indices because the similarity differs among taxa. These can be included by indices of quadratic and taxonomic entropy.

Space and time Prediction

The clorpt approach

As previously stated, the ‘clorpt’ methods are based on the empirical-deterministic models that originated from Hans Jenny's *Factors of Soil Formation*. Jenny's state-factor equation, in its extended form, is expressed as (McBratney et al. 2003):

$$S = f(s, c, o, r, p, a, n), \qquad [3]$$

where S is some soil properties or soil type as a function f of state factors: s as some other soil property at a point, c as climate, o as organisms, r as relief, p as parent material, a as time or age, and n as space or spatial position. Soil spatial variability is therefore considered as being causative realizations of the complex combinations of soil-forming processes as influenced by the soil-forming factors. The scorpan function (Eqn [3]) – in its original form, the clorpt – earlier in the twentieth century stimulated numerous studies, mainly quantitative prediction of soil attributes, which we shall treat in a later section. Until recently, because data on soil classes were non-quantitative, the scorpan methods were only restricted to predicting soil attributes measured on a continuous scale.

Models such as that expressed in Eqn [3] can be derived for predicting soil classes or soil types using various techniques, for example logistic regression, classification trees and other machine learning models. Logistic regression is designed specifically for situations in which we have a dichotomous (nominal or ordinal) dependent variable (e.g., soil classes), in comparison to classical linear regression typically used for continuous dependent variables (e.g., soil pH value). Several problems arise when a dependent variable is binary: the error terms are non-normal and their variance is non-constant. Although the errors are not normal, this method still provides unbiased regression estimators that are approximately normal if the sample size is large. Classification tree algorithms and other machine learning models, on the other hand, search for combinations of values of independent variables (e.g., Jenny's state factors or attributes derived from them) that best predict the value of the dependent variable (e.g., unordered factors such as soil classes). Prediction quality is based on criteria such as the within-group noises, variance, or statistical significance of the conditional frequency distribution of values for the dependent variable, conditioned on the answers to questions asked. A digital map of soil suborder classes produced by the classification tree model is shown in Figure 4.

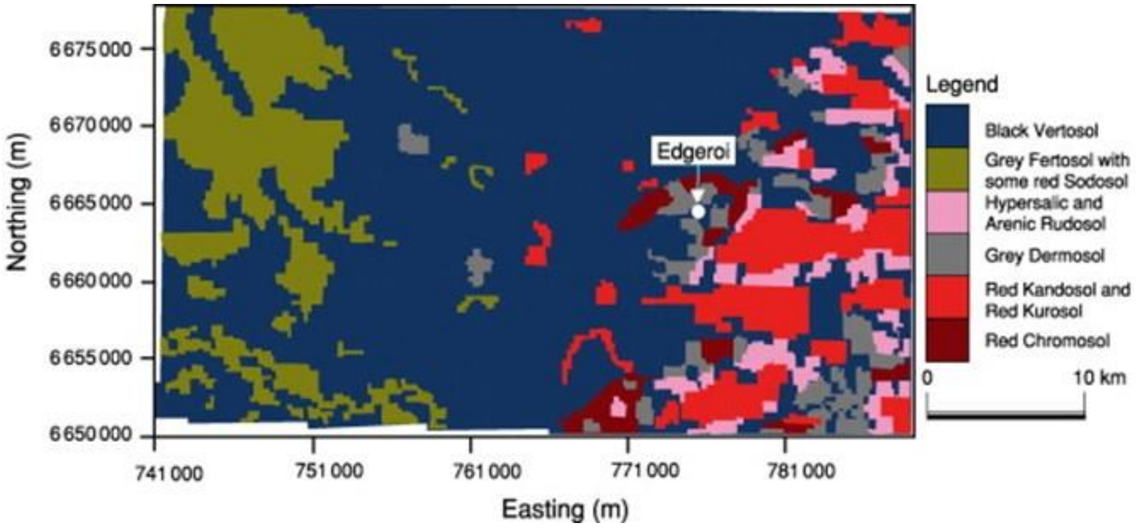

Figure 4: Predicted soil classes for the Edgeroi area in accordance with Australian Soil Classification. (Reproduced with permission from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. Geoderma 97: 293–327).

The scorpan approach

The ‘scorpan’ approach (Eqn [3]) earlier in the twentieth century stimulated numerous studies in quantitatively predicting soil attributes measured on a continuous scale. Many of the earlier studies, and indeed some recent examples, were based on general and bivariate-simple linear regression, although multiple polynomial regression models were occasionally applied. But due to nonlinearity in the correlations among many soil variables, and indeed of many of the soil variables with ancillary variables, robust methods such as generalized linear models (GLMs), generalized additive models (GAMs), and random forest (RF) have been developed and applied. Another development is the artificial neural network models, which are nonparametric modelling techniques that mimic the neural networks of the brain. The networks are composed of processing units, or neurons, which are organized into layers, i.e., input, hidden, and output layers. We shall describe mapping based on machine learning in the next subsection.

The problem is that while the classic models or the more robust methods may take care of the deterministic relations, they do not account for spatial autocorrelations of soil properties, especially at the local level. To solve this problem, the pioneer pedometricians initiated the application of geostatistics (which was primarily developed for the mining industry).

Geostatistics

Matheronian geostatistics is based on the theory of regionalized variables, which allows us to consider spatial variability of a soil property or even soil types, if quantified, as a realization of a random function represented by a stochastic model. The generic geostatistical method of spatial interpolation is termed ‘kriging’ in its various forms: simple, ordinary, lognormal, and disjunctive kriging (Webster and Oliver, 2007).

In geostatistics, the variogram is a primary requirement for spatial prediction or kriging of a target geographical feature. The variogram can be obtained by computing the spatial correlation or covariance between pairs of samples at certain distances apart in a data set to produce empirical semivariances. The plot of the semivariance and the corresponding distance or lag at which the pairs are separated produce the variogram. The latter describes the magnitude, spatial scale, and the general form of variation of a given variable (Heuvelink and Webster, 2001). An example of an experimental variogram is shown in Figure 5, together with several exponential models based on different fitting methods.

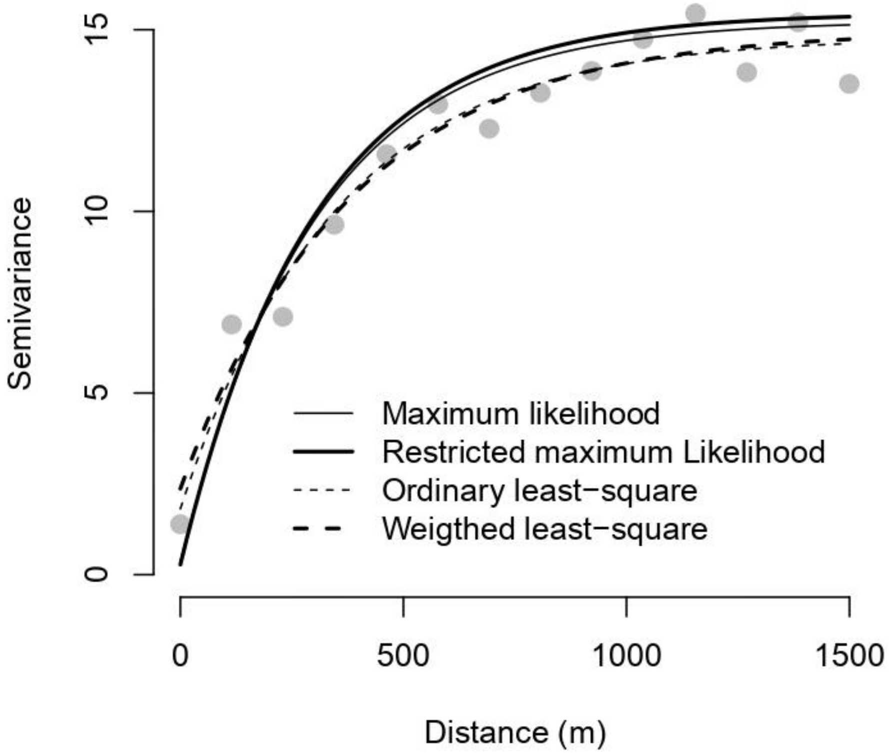

Figure 5: An experimental variogram fitted with exponential models based on different fitting techniques.

The first major applications of kriging, in its ordinary and univariate form, for soil studies were in the early 1980s. Since then ordinary kriging has been widely used in various subfields of soil science, including soil reclamation, soil classification, and soil pollution. Major limitations of the univariate geostatistics technique of kriging are due to the assumptions of stationarity, which are not often met by the field-sampled data sets and, of course, the often-cited requirement of large amounts of data to define the spatial autocorrelation. However, with increasing availability of ancillary information, the lack of adequate samples has been partially solved. The univariate use of kriging is also limiting in situations of complex terrain where the soil-forming processes are themselves complex. In such situations there is the need to model both the structured and the spatially dependent components of the soil variable. Also there are economic and logistic reasons for including the ancillary variables influencing soil variability, especially if the latter are more readily and cheaply available. As both the soil and exogenous factors are multivariate, the most obvious choices are appropriate combinations of multivariate or univariate analysis using the scorpan factors and geostatistical methods. These combinations constitute the hybrid techniques.

The hybrid of scorpan and geostatistics

The hybrid techniques for soil survey and mapping are based on various combinations of the geostatistical and multivariate or univariate scorpan methods. Let us suppose that a data vector describing a soil property is a random variable Z, determined at locations $\boldsymbol{s}$ in a region $\mathfrak{A}$, $\boldsymbol{s}_i(i = 1, \dots, N; \boldsymbol{s}_i \in \mathfrak{A})$, and consisting of three components as:

$$Z(\boldsymbol{s}) = m(\boldsymbol{s}) + Z_1(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}), \qquad [4]$$

where m is the local mean for the region, Z_1 is the spatially dependent component, and ε the residual error term, spatially independent. Now there may be situations where m varies and is dependent on some exogenous factors such as the scorpan factors at spatial location $\boldsymbol{s}$. In other words it is deterministically related to some causative factors (in geostatistics parlance, the variable is said to exhibit a trend). Wherever trend exists, ordinary univariate kriging is inappropriate. Several methods have been designed to accommodate the trend.

Universal kriging has been the commonly used method to accommodate the trend or 'changing drift,' as it is sometimes known, in a soil variable. Universal kriging is a combination of the standard model of multiple-linear regression and the geostatistical method of ordinary kriging, which is also analogous to combining scorpan methods with univariate kriging, but only using the geographic coordinates for determining the drift (the n factor of the scorpan model). Another approach is to use an intrinsic random function of order k (IRF-k), which has been used to accommodate the varying nature of the trend in a regionalized soil variable. The term 'k' represents the order of polynomial trends: $k = 0$ means constant drift, and the IRF-k is equivalent to the ordinary kriging system of equations. If $k = 1$, we have linear drift; $k = 2$ yields quadratic drift; but where there is no trend but deterministic relationships are with some known or readily available and inexpensive covariates (scorpan factors) or other easy-to-measure soil variables, co-kriging has played a major role in efficiently predicting the target soil variable. Universal co-kriging is also possible when considering the trend and covariation with one or more secondary variable.

Co-kriging is the multivariate extension of kriging that allows the inclusion of more readily available and inexpensive attributes in the prediction process. There are many instances in soil survey where the scorpan factors such as topography, time, and variable parent material, are easily discernible or are either readily available and/or are cheap to obtain. The most efficient way to predict the expensive-to-measure target soil variable, the variation of which is affected by the scorpan factors, is to supplement the information of the sparsely sampled target variable with more densely available information from a cheap to obtain variable.

Regression-kriging (RK) is another hybrid method that combines either a simple/multiple-linear regression model or machine learning (for example a variant of GLM, GAM, RT or RF) with ordinary, or simple, kriging of the regression residuals. The assumption here is that the deterministic component (m in Eqn [4]) of the target (soil) variable is accounted for by the regression model, while the model residuals represent the spatially varying but dependent component (Z_1 in Eqn [4]). If the exogenous variables used in the regression equation are available at more dense locations than the target variable, the equation can then be used to predict m on to those locations. The Z_1 can also be predicted to the same locations by the simple kriging system of equations, and then added to the m to obtain Z^*. A variant of RK is kriging with uncertainty, which introduces regression residuals (as representing model uncertainty) into the kriging system used to predict the target soil variable. This reduces the extrema of the target soil variable and therefore produces a smoother function of the predicted values. Another variant of RK, kriging with external drift (KED), is a hybrid technique that integrates the universality conditions into the kriging system using one or more of the ancillary drift variables. It is similar to universal kriging, but using an ancillary variable to represent the trend. As shown in Figure 7a, a regional digital map of cation exchange capacity (CEC) has been produced using KED with elevation as the external drift. Comparing this map with the one produced using RK (Figure 7b) indicates a slightly more smoothed map produced by KED than RK.

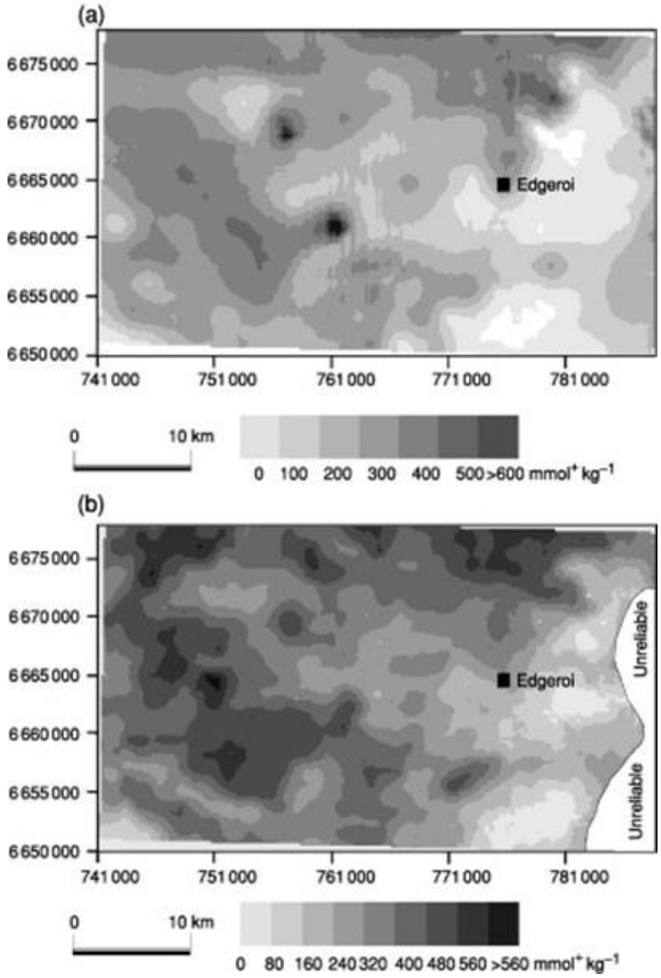

Figure 6: Predicted cation exchange capacity (CEC; millimoles per kilogram) for the Edgeroi area using (a) kriging with elevation as external drift and (b) regression-kriging. (Reproduced with permission from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. Geoderma 97: 293–327).

Machine learning

Machine learning (ML) has emerged since the 1990s as a set of algorithms useful for soil mapping, but it is in the past decade only that the use of ML algorithms in soil science have burgeoned. Machine learning refers to a large set of non-parametric and data-driven algorithms developed for data mining and pattern recognition purposes, and now frequently used in all fields of soil science for classification and regression purposes (Wadoux et al. 2020). The increase in spatial prediction with ML corresponds with the recent availability of large digital soil databases and computer resources. In the previous Section we discussed some limitations of geostatistical mapping of soil, such as stationarity, normal distribution of the residuals, and difficulties for non-linear modelling of soil properties using numerous and cross-correlated covariates, among others. Machine learning can handle most of the limitations of geostatistical models when the amount of data to calibrate the model is sufficiently large. Indeed, ML algorithms do not make assumptions on the distribution of the soil properties and can accommodate numerous, categorical and continuous, and correlated covariates as predictors.

Examples of ML algorithms are, for example, regression tree, random forest (RF), support vector machine, genetic algorithms, and, perhaps the most popular of all, artificial neural networks (ANN). Variants of ANN are the convolutional neural network, which uses images as input and long short-term memory for space-time prediction. Despite the wide adoption of such models for mapping, it is still unclear if interpretation of these complex models is possible, and if it could reveal the importance of the environmental covariates used as predictors. Conversely, including pedological knowledge on the soil processes into these highly empirical models is, to date, very limited. These constitute current challenges in pedometric research (Wadoux et al. 2021).

Space-time prediction

Space-time prediction is concerned with the prediction of soil properties that vary both in space and time, such as soil moisture, soil available water capacity or soil organic carbon. Space-time prediction was first developed as an extension of classical geostatistics that considers spatial variation, but non-parametric models such as machine learning models also have potential extensions for space-time prediction. Extension to space-time prediction of a soil property is a challenge because variation in space might be different from variation in time. Two approaches are usually considered to extend spatial prediction to space-time. The first adds a time dimension by mapping a spatial variable repeatedly over time. Often this approach is preferred when the temporal correlation between time steps is weak, such as when mapping short-term change in soil texture. The second approach treats time as an additional dimension to be considered. This approach is taken in space-time geostatistics by extending the methods of classical geostatistics to the space-time domain, such as the variogram. Many contributions from statisticians have been made to develop space-time variogram functions. Consider the model of spatial variation in Eqn 4, it is extended to the space-time domain by (Heuvelink et al., 2017):

$$Z(\boldsymbol{s},\boldsymbol{t}) = m(\boldsymbol{s},\boldsymbol{t}) + Z_1(\boldsymbol{s},\boldsymbol{t}) + \varepsilon(\boldsymbol{s},\boldsymbol{t}), \qquad [5]$$

where $Z(\boldsymbol{s},\boldsymbol{t})$ is a spatio-temporal random field, m is a space-time trend, often modelled as a function of known environmental covariates, and $Z_1(\boldsymbol{s},\boldsymbol{t})$ is a spatially and temporally dependent component usually assumed to be (multivariate) normally distributed and characterized by a covariance function. An example of variogram modelling the spatio-temporal covariance function is presented in Figure 7. In space-time geostatistics, measurement can be repeated in time at the same spatial locations or vary between time steps. Predictions are made at unobserved locations in space and time.

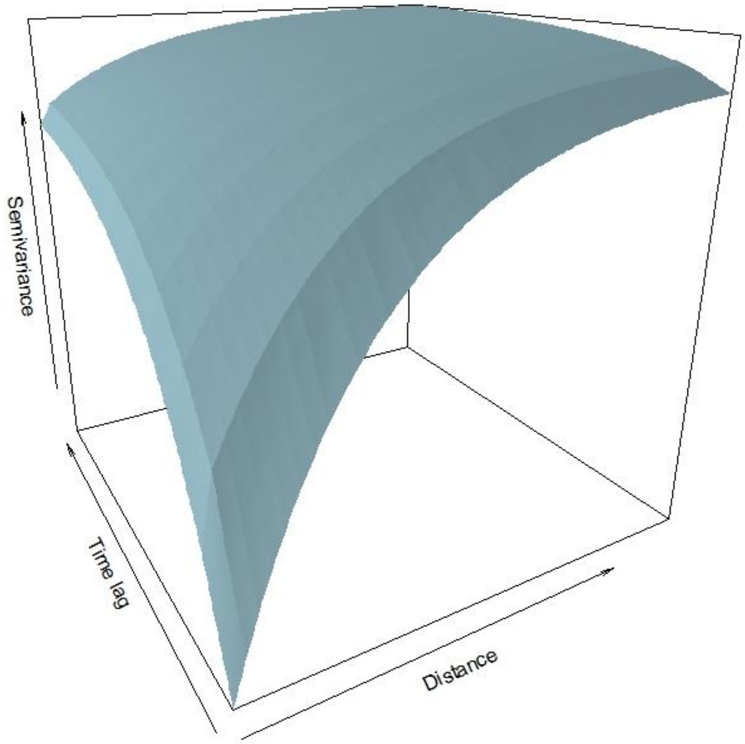

Figure 7: A fitted spatio-temporal variogram using a separable function. This is the space-time extension of the spatial variogram presented in Figure 5.

Extension of machine learning models to the space-time domain is also possible but is still at the preliminary stages of academic research and has not yet reached a large audience. We expect this to change in the future with the development of new software implementations.

Sampling design

The sampling design used to calibrate and validate the spatial or space-time prediction models play a key role in the final accuracy. Pedometricians have contributed to the development of sampling approaches for mapping. Sampling strategies can be divided into two main approaches, namely the design- or model-based approaches (De Gruijter et al. 2006). The design-based approach relies on sampling theory where the sampling locations are selected randomly, either with simple random sampling, or by more advanced probability sampling designs such as stratified sampling. Generally, non-probability sampling approaches (i.e. model-based sampling) are more suitable for local mapping of the soil. Examples of such designs are regular grid sampling, feature space coverage or spatial coverage sampling. A method much used in digital soil mapping is conditioned Latin hypercube sampling (cLHS). Geostatistical techniques also enable sampling optimization, in which the fitted variogram and optimization algorithms enable the optimal spatial locations for mapping to be found. Much research has been done in this area, and developments along these lines are being made for machine learning. All these techniques apply for selecting the calibration sample, but it has been acknowledged that whenever possible the soil maps should be validated by independent, probability sampling (Brus et al. 2011).

Soil Utility and Quality

Land Evaluation

Pedometrics has played a major role in advancing the role of land evaluation and, more recently, the quantification of soil quality for land management and sustainable use of land resources (Rossiter et al. 2018). Although 'soil quality assessment' is often used as a misnomer for 'land evaluation,' both could be regarded as the interpretative phase of soil survey. While land evaluation is concerned with the assessment of land performance when used for specified purposes, soil quality is defined as 'the capacity of a specific kind of soil to function, within natural or managed ecosystem boundaries, to sustain plant and animal productivity, maintain or enhance water and air quality, and support human health and habitation'. In considering productivity, environmental quality, and human health as major functions of soil, this definition requires that values be placed on specific soil functions as they relate to the overall sustainability of alternate land-use decisions. Pedometric techniques (based on 'hard' and/or fuzzy techniques) are being used to quantify soil quality and associated uncertainty. For example soil quality can be assessed in terms of requirement for different specific uses such as shown in Figure 8. Many of the soil quality indicators are required at various scales and spatial extents in which pedometrics is playing a major role in modelling them spatially for incorporation into regional, catchment, and field-scale modeling. Another approach to soil quality analysis may be based on the concept of pedodiversity (see also previous section).

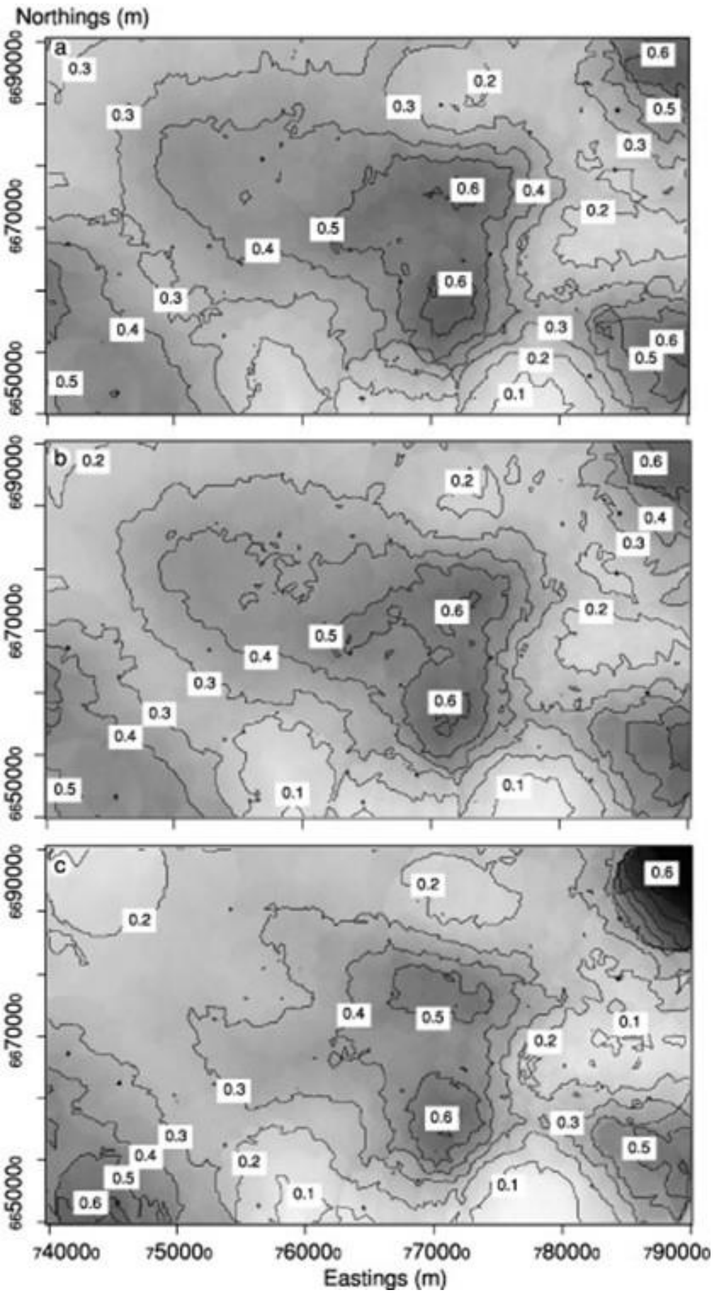

Figure 8: Interpolated suitability of (a) wheat, (b) sorghum and oats, and (c) dryland cotton. (Based on Triantafilis J, Ward WT, and McBratney AB (2001) Land suitability assessment in the lower Namoi Valley of Australia using a continuous model. Australian Journal of Soil Research 39: 273–290).

Soil functions

Pedometrics is also playing a role in quantifying and mapping soil functions. Soils provide services in terms of agricultural production, water protection and filtration, carbon storage or biodiversity habitat, among others. The ability of a soil to provide a set of ecosystem services is determined by its functions, which can be evaluated in various ways. One way is to use empirical models based on a set of indicators (i.e. basic soil properties) and combine them into soil functions. The choice of indicators is based on correlation statistics between properties, on pedological knowledge as well as on a context in which the functions will be used. Another way to model functions is to use biophysical models that better account for the soil management practices and environmental factors. In both ways, the different soil functions are interdependent since they are based on similar basic soil indicators. A change in agricultural practices to improve one function (e.g. production) can simultaneously affect other functions positively or negatively (e.g. the habitat of microorganisms). Because of these synergies and antagonisms between functions, a soil can never mobilize to its full potential each of the functions it can support. The need to study multifunctionality was the basis for the concept of soil security, or other concepts such as soil health. Converting basic properties into a product that better matches end-user demand was also the objective of land evaluation. Pedometric techniques are being used to quantify soil functions, to account for their spatial and temporal variation and to quantify and propagate their uncertainty. For example, in Figure 9 two maps of soil functions are shown.

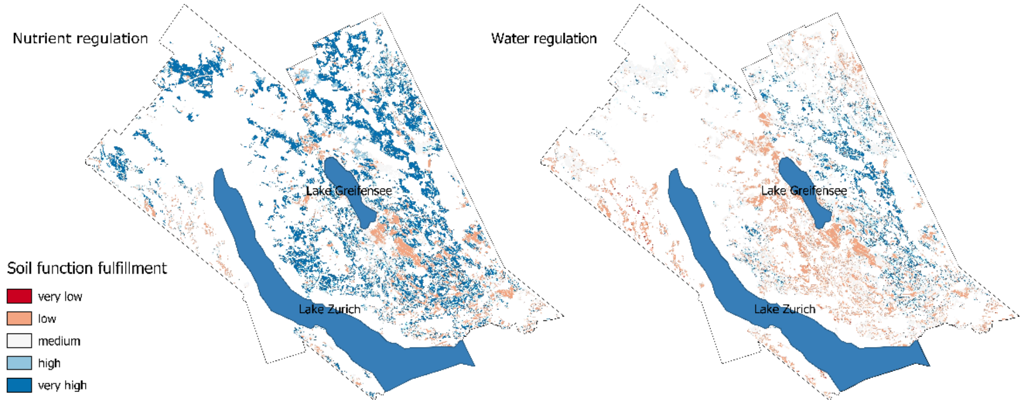

Figure 9: Soil function maps of water regulation function and nutrient storage capacity for the agricultural land in the Greifensee study area, Switzerland. (Reproduced with permission from Greiner, L, Nussbaum, M, Papritz, A et al. (2018). Assessment of soil multi-functionality to support the sustainable use of soil resources on the Swiss Plateau. Geoderma Regional, 14, e00181).

Quantitative Pedogenesis

As stated above, soil variability is a function of factors of soil formation. Since the publication of Jenny's state factors, pedologists have attempted to develop pedogenetic (mathematical) functions that could explain soil variability and, in many instances, predict or even simulate the soil. This has resulted in models of pedogenesis based on chemical and physical processes. The formulation of the process models and their applications depend on the scales in both the temporal and spatial dimensions. This modelling of the space-time continuum is determined by the complexity of the modelling process, which can be characterized in several ways. The first is based on computational complexity of the model, which ranges from purely qualitative (mental models) to highly quantitative, with the latter involving complex computer coding; the second is based on the complexity of the model structure, which distinguishes between mechanistic (highly complex) and empirical (simplified functional) models; and the last, but not least, is based on organizational hierarchy, which determines at which level a model is used to simulate the soil system. The hierarchical levels range from $i - 4$ (molecular-level processes) to $i + 6$ (global level). Each level is a subsystem of the level above it, therefore allowing room for upscaling.

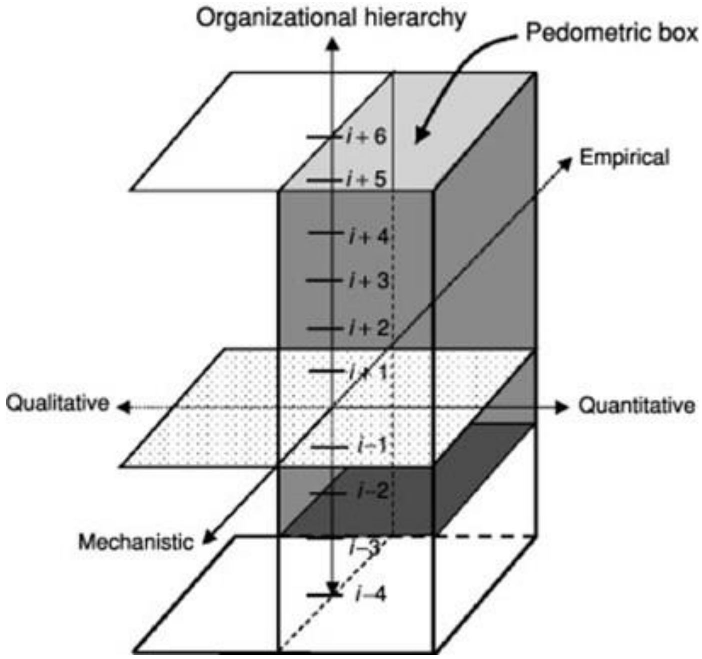

Figure 10: Pedometric box in the organizational hierarchy of pedogenetic modelling approach. (Adapted from Hoosbeek MR and Bryant RB (1992) Towards quantitative modeling of pedogenesis – a review. Geoderma 55: 183–210, with permission).

The three characterizations above are well illustrated in Figure 9. It should be noted that pedometrics has played a dominant role in the quantitative half of the organizational hierarchy, particularly from and above the $i - 3$ level, which is governed by material fluxes in the pores between the primary particles. Recent developments in process modelling focus on mechanistic stochastic simulations, particularly at levels $i - 3$ to i (Figure 9) and on how the latter can be upscaled to any of the higher hierarchies.

Soil processes

Soil processes that are responsible for differentiation of the soil profile have been modelled in various ways (Stockmann et al. 2018). Soil transformation processes such as soil physical and chemical weathering are modelled through empirical chronofunctions or by mass balance models linking the chemical composition of bedrock to soil profile properties. Translocation processes such as eluviation/illuviation and soil mixing are quantified by mathematical functions based on soil textural properties. Overall, these models require considerable input of measurement data, for example radionuclide data, and have many parameters to estimate. Other problems are that these models do not provide uncertainty, and have been derived using data from soil profiles or at the pedon scale (i.e. they have scaling issue to larger area). Pedometricians have contributed to solving some of these challenges by the spatialization of soil profile process models.

Soil-landscape evolution

Process modelling has enabled the development of soil–landscape evolution models. These models are quantitative models of the short- and long-term processes of soil formation and which predict the present-day variation of soil properties. Often these models are a combination of the understanding of process and empirical modelling. For example: a mechanistic pedogenetic model, based on a digital elevation model (DEM, Burrough and McDonnell, 2000), was used to simulate pedogenesis by a combination of several submodels: (1) rate of physical weathering as an exponential decline function of soil thickness; (2) rate of chemical weathering represented as a negative exponential function of both soil thickness and time; and (3) the movement of material as characterized by the diffusion transport model. The result of upscaling such an analysis is illustrated in Figure 10, whereby, after 10 000 years, soil accumulation is predominant in the gullies compared with the ridges, where soil erodes. See also Minasny et al. (2015) for further reading.

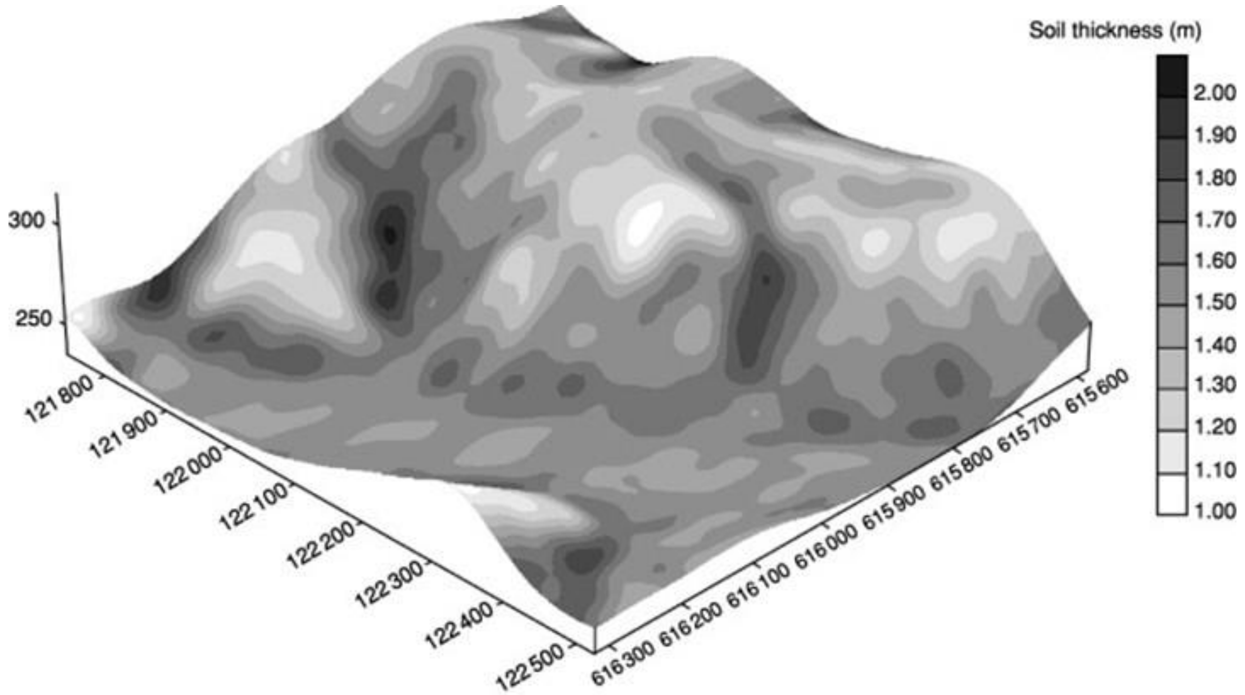

Figure 11: Soil formation in a landscape after 10 000 years. (Reproduced with permission from Minasny B and McBratney AB (2001) A rudimentary mechanistic model for soil formation and landscape development: II. A two-dimensional model incorporating chemical weathering. Geoderma 103: 161–179).

Recent research has shown that the main impediments of soil process modelling discussed above (i.e. data requirement, mechanistic understanding, number of parameters to estimate, scaling problems, uncertainty) may be tackled using statistical (data-driven) modelling, for example with state-space modelling and the space-time Kalman filter (Webster and Heuvelink, 2006). With increasing environmental concerns, the development and applications of dynamic landscape models under different land-use scenarios may well dominate pedometric research in the next few decades.

Conclusion and Outlook

Stepping back in time, pedometrics appeared from a need for quantitative and digital methods in soil science and emerged from research on numerical soil classifications and geostatistical modelling of spatial variation. During the past two decades, the considerable demand for soil information has motivated pedometricians to address other soil questions from a quantitative point of view: digital soil mapping, soil monitoring, quantitative pedogenesis and soil utility, soil functions and security and to diversify the tools and techniques that they use to carry out such research: soil spectroscopy, machine learning, spatio-temporal soil inventories and quantitative process-based soil-landscape models, among others. These efforts go beyond the scope of this short introductory chapter, but were described in 2018 in a book (McBratney et al. 2018) that attempts to cover almost all aspects of pedometric research.

While pedometric research is undeniably in an expansion phase and has now many areas of application, some gaps in knowledge and challenges need to be tackled. Wadoux et al. (2021) describe ten recent and longstanding pedometric challenges. They are classified into three aspirational categories, i) a better understanding of soil formation, ii) an improvement in methods for obtaining relevant digital soil data, and iii) an improvement of our ability to address demands by soil users. These challenges show that pedometrics can contribute to many areas of soil science, and is at the interface with many questions relevant to the sustainable growth of our societies.

References

Burrough A. B. and McDonnell R. A. (2000) *Principles of Geographical Information Systems*. Oxford university press, New York.

Brus, D. J., Kempen, B. and Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science*, *62*(3), 394–407.

De Gruijter, J. J., Brus, D. J., Bierkens, M. F. and Knotters, M. (2006). *Sampling for Natural Resource Monitoring*. Springer, Berlin.

Heuvelink G. B. M., Pebesma E. and Gräler B. (2017) Space-Time Geostatistics. In: Shekhar S., Xiong H., Zhou X. (eds.) *Encyclopedia of GIS*. Springer, Cham

Heuvelink, G. B. M. and Webster, R. (2001). Modelling soil variation: past, present, and future. *Geoderma*, 100(3-4), 269–301.

Hughes, P., McBratney, A. B., Huang, J. et al. (2018). A nomenclature algorithm for a potentially global soil taxonomy. *Geoderma*, 322, 56–70.

Ibáñez, J. J. and Bockheim, J. G. (2013). *Pedodiversity*. CRC Press, Boca Raton.

McBratney A. B., Mendonça Santos M. L. and Minasny B. (2003) On digital soil mapping. *Geoderma* 117, 3–52.

McBratney A. B., Odeh I. O. A., Bishop T. F. A., Dunbar M. S. and Shatar T. M. (2000) An overview of pedometric techniques for use in soil survey. *Geoderma,* 97, 293–327.

McBratney, A. B. and Lark, R. M. (2018). Scope of pedometrics. In B. Minasny, A. B. McBratney and U. Stockmann (eds.), *Pedometrics* (pp. 7–39). Springer, Cham

McBratney, A. B. and Odeh, I. O. (1997). Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma*, 77(2–4), 85–113.

McBratney, A. B., Budiman M. and Stockmann, U. (2018) *Pedometrics*. Springer, Cham.

Minasny, B., Finke, P., Stockmann, U., Vanwalleghem, T. and McBratney, A. B. (2015). Resolving the integral connection between pedogenesis and landscape evolution. *Earth-Science Reviews*, *150*, 102-120.

Nocita, M., Stevens, A., van Wesemael, B. et al. (2015). Soil spectroscopy: An alternative to wet chemistry for soil monitoring. *Advances in agronomy*, 132, 139-159.

Rossiter, D. G., Hewitt, A. E. and Dominati, E. J. (2018). Pedometric valuation of the soil resource. In B. Minasny, A. B. McBratney and U. Stockmann (eds.), *Pedometrics* (pp. 521–546). Springer, Cham.

Stockmann, U., Salvador-Blanes, S., Vanwalleghem, T., Minasny, B. and McBratney, A. B. (2018). One-, Two- and Three-Dimensional Pedogenetic Models. In B. Minasny, A. B. McBratney and U. Stockmann (eds.), *Pedometrics* (pp. 555-593). Springer, Cham.

Van Looy, K., Bouma, J., Herbst, M. et al. (2017). Pedotransfer functions in Earth system science: challenges and perspectives. *Reviews of Geophysics*, 55(4), 1199–1256.

Wadoux, A. M. J.-C., Minasny, B. and McBratney, A. B. (2020). Machine learning for digital soil mapping: applications, challenges and suggested solutions. *Earth-Science Reviews*, 210: 103359.

Wadoux, A. M. J.-C., Heuvelink, G. B. M., Lark, R. M. et al. (2021). Ten challenges for the future of pedometrics. Geoderma. 401: 115155.

Webster R. (1994) The developments of pedometrics. Geoderma 62, 1–15.

Webster, R. and Heuvelink, G. B. M (2006). The Kalman filter for the pedologist's tool kit. *European Journal of Soil Science*. 57, 758–773

Webster R. and Oliver M. A. (2007) *Geostatistics for Environmental Scientists*. (2nd Edition) John Wiley & Sons, Chichester.