Chapter 1
Scope of Pedometrics

Alex. B. McBratney and R. Murray Lark

"Those are my principles, and if you dont't like them well, I have others."

Groucho Marx

1.1 The Agenda of Pedometrics

Why do we need pedometrics? What is its agenda? Pedometrics addresses certain key soil-related questions from a quantitative point of view. The need for the quantitative approach arises from a general demand for quantitative soil information for improved economic production and environmental management. Pedometrics addresses four main areas which are akin to the problems of conventional pedology:

1. Understanding the pattern of soil distribution in character space – soil classification
2. Understanding soil spatial and temporal variation
3. Evaluating the utility and quality of soil
4. Understanding the genesis of soil

A.B. McBratney (✉)
Sydney Institute of Agriculture & School of Life and Environmental Sciences,
The University of Sydney, Sydney, NSW 2006, Australia
e-mail: alex.mcbratney@sydney.edu.au

R.M. Lark
British Geological Survey, Keyworth, Nottingham NG12 5GG, UK
e-mail: mlark@bgs.ac.uk

A.B. McBratney et al. (eds.), *Pedometrics*, Progress in Soil Science,
DOI 10.1007/978-3-319-63439-5_1

1.1.1 Unravelling the Structure of the Soil Character Space: Soil Classification

We all search for organising generalisations in the face of complexity. Many soil properties need to be observed to characterise a soil. The organisation of this soil complexity through classification has been problematic, probably because classification is an innate human faculty. Leeper (1956, p. 59) noted "*When scientists discuss methods of analysing a solution for phosphate, they are practical, reasonable, and unemotional. When the same men discuss the classification of soils, these virtues are likely to evaporate*". Pedometrics attempts *to resolve some of the polemic of soil classification by a* search for insight into the structure of soil character space. Multivariate and numerical taxonomic methods (Webster 1977a) attempt to unravel this structure, with a view to better prediction and understanding. Chapters 4, 8 and 9 explore these issues in more detail.

1.1.2 Understanding Soil Spatial and Temporal Variation

The soil skins the land, and its attributes vary spatially and temporally in a sometimes continuous, sometimes discrete and sometimes haphazard fashion. Using current technologies, we can still only measure most attributes of the soil at a finite number of places and times on relatively small volumes, and therefore statements concerning the soil at other places or times involve estimation and prediction and an inevitable uncertainty. Such prediction and estimation are required for inventory, assessment and monitoring.

One of the tasks of pedometrics is to quantify this inexactitude in order that it can be known and managed accordingly (Heuvelink and Webster 2001). Pedometrics also seeks insight into such spatio-temporal patterns using geostatistical and other spatial and temporal description and prediction tools. Chapters 10, 11 and 12 explore these issues in more detail.

1.1.3 Evaluating the Utility and Quality of Soil

The importance of knowledge and awareness of soil resources ranging from individual fields to the global scale is axiomatic. Over and above the problems of spatial and temporal variation, we must put objective value judgments on the utility of soil for specified purposes. This may involve encapsulating practical experience and developing objective rules for management. Conventionally, this has been called land evaluation, and quantification has been underway since Storie (1933) or earlier and has been developed further by inter alia Rossiter (1990).

More recently, the focus has moved to the environmental performance of soil (Adhikari and Hartemink 2016). The somewhat problematic concept of soil quality includes assessment of soil properties and processes as they relate to the ability of

soil to function effectively as a component of a healthy ecosystem (Doran and Parkin 1994). The quality of soil as part of an ecosystem may depend on its pedodiversity (Ibáñez et al. 1995). Soil can provide a number of valuable ecosystem services (Dominati et al. 2010) that require to be quantified. This is also reflected in the newly emerging multidimensional concept of soil security where soil plays an integral part in the global environmental sustainability challenges (McBratney et al. 2014). In recent years, pedometrics has contributed more and more to these areas, with the use of pedometric products for land suitability analysis (e.g. Kidd et al. 2015) and assessment of the change in the soil resource in space and time (e.g. Stockmann et al. 2015), for example. These and related issues are explored in Chap. 17.

1.1.4 Understanding the Genesis of Soil

Ultimately, pedometrics would attempt to provide quantitative models of soil formation. This is achieved by encapsulating knowledge of pedological processes in mathematical forms as an alternative to purely statistical approaches. The success of such a modelling approach depends, however, on pedological knowledge and the non-linearity of processes (Phillips 1998). The advantages of a successful modelling of soil formation are substantial and manifold. The three previously discussed agenda items of pedometrics, namely, soil classification, spatial and temporal variation and soil utility and quality, would be predictable from such a model. Therefore, attempts at such an approach have emerged in recent years. Hoosbeek and Bryant (1992) perhaps first outlined the problem. Minasny and McBratney (1999, 2001), Cohen et al. (2010) and Vanwalleghem et al. (2013) have provided the first substantive, but still rudimentary models, followed by more sophisticated models on the soil profile scale and soil landscape scale. This intriguing approach is discussed further in Chaps. 18 and 19.

1.2 Types of Models and Their Evaluation

Models are abstractions of reality. Harvey (1969) gives an early general discussion of models and quantification. Dijkerman (1974) discusses the kinds of models used in soil science generally. Because the soil is a complex and variable system, we represent it with a simpler or more abstract model. Originally, these models were descriptive mental models, but over time they have become more quantitative (Dijkerman 1974). In the context of pedometrics, with its quantitative impulse, a model generally corresponds to one of Webster"s dictionary definitions, i.e. *a system of postulates, data, and inferences presented as a mathematical description of an entity or state of affairs*. Underlying this general statement, there is a variety of model types which we shall now consider and then go on to discuss how to judge when models are performing adequately.

1.2.1 Types of Models

Models can be described in various ways by dividing them into a number of binary categories.

Models may be *qualitative* or *quantitative*. In traditional pedology models tend to be qualitative, whereas in pedometrics the tendency is for quantification. The main advantage of quantitative models seems to be that of reproducibility. Some would argue that quantitative models are more objective, but this is a difficult area from a philosophical point of view. So qualitative models may be considered inferior; however, they do contain knowledge and that knowledge may be captured in some objective way. Qualitative models can be made quantitative. So a qualitative model may be a good starting point for subsequent quantitative investigation, e.g. qualitative models expressing the expert knowledge of a group of experienced scientists and practitioners may be formally written down and used as hypothesis for quantitative testing.

Models may be *static* or *dynamic*. Pedometrics tends to focus on static models, whereas the models of soil physics are dynamic. In the future, pedometric models probably will be increasingly dynamic.

Models may be *empirical* or *mechanistic*. By empirical models we simply describe a phenomenon with as few parameters as possible without necessarily seeking to describe the mechanism underlying the phenomenon, whereas mechanistic models attempt to describe the mechanism (at some scale). For example, we could model the changes in moisture content at some location as a function of time using a purely empirical time series or transfer function model. Alternatively, we could describe the same observations using Richards' equation (Pachepsky et al. 2003) and various other soil physical mechanisms. The latter requires knowledge of basic soil properties, whereas the former does not. Whether one uses a mechanistic or an empirical model will depend on our level of process knowledge and the relative predictability of the models. Another way of describing empirical versus mechanistic models is to term them functional and physical models (Addiscott and Wagenet 1985). Functional is largely synonymous with empirical. Sometimes the empirical and mechanistic categorisation is confused with static and dynamic. Mechanistic models are usually dynamic but as an example given above, empirical models can be dynamic as well as static.

A very important distinction is to consider *deterministic* or *stochastic* models. Underlying Laplace's (1749–1827) statement, "*All the effects of Nature are only the mathematical consequences of a small number of immutable laws*", is the origin of determinism described by Laplace and quoted by Addiscott and Mirza (1998), "*An intellect which at any given moment knew all the forces which animate nature and the mutual positions of all the beings that comprise it, if this intellect was vast enough to submit its data to analysis, could condense into a single formula the movement of the greatest bodies of the universe and that of the lightest atom: for such an intellect nothing could be uncertain, and the future just like the past would be present before his eyes*". Stochastic models concern some

phenomenon that has randomness innate to its structure, whereas deterministic models have not. Many earth scientists do not readily accept that such randomness is inherent to nature. This unease recalls Einstein's famous saying "*The Lord does not play with dice*". No one has suggested that any soil process is constitutionally random unless we consider them at the quantum level. The ensemble averages are considered to be deterministic. So, for example, at the so-called Darcy scale, soil water phenomena can be described by a deterministic equation. However, given the inevitable lack of understanding and particularly lack of knowledge of soil processes, then randomness in models is a way of incorporating our ignorance. So in this sense, stochastic models may be seen as pragmatic. Up until the 1980s, stochastic models were rarely recognised or used in soil science. Dexter (1976) was one of the first examples. In hydrology on the other hand, stochastic models are abound (Yevjevich 1987). Pedometrics will probably make an ever-increasing use of stochastic models.

More recently, scientists have realised that deterministic models may have outcomes that look for all intents and purposes like stochastic ones – the outcomes appear random. These models describe processes which have a sensitive dependence on initial conditions. These are the so-called non-linear dynamic models, the basis of chaos theory (Gleick 1987). Such chaotic models blur the distinction between determinism and stochasticity (Addiscott and Mirza 1998). Phillips (1993, 1998) has suggested this kind of determinism as the basis for soil variability itself, i.e. soil formation is a chaotic process. Non-wetting phenomena in soil physics (Persson et al. 2001) are probably chaotic processes that cause soil variation. Minasny and McBratney (2001) give an example of a chaotic model and suggest how a chaotic deterministic model may be distinguished from a random one (Sugihara and May 1990).

As described above, models may be (1) qualitative or quantitative, (2) static or dynamic, (3) empirical or mechanistic (functional or physical) and (4) deterministic or stochastic. If we consider these four distinctions, then there are 16 combinations. Not all of them are sensible ones. In pedometrics, it is certainly worth considering the quantitative combinations and particularly those relating to stochastic and deterministic and empirical and mechanistic as shown in Table 1.1. So these four kinds of models are rather important in soil science. Another category has been

Table 1.1 Examples of quantitative models

	Deterministic		
	Certain	Uncertain	Stochastic
Empirical	Functional leaching model	Pedotransfer functions Fuzzy models	Regionalised variables geostatistics Markov chain models
Mechanistic	Richards' equation	Non-wetting phenomena Soil-landscape model	Molecular-scale diffusion Soil mechanics

added to the table, and we distinguish between certain and uncertain deterministic models. These models lie between purely deterministic and purely stochastic models. Examples of soil phenomena modelled in the various categories are given.

Pedometrics deals particularly with uncertain deterministic models and stochastic ones, i.e. models where randomness has to be introduced to deal with uncertainty (pedotransfer functions, Wösten et al. 2001) and/or randomness is seen as part of nature itself (soil as a realisation of random field – geostatistics). As such many pedometric models have a statistical basis. The uncertain deterministic, mechanistic and the stochastic, mechanistic models (e.g. soil mechanics, Manolis 2002) are not well developed in soil science or pedometrics.

One class of model that does not readily fit in to this scheme are the fuzzy models. They can be possibly thought of as uncertain deterministic, empirical models. They are probably best thought of as another categorisation, that of continuous or discrete states. Bárdossy et al. (1995) modelled soil water movement using the fuzzy approach as described in McBratney and Odeh (1997). Dou et al. (1999) modelled solute movement in a similar way. Pedotransfer functions (Wösten et al. 2001) which are largely linear or non-linear regression equations are empirical deterministic uncertain models, which will be discussed further in Chap. 7 of this book.

1.2.2 Critical Evaluation of Models and Their Parameters

Critical evaluatory procedures for models are needed to maintain the integrity of modelling and to ensure that the increasingly widespread use of models does not result in the propagation of misleading information. Generally speaking, models particularly quantitative pedometric ones, should be good predictors, while at the same time, they should not have too many parameters. The quality of prediction has to be tested with independent data sets by comparing predicted and observed values (Van Kuilenburg et al. 1982) or less satisfactorily with some kind of cross-validation procedure (Solow 1990).

The Franciscan friar, William of Ockham, was concerned about the number of parameters; he wrote in 1332, "*Pluralitas non est ponenda sine necessitate*", which roughly translates as "*Don't make things more complicated than you have to*". This is called Ockham's razor. The two ideas of quality prediction and minimising the number of parameters quantitatively are reflected in measures such as the Akaike information criterion (Webster and McBratney 1989).

Addiscott et al. (1995) pointed out that no model can be validated in the sense that it has been unequivocally justified. All that can be achieved is to show how small the probability is that the model has been refuted. Whether this probability is acceptable remains a subjective decision. In general, the further the data used for parameterisation are removed from the data to be simulated, the better. Problems can arise in both parameterisation and validation if the model is non-linear with

respect to its parameters, and the latter have appreciable variances. Parameterisation and validation become more difficult as the complexity of the model or the scale at which it is used increases.

1.3 Methods × Problems

The gamut of pedometrics can also be thought of as a two-way table. The rows are the problems that pedometrics can possibly address, and the columns are the potential mathematical and statistical methods that could be applied to such problems. Some cells of the table will be richly endowed, whereas others will require thought and invention to come up with new approaches. This book, which is essentially a contribution to soil science, and thereby earth, environmental and ecological sciences, is arranged with a problem focus rather than on the methods themselves which would be more appropriate to an applied statistics text.

1.3.1 The Problems of Pedometrics

"Sir Walter Blunt, new lighted from his horse.
Stain'd with the variation of each soil
Betwixt that Holmedon and this seat of ours;"

William Shakespeare, Henry IV Part 1 Act 1 Scene 1.

Pedometrics arises principally from this common observation that the soil is spatially variable. The soil varies laterally and with depth, and this variation has implications for its use and management. This is not a new insight. The rabbinical biblical commentators on the book of Genesis discussed the question of how much water is needed to sustain plant growth:

"How much rain must descend that it may suffice for fructification? As much as will fill a vessel of three handbreadths. This is Rabbi Meir's opinion. Rabbi Judah said: in hard soil, one handbreadth; in average soil, two; in humid soil three."

Midrash Genesis Rabbah XIII, 13.

Later in history, Thomas Tusser, in his *One Hundred Points of Good Husbandry* published in England in the sixteenth century, noted pithily that *each divers soil hath divers toile*. More recently, the development of agronomy, soil management and engineering has shown that the spatial variation of the soil means that its suitability for different purposes will vary in space. This is important for the planning of infrastructure, the government agencies making planning decisions about land use and, in recent years, the individual farmer trying to improve the efficiency of cropping systems by managing inputs in response to fluctuations in crop requirements at within-field scales.

We can think of other problems where the spatial variation of the soil will be of practical significance. Throughout the industrialised world, it is recognised that the soil is at risk of pollution from unregulated inputs and emissions. If we are to monitor this problem and focus protection and remediation where the problem is most severe, or the soil most vulnerable, then we must be able to detect changes in pollutants, which arise from complex processes of deposition and transport, against a background of complex intrinsic variation of the soil. A similar problem arises when we set out to monitor changes in the carbon content of the soil to evaluate its significance as a source and sink for greenhouse gases.

Soil variation is linked to some pressing problems, but it is also pertinent to other questions. Faced with variation in the soil cover of any terrain, the natural response is to ask how and why it arose. Many soil scientists would not have become interested in their subject if the soil was more or less uniform in space. To understand the variation of the soil at scales from the aggregate to the continent and the relationship of this variation to that of the vegetation which it supports is a basic scientific challenge. The better we understand these basic questions, the better equipped we will be to address the practical ones. Pedometrics is concerned both with meeting the practical requirements of soil management and also with generating insight into how and why the soil material varies as it does.

These, in summary, are the problems of pedometrics. In the remainder of this section, we want to set them in a general framework, with examples, as a prelude to introducing the methods which are used in their solution.

1.3.1.1 Prediction

The first type of problem is prediction. The basic question requiring a prediction has the form:

1. What soil conditions pertain at position **x**?

Vector **x** may contain two or three Cartesian coordinates which define a location in space and possibly a further number which defines a time. This very general question may be refined in different ways:

1.a What is the value of soil variable s at **x**, $s(\mathbf{x})$, *given a set of observations of the variable at other locations?*

The variable $s(\mathbf{x})$ might be the concentration of available potassium in the topsoil at a location in space $\{x,y\}$. Note that some finite volume of the soil is implicitly of interest. We call this volume the geometric support of $s(\mathbf{x})$.

This problem arises because very few if any soil properties of direct practical interest will have been measured exhaustively across a region. The soil is a continuum, and the constraints of costs and time mean that the soil may only be sampled and measured at a few sites. Information will inevitably be required about soil which has not been measured directly.

Reflection on this problem makes it clear that the best answer we can obtain to our question, short of actually sampling the soil at **x**, will be an estimate of $s(\mathbf{x})$,

$\widehat{s}(\mathbf{x})$, a number with attendant uncertainty. The way in which this uncertainty is quantified and controlled is a critical issue in pedometrics, to which we will return repeatedly.

A second refinement of the question is possible:

1.b. What is the value of, $y(\mathbf{x})$*, a mathematical function of soil property s at* $\mathbf{x}$*?*

This may seem at first sight like a pedantic variant of question *1.a.*, but it raises an important issue, and a practical one. We may be able, for example, to obtain a reasonable estimate of the clay content of the soil at $\mathbf{x}$, but are really interested in its available water capacity. A reasonable estimate of this latter quantity might be obtained as a function of the clay content. If y is a linear function of s, then the transformation of the estimate $\widehat{s}(\mathbf{x})$ to an estimate $\widehat{y}(\mathbf{x})$ is simple. If the relationship is not linear, then the uncertainty of the estimate $\widehat{s}(\mathbf{x})$ must be accounted for.

Other variants on the simple prediction problem are possible. For example:

1.c. What is the difference between $s\{x, y, t_i\}$ *and* $s\{x, y, t_j\}$*?*

where the third term in the vector denotes a time. This question will arise in environmental monitoring. If $s\{x, y, t_i\}$ is the concentration of a pollutant at x at time t_i, then the difference may be a measure of the success of a soil remediation campaign or the environmental impact of a change in regulations.

This problem will vary in form. A critical question is whether it is possible in principle to measure both variables – given the support of $s\{x, y, t_i\}$, does the disturbance of the initial sampling prevent a meaningful measurement from being made at $\{x, y, t_j\}$? This problem presents pedometricians with interesting challenges. Since politicians are increasingly interested in reliable estimates of the change of organic carbon in the soil, in response to the Kyoto protocol, the problem is also timely and topical.

So far, we have considered the soil over a small volume about a notional two- or three-dimensional location. In practice, we may be concerned more often with predictions about larger parcels or blocks of land, regularly or irregularly shaped. Such parcels may constitute management units, for example. The general problem, then, is

1.d. What is the value of $s(\mathbf{X}) = \int_{x \subset X} s(x)\mathrm{d}x$?

The integral implies that the new variable is effectively the average of all the notional point values, $s(\mathbf{x})$, in the region. Thus, we might be asked, 'what is the mean concentration of lead in the soil at this former factory site' or 'what is the mean concentration of available phosphorous in the soil of this field'? Since our prediction will be based on observations which are effectively point samples on volumes of soil, very small by comparison to the block $\mathbf{X}$, the problem involves generalisation from one spatial scale to a coarser scale. This is sometimes called 'upscaling' or 'aggregation'.

The change of scale can cause problems. Consider the following:

1.e. What is the value of $y(X)$ where y is a mathematical function of soil variables w and z at the scale of point observations $s(x)$ and $w(x)$?

Again, y might be the available water capacity of the soil and w and s bulk the density and clay content, respectively. If w and s are measured at all locations, then we might evaluate y at all these locations then aggregate these to the coarser unit. If the two variables have been measured independently, then we can combine two aggregated values, $s(X)$ and $w(X)$, but only if the functional relationship is linear. Otherwise, it is necessary to make some inference about the joint variation of variables s and w at the original point scale; this is sometimes called 'downscaling' or 'disaggregation'.

We have raised the issue of uncertainty already in the context of simple prediction, but uncertainty may sometimes be addressed directly in a pedometric problem. Consider the following:

1.f. What is the risk that $s(\mathbf{x})$ exceeds at threshold t?

The variable s might be the concentration of a pollutant and t a regulatory threshold. For example, the limit for concentration of lead in the soil set by the ANZECC guidelines is 300 mg/kg.

A type 1a question requires a simple prediction, $\widehat{s}(\mathbf{x})$, but in the context of the present problem, this is generally not adequate. If we act on the prediction, there remains the risk that contaminated soil is left untreated (because $\widehat{s}(\mathbf{x}) < t$ but $s(\mathbf{x}) > t$) or that expensive remediation is applied to land unnecessarily (because $\widehat{s}(\mathbf{x}) > t$ but $s(\mathbf{x}) < t$). If our predictions are unbiased, i.e. on average $\widehat{s}(\mathbf{x}) = s(\mathbf{x})$, and the costs of an error in one direction are more or less equal to the costs of an error in the other direction, then all we can do is try to reduce the uncertainty of our predictions as far as possible. Often, however, the costs of an error in one direction are much steeper than the other (e.g. the fines for leaving land unremediated may be large compared to the costs of remediation). In these circumstances, the best decision must account for the risk that $s(\mathbf{x}) > t$, given all available knowledge implicit in the prediction $\widehat{s}(\mathbf{x})$.

This latter problem was expressed in terms of a near-point support at $\mathbf{x}$. Answering the final prediction-type problem:

1.g. What is the risk that the mean of s over region $\mathbf{X}$ exceeds threshold t?

raises further problems for the pedometrician. Since management decisions will generally be made about a parcel of land, practical questions will often be framed this way.

1.3.1.2 Inventory and Allied Problems

A second category of pedometric problems can be recognised. If the problems in category 1 are variants of the question 'what conditions pertain at $\mathbf{x}$', then class 2 consists of variants on:

2. Over what subregion does conditions pertain?

At its simplest, the problem may be one of inventory – enumeration of the sites over which certain conditions are found. Thus, for example:

2.a. Over what region, X', does the value of soil variable s fall within the range $s_1 < s < s_2$?

As with *questions of* type 1.a., there is uncertainty attendant on any answer to this question since the soil properties are not measured exhaustively.

This problem might be posed by the land manager who wants to know where the depth of the soil over the underlying rock is large enough to permit the growth of a tree crop. It might also be asked by the farmer who wants to know where soil pH is likely to limit certain crops. In reality, a more complex problem might be posed:

2.b. Over what region, X', does some function of soil variables $y = f(s, w)$ fall within the range $s_1 < y < s_2$?

If we want to identify areas where the likely erosion losses exceed some threshold, then we may have to compute some non-linear function of soil properties like the universal soil loss equation (Wischmeir and Smith 1978). This poses similar problems to the analogous prediction problem 1.b.

Temporal monitoring of the soil may also generate problems with an inventory flavour such as:

2.c. Over what region, X', does the change in soil property s from time t_1 to time t_2 fall within the range $s_1 < s < s_2$?

This problem arises when the policy maker wants to know over what proportion of a landscape the concentrations of pollutants in the soil are diminishing or where the organic carbon content is increasing.

1.3.1.3 Decision Support

Prediction and inventory are tools. Managers require that the answers to these basic questions are integrated in a way which aids decision-making directly. This is the sphere of decision-support systems, and pedometric problems arise. There are two general types of problems:

3.a. What is the optimum management strategy over region X' to achieve goals A subject to constraints B?

So, for example, what is the optimum nitrogen rate to prescribe for a particular parcel of a field to maximise the economic return to the producer subject to the constraints that emissions to the environment through leaching and denitrification do not exceed some threshold? The problem requires process models of an appropriate level of sophistication to describe the whole system under different scenarios. Information on soil properties within the parcel of concern will also be required – subproblems of type 1.c., for example. The optimisation subject to a constraint

may be done numerically – i.e. by computing several runs of the system model in an ordered way to find the scenario which best meets the goals subject to the constraints.

In practice, decisions have to be made under uncertainty about many critical conditions, e.g. weather in the case of crop management. We then have a problem;

3.b. What is the optimum management strategy over region X′ to achieve goals A subject to constraints B and given that the conditions C have a particular statistical distribution?

1.3.1.4 In Pursuit of Insight

The problems so far have had a strongly practical flavour, but pedometric problems include basic scientific questions where a quantitative account of the spatial variation of the soil is required. We can identify such a problem:

4.a. Which factors appear to determine the lateral and horizontal spatial variation of soil property s?

This question implies a spatial scale – are we concerned with variations at the scale of microbial activity or geomorphic processes or some range between? This general problem may be addressed using analytical methods, but these always rest on assumptions which may not be realistic. An approach to the problem can never be driven purely by data analysis, however. Our investigation will be most fruitful if it is structured around a hypothesis. The approach may be statistical:

4.b. Is hypothesis H *about the causes of soil variation supported by the given observations of properties s...?*

or structured around a mechanistic model:

4.c. Given a process model linking input variables s, w and output y, can this particular set of observations of the variables be held to validate the model?

1.3.2 The Methods Pedometrics Uses

These questions are addressed by pedometrics. We now offer an overview of the pedometric methods that are treated in more detail in later chapters.

1.3.2.1 Statistical Prediction and Modelling

Random Variables

Ideally, soil scientists would like to base pedometric methods on quantitative understanding of soil processes. The ideal way of predicting the value of a soil property at location x would be to enumerate factors of soil formation which

pertain at x, i.e. the climate, the organic influences, the relief and parent material and the development of these factors over the time in which the soil material has developed. This information would then be combined with knowledge of the relevant soil processes in order to predict the soil property of interest. This recalls the determinism of Laplace, who held that if we know the momentum and position of all particles in the Universe at a given time, then all future changes are predictable. But modern physics has had to abandon Laplace's ideals, and neither is it an option for pedometrics. First, it is clear that our knowledge of the soil-forming factors operating over time at any location is very incomplete, and some of these factors will be unpredictable in principle. For example, a grazing *Megaloceros* in the Pleistocene disturbing a patch of metastable soil on a periglacial hillslope might influence the course of local solifluction events and play a large part in determining the clay content of the topsoil at positions downslope which a pedometrician later wants to predict. Second, many soil-forming factors have a complex and non-linear effect on the development of the soil and will interact. As a result small errors in our information about critical soil-forming factors may render our predictions quite inaccurate. This sensitivity to initial conditions recalls chaos theory. The solution to the problem is usually to replace our notional deterministic model of the soil with a statistical model. This latter model may have a structure which reflects the knowledge of soil-forming processes contained in the deterministic model, while representing the relationship between these factors and soil properties statistically.

At its simplest, we regard the value of a soil property s at location $\mathbf{x}$ as the outcome of a random process s. To do this is to treat the soil property as if it were the outcome of a process such as a toss of a coin or a roll of a die. At first glance, this is a deeply paradoxical approach for the scientist. By definition, the outcome of a random process is uncaused, and yet we know that soil properties are caused by processes which we can list and of which we have, more or less, a sound scientific understanding.

Treating a soil property as a random variable is an assumption. That is to say, we know that the soil property is not strictly the outcome of a random process, but that certain conditions permit us to treat it as such for certain purposes. More generally, we do not necessarily treat all the variation of the soil as random, but rather may limit this assumption to a component of the variation which a simplified mathematical description of the variation as a whole cannot account for.

When statistical models are used to describe the variation of the soil, there are broadly two conditions which justify the assumption that data on the variables of interest are outcomes of a random process. In the first instance, we may treat a set of measurements of a soil property as random variables if the selection of the sample has been done in a random way. This is the basis of design-based statistical analysis where data are obtained in accordance with a design which does not specify where the soil is sampled to collect a specimen for analysis, but which allows us to state in advance only the probability that a particular location will be sampled. The second situation is subtly different. Here, the assumption is that a set of measurements of the soil in space and/or in time may be regarded as a realisation of an assemblage of random variables which have a structure in space and/or time which has sufficient

complexity to encapsulate at least some of the features of the spatial variations of the real variables. In this case, we assume an underlying random process and that all observed and unobserved values of the soil within the region of interest constitute a single realisation of this process. This approach allows us to predict the values of the soil at unobserved locations from the values which have been observed in a way which is in some sense optimal. This approach is the basis of the so-called model-based analysis.

We must review some properties of random variables. A random variable is characterised by an underlying probability density function, (pdf) $p(S)$, such that the probability that a given value lies in the interval S_1 to S_2 is given by the integral

$$\int_{S_1}^{S_2} p(S)\mathrm{dS}.$$

It follows that the integral over the interval $-\infty$ to $+\infty$ is exactly 1. From the pdf, we may also define the distribution function for $S, f(S)$:

$$f(S) = p\,[S \leq s] = \int_{-\infty}^{s} p(S)\mathrm{dS}.$$

If all we know about a particular value of a soil property is that it is a realisation of a particular random variable, then our best estimate of the value is the statistical expectation $E[S]$, where

$$E\,[S] = \int_{-\infty}^{\infty} \mathrm{S}\ \ \mathrm{p}(S)\mathrm{dS}.$$

The expectation of a random variable is also known as its mean or first-order moment. Higher-order moments may also be defined. For example, the variance σ^2

$$\sigma^2 = E\left[\{S - E\,[S]\}^2\right] = \int_{-\infty}^{\infty} S\{S - E\,[S]\}^2 p(S)\mathrm{dS}.$$

If we assume, as is often done, that our soil property is a realisation of a normally distributed random variable, then these two moments alone are sufficient to characterise it and may be estimated from data.

The Linear Model

Simple random variables alone are of relatively little use for pedometric purposes. Ideally, we incorporate them into a statistical model to account for that variation which a simplified predictive equation cannot explain. Consider as an example the soil property organic carbon content. Many factors will determine this variable. One

such factor may be clay content, since the clay fraction of the soil may protect some of its organic matter. In addition, heavier soils which retain more moisture may have a larger input of fresh organic material in litter, roots and detritus from more vigorous vegetation. These and other considerations may lead us to expect that larger organic carbon contents of the soil may be associated with larger clay contents, although other factors will also be important and may mean that a good deal of observed variation in organic carbon content cannot be explained by a predictive relationship to clay content.

At its simplest, such a relationship might be a linear one of the form

$$\mathrm{OC} = a + b.\mathrm{Clay} + \varepsilon.$$

Here b is the slope of the linear relationship and a the intercept. The last term is the error or residual term, which sums up the other factors which determine organic carbon content of the soil. In statistical modelling, it is this term which we treat as a random variable. If the variability of this last term is small relative to the variability of organic carbon overall, then the model may be useful for predictive purposes. Under certain circumstances, we may estimate the parameters a and b from observations of the organic carbon and clay content of the soil by finding the values which minimise the deviations:

$$\mathrm{OC} - (a + b.\mathrm{Clay}),$$

in effect by minimising the variance of the error term.

The reader may have recognised that we have just described the ordinary least-squares regression model. This is a special case of the general linear model:

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is a vector of observations of a variable (which we may wish to predict) and matrix $\mathbf{X}$ contains a set of variables on which we assume the variable in y to be dependent. Vector $\mathbf{b}$ contains a set of coefficients, and $\boldsymbol{\varepsilon}$ contains realisations of a random variable, the error term.

Each column of matrix $\mathbf{X}$ corresponds to a predictor variable. These may be continuous variables (like clay content) or indicator variables which take the value 0 or 1 indicating whether or not an observation belongs to a particular class.

Consider a case where we wish to model the effect of parent material on soil properties. If clay content is the variable of interest, then different classes of parent material may weather to produce material of different textures. If g parent material classes have been defined, so that any one observation belongs to exactly one of the g classes, then the solution of the general linear model will be given by a vector of coefficients equal to the class means. Other variables, continuous or indicator, may be combined in the model. This general linear model (GLM) (McCullagh and

Nelder 1989) is equivalent to the familiar analysis of variance on the clay data testing the differences among the class means. In a similar vein, statisticians have recognised that nature is often not linear and have developed a class of non-linear models analogous to the GLM called the general additive models (GAMs) (Hastie and Tibshirani 1986, 1990).

Random Functions

A random function is an assemblage of random variables and is defined as a function of the location in space, $S(x)$. This function is assumed to have certain properties which characterise its behaviour over a set of all possible locations within some region of interest. A set of observations of a soil property at n locations in this region, $s(x_1)\,,\ \ldots\ s(x_n)$, is a sample of one realisation of this random function. We can only make progress by assuming that there are some restrictions on the joint distribution of $S(x_1)\,,\ \ldots\ S(x_n)$ which allows us to make inferences about it from the assemblage of observations.

We have introduced the concept of a pdf and a df of a random variable. We may similarly define the joint df of a random function over a set of locations:

$$F_{\{x_1,\ldots,x_n\}}\left(s_1,\ldots,s_n\right) = \text{Prob}\left[S\left(x_1\right) \leq s_1,\ldots,S\left(x_n\right) \leq s_n\right].$$

The simplest assumption which we might make about a random function is that the joint df for a configuration of sample points depends only on their relative distribution in space and not on their absolute position. Thus,

$$F_{\{x_1,\ldots,x_n\}}\left(s_1+\mathbf{h},\ldots,s_n+\mathbf{h}\right) = F_{\{x_1,\ldots,x_n\}}\left(s_1,\ldots,s_n\right),$$

where $\mathbf{h}$ is a vector defining a spatial separation or lag.

This is the assumption of strict stationarity. In practice, we work with less restrictive assumptions. One is to assume that only first- and second-order moments of the joint distribution are invariant with a shift in position. Thus, the mean is assumed to be constant:

$$E\left[S\left(x_i\right)\right] = \mu \text{ for all } x_i.$$

The covariance of any $S(x_i)$ and $S(x_j)$:

$$E\left[\left\{Z\left(xi\right) - E\left[Z\left(xi\right)\right]\right\}\left\{Z\left(xj\right) - E\left[Z\left(xj\right)\right]\right\}\right]$$

is also assumed to depend only on the interval $(x_i - x_j)$. If the joint df of the random function is multivariate normal, then strict stationarity is equivalent to this second-order stationarity since there are no higher-order moments.

Even this weaker version of stationarity may be too restrictive. In pedometrics, we often work with the assumption of intrinsic stationarity. This has a first-order assumption:

$$E\left[S\left(x_i\right) - S\left(x_j\right)\right] = 0 \text{ for all } x_i \text{ and } x_j,$$

and a second-order assumption that

$$E\left[\left\{S\left(x_i\right) - S\left(x_j\right)\right\}^2\right] = \gamma\left(x_i - x_j\right),$$

where γ is a function of the separation vector $x_i - x_j$. This function is the variogram and is widely used in pedometrics. It may be estimated from data and modelled with an appropriate parametric function. It is also possible to model the joint spatial variation of two or more variables in terms of random functions with variograms and cross-variograms. This is discussed in more detail in Chaps. 10 and 21. Having obtained a variogram model, this may be used in determining an optimal estimate of the value of a soil variable at an unsampled site as a weighted average of the values at neighbouring sites. This is the technique known as kriging.

State-Space Models

Consider a soil property which evolves over time. We may be able to express this change quantitatively by a model of the form

$$s_t = f\left(s_{\{t-i\}}\right)_{\{i=1,\ldots,n\}} + \varepsilon_t.$$

This states that the observed value is a function of past values, with an error term – a random variable which describes other factors which the model cannot account for. This is the so-called state equation. We recognise that our measurement of s, which we may make on several occasions, is made with error. The measurement process is described by a linear measurement relation:

$$u_t = g_t\left(s_t\right) + \varepsilon_t$$

where ε_t is an error term. The technique of Kalman filtering combines these two equations so that s_t may be estimated from a series of measurements. The state-space approach is of particular interest because the state equation can be a process model which incorporates our best scientific understanding of the processes whereby s is evolving over time. This state-space model may also be incorporated into a spatial model. In recent years, this approach is increasingly used by soil scientists in alliance with computational advances (e.g. Heuvelink et al. 2006; Huang et al. 2017).

Point Processes

There is another kind of process in the soil which we may usefully describe by or compare with probabilistic models of spatial variation. This is the distribution in space of discrete events. Consider, for example, the distribution in the soil of eggs of a nematode such as *Heterodera*. Soil scientists and agronomists might ask the pedometrician how quantitative analysis might illuminate the process by which the eggs are distributed. Does the distribution appear to be controlled by soil variation? Is it 'patchy'? We may make useful progress by comparing the observed distribution of eggs with what might be expected if they were scattered at random. In general terms, two departures from simple spatial randomness (SSR) are possible. The first is a so-called contagious distribution where occurrences of the event are more 'clumped' than SSR. This might occur if the nematodes grow and reproduce preferentially in areas of the field where certain soil conditions prevail and so tend to be aggregated here. The second departure from SSR is called overdispersion where the events are distributed more evenly than is expected of a random process. This might occur if competition between the adult nematodes is so intense that they tend to disperse in search of resources. Distinguishing SSR from contagious or overdispersed spatial patterns is not always simple. It is necessary to allow for the fact that we might only observe a fraction of the events in a field by sampling. Furthermore, the spatial distribution of a process might be complex with features such as patchiness being observed at particular spatial scales. Walter et al. (2005) used spatial point-process statistics to model lead contamination in urban soil. More conventional approaches are discussed in Chaps. 10 and 12. Glasbey et al. (1991) modelled the three-dimensional pore space within soil aggregates with randomly positioned overlapping spheres. Pore structure modelling is discussed further in Chap. 6.

Basis Functions and Decompositions

Any ordered set of n measurements on soil can be written down in a $n \times 1$ array or vector. Two instances are considered here. In the first, the vector contains measurements of one soil variable at n locations; here we consider regularly spaced locations on a linear transect. Location determines the order in this vector. In the second, the vector contains measurements of n soil properties from one location. A standard, though arbitrary, order of the properties is determined in advance.

It is easy to imagine a 3×1 vector as a point in the three-dimensional rectilinear space of our everyday experience, the three values being Cartesian coordinates relative to some arbitrary origin at {0,0,0}. In fact this is a special case of a general class of vector spaces. A vector space is a set of vectors with particular properties, so the set of all possible vectors of measurements of the volumetric water content, pH and clay content of the soil constitute one vector space with particular properties. All these soil variables are real numbers and vary continuously. This vector space is normed, i.e. we can define a distance between two vectors $\mathbf{x}$ and $\mathbf{y}$; in this case, a natural norm is the Euclidean distance $[\mathbf{x}\mathbf{y}^T]^{1/2}$.

Vectors of the first kind, where the soil properties vary continuously and with finite variance, and of the second kind, where the soil properties are continuously varying real numbers, occupy a subset of Hilbert space. The reader is referred to mathematical textbooks for a rigorous account of Hilbert space. For our purposes, the key fact is that Hilbert spaces can be described in terms of basis vectors. If $\mathbf{x}$ is any vector in such a space $\mathbf{X}$ and the basis vectors of $\mathbf{X}$ are $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, then we may express $\mathbf{x}$ by

$$\mathbf{x} = \alpha\mathbf{a}_1 + \beta\mathbf{a}_2 \ldots$$

If the basis vectors are orthogonal, i.e. $\mathbf{a}_i\mathbf{a}_j^T = 0$ for any $i \neq j$, then n such vectors provide a complete basis for an n-dimensional space $\mathbf{X}$. As an intuitive example, note that we can characterise the position of any point in the 3D space of daily experience in terms of any three coordinates from a system where the axes are mutually perpendicular.

The reason for this mathematical digression is that pedometricians can often gain insight into data expressed as one or more vectors by expressing them as combinations of basis vectors. There are two general approaches. First, we may find the coefficients which express our data in terms of certain predetermined basis vectors. The analysis of the resulting coefficients may be informative. Second, we may determine both the basis vectors and the coefficients in our analysis where the former are chosen to illuminate how the data are distributed in vector space.

The first of these approaches is exemplified by Fourier analysis and wavelet analysis. The vectors here are of the first type (i.e. one vector represents one soil property measured at different locations). In Fourier analysis, we decompose a continuous variable (or its discrete sampling) into additive combinations of sine functions of different frequencies (where spatial frequency has units of distance^{-1}). The Fourier coefficients are complex numbers and so can convey two pieces of information, in this case, the amplitude of the particular component (the height of a peak over a trough) and the phase (how the peaks are aligned relative to our sample points). Analysis of these coefficients can show us how the variation of the soil property is partitioned between spatial frequencies, that is to say, between fluctuations over short distances in space and longer-range fluctuations. This may be informative. However, if our data show changes in the nature of their variability in space, non-stationarity in the sense of our discussion above, or show marked singular features or discontinuities, then the decomposition of the data on a basis of uniformly oscillating sine functions is not natural, and the coefficients will not yield a simple interpretation. For this reason, Fourier analysis has not been widely used in pedometrics. Where it has been found useful, the soil variation has been dominated by a more or less uniform periodic component such as the gilgai pattern analysed by Webster (1977a, b), McBratney (1998) and Milne et al. (2010).

Wavelet analysis is related to Fourier analysis in that data are decomposed on a basis of oscillating functions which respond to soil variation at different spatial frequencies. However, wavelets only oscillate locally and damp rapidly to zero so a single wavelet coefficient only describes the soil variation in a particular neighbourhood (the size of which depends on the spatial frequency).

A complete wavelet basis consists of dilations of a basic (mother) wavelet function which respond to different spatial frequencies and translations (shifts) of these which describe different neighbourhoods. The wavelet transform is a new tool in pedometrics, but a very promising one, and we discuss it further in Chap. 15.

When we have several data vectors of the second kind, observations on several soil variables at different locations, then we can define an orthogonal basis for the vector space in which these are distributed. An orthogonal basis is effectively a new set of coordinates for the vector space obtained by rotating the coordinate system of the original space. It is possible to find a basis such that the new variables which are defined are uncorrelated and so that one is the most variable such transformation of the original data which is possible, the second is the next most variable transformation orthogonal with the first and so on. These new variables are called principal components. If some or all of the original variables are substantially correlated, then it may be found that much of the original variability in n-dimensional space may be accounted for by fewer than n of the principal components. This may help the exploration of the multivariate structure of large and complex data sets and give insight into the joint variation of many soil variables. This is explored further in Chap. 4.

Classification and Allocation

Faced with many variable objects, the instinctive human response is to group them together into a manageable number of classes about which meaningful generalisations are possible. Soil classes may be formed by interpretation of profile characteristics in terms either of the evidence which they offer of the processes of soil genesis or the behaviour of the soil for practical purposes (e.g. drainage classes). This is principally a qualitative activity in which the expertise of the soil surveyor or pedologist is brought to bear on the problem. Pedometrics can offer two principal aids to the process:

1. Clustering. Consider again the vector space defined by several soil variables. If the space which is occupied by observations is more or less evenly populated, then it is clear that the recognition of classes defined on these properties will be difficult and somewhat arbitrary. We could, for examples, divide our observations into a number of more or less similar subvolumes of the vector space, but there are probably very many possible subdivisions which are of similar compactness as measured by the internal variability of the classes. A vector space occupied by data on soil which consists of a few distinct ‘types’ is likely to contain distinct ‘lumps’ or clusters of observations, since observations which are similar will be near to each other in the vector space. In the basic conceptual model which underlies classification, observations resemble typical or central members of the class to which they belong, although not exactly, and so should be clustered in the vector space defined on important soil properties.

Finding clusters in a vector space of many dimensions is a task for the computer. We will discuss the problem in more detail elsewhere. Suffice it to say that the key to the task is to define a model of within- and between-class variation in terms of a norm in vector space which is appropriate to the actual multivariate variability of the data.

2. Allocation. Having formed a classification, we may wish to allocate a new observation to the appropriate class. This is relatively straightforward where the classes have been defined (perhaps by clustering) in an n-variate vector space, and we have a new n-variate vector to allocate. This is often not the case. If we have measurements of one or more secondary variables which differ between the classes even though they are not definitive, then it may be possible to allocate an observation to a class using these. For example, we may have measurements of the reflectance of the soil surface in several frequency bands obtained by a remote-sensing device. Using several sites where the reflectances are known and the soil class, rules may be determined for predicting the soil class at sites where only the reflectances are known. Discriminant analysis due to Fisher is one technique for deriving these rules, which are formulated in terms of a norm in the vector space. In recent years, methods such as neural networks have been developed to solve the same problem.

   Classification and allocation are discussed further in Chaps. 8 and 9.

Modelling

A soil scientist may propose a model of some soil process which describes it quantitatively. The pedometrician may be interested in using this model to predict the behaviour of the soil in space and time. This entails four principal tasks:

1. Calibration. A model will often contain parameters – quantities which are constant in any one instance. Ideally, the values of parameters will be deduced from first principles, but usually it is necessary to estimate some or all of a set of parameters, some of which may bear a complex relationship to underlying soil properties (e.g. the 'tortuosity' of pathways for solute movement through the soil which depends on soil structure at different scales) or are somewhat artificial, not bearing a simple physical interpretation (e.g. the 'permeability' parameter of some functional models of solute leaching (Addiscott and Bailey 1991). Calibration requires that some data are available on the basic soil variables which are inputs to the model and those which are outputs. The temptation to estimate parameters by least squares must (generally) be resisted because this assumes that only the output variable is subject to random error, and this may often not be the case.
2. Validation. A model which generates quantitative predictions of a soil process must be measured against reality before it is relied on for practical purposes, and scientists will be interested in assessing model performance to identify

gaps in their understanding of the underlying processes. Model validation is often difficult. The outcome of a modelled process may often be difficult if not impossible to measure in the field, particularly when the model describes processes at coarse spatial scales. Where the modelled output can be measured directly, then ideally, we would wish to measure outputs from several locations or instances where the values of input variables are all the same. This allows us to estimate the variability of the output variable which is not explained by the model (since the model output will be the same in all instances) and so to estimate from the variability of the deviations between the measured and predicted outcomes how much of this can be attributed to 'lack of fit' of the model (Whitmore 1991).

3. Assimilation. Ideally, a model of a soil process, driven by some readily measured variables, will substitute for direct measurement of the process. In practice, we may wish to combine a model with limited measurements in order to predict outcomes or monitor a process. In this way, a purely statistical approach to estimating values is supplemented with understanding of how the system is likely to behave. The incorporation of measurements of a process into a model is known as data assimilation. It is widely used in oceanography and atmospheric science and has lately become fruitful in soil science. The state-space models described above are useful to this end; see, for example, Huang et al. (2017).
4. Error propagation. In practice, data with attendant errors are used as input to models with parameters estimated with error. Often the output of one model might function as input to another (e.g. modelled production of nitrate by mineralisation may be part of the input to a leaching model). It is clear that errors will propagate through such a system. If the models are linear, then the errors accumulate relatively slowly, and if all the errors have zero mean (i.e. are unbiased), then so is the final outcome. Neither is true of non-linear models, which may rapidly magnify the error and where the output may be biased even if the mean error of the inputs is zero. The error propagation of a set of models might be investigated by a Taylor series approximation where an analytical form of the model exists. For example, if an input variable, s, has a mean of $\widehat{s}$ and a variance of σ_s^2, and a model for an output variable y may be described by the non-linear function $y = f(s)$, then the approximate mean value of the output according to a second-order Taylor series approximation is

$$f\left(\widehat{s}\right) + \sigma_s^2 \frac{\delta^2 s}{\delta s^2} f\left(\widehat{s}\right).$$

A similar approximation to the variance of the output may be written.

A more generally applicable approach is to use Monte Carlo simulation. This requires that we have a reasonable knowledge of the variances and correlations of the input variables. We may then simulate many sets of inputs and obtain the mean and variance of the modelled outputs.

Heuvelink et al. (1989) used both approaches to study error propagation in the widely applied universal soil loss equation which combines several input variables

in a strongly non-linear way. They showed that the propagation of error was such that quite small uncertainty in the input variables could make the resulting predictions of little practical use. This is discussed further in Chap. 14.

1.4 Soil Attributes and Objects

It is worth considering the kind of objects that pedometrics studies. Pedometrics deals mainly with data on attributes observed on soil objects.

1.4.1 Soil Attributes

The attributes can either be direct or indirect (proximally or remotely sensed, including humanly sensed) measurements of physical, chemical or biological soil properties in the field or on specimens taken back to the laboratory or multivariate soil classes derived from them. The word attribute is used because the properties are *attributed* to the soil. Attributes are a function of the soil and the measurement process. The attributes can be discrete or continuous; see Chaps. 2, 3, 4 and 5.

1.4.2 Soil Objects

1.4.2.1 Geometric Objects

The objects on which attributes are observed can take a variety of forms, but most usually it is from some more-or-less fixed volume such as a soil core or pattern of soil cores or an auger boring or profile pit. Some workers have recognised the need for this volume to be of a minimum size in order that a good estimate of particular properties can be made. The measurement of water content of a saturated uniformly packed sand is a simple example of the dependence on scale of soil measurements. Sampling a very small volume gives a variation in the volumetric water content between zero and unity. Sampling a larger volume reduces the range of results, until, when a large enough volume is sampled, each measurement gives effectively the same result. The smallest volume at which this occurs is the representative elementary volume (REV). Buchter et al. (1994) effectively demonstrated the concept for stone content in a Swiss soil (Fig. 1.1). The range in percentage stones decreases exponentially from 100% for small sample lengths (volumes) to an asymptote of about 10%, which represents some larger-scale or macroscale variability. Figure 1.1 suggests that in this soil, the representative elementary volume is about $0.3 * 0.3 * 0.3 = 0.0027$ m^3 or about 30 l.

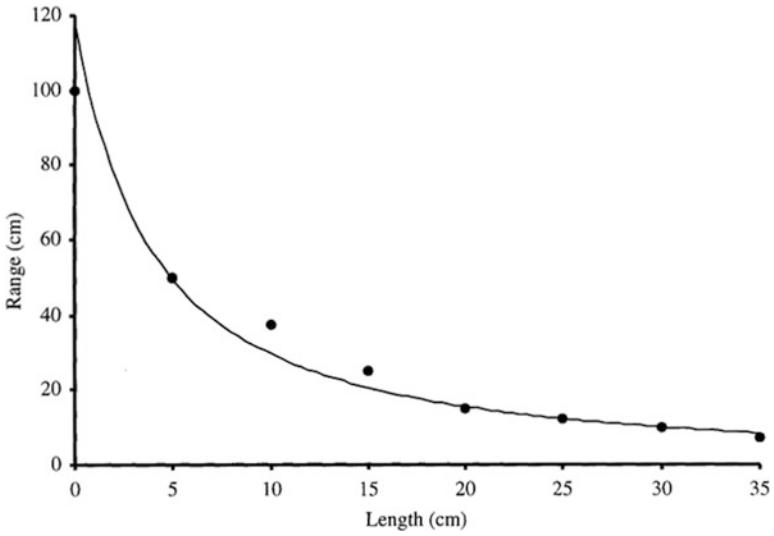

Fig. 1.1 The representative elementary 'length' of stone content in a Swiss soil. The range in percentage stones decreases exponentially from 100% for small sample lengths to an asymptote of about 10%. The graph suggests the representative elementary volume is about 0.03 m^3 (30 l) (Figure 6 from McBratney 1998)

The REV concept does not seem to have been applied to chemical and biological attributes or to soil classes for that matter. In a related way, geostatisticians (Isaaks and Srivastava 1989) recognise the need to define the so-called geometric support of observations. Geostatistics tells us that observations on larger supports or volumes within a fixed area will have smaller variances. The stabilisation of this variance is a way of defining the representative elementary volume. So, the REV concept and the geostatistical concept of geometric support are closely related.

1.4.2.2 'Natural' Objects?

These statistical concepts are not related to natural soil individuals, however. Soil science has always been troubled by a lack of clear definitions of individuals. When we study botany, we can almost always define and recognise an individual plant but in soil science, the definition of an individual soil is problematic. Pedologists recognise individuals such as the soil horizon (FitzPatrick 1993), the soil profile and the pedon (Johnson 1963). The problem with these is the lack of clear boundaries or of a fixed geometry. Russian pedologists recognise larger individuals such as the elementary soil area, the tessera and the tachet (Fridland 1976). Once again these are geometrically problematic.

1.4.2.3 Operational Objects

Holmgren (1988) sees the definition of soil individuals as a difficult problem and wrong-minded. He says 'that such distinct entities do not exist in soil is obvious'. He therefore suggests we should not seek such natural objects and argues for what he calls a point representation. He proposes a concept of a pedon that comprises a set of observational procedures which will lead to a set of attributes, including measurements of spatial and temporal variation, centred on a point. Holmgren's (1988) definition is:

> *A pedon is the possibility for soil observation in respect to a geographic point location. It can be realised by a set of observational propositions, each spatially and temporally specified in relation to that location.*

This concept seems to be more related to the practice in soil physics, chemistry and biology. This representation may be aligned with the geostatistical and REV concept above. The space-time geometric support of this operationally defined object is not clear, however. The two concepts could be combined, however. This is discussed further in Chaps. 8, 9, 10 and 21.

Ultimately, pedometrics is more a study of soil information (about given locations) than a study of natural or unnatural soil bodies and their attributes.

1.5 Scale Hierarchy

The word 'scale' is widely used in contemporary soil science, often confusingly. Conventionally, cartographers use the word 'scale' to refer to the ratio of a distance on a map to the corresponding distance on the ground. This simple descriptor denotes the detail which a particular map can convey. Thus, a 'large-scale' map (e.g. 1:5,000) is a relatively detailed representation of a relatively small area; a 'small-scale' map (e.g. 1:250,000) generalises over a large area. This notion of generalisation is important in soil science. It is obvious that we can describe the soil in terms of processes and phenomena across continents and landscapes or within horizons and aggregates. Confusingly, the former generalisation (which we would represent on a small scale map) is generally called large scale, and the generalisation over smaller intervals in space is referred to as small scale. Note that wavelet analysis also uses the term 'scale' in an opposite sense to the cartographers. We will use the term 'coarse scale' to refer to generalisations over large distances, represented at small cartographic scales, and 'fine scale' to refer to generalisations over short distances. This section aims to identify some of the key issues in current discussion of spatial and temporal scales in soil science which are relevant to pedometricians. It draws on an admirable review of the subject by Wagenet (1998).

In soil science, we may generalise over coarse or fine scales in both space and time. So, we may consider the soil system in terms of variations over a landscape

(coarse spatial scale) or variations within a profile (fine spatial scale). Similarly, we may consider coarse temporal scales (e.g. the decarbonation of a soil on an old land surface) or fine scales (e.g. hourly fluctuations in the water content of a cropped soil). Scale in space and time becomes of practical importance to soil scientists in general, and pedometricians in particular, when we find that some scales of generalisation are relatively rich in terms of our information about the soil and our ability to describe soil processes with confidence, but these do not necessarily coincide with the scales at which a manager might require predictions of soil behaviour under different scenarios.

The notion of a scale hierarchy is useful here and has been developed by ecologists who have recognised that ecosystems may be characterised as open systems with self-organising processes and that the flux of energy and matter through such a system tends to result in organisation over a hierarchy of spatial scales. This may be seen in soil systems too, where structure can be recognised in the organisation of clay particles by flocculation, the organisation of this material into aggregates, the structure of aggregates and macropores in horizons and so on up to the broad zonal patterns of soil variation at continental scale. Similarly, soil variation over time is organised in a hierarchy of scales from hourly fluctuations driven by rainfall events, the diurnal cycles of temperature, annual weather cycles and coarser scale variations still, associated with long-term cycles of climate and long-term processes of geomorphological change.

Bergkamp (1995) has identified concepts of scale hierarchy in the ecological literature which are relevant to soil scientists. Key to this development is the concept of a hierarchical structure of a complex system with explicit scales in time and space. The basic stable element of such a hierarchy is called the holon. A holon is essentially a distinct subsystem defined principally in terms of processes. It may therefore be defined in line with the concerns of a particular investigation. The principle interactions within a system take place within holons over particular scales of time and space. However, there are interactions within the hierarchical system of holons (holarchy) whereby processes at one scale of time and/or space constrain processes at other scales. Thus, for example, the process of aggregation at one scale both constrains and is constrained by the processes at coarser spatial and temporal scales which cause the differentiation of horizons. In general, however, holons at coarse temporal/spatial scales are more susceptible to the effects of events at comparable scale and may only respond to finer scale processes under conditions which are otherwise stable. Thus, processes of soil development in the profile – illuviation, development of clay skins and formation of pans – are likely to be overridden by coarse (temporal and spatial)-scale effects such as isostasy or climate change.

What are the implications of this analysis for soil research in general and pedometrics in particular? Wagenet (1998) identified three particular issues for research on a particular holon or subset of a holarchy at particular spatial and temporal scale. First, it is important to ensure that the basic unit of soil which is measured in such a study corresponds to the representative elementary volume

(REV) for the scale of interest. That is to say, the basic unit must be large enough to encompass the holons at finer scale which constitute the holon of interest, but not so large as to embrace the variability which occurs at this scale. The second point is that this variability must be adequately characterised. These are important issues for pedometricians and imply the careful choice of sample support and the selection of an appropriate sampling design. Since, as Wagenet (1998) recognises, the precise scale of a study may be constrained as much by technical limitations as the problem of interest, the pedometrician must be involved at all stages in the planning of the research activity. The third point made by Wagenet is that the process models which are used in a study must be appropriate to the scale. Thus, Darcy-type models must give way to two-domain models at some spatial scale.

We have discussed spatial and temporal scale as independent, but in fact they will generally be correlated, coarse spatial scale factors (topography) being correlated with coarse temporal scale factors (isostasy), while fine-scale processes (e.g. at molecular level) may tend to occur over fine temporal scales (Fig. 1.2). This assumption is implicit in Hoosbeek and Bryant's (1992) well-known figure linking scale (explicitly spatial) to appropriate modelling strategies (Fig. 1.3).

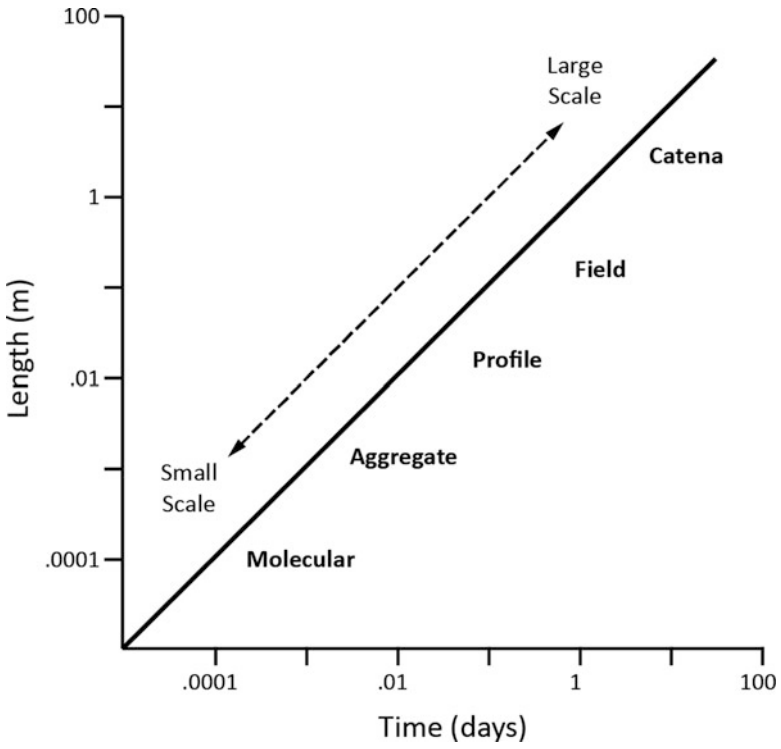

Fig. 1.2 Scales of soil processes (Figure 4 redrawn from Wagenet 1998)

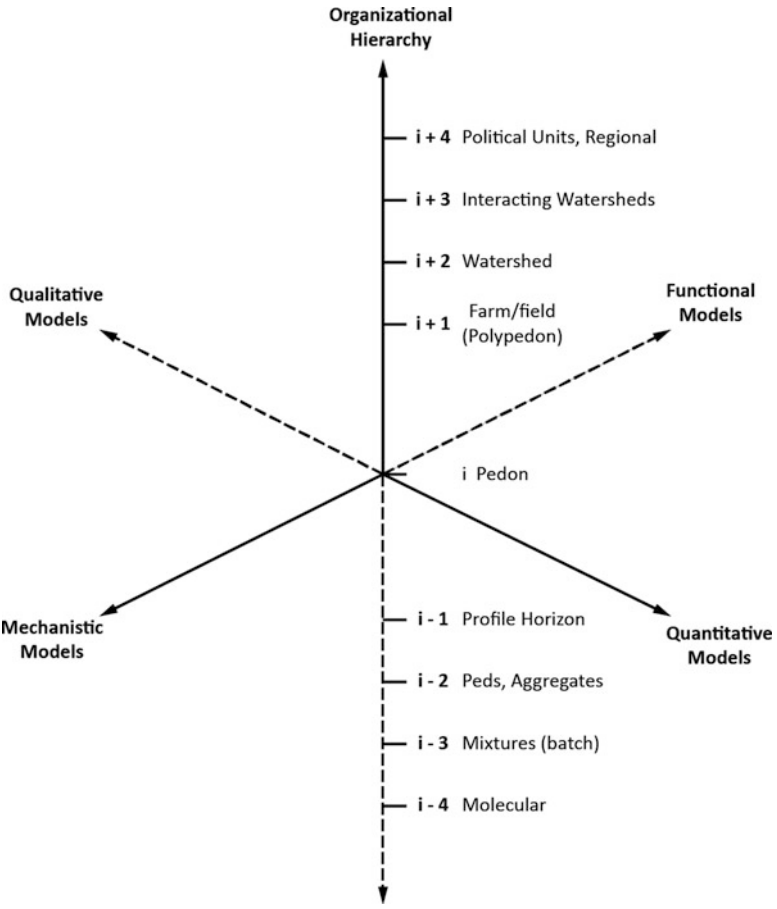

Fig. 1.3 A hierarchical representation of spatial scales of soil processes. At each scale, there is a plane of possibilities of research modelling approaches. The front half of each horizontal level corresponds to Table 1.1 (Redrawn from Hoosbeek and Bryant 1992)

An understanding of this relationship is important, particularly, to return to the key problem of scale, when we wish to translate knowledge and research findings between scales. McBratney (1998) discussed some of these issues and Bierkens et al. (2000) offer a detailed discussion, and we return to this issue in Chap. 15. For the present, however, it is worth noting that important processes may occur off the diagonal of Fig. 1.2. For example, non-equilibrium processes of pesticide-soil interaction occur over short distances (fine spatial scale) but long periods (coarse temporal scale), and this type of process is poorly modelled (Wagenet 1998). Similarly, the weathering of rock occurs at fine spatial scales but over long time periods (McBratney 1998). Rapid (fine temporal scale) processes may also occur

over relatively long distances, e.g. preferential flow through macropores (Wagenet's 1998 example) or global scale but rapid processes such as the Chernobyl incident (McBratney's 1998 example).

In short, pedometricians must be aware of the issues of spatial and temporal scale when collecting and analysing data on soil. A good summary of the key issues is due to Bergkamp (1995) who noted (1) that spatial patterns differ at different scales, (2) changes of spatial patterns differ at different temporal scales, (3) processes controlling changes in spatial pattern differ at different scales and (4) at different scales, different variables are often needed to describe similar processes. These considerations all pose mathematical and statistical challenges to the pedometrician. Some of these are met by the techniques which we describe in this book. Others are challenges for the future.

1.6 Pedometric Principles and Hypotheses

As a summary of this introductory chapter, we outline a few general and specific ideas that underlie the pedometric approach.

1.6.1 General Principles

All the 'metric' disciplines (e.g. biometrics, chemometrics, envirometrics), attempting a quantification of the natural environment and phenomena, will recognise at least three principles:

1. *A quantification principle*
   This suggests that *wherever possible*, *observations should be quantitative rather than qualitative*. The main rationale for this is that the difference between two observations can be calculated. For example, if we say the texture of a soil horizon is 'light' or 'heavy', we might not be able to easily understand how different 'light' is from 'heavy'. If however we describe one horizon as having a clay content of 10 dag/kg and another as 70 dag/kg, we know immediately that the difference is 60 dag/kg. This principle implies that vague descriptors should be quantified.
2. *An uncertainty principle*
   This principle is not the Heisenberg one, but simply further suggests that *wherever possible*, *a measure of uncertainty should accompany quantitative observations*. This may be a statistical uncertainty. For example, if we have a field estimate of horizon texture, then the uncertainty might have a standard deviation of ± 5 dag/kg, whereas if this was a laboratory measurement, the associated standard deviation might be ± 1 dag/kg. Further, if the soil attribute is predicted rather than measured, then the one which is least uncertain is to be preferred.

3. *A modelling parsimony principle (Ockham's razor)*
   This old principle described 700 years ago by William of Ockham states that *when comparing models for describing a phenomenon*, *choose the simplest one*. For example, if we can describe the particle-size distribution of a soil material equally well with a two- or a three-parameter distribution, then the model with two parameters is to be preferred.

We could suggest some principles from other disciplines, e.g. economic or sociological ones, but they do not seem to be in the domain of quantifying the natural environment. They may apply when various uses are made of soil information.

1.6.2 Specific Principles

In addition to these general principles, there are some principles or hypotheses specific to the domain of pedometrics. These may be summarised by a single all-embracing statement:

Soil is a more-or-less continuous, anisotropic, non-stationary, multivariate phenomenon in various states of equilibrium with its environment.

Attempts to quantify soil should recognise these features. They lead to various hypotheses that may be tested or to assumptions commonly underlying pedometric analysis. Therefore, various null hypotheses that could or should be tested by empirical observation at any location include:

Soil is a spatially discontinuous phenomenon.
Soil variation is constant with direction.
The mean of a soil property is independent of position in space or time.
The soil can be adequately described by a single property, e.g., *pH.*
Soil is a function of its current environment.

Finally, it is a difficult philosophical point to contend that these principles and hypotheses lead to a more objective treatment of reality. They are all human constructions. Commonsensically, it would seem that they probably lead to a legitimate empirical testing of humanity's hypotheses about the nature of soil. Nevertheless, as the famous biologist J.B.S. Haldane once wrote, 'If you are faced by a difficulty or a controversy in science, an ounce of algebra is worth a ton of verbal argument'.

References

Addiscott TM, Bailey NJ (1991) Relating the parameters of a leaching model to the percentages of clay and other components. In: Roth K, Flühler H, Jury WA, Parker JC (eds) Field-scale solute and water flux in soils. Birkhaüser Verlag, Basel, pp 209–221

Addiscott TM, Mirza NA (1998) New paradigms for modelling mass transfers in soils. Soil Tillage Res 47:105–109

Addiscott TM, Wagenet RJ (1985) A simple method for combining soil properties that show variability. Soil Sci Soc Am J 49:1365–1369
Addiscott T, Smith J, Bradbury N (1995) Critical evaluation of models and their parameters. J Environ Qual 24:803–807
Adhikari K, Hartemink AE (2016) Linking soils to ecosystem services – a global review. Geoderma 262:101–111
Bárdossy A, Bronstert A, Merz B (1995) 1-, 2- and 3-dimensional modeling of water movement in the unsaturated soil matrix using a fuzzy approach. Adv Water Resour 18:237–251
Bergkamp G (1995) A hierarchical approach for desertification assessment. Environ Monit Assess 37:59–78
Bierkens MFP, Finke PA, de Willigen P (2000) Upscaling and downscaling methods for environmental research. Kluwer, Dordrecht
Buchter B, Hinz C, Fluhler H (1994) Sample size for determination of coarse fragment content in a stony soil. Geoderma 63:265–275
Cohen S, Willgoose G, Hancock G (2010) The mARM3D spatially distributed soil evolution model: three dimensional model framework and analysis of hillslope and landform responses. J Geophys Res 115:F04013
Dexter AR (1976) Internal structure of tilled soil. J Soil Sci 27:267–278
Dijkerman JC (1974) Pedology as a science: the role of data, models and theories in the study of natural soil systems. Geoderma 11:73–93
Dominati EJ, Patterson M, Mackay AD (2010) A framework for classifying and quantifying the natural capital and ecosystem services of soils. Ecol Econ 69:1858–1868
Doran JW, Parkin TB (1994) Defining and assessing soil quality. In: Doran JW, Coleman DC, Bezdicek DF, Stewart BA (eds) Defining soil quality for a sustainable environment, vol 35. Soil Sci Soc Am, Madison, pp 3–21
Dou C, Woldt W, Bogardi I (1999) Fuzzy rule-based approach to describe solute transport in the unsaturated zone. J Hydrol 220:74–85
FitzPatrick EA (1993) Principles of soil horizon definition and classification. Catena 20: 395–402
Fridland VM (ed) (1976) Soil combinations and their genesis. Amerind Publishing Co, New Delhi
Glasbey CA, Horgan GW, Darbyshire JF (1991) Image analysis and three-dimensional modelling of pores in soil aggregates. J Soil Sci 42:479–486
Gleick J (1987) Chaos. Making a new science. Sphere Books, London
Harvey D (1969) Explanation in geography. Edward Arnold, London
Hastie T, Tibshirani R (1986) Generalized additive models. Stat Sci 1:297–318
Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman and Hall, London
Heuvelink GBM, Burrough PA, Stein A (1989) Propagation of errors in spatial modelling with GIS. Int J GIS 3:303–322
Heuvelink GBM, Webster R (2001) Modelling soil variation: past, present and future. Geoderma 100:269–301
Heuvelink GBM, Schoorl JM, Veldkamp A, Pennock DJ (2006) Space–time Kalman filtering of soil redistribution. Geoderma 133:124–137
Holmgren GGS (1988) The point representation of soil. Soil Sci Soc Am J 52:712–715
Hoosbeek MR, Bryant RB (1992) Towards the quantitative modeling of pedogenesis – a review. Geoderma 55:183–210
Huang J, McBratney AB, Minansy B, Traintafilis J (2017) Monitoring and modelling soil water dynamics using electromagnetic conductivity imaging and the ensemble Kalman filter. Geoderma 285:76–93
Isaaks EH, Srivastava RM (1989) Applied geostatistics. Oxford University Press, New York
Ibáñez JJ, De-Alba S, Bermúdez FF, García-Álvarez A (1995) Pedodiversity: concepts and measures. Catena 24:215–232
Johnson WM (1963) The pedon and the polypedon. Soil Sci Soc Am Proc 27:212–215
Kidd D, Webb M, Malone B, Minasny B, McBratney A (2015) Digital soil assessment of agricultural suitability, versatility and capital in Tasmania, Australia. Geoderma Reg 6:7–21

Leeper GW (1956) The classification of soils. J Soil Sci 7:59–64
Manolis GD (2002) Stochastic soil dynamics. Soil Dyn Earthq Eng 22:3–15
McBratney AB (1998) Some considerations on methods for spatially aggregating and disaggregating soil information. Nutr Cycl Agroecosyst 50:51–62
McBratney AB, Odeh IOA (1997) Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma 77:85–113
McBratney AB, Field DJ, Koch A (2014) The dimensions of soil security. Geoderma 213:203–213
McCullagh P, Nelder J (1989) Generalized linear models, 2nd edn. Chapman and Hall, London
Milne AE, Webster R, Lark RM (2010) Spectral and wavelet analysis of gilgai patterns from air photography. Aust J Soil Res 48:309–325
Minasny B, McBratney AB (1999) A rudimentary mechanistic model for soil production and landscape development. Geoderma 90:3–21
Minasny B, McBratney AB (2001) A rudimentary mechanistic model for soil production and landscape development, II. A two-dimensional model incorporating chemical weathering. Geoderma 103:161–179
Pachepsky Y, Timlin D, Rawls W (2003) Generalized Richards' equation to simulate water transport in unsaturated soils. J Hydrol 272:3–13
Persson M, Yasuda H, Albergel J, Berndtsson R, Zante P, Nasri S, Öhrström P (2001) Modeling plot scale dye penetration by a diffusion limited aggregation (DLA) model. J Hydrol 250:98–105
Phillips JD (1993) Stability implications of the state factor model of soils as a nonlinear dynamical system. Geoderma 58:1–15
Phillips JD (1998) On the relations between complex systems and the factorial model of soil formation (with discussion). Geoderma 86:1–42
Rossiter DG (1990) ALES: a framework for land evaluation using a microcomputer. Soil Use Manag 6:7–20
Solow AR (1990) Geostatistical cross-validation: a cautionary note. Math Geol 22:637–639
Stockmann U, Padarian J, McBratney A, Minasny B, de Brogniez D, Montanarella L, Hong SY, Rawlins BG, Field DJ (2015) Global soil organic carbon assessment. Glob Food Sec 6:9–16
Storie RE (1933) An index for rating the agricultural value of soils. University of California Coop. Ext. Bull. 556
Sugihara G, May RM (1990) Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature 344:734–741
Van Kuilenburg J, de Gruijter JJ, Marsman BA, Bouma J (1982) Accuracy of spatial interpolation between point data on soil moisture capacity, compared with estimates from mapping units. Geoderma 27:311–325
Vanwalleghem T, Stockmann U, Minasny B, McBratney AB (2013) A quantitative model for integrating landscape evolution and soil formation. J Geophys Res Earth 118:1–17
Wagenet RJ (1998) Scale issues in agroecological research chains. Nutr Cycl Agroecosyst 50:23–34
Walter C, McBratney Alex B, Viscarra Rossel RA, Markus JA (2005) Spatial point-process statistics: concepts and application to the analysis of lead contamination in urban soil. Environmetrics 16:339–355
Webster R (1977a) Quantitative and numerical methods in soil classification and survey. Oxford University Press, Oxford
Webster R (1977b) Spectral analysis of gilgai soil. Aust J Soil Res 15:191–204
Webster R, McBratney AB (1989) On the Akaike information criterion for choosing models for variograms of soil properties. J Soil Sci 40:493–496
Whitmore AP (1991) A method for assessing the goodness of computer simulation of soil processes. J Soil Sci 42:289–299
Wischmeier WH, Smith DD (1978) Predicting rainfall erosion losses – a guide to conservation planning. US Department of Agriculture, Agriculture Handbook No 537, Washington, DC

Wischmeier WH, Smith DD 1978b Predicting rainfall erosion losses: guide to conservation planning. USDA, agriculture handbook 537. U.S. Government Printing Office, Washington, DC

Wösten JHM, Pachepsky YA, Rawls WJ (2001) Pedotransfer functions for estimating saturated hydraulic conductivity: implications for modeling storm flow generation. J Hydrol 251:202–220

Yevjevich V (1987) Stochastic models in hydrology. Stoch Hydrol Hydraul 1:17–36