Available online at www.sciencedirect.com
SCIENCE DIRECT®
Geoderma 117 (2003) 3–52
GEODERMA
www.elsevier.com/locate/geoderma

On digital soil mapping

A.B. McBratney[a,*], M.L. Mendonça Santos[b], B. Minasny[a]

[a] *Australian Centre for Precision Agriculture, Faculty of Agriculture, Food and Natural Resources, McMillan Building A05, The University of Sydney, Sydney, New South Wales 2006, Australia*
[b] *EMBRAPA-Centro Nacional de Pesquisa de Solos, Rua Jardim Botânico 1024, 22.460-000, Rio de Janeiro, RJ, Brazil*

Received 19 November 2002; received in revised form 14 May 2003; accepted 5 June 2003

Abstract

We review various recent approaches to making digital soil maps based on geographic information systems (GIS) data layers, note some commonalities and propose a generic framework for the future. We discuss the various methods that have been, or could be, used for fitting quantitative relationships between soil properties or classes and their 'environment'. These include generalised linear models, classification and regression trees, neural networks, fuzzy systems and geostatistics. We also review the data layers that have been, or could be, used to describe the 'environment'. Terrain attributes derived from digital elevation models, and spectral reflectance bands from satellite imagery, have been the most commonly used, but there is a large potential for new data layers. The generic framework, which we call the scorpan-SSPFe (soil spatial prediction function with spatially autocorrelated errors) method, is particularly relevant for those places where soil resource information is limited. It is based on the seven predictive scorpan factors, a generalisation of Jenny's five factors, namely: (1) *s*: soil, other or previously measured attributes of the soil at a point; (2) *c*: climate, climatic properties of the environment at a point; (3) *o*: organisms, including land cover and natural vegetation; (4) *r*: topography, including terrain attributes and classes; (5) *p*: parent material, including lithology; (6) *a*: age, the time factor; (7) *n*: space, spatial or geographic position. Interactions (*) between these factors are also considered. The scorpan-SSPFe method essentially involves the following steps:

(i) Define soil attribute(s) of interest and decide resolution ρ and block size β.
(ii) Assemble data layers to represent Q.
(iii) Spatial decomposition or lagging of data layers.
(iv) Sampling of assembled data (Q) to obtain sampling sites.
(v) GPS field sampling and laboratory analysis to obtain soil class or property data.
(vi) Fit quantitative relationships (observing Ockham's razor) with autocorrelated errors.
(vii) Predict digital map.

* Corresponding author. Tel.: +61-2-9351-3214; fax: +61-2-9351-3706.
E-mail address: alex.mcbratney@acss.usyd.edu.au (A.B. McBratney).

doi:10.1016/S0016-7061(03)00223-4

(viii) Field sampling and laboratory analysis for corroboration and quality testing.
(ix) If necessary, simplify legend or decrease resolution by returning to (i) or improve map by returning to (v).

Finally, possible applications, problems and improvements are discussed.

Keywords: Soil map; Soil survey; Digital map; Classification tree; DEM; DTM; GAM; Generalised linear model; Geophysics; Geostatistics; GIS; Neural network; Pedometrics; Pedotransfer function; Regression tree; Remote sensing; Soil spatial prediction function; Wavelets

> Greenhough's basic notion that his map should be entirely empirical, and that its drawing should be based only on observations, and not tied to any particular theory about which rocks were where, and why and when and how they had been laid down. Rumination, Greenhough reasoned, had no place in a geological map: what should appear on the finished sheets of his charts should reflect facts that were quite unsullied by any theoretical presuppositions. (Simon Winchester, 2001. The Map that Changed the World. Viking, London. 338 pp. p. 231.)

1. Introduction

With the great explosion in computation and information technology has come vast amounts of data and tools in all fields of endeavour. Soil science is no exception, with the ongoing creation of regional, national, continental and worldwide databases. The challenge of understanding these large stores of data has led to the development of new tools in the field of statistics and spawned new areas such as data mining and machine learning (Hastie et al., 2001). In addition to this, in soil science, the increasing power of tools such as geographic information systems (GIS), GPS, remote and proximal sensors and data sources such as those provided by digital elevation models (DEMs) are suggesting new ways forward. Fortuitously, this comes at a time when there is a global clamour for soil data and information for environmental monitoring and modelling.

Consequently, worldwide, organisations are investigating the possibility of applying the new spanners and screwdrivers of information technology and science to the old engine of soil survey. The principal manifestation is soil resource assessment using geographic information systems (GIS), i.e., the production of digital soil property and class maps with the constraint of limited relatively expensive fieldwork and subsequent laboratory analysis.

The production of digital soil maps ab initio, as opposed to digitised (existing) soil maps, is moving inexorably from the research phase (Skidmore et al., 1991; Favrot and Lagacherie, 1993; Moore et al., 1993) to production of maps for regions and catchments and whole countries. The map of the Murray–Darling basin of Australia (Bui and Moran, 2001, 2003) comprising some 19 million 250 × 250 m pixels or cells and the digital Soil Map of Hungary (Dobos et al., 2000) are the most notable examples to date.

McBratney et al. (2000) reviewed pedometric methods for soil survey and suggested three resolutions of interest, namely >2 km, 20 m–2 km and <20 m corresponding to national to global, catchment to landscape and local extents. Table 1 provides a slightly more detailed overview with five resolutions of interest. The third one (D3) which deals with subcatchments, catchments and regions is the one which attracts the most attention. In the language of digital soil maps (Bishop et al., 2001), different from that of conventional cartography, scale is a difficult concept, and is better replaced by resolution and spacing. D3 surveys, which in conventional terms, have a scale of 1:20,000 down to 1:200,000, have a block or cell size from 20 to 200 m, a spacing also of 20–200 m and a nominal spatial resolution of 40–400 m (see Table 1).

The Netherlands has complete coverage at a nominal spatial resolution of 100 m. In France, on the other hand, a highly developed western economy, but with a large land area, only 26% of the country is covered at a nominal spatial resolution of 500 m and 13% at a nominal spatial resolution of 200 m (King et al., 1999). One-third of Germany is covered with soil maps at a nominal spatial resolution of 10 m (1:5000), but most of these are not yet digital (Lösel, 2003). On the other hand, complete coverage of Germany at coarser reso-

Table 1
Suggested resolutions and extents of digital soil maps

Name	Approximate USDA survey order[a]	Pixel size and spacing[b]	Cartographic scale[b]	Resolution 'loi du quart'[c]	Nominal spatial resolution[b]	Extent[d]	Cartographic scale[b]
D1	0[e]	<(5 × 5) m	>1:5000	<(25 × 25) m	<(10 × 10) m	<(50 × 50) km	>1:5000
D2	1, 2	(5 × 5) to (20 × 20) m	1:5000–1:20,000	(25 × 25) to (100 × 100) m	(10 × 10) to (40 × 40) m	(500 × 500) to (200 × 200) km	1:5000–1:20,000
D3	3, 4	(20 × 20) to (200 × 200) m	1:20,000–1:200,000	(100 × 100) to (1 × 1) km	(40 × 40) to (400 × 400) m	(2 × 2) to (2000 × 2000) km	1:20,000–1:200,000
D4	5	(200 × 200) to (2 × 2) km	1:200,000–1:2,000,000	(1 × 1) to (10 × 10) km	(400 × 400) to (4 × 4) km	(20 × 20) to (20,000 × 20,000) km	1:200,000–1:2,000,000
D5	5	>(2 × 2) km	<1:2,000,000	>(10 × 10) km	>(4 × 4) km	>(200 × 200) km	<1:2,000,000

[a] Soil Survey Staff (1993).

[b] Digital soil maps are partly defined by their block size and spacing (refer to Bishop et al., 2001, Fig. 3), which, here, we equate with pixel size. The cartographic scale, calculated as 1 m/(side length of 1000 pixels), assumes that the smallest area discernible is 1 × 1 mm. Conversely, the pixel size (p) of a 1:100,000 conventional map can be calculated as $p = 1/\chi^{*}\lambda = 100{,}000 \times 0.001 = 100$ m if we consider the smallest area resolvable on a map (λ), with representative fraction χ to be 1 × 1 mm. Following notions in microscopy, and the Nyquist frequency concept from signal processing, it may be argued that the minimum resolution is the size of 2 × 2 pixels. We define this here as the nominal spatial resolution. Small pixel sizes correspond to fine resolutions and large pixel sizes correspond to coarse resolutions.

[c] According to Boulaine (1980), the smallest area discernible on a map is 0.5 × 0.5 cm or one quarter of a square centimetre, hence, the term 'loi du quart'. The USDA Soil Survey Field Handbook (Soil Survey Staff, 1993, Table 2-2) quotes 0.6 × 0.6 cm. Both of these really refer to conventional map delineations, and resolution estimates based on these minimum areas should be regarded as very conservative.

[d] Calculated as minimum resolution times 100 (pixels) up to maximum resolution times 10,000 pixels.

[e] This order was suggested by Dr. Pierre C. Robert, University of Minnesota, for applications in precision agriculture.

lutions (nominally 100 and 400 m) is available. The situation in larger countries such as Australia and Brazil is much worse. In Australia, for example, prior to Moran and Bui's (2002) work, the Murray–Darling Basin, Australia's most important agricultural region comprising some 14% of the land area, had 50% coverage at 500 m and 3% at 200 m. In Brazil, the country is uniformly covered by the Soil Map of Brazil and the Agricultural Suitability Map of Brazil at a nominal spatial resolution of 10 km, exploratory soil maps by the RADAM/EMBRAPA Solos project (1:1,000,000 or nominally 2 km) and Agroecological Zoning (diagnosis of environmental and agro-socio-economic features, nominally 4 km or 1:2,000,000). The main reason for this lack of soil spatial data infrastructure worldwide is simply that conventional soil survey methods are slow and expensive. This paper addresses this worldwide problem head on.

GIS, a tool for collating all kinds of spatial information (Burrough and McDonnell, 1998), in itself is incapable of soil mapping; it requires an intellectual framework. Indeed, most of the real computational work so far has been done outside the framework of formal GIS packages. We feel it is timely to outline a general intellectual and operational framework for digital mapping of soil properties and classes for D3 surveys, in the form of an empirical partially deterministic partially stochastic model—the so-called soil spatial prediction functions (SSPFe) with spatially autocorrelated errors. We review various approaches with numerous examples from the literature, which are largely seen as special cases of the approach suggested here. First, we trace the development of the quantitative ideas and methods over the last 60 or so years. An independent review of digital soil mapping under the name of 'predictive' soil mapping has appeared (Scull et al., 2003a,b).

2. Brief review of approaches to soil spatial prediction

Hudson (1992) contended that soil survey is a scientific strategy based on the concepts of factors of soil formation coupled with soil–landscape relationships. Hewitt (1993) pleaded for the need for explicitly stated, but not necessarily quantitative, models for soil survey. The models may be knowledge-based (Bui, 2003). In this view, soil maps are representations of soil surveyors' knowledge about

soil objects. In this paper, we argue towards quantitative predictive models.

2.1. Jenny

Recalling Jenny's (1941) famous equation, which he intended as a mechanistic model for soil development,

$$S = f(c, o, r, p, t, \ldots).$$

Implicitly, S stands for soil, c (sometimes cl) represents climate, o organisms including humans, r relief, p parent material and t time. Some have asserted its insolubility. Nonetheless, since Jenny published his formulation, it has been used by innumerable surveyors all over the world as a qualitative list for understanding the factors that may be important for producing the soil pattern within a region. Numerous researchers have taken the quantitative path and have tried to formalise this equation largely through studies of cases where one factor varies and the rest are held constant. Therefore, quantitative climofunctions, topofunctions, etc., have been developed. Much of this work was done before sophisticated numerically intensive statistical methods became available. Here are some brief examples.

(c) Sometimes (cl). Climofunctions were the ones most developed by Jenny in his 1941 book. Jones (1973) found relationships between carbon, nitrogen and clay and annual rainfall and altitude in W. African savanna using linear and multiple linear regression. Simonett (1960) found a power–function relationship between mineral composition of soil developed on basalt in Queensland and annual rainfall.

(o) There seems less development of organofunctions, many believing that the principal organofunction or biofunction, that of vegetation, is dependent on soil rather than the converse. Noy-Meir (1974) found relationships between vegetation and soil type in S.E. Australia. The other principal organofunction, the anthropofunction, has only been recently quantified. Much of the work on soil degradation and soil quality are evidence of the effect of humanity on soil. The classic work of Nye and Greenland (1960) is an early quantitative example.

(r) The relationship between soil and topographic factors has been evident at least since Milne's (1935) paper. Quantitative topofunctions are manifold. For example, Furley (1968) and Anderson and Furley (1975) found a piece-wise linear relationship between organic carbon, nitrogen and pH of surface horizons and slope angle for profiles developed on calcareous parent materials around Oxford in England.

(p) Quantitative lithofunctions have not been developed often, perhaps due to a difficulty in recognising and quantifying the dependent and independent variables. Barshad (1958) quantified mean clay content as a function of rock type.

(t) Some consider this the only truly independent variable (but if that is the case, why is space not also included?). Chronofunctions are often theoretical or hypothetical rather than observed. Hay (1960), however, found an exponential relationship between clay formation and time for soil developed in volcanic ash on the island of St. Vincent, as would be expected from first-order kinetics.

A lot of this early quantitative work was very well summarised by Yaalon (1975). Many of the relationships found are not linear. It should be remembered that the aim of these investigations was to understand soil formation and not necessarily to predict soil from the other factors.

Recognition of interactions between the soil-forming factors is potentially important because it is one possible source of detailed soil pattern. It is difficult to find work that considers such interactions explicitly. Webster (1977) perhaps came closest with his canonical correlation studies of sets of soil properties and environmental factors. From this work, he suggested, for example, that soil will reflect a strong interaction between topography and lithology particularly on upper slope positions but this will be time-dependent. Odeh et al. (1991) using closely related methods made similar findings.

2.2. Geographic and neighbourhood (or purely spatial) approaches

Since the late 1960s, there has been an emphasis on what might be called geographic or purely spatial approaches, i.e., soil attributes[1] can be predicted from spatial position largely by interpolating between soil

[1] Soil attributes is a general term to mean that which can be attributed to the soil by measurement or inference, e.g., soil properties like pH, or classes like a soil profile class, or the presence or absence of a soil horizon class.

observation locations. Another way of thinking about this is as a "neighbourhood law", expounded first perhaps by Lagacherie (1992), but is the basis underlying the soil combinations of Fridland (1972), and also of soil geostatistics (Giltrap, 1977), etc. Generally, we can consider the soil at some location (x,y) to depend on the geographic coordinates x,y and on the soil at neighbouring locations $(x+u, y+v)$, i.e.,

$$\underline{s}(x,y) = f((x,y), s(x+u,\ y+v))$$

the dependence usually being some decreasing function of the magnitude of u and/or v.

This approach arose originally out of the need for spatial prediction to make soil maps, and because of a failure to obtain prediction from the soil-forming factors largely because the quantitative variables describing these factors were not readily available to do such predictions. These purely spatial approaches are almost entirely based on geostatistics and its precursor trend-surface analysis, although thin-plate smoothing splines have been suggested and used occasionally (Laslett et al., 1987; Hutchinson and Gessler, 1994). Exact-fitting splines do not perform well (Laslett et al., 1987; Voltz and Webster, 1990).

2.2.1. Geostatistics and related methods

2.2.1.1. Trend surfaces—s(x,y) = f(x,y). Trend surfaces are low-order polynomials of spatial coordinates. Several applications have been reported in the literature. Davies and Gamm (1969) applied this technique to soil pH values from the county of Kent in England. Edmonds and Campbell (1984) described the average annual soil temperatures at locations within a network of stations from Virginia and neighbouring states with a third-degree polynomial that explained 71% of the observed variation. On the other hand, Kiss et al. (1988) found the spatial pattern of ^{137}Cs activity in well-drained, native noneroded soil in the agricultural portion of Saskatchewan was complex, and could not be adequately described by a second-order trend surface. There appears to be no literature on trend surfaces for soil classes. Nevertheless, Wrigley (1978) has made an attempt to map the probability. Spatially, trend surfaces are rather simplified 'unnatural' representations and more complex spatial patterns often need to be modelled.

2.2.1.2. Kriging—s(x,y) = f(s(x+u, y+v)). It was recognised that more complex spatial patterns could be accommodated by treating soil variables as regionalised variables using the methods of geostatistics, particularly, various forms of kriging. The papers by Burgess and Webster (1980a,b) and Webster and Burgess (1980) are probably regarded as the most seminal. These kriging methods, reviewed by Burrough (1993), Goovaerts (1999) and Heuvelink and Webster (2001), could deal with continuous soil properties and classes, give estimates for blocks or pixels of varying size and, moreover, estimate uncertainty.

2.2.1.3. Co-kriging—s(x,y) = f(s(x+u, y+v), {c,o,r,p,t}(x,y)). It was recognised early in the development of soil geostatistics that soil could be better predicted if denser sets of secondary variables (spatially cross) correlated with the primary variable were available. This technique is called co-kriging. In the early co-kriging studies (McBratney and Webster, 1983; Vauclin et al., 1983; Goulard and Voltz, 1992), these ancillary variables were other soil variables, indicating that other soil variables are themselves useful predictors of soil. Later, with the advent of GIS and improved technology, co-kriging was performed with detailed secondary data sets of environmental variables derived from digital elevation models and satellite images (Odeh et al., 1994, 1995).

2.2.2. Jenny and geography—corpt or clorpt—s(x,y)=f({c,o,r,p,t}(x,y))

An alternative spatial prediction strategy to the purely geographic approaches was developed in the early 1990s, although there were precursors. In these studies, the state-factor equation was put explicitly into a spatial framework and the factors were also observed in the same spatial domain. This approach probably resulted from the advent of the first geographic information systems, and also possibly as a pedological response to geostatistics. It seems to be based on a much earlier one-dimensional example of using environmental (terrain, representing r) attributes for soil prediction, namely that of Troeh (1964) and Walker et al. (1968). Probably, the first of its kind, Troeh (1964) analysed the elevation data from two catenas and derived slope and profile curvature. He then plotted the slope gradient and profile curvature and found that the soil drainage classes could be

distinguished by paraboloid of revolution equations. Walker et al. (1968) used slope, curvature, aspect and distance from the local summit, in combination with multiple linear regression to predict soil morphological properties such as A horizon depth, depth to mottling and carbonates along a transect. An early, perhaps the first, two-dimensional example is Legros and Bonneric (1979), based on earlier work by Legros (1975). They described a soil–environment relationship using various factors (altitude, slope, exposure, parent material) which were observed on a 500-m grid-cell basis to predict the degree of podzolisation in Massif du Pilat of France, and mapped it digitally at a resolution of 500 m. The prediction was achieved by a kind of taxonomic distance relative to reference sites. This was done well before the advent of formal GIS.

The GIS-based studies started at the beginning of the 1990s. Terrain analysis had improved and secondary rasterised layers providing a kind of complete enumeration of the area could be put in GIS. The soil observation points were intersected with the layers of secondary data, a model fitted by various means, and then the model was used to predict all other locations on the raster. Moore et al. (1993) gave the first two-dimensional example using a set of terrain attributes derived from a digital elevation model on a 15-m grid to predict continuous soil properties such as A horizon thickness and pH for a small catchment in Colorado. Odeh et al. (1994) did a similar study in South Australia. Skidmore et al. (1991) predicted forest soil classes in New South Wales from layers of natural vegetation data (representing o), and terrain attributes on a 30-m grid. Bell et al. (1992, 1994) predicted soil drainage class from terrain data, and Lagacherie and Holmes (1997) predicted soil classes in the Languedoc using layers of lithological and terrain data. Favrot and Lagacherie (1993) foreshadowed this as a general approach for making soil class maps.

For quantitative prediction purposes, this has been called the ‘clorpt’ or ‘corpt’ equation (McBratney et al., 2000). Some people have termed the approach ‘‘environmental correlation’’ (McKenzie and Austin, 1993). McKenzie and Ryan (1999) used environmental correlation associated with stratigraphy, digital terrain models and gamma radiometric survey, respectively, to predict soil properties in Australia. Ryan et al. (2000) reviewed the concepts and applications of spatial modelling using the ‘‘environmental correlation’’ approach and used it to predict forest soil properties at the landscape level.

For predicting soil classes, S_c, or soil properties, S_p, often only a subset of the five soil-forming factors has been used, e.g., when information from a digital elevation model is available: $S_c = f(r)$, e.g., Bell et al. (1992), or $S_p = f(r)$, e.g., Moore et al. (1993) or relief and a lithology map, $S_c = f(r,p)$, e.g., Lagacherie and Holmes (1997), or relief and vegetation, $S_c = f(r,o)$, e.g., Skidmore et al. (1991).

2.2.3. Combinations—clorpt (or corpt) and kriging

Alert readers will have noted that there has been a certain similarity and convergence between the co-kriging and the environmental correlation approach. Some workers recognised this in the mid-1990s and combined the two in what is generically known as regression kriging (Knotters et al., 1995; Odeh et al., 1995). In this approach ‘clorpt’ is used to predict the soil property of interest from environmental variables and kriging is used on the residuals. Bourennane et al. (1996) used kriging with external drift, which is related to regression kriging but only allows for linear relationships between the variable of interest and the environmental variables (the external drifts).

2.3. Predicting soil attributes from other soil attributes—$s_1 = f(s_2)$

As noted above, some of the co-kriging studies (McBratney and Webster, 1983; Vauclin et al., 1983) showed that soil could be predicted from other soil attributes. This observation in itself is not very useful unless there are much denser secondary variables available. Remote (e.g., gamma radiometrics) and proximal sensing (e.g., electromagnetic induction) offer this possibility. This becomes increasingly important because Phillips (2001) gives several examples where ‘clorpt’ apparently does not work, particularly at fine resolutions. This suggests that for predictive purposes s (for soil) should be added to the ‘corpt’ list.

2.4. Some brief conclusions

From this brief review, we see that:

1. Quantitative relationships have generally been most easily found between soil topography but

there is evidence of quantitative relationships with the other four soil-forming or soil-altering factors.
2. In general, the relationships cannot be assumed to be linear.
3. Little work has been done on interactions between factors.
4. Soil can be spatially predicted from geographic position using a variety of techniques.
5. Soil can be predicted from other soil attributes at the same location.
6. Soil can be predicted from itself, other soil attributes and environmental attributes at neighbouring locations.

We now go on to incorporate these points in a more generic framework for soil spatial prediction.

3. The scorpan model

Here, we generalise and formalise the approach that has begun to emerge in papers published lately. We use a Jenny-like formulation not for explanation but for empirical quantitative descriptions of relationships between soil and other spatially referenced factors with a view to using these as soil spatial prediction functions. We consider seven factors:

- *s*: soil, other properties of the soil at a point;
- *c*: climate, climatic properties of the environment at a point;
- *o*: organisms, vegetation or fauna or human activity;
- *r*: topography, landscape attributes;
- *p*: parent material, lithology;
- *a*: age, the time factor;
- *n*: space, spatial position.

We have included soil as a factor because soil can be predicted from its properties, or soil properties from its class or other properties. We shall call this the scorpan model, which can be written as:

$$S_c = f(s,c,o,r,p,a,n) \text{ or } S_a = f(s,c,o,r,p,a,n)$$

where S_c is soil classes and S_a is soil attributes. The *s* refers to soil information either from a prior map, or from remote or proximal sensing or expert knowledge. Implicit in this are the spatial coordinates *x*,*y* (and probably not *z*) and an approximate or vague time coordinate $\sim t$. This time coordinate can be expressed as 'at about some time *t*'. So explicitly, e.g.,

$$S_c[x,y,\sim t] = f(s[x,y,\sim t], c[x,y,\sim t], o[x,y,\sim t], r[x,y,\sim t], p[x,y,\sim t], a[x,y], [x,y]).$$

Each factor will be represented by a set of one or more continuous or categorical variables, e.g., *c* by average annual rainfall and average annual temperature or climate class.

We shall not consider the direction of causality. For example, many reckon vegetation to be dependent on soil and we could write $o = g(S)$, where *o* is set of vegetation classes or percentage cover of a species, *g* is some arbitrary function and *S* is a set of soil classes or attributes. For our purpose, we could write $S = g^{-1}(o)$, where g^{-1} is the inverse function of *g*, $S = g^{-1}(o) = f(o)$. We stress that the approach in general is not theoretical; it is largely empirical—where there is evidence of a relationship we use it. Clearly, although we do not require causality, we should be mindful of potential problems of nonuniqueness if *g* is not a monotonic function, as it might if *S* is say topsoil pH and *o* is the number of plants of a particular species per unit area.

A general soil prediction model would be

$$S(x,y,z,t) = f(Q)$$

where *Q* is predictor variable(s). Here, we will consider some restrictions in cases where *S* stands for $S(x,y,(z),t)$, i.e., the soil class or attribute at some spatial location *x*, *y*, (*z*) and at some time *t*.

3.1. What is S? Soil classes S_c or individual soil attributes S_a

The model must be able to predict the probability of a set of classes, e.g., for the case of five classes, say, A, …, E, the model would predict the probability vector (p[A], p[B], p[C], p[D], p[E]), e.g., $S_c[x,y]$=(0.01, 0.72, 0.01, 0.02, 0.25) along with some measure of uncertainty. The problem will generally consist of a preexisting soil class label (from some soil classification system) at each soil observation location and a set of colocated environmental variables. These are called the training data. This represents a supervised classification or supervised learning problem.

More rarely, unsupervised learning, also known as numerical classification, may be used on observed soil attributes to first generate the class labels. The supervised learning rules are fitted using the training data, and then used at other locations where only environmental variables are observed.

Alternatively, the model should be able to predict individual soil attributes S_a along with a measure of uncertainty. The S_a might be the value of a given soil attribute at a certain depth, e.g., the clay content at 60 cm, i.e., $S_a[x,y] = 310$ g/kg, along with an uncertainty measure. Similarly to the class problem, this will generally consist of a measured soil attribute at each soil observation location and a set of colocated environmental variables. These are the training or calibration data. This represents a generic (multiple) regression problem. The generic regression equations or rules are fitted using the calibration data and are then used at other locations where only environmental variables are observed.

We shall not consider the case $S_a = f(s)$, with no spatial consideration—these are the so-called and very useful pedotransfer functions (PTF). Much work has been done on these and it has been reviewed elsewhere (Wösten et al., 2001; McBratney et al., 2002). This point is further elaborated in Section 5.3.1.) As we shall see the form of f for pedotransfer functions and our soil spatial prediction functions (SSPF) are not unrelated.

Heuvelink and Webster (2001) have discussed the merger of discrete and continuous models of spatial variation. Heuvelink (1996) suggested the mixed model of spatial variation, in which the soil property is treated as the sum of a global mean, a class-dependent deviation from the mean and a spatially correlated residual. Prediction with this model boils down to kriging with an external drift (Delhomme, 1978), which in this case is a classification. Its main advantage is that it performs well over the whole range of spatial variation, from exclusively discrete realities. A more general interpretation of this kind of idea, and the one we use here is that the external drift represents $f()$ and can be any kind of function. The discreteness or continuity of S will depend on the magnitude and form of $f()$. In the Heuvelink (1996) case, the $f()$ is a one-way analysis of variance model, a special case of a generalised linear model (McCullagh and Nelder, 1983; Lane, 2002).

3.2. The general approach

If we write the equation as $S = f(Q) + e$, then the general approach we shall use is to take some observations of S in the field at known locations $[x,y]$ and fit some kind of function a set of pedologically meaningful predictor variables Q which will be generally raster data layers of size M in a GIS. Once the model is fitted at the m observation points, the prediction can be extended to the M points or cells in the raster thereby giving a digital map. The efficiency of the method relies on the fact that hopefully $m \ll M$, and because S is much more difficult and expensive to measure than the Q. The success will depend on:

1. Having sufficient predictor variables observed everywhere or at least with a relatively high data density.
2. Having enough soil observations (data points) to fit a relationship.
3. Having functions $f()$ flexible enough to fit a nonlinear relationship.
4. Having a good relationship between the soil and its environment.

Followed by a discussion of quantitative procedures for fitting $f()$ in Section 3.3, we present some considerations concerning e in Section 3.4, and a review of previous studies in Section 3.5.

3.3. Form of f()

$f()$ is some form of empirical quantitative function f linking S to the scorpan factors (). There are different combinations of predictors and predicted variables that can be summarised in Table 2. If soil classes S_c, are to be predicted, they can be hard or fuzzy. For the

Table 2
Useful combinations of predictor and predicted attributes (*)

Predicted S	Predictor				
	Class		Continuous	Fuzzy	Mixed
	Hard	Fuzzy			
Hard class, S_{ch}	*				
Fuzzy class, S_{cf}	*	*			
Continuous, S_{ph}	*	*	*	*	
Fuzzy, S_{pf}	*	*	*	*	
Mixed, S_{pm}	*	*	*	*	

prediction of soil attributes, the variables can be continuous, fuzzy or mixed. The * symbols in the table represents *efficient* prediction of S from the given predictors (McBratney et al., 2002).

We will now discuss some forms of $f()$, most, but by no means all, of which have been or can be used for this kind of problem. For the sake of brevity, we shall not delve deeply into the mathematics of the methods. The advance in statistical learning techniques, enhanced by the growing need in data mining, has aided the use of different forms of $f()$ in soil science. Recent developments and technical details of the statistical modelling have been recently and extensively reviewed by Hastie et al. (2001). When predicting soil classes some kind of *supervised classification* will be used, and for soil attributes some kind of *generic regression* will be used. These are now discussed.

3.3.1. Linear models

Linear models include regression for predicting soil attributes, and classification for predicting soil classes. Linear regression included in this section is linear models using ordinary or generalized least squares. Linear methods for classification include discriminant analysis. The theory can be found in Hastie et al. (2001).

3.3.1.1. Ordinary least squares. For multiple linear regression, the model is written as:

$$\mathbf{s} = \mathbf{Q}\mathbf{b} + \mathbf{e}$$

where **s** is the vector of response (predicted soil attribute), **Q** is the matrix of predictor variables and **b** is parameter vector of the linear function. The error component, **e** represents of the deviations of the model to the observed value. The parameter is usually solved using ordinary least squares (OLS), with assumptions that **e**:

1. is independently and identically distributed (independence assumption),
2. have zero mean and finite variance (homoscedasticity assumption) and
3. is normally distributed (normality assumption).

OLS has been used widely in prediction of soil attributes because of the easiness and wide availability. The predictors are usually continuous variables. However, qualitative factors or discrete variables can also be integrated. This involves coding the factor of K levels into $K-1$ variables. Such coding is automatically generated in statistical packages such as S-Plus.

3.3.1.2. Principal component regression and partial least squares. When large number of correlated predictor variables are present (such as electromagnetic spectra), principal component analysis is usually used to produce linear combinations of the original inputs. Selected principal components are then used in place of the original predictors. Alternatively, partial least squares (PLS) (Martens and Naes, 1989) are developed which constructs a new set of components as regressor variables which are linear combination of the original variables. Unlike principal component regression which only used the combination of the predictors, the components in partial least squares are determined by both the response variable(s) and the predictor variables.

Principal component regression and PLS have been used quite extensively in predicting soil attributes from electromagnetic spectrum, especially in the near- and mid-infrared ranges (such as Chang et al., 2001). This method may be necessary if the environmental covariates consist of hyperspectral imagery.

3.3.1.3. Linear discriminant analysis. Discriminant analysis (Fisher, 1936) is the seminal supervised learning technique. It has been applied in soil science for more than 60 years. The first application was by Cox and Martin (1937) in which discriminant analysis was used to determine whether chemical properties give significant information on the presence of *Azotobacter* in soil. Webster and Burrough (1974) used the method to allocate soil observations into existing classes. Henderson and Ragg (1980) employed a multivariate logistic method to assess the usefulness of soil properties for distinguishing between taxonomic units. The method was perhaps first used for digital soil mapping by Bell et al. (1992, 1994) who related soil drainage classes to landscape parameters, and used the resulting discriminant functions for spatial predictions. Other examples can be seen in Table 3.

The theory is readily accessible in Webster and Oliver (1990) and Hastie et al. (2001). Triantafilis et al. (2003) have generalised the theory to a fuzzy linear

Table 3
Summary, in chronological order, of previous quantitative scorpan-like studies in which soil classes and/or attributes were spatially predicted

Soil		Predictive model (f)	Predictive factors							Study area spatial extent	No. of observations	Grid distance (m)		Location	Authors	Scale of map produced (1:x)	Study area (km^2)
S_{class}	$S_{attribute}$		s	c	o	r	p	a	n			Soil sample	Image				
Soil drainage classes		Linear regression				×				D3				USA	Troeh (1964)		
	Soil horizon thickness, subsoil mottle, depth to mottle	Linear regression				×				D1	90		10	USA	Walker et al. (1968)		0.007
Soil classes		Discriminant analysis	×							D3	30	1000		USA	Pavlik and Hole (1977)		450, 250
Soil classes	Degree of podzolisation	Modified principal component analysis (Escoufier, 1970)				×	×			D4	38		500	France	Legros and Bonneric (1979)	500,000	624
	Thickness of A horizon, depth to $CaCO_3$	Discriminant analysis, linear regression				×				D2	522	10, 50	10, 50	Canada	Pennock et al. (1987)		
Soil classes		Clustering	×			×				D2				USA	Lee et al. (1988)		
	Organic C, Fe/C	Clustering and regression			×					D2	32			USA	Frazier and Cheng (1989)		500
	Organic C, P	Regression, kriging			×				×	D2	172	15	15	USA	Bhatti et al. (1991)		0.26
	Soil morphological, physical and chemical properties	Ordination techniques	×			×				D2	194	2, 8	10	Australia	Odeh et al. (1991)	10,000	0.26
Soil classes			×			×				D2	194	2, 8	10	Australia	Odeh et al. (1992)	10,000	0.26
Soil drainage classes		Discriminant analysis				×	×			D3	305			USA	Bell et al. (1992, 1994)		
	Clay content, CEC, EC, pH, bulk density, COLE, θ at -10 and -1500 kPa	Ordination, GLM				×	×			D3	224	300	100	Lower Macquarie Valley, Australia	McKenzie and Austin (1993)	100,000	500

	Horizon thickness, OM, pH, extract. P, silt, sand	Linear regression			×			D2	231	15	15	Australia	Moore et al. (1993)	15,000	0.054
Soil series		Rule induction			×	×		D2	Digital soil map		25	New Zealand	Dymond and Luckman (1994)	15,000	
	Depth to solum, depth to bedrock, topsoil gravel, subsoil clay	Linear regression, kriging, co-kriging, regression kriging			×		x	D2	194	2, 8	10	Australia	Odeh et al. (1994, 1995)	10,000	0.26
	OM	Linear regression	×	×	×			D3	194		1000	France	Arrouays et al. (1995)	1,000,000	
	Horizon depth, presence of E horizon	GLM			×			D3	60	500	10	Australia	Gessler et al. (1995)	100,000	100
	Horizon depth	Kriging, co-kriging, regression kriging	×				x	D2	539, 117	35		Netherlands	Knotters et al. (1995)		0.97
	Soil properties	Expert/ rule-based system	×					D2	231	15	15	Australia	Cook et al. (1996a)	15,000	0.054
	Soil parent material	Clustering				×		D3			70	Australia	Cook et al. (1996b)		200
Soil units		Bayesian and expert system rules	×		×	×		D2	53	95, 110	10	Australia	Skidmore et al. (1996)	15,000	0.79, 1.28
	Soil available water capacity	Linear regression				×		D2, D3				USA	Zheng et al. (1996)		
Soil drainage classes		Classification tree			×	×		D2	Digital soil map	12	10	USA	Cialella et al. (1997)	12,000	24
Soil fertility classes		Fuzzy logic	×					D3	384	750		Philippines	Dobermann and Oberthur (1997)	250,000	192
Soil classes		Classification tree				×	×	D2	Digital soil map	50	50	France	Lagacherie and Holmes (1997)	50,000	35
	pH, EC, av.P, ex. K, ex. Ca, ex. Mg, Total N, total P	Linear regression			×	×		D2	103	50	10	Australia	Skidmore et al. (1997)	15,000	0.18
	Hydromorphic Index	Linear regression				×		D2	143	10–20	10	USA	Thompson et al. (1997, 2001)		

(continued on next page)

Table 3 (*continued*)

Soil		Predictive model (f)	Predictive factors							Study area spatial extent	No. of observations	Grid distance (m)		Location	Authors	Scale of map produced (1:x)	Study area (km^2)
S_{class}	$S_{attribute}$		s	c	o	r	p	a	n			Soil sample	Image				
	Wilting point	Kriging	×						×	D3	426	100, 141, 200	200	France	Voltz et al. (1997)	100,000	17, 36
Soil series	A horizon depth	Fuzzy logic and expert system			×	×	×			D2	64	30		USA	Zhu and Band (1994); Zhu et al. (1996, 1997, 2001)	30,000	36
	Transmissivity	Fuzzy logic and expert system				×	×			D1	32			Australia	Zhu et al. (1997)		5.5
Soil classes		Decision trees, Bayesian model				×	×			D4	Digital soil map	100, 2000	250	Australia	Bui et al. (1999)	250,000	1300
	Organic C	Look-up table, Bayesian rule				×				D3	72	2000			Lilburne et al. (1998)	130,000	260
	Soil depth, P total, C total	GLM, Regression tree		×		×	×			D3	165		25	Australia	McKenzie and Ryan (1999)		500
	Soil texture classes	Ordinary kriging, indicator kriging, indicator kriging with soft information	×						×	D3	384, 208	750, 2000	250, 250	Philippines, Thailand	Oberthur et al. (1999)	100,000	192, 390
	Soil depth	Discriminant analysis	×			×				D2	2448	50	12.5	Germany	Sinowski and Auerswald (1999)		1.5
Soil classes		GLM (Discriminant Analysis)				×	×			D3	120, 100		50	France	Thomas et al. (1999)	100,000	60
	Soil organic C	Linear regression	×		×	×					745	10	25	USA	Bell et al. (2000)		
	Thickness of horizon	Kriging with external drift				×			×	D2	219		20	France	Bourennane et al. (2000)	20,000	0.38
	Hydromorphic Index	Kriging, co-kriging				×			×	D1	182	10	10	France	Chaplot et al. (2000)		0.02
Soil classes		GLM (discriminant analysis)			×	×				D4	1236		1000, 100	Hungary	Dobos et al. (2000)	500,000, 1,000,000	93,000

	Wilting point	Conditional probabilities, kriging			×		×	D3	374	200	10	France	Lagacherie and Voltz (2000)	100,000	20
	Clay content, CEC	GLM, GAM, regression trees, neural networks, kriging, co-kriging, regression kriging	×		×	×		D1, D2, D3	95, 180, 734	2800	2, 200, 500	Australia	McBratney et al. (2000)	2000, 200,000, 500,000	0.42, 1100, 45,600
	Horizon depth and chemical properties	ANOVA			×			D3	72	2000		New Zealand	McIntosh et al. (2000)	130,000	260
	Clay content	Linear regression, regression kriging			×		×	D3			1000	Australia	Odeh and McBratney (2000)		
	Soil depth, C, P density, site water capacity	Regression trees and GLM (multiple regression)		×	×	×	×		D3	50, 165			Australia	Ryan et al. (2000)		
Soil series	A horizon depth	ANN			×	×	×		D2	64	30		USA	Zhu (2000)		36
	CEC	GAM, regression tree, linear regression, kriging	×		×	×		×	D1	113		5	Australia	Bishop and McBratney (2001)	5000	0.74
Soil classes		Decision tree	×			×	×		D4	Digital soil maps		250	Australia	Bui and Moran (2001)	250,000	1,058,000
Soil units		Discriminant analysis			×	×					1000	1000	Part of Europe	Dobos et al. (2001)	1,000,000	1,650,000
	Thickness of horizon	Correlation				×			D2	160	10–20	10	USA	Park et al. (2001)		0.9
	Organic C, N_2O emission	Landform segmentation				×			D2	99	25	10	Canada	Pennock and Corre (2001)		
Soil classes		Fuzzy classification	×						D4	600		200	Czech Republic	Boruvka et al. (2002)	200,000	1327
Soil drainage classes		Logistic regression			×	×			D3	295 + 72		25	Nigeria	Campling et al. (2002)	50,000	589
Soil horizon and soil classes		Fuzzy logic	×			×	×		D3	Rast. soil map	10	20, 25	France	Carre and Girard (2002)	100,000	1054
	Hydromorphic Index	Logistic regression				×			D3	141 + 41 + 54 + 162 + 308		30	France	Chaplot and Water (2002)		30,000

(continued on next page)

Table 3 (*continued*)

Soil		Predictive model (f)	Predictive factors							Study area spatial extent	No. of observations	Grid distance (m)		Location	Authors	Scale of map produced (1:x)	Study area (km^2)
S_{class}	$S_{attribute}$		s	c	o	r	p	a	n			Soil sample	Image				
	Soil moisture, residual P, solum thickness, depth to CaCO3, OC	Linear regression				×				D2	210	21	15	Canada	Florinsky et al. (2002)		0.67
	Soil OC content	Linear regression				×				D2, D3, D5			15, 1000, 5 min	Canada	Florinsky and Eilers (2002)		0.64, 16,206, 2,450,000
	pH, organic matter	Linear regression		×	×	×				D4	2350	10,000	1000	Croatia	Hengl et al. (2002)	1,000,000	56,000
Soil drainage classes		Discriminant analysis, logistic discriminant	×			×					107	50	10	USA	Kravchenko et al. (2002)		0.02
Soil subgroups and groups		Lattice graphs, ANN				×	×			D3	2294	25	50	UK	Mayr et al. (submitted for publication)	50,000	100
Soil classes		'Boosted' classification tree			×	×				D4		Digital soil map	250	Australia	Moran and Bui (2002)	250,000	1300
	Silt, ECEC, TEB, Mn, oxided	ANN, regression tree, GLM	×		×	×				D1	502	25	10	UK	Park and Vlek (2002)		0.03
Soil drainage class		Supervised classification	×		×	×				D2	49		10	USA	Peng et al. (in press)		0.057
Surface gravel content		GLM, GAM, Reg.Tree, combined with regression kriging	×		×	×				D3				USA	Scull et al. (in press)		800
	Available water capacity	Rule-based	×		×	×	×			D2			5	Germany	Sommer et al. (2003)		0.1
	Topsoil thickness, pH, organic matter	GLM and regression-kriging				×				D4	135	10,000	1000	Croatia	Hengl et al. (in press-b)	1,000,000	56,000

discriminant analysis. This considers the a priori membership of each individual to each of the classes.

3.3.2. Generalised linear models

Generalised linear models (GLMs) extend the linear regression models to accommodate with the nonnormal response distributions (Hastie and Pregibon, 1992). The theory and applications in soil science has been reviewed by Lane (2002).

Usually, to accommodate for nonlinearity, transformation of variable is introduced, GLM attempt to modify the model rather than transforming the data (Lane, 2002). GLMs have the assumption of independence between the response and predictor variables. The predictor variables q may influence the distribution of the predicted soil attribute S through a single linear function called linear predictor: $\eta = \sum_{j=1}^{p} \beta_j q_j$.

GLMs consist of two functions:

1. (i) A link function that describes how the mean depends on linear predictors.

$$\mu = m(\eta)$$

$$\eta = m^{-1}(\mu) = \ell(\mu)$$

where $\ell(.)$ is the link function.

2. (ii) A variance function that captures how the variance of the response depends upon the mean:

$$\mu = \int \eta.$$

The form of the link and variance function depends on the distribution of the response (McCullagh and Nelder, 1983).

The response distribution could be Gaussian, binomial, poisson or others, and each distribution function allows a variety of link functions, such as logit, probit, inverse (Venables and Ripley, 1994).

Various models can be derived from this generalisation by specifying the appropriate link function. For example, multiple linear regression corresponds to an identity link function, constant variance and a normal distribution. Logistic regression is a form of GLM where we wish to model the posterior probabilities of the K classes via linear functions of the predictors x and also ensure that they sum to unity within the range [0,1]. The model has the form:

$$\Pr(G = c \mid X = q) = \frac{\exp(\beta_{c0} + \beta_c^T q)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T q)},$$

$$c = 1, \ldots, K-1.$$

The model is usually fitted by maximum likelihood (Hastie et al., 2001).

Co-kriging, which is a case of general spatial linear model, is discussed under its own heading in Section 3.4.2.2 below.

3.3.2.1. Prediction of continuous soil attributes S_a. McKenzie and Austin (1993) used generalised linear models to predict soil attributes (clay content, CEC, EC, pH, bulk density and COLE) using environmental variables (geomorphic unit, local relief, etc.) as predictors. Other examples include Odeh et al. (1995, 1997) and McKenzie and Ryan (1999), who used GLMs to predict nonnormally distributed continuous variables. In both of these cases, the GLM was used in preference to standard linear regression because of the nonnormal distribution of the response variable. Park and Vlek (2002) found that GLMs performed better than neural networks and regression trees in predicting soil attributes from environmental variables.

3.3.2.2. Prediction of soil classes S_c. Gessler et al. (1995) used GLMs to predict the presence or absence of a bleached A2 horizon using digital terrain information. In this case, a logit link function was used due to the binomial distribution. Campling et al. (2002) used (multiple) logistic regression to model soil drainage classes from terrain attributes and vegetation indices as calculated from a Landsat TM image.

3.3.3. Generalised additive models

Generalised additive models (GAMs) attempt to characterise the nonlinear effect which is not considered in generalised linear models. GAMs have the form:

$$S = \alpha + \sum_{j=1}^{p} f_j(q_j) + e$$

where *f* are nonparametric 'smoothing' functions. The smoothing functions can be splines, loess, kernel and other smoothers (Venables and Ripley, 1994). Smoothers fit the data locally and crucial to the fit is the size of the local neighbourhood, which is generally controlled by a smoothing parameter. The smoothing parameter controls the variance-bias trade off. A large neighbourhood produces estimates with low variance and potentially high bias, a small neighbourhood produces the reverse effect (Hastie and Tibshirani, 1990).

The use of GAMs in the soil science literature has been minimal. One of the few studies has been Odeh et al. (1997), who compared GAMs with GLMs and linear regression for the prediction of organic carbon with digital terrain information as secondary information. GAMs were found to be superior. Bishop and McBratney (2001) used GAMs for the mapping of soil cation exchange capacity from environmental factors (terrain attributes, bare soil colour aerial photograph, bare soil LANDSAT imagery, crop yield data and soil apparent electrical conductivity).

3.3.4. Tree models—classification and regression

The term machine learning has been identified with

(a) an expert system whereby rule-based software is built from sample cases volunteered interactively (Section 3.3.10) and
(b) as a method of data analysis whereby rule-structured classifiers is built from a "training set" of preclassified cases (Feng and Michie, 1994).

We discuss the latter here. An early practical application is by Michalski and Chilauski (1980), who generated an automatic rule-based classifier for identifying soybean diseases from plant morphological description.

Rather than fitting a model to the data, a tree structure is generated by partitioning the data recursively into a number of groups, each division being chosen as to maximise some measure of difference in the response variable in the resulting two groups. It can handle both categorical and continuous data, for prediction of discrete soil classes or continuous soil attributes. The advantage of regression trees over linear model is the ability to deal with nonlinearity. In addition, they require no assumptions about the data and able to deal with nonadditive behaviour, while other regression methods require interactions to be prespecified (Breiman et al., 1984). Using a decision algorithm, the tree model decides automatically the splitting variables and splitting points, and also the shape (topology) of the tree. A popular method for classification and regression trees is called CART (Breiman et al., 1984).

The main advantage of tree models is they are easy to interpret as opposed to methods like GLMs, GAMs and neural networks (Clark and Pregibon, 1992). Because of this, regression trees have been widely used for the prediction of soil attributes and more recently for prediction of crop yield for site-specific management (Shatar and McBratney, 1999).

3.3.4.1. Prediction of continuous soil attributes S_a. This is called a regression tree. Pachepsky et al. (2001) used regression tree to predict sand and silt contents and water retention from terrain attributes (slope, curvature). Lapen et al. (2001) study the relationships between maize grain yields and management practice, soil strength/compaction and soil nutrient status.

One of the limitations in regression tree is the discrete predictions from each terminal node, which resulted in the lack of smoothness of the prediction surface. This can result in unrealistic representations of soil variability if the tree has a small number of terminal nodes (McKenzie and Ryan, 1999). An improvement to the regression trees is to build multivariate linear models in each node (leaf). This type of model, which is analogous to using piecewise linear functions, has been implemented in the program Cubist (RuleQuest Research, 2000). To further improve the prediction, we could apply fuzzy memberships to allow a smoother transition from one class to another (Jang, 1993, 1997).

3.3.4.2. Prediction of soil classes S_c. This is called a decision tree or a classification tree. The binary decision-tree algorithm uses a binary split which has exactly two branches at each internal node. There are different decision trees methods. The most commonly used is CART (Breiman et al., 1984). Lagacherie and Holmes (1997) discussed the application of CART for soil classification and its sensitivity to error. Another

popular algorithm is C4.5 (Quinlan, 1992) and its later version See5 (Rulequest Research, 2000). Bui and Moran (2003) utilised this program for mapping soil classes across the Murray–Darling Basin in eastern Australia. Moran and Bui (2002) refined the analysis of Bui et al. (1999) by using a 'boosted' tree (see Section 3.3.8) to reduce the classification error.

Zighed and Rakotomalala (2000) generalise decision trees into decision graphs using the SIPINA algorithm (Zighed, 1985) and the Fusbin and Fusinter discretisation methods (Zighed et al., 1996) where a new operator, called "merge between leaves", is introduced during the growing process. This approach was employed by Mayr et al. (2003).

3.3.5. Neural networks

Neural networks attempt to build a mathematical model that supposedly works in an analogous way to the human brain. Neural networks have a system of many elements or 'neurons' interconnected by communication channels or 'connectors' which usually carry numeric data, encoded by a variety of means and organised into layers. Neural networks can perform a particular function when certain values are assigned to the connections or 'weights' between elements. To describe a system, there is no assumed structure of the model, instead the networks are adjusted or 'trained' so that a particular input leads to a specific target output (Gershenfeld, 1999). The mathematical model of a neural network comprises of a set of simple functions linked together by weights. The network consists of a set of input units, output units, and hidden units, which link the inputs to outputs. The hidden units extract useful information from inputs and use them to predict the outputs.

Neural networks are now widely used in the soil science literature, mainly for predicting soil attributes. The application of neural networks as pedotransfer functions for predicting soil hydraulic properties is the most common.

3.3.5.1. Prediction of continuous soil attributes S_a. The application in predicting soil hydraulic properties in the form of pedotransfer functions can be found in many studies such as Minasny and McBratney (2002). Chang and Islam (2000) predict soil texture from multitemporal remotely sensed brightness temperature and soil moisture maps.

3.3.5.2. Prediction of soil classes S_c. Neural networks can be used to predict the probability of classes using multi-logit transformation of the output. Another type of network is called self-organising maps (SOM, in this case, not soil organic matter) (Kohonen, 1982). Kohonen's network is an unsupervised classification splitting input space into patches with corresponding classes. It has the additional feature that the centres are arranged in a low dimensional structure (usually a string, or a square grid), such that nearby points in the topological structure (the string or grid) map to nearby points in the attribute space.

Zhu (2000) used neural networks to predict the probability of soil classes from soil environmental factors. Fidêncio et al. (2001) applied two types of neural networks (radial basis function networks and self-organising maps) to classify soil samples from different geographical regions in Sao Paulo, Brazil by means of their near-infrared (diffuse reflectance) spectra.

3.3.6. Fuzzy systems

Fuzzy systems attempt to represent the uncertainty in the predictor and predicted attributes or classes. Fuzzy inference systems map a given input to an expected output using fuzzy logic. The most commonly inference system used are the Mamdani and the Sugeno type, which are described in Jang (1997). The steps usually involve fuzzifying the 'hard' input variables, define the rule or fuzzy operator, apply an implication method, aggregate the outputs, and defuzzify the outputs. Dobermann and Oberthur (1997) used fuzzy logic to produce a soil fertility map for rice from soil variables. Zhu (1997) represent a soil at a given location with a vector of membership values (or so called Soil Similarity Vector) to the existing soil classes. Zhu et al. (1996) and Zhu et al. (1997) used fuzzy logic to infer the membership of a soil to particular classes from the environmental variables, such as parent material, elevation, aspect, gradient, profile curvature and canopy coverage.

The model Adaptive Neuro-Fuzzy Inference Systems (ANFIS) (Jang, 1993) is analogous to neural networks and can be used to predict continuous variables. McBratney et al. (2002) give an example for predicting hydraulic conductivity.

3.3.7. Other methods

Genetic algorithms (GA) (Goldberg, 1989) are randomised search and optimisation techniques guided by the principles of biological evolution and natural genetics. They have been used mainly in optimisation of large multidimensional problems. Pal et al. (1998) developed a GA classifier and applied it to satellite imagery (Pal et al., 2001). It attempts to approximate the class boundaries of a data set with a fixed number of hyperplanes in such a manner that the associated misclassification of data points is minimised. Maulik and Bandyopadhyay (2000) proposed a genetic algorithm for unsupervised classification.

Various models used in data mining are also available such as Multivariate Adaptive Regression Splines (MARS) to model continuous variables (Friedman, 1991; Hastie et al., 2001). MARS has been used by Shepherd and Walsh (2002) to build prediction equations for eastern Africa for a number of soil properties from NIR diffuse reflectance spectra. Another data mining tool made available recently is TreeNet (Friedman, 1999; http://www.salford-systems.com/treenet.html), which forms a network with several dozen to several hundred small trees, each typically no larger than two to eight terminal nodes. The model is analogous to a long series expansion, such as a Fourier or Taylor's series, where a sum of factors becomes progressively more accurate as the expansion continues.

3.3.8. Strengthening models: bagging, boosting

There has recently been empirical evidence that the accuracy of $f()$ prediction can be enhanced by generating multiple models and aggregating them to produce an estimate. There are two renowned approaches for producing and using several models that are applicable to a wide variety of statistical learning methods. Bootstrap aggregating or bagging (Breiman, 1996) and boosting (Freund and Schapire, 1996) manipulate the training data in order to generate different models. These methods arise more naturally in the supervised classification problem, but they can be extended to generic regression.

Bootstrap methods (Efron and Tibshirani, 1993) assess the accuracy of a prediction by sampling the training data with replacement. Suppose the training data is composed of predictors Q and response S of size N, we draw B datasets each of size N of the training data by sampling with replacement. For each of the bootstrap dataset Z^b, $b=1, 2, \ldots, B$, we fit model $\hat{f}^b(q)$. The bagging estimate is calculated as:

$$\hat{f}_{\text{bag}}(q) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(q).$$

Boosting combines the outputs of many "weak" models to produce a powerful "committee". Boosting uses all the data at each repetition, but maintains a weight for each instance in the training set that reflects its importance. Adjusting the weights causes the model to focus on different data, hence leads to different models. The multiple models are then aggregated by voting to form a composite model. In bagging, each component model has the same vote, while boosting assigns different voting strengths to component classifiers on the basis of their accuracy. Moran and Bui (2002) used boosting to improve their digital soil map of the Murray–Darling basin. A popular algorithm for supervised classifiers is known as AdaBoost (Freund and Schapire, 1997).

3.3.9. Expert (knowledge-based) systems

Expert systems (Dale et al., 1989) are ways of harvesting and engineering knowledge. Bui (2003) argues that soil maps and their legends are representations of structured knowledge, namely the soil surveyor's mental soil–landscape model. Methods which can make such models explicit from previously surveyed areas are potential ways of producing $f()$ for classes. Bui (2003) and Wielemaker et al. (2001) suggest methodological frameworks to formalise the landscape knowledge of the soil surveyor is by structuring terrain objects in a nested hierarchy followed by inference and formalisation of knowledge rules. The principal attempts to formally make such knowledge rules are *Prospector* (Duda et al., 1978), *Expector* (Skidmore et al., 1991; Cook et al., 1996a), and *Netica* (http://www.norsys.com/netica.html).

The *Expector* approach described by Cook et al. (1996a) and Corner et al. (1997). *Expector* builds on existing soil surveyor knowledge to construct quantitative statements about individual soil properties via the development of a network of rules. These rules operate within a system of Bayesian inference to

assign the varying probability of occurrence of a soil property of interest within an area, given evidence that relates to it in a known way. Permissible evidence includes the range of attributes normally used by a soil surveyor, such as landform, vegetation, land use, or parent material, and can also include remotely sensed digital data. Evidence is weighted according to the uncertainty associated with it, and combined to produce a single estimate of probability of a given attribute. The relationship between the evidence and prediction is stated explicitly at each stage of the procedure and is, thus, repeatable in a consistent manner. *Expector* generates maps showing the probabilities of each hypothesis class. *Expector* was used by Lilburne et al. (1998) to predict topsoil carbon class from elevation (*r*), plan curvature (*r*), solar radiation (*c* derived from *r*), soil order (*s*) and vegetation class (*o*). Clearly these methods grade into some of the ones described above. *Expector* is probably closer to the methods above than it is to a pure soil surveyor knowledge-based approach.

3.3.10. Unsupervised classification

In the previous sections, when we were discussing classes, we were considering 'supervised classification'. This is also known as allocation or identification. This is where we wish to produce prediction equations for placing soil existing soil classes, such as a particular categorical level in a national or international classification system. However, we may first wish to make new classes from the observed soil properties. This is known as unsupervised classification. Much of the early work on pedometrics, in the 1960s, focussed on this topic. The numerical classification methods that have been used quite extensively in soil science more recently are *k*-means and fuzzy *k*-means (Odeh et al., 1992; de Bruin and Stein, 1998; Triantafilis et al., 2001). There is also a semi-supervised classification considering classification in the presence of some labelled data (Pedrycz and Waletzky, 1997).

Unsupervised classification is an option for making digital soil class maps, particularly where the national or international scheme does not project well onto the soil–landscape. However, once the new classes have been established at the soil observation locations, then one of the previous methods, inter alia discriminant analysis, multiple logistic regression, regression trees, needs to be applied to fit equations and then make predictions from environmental covariates at the other locations where no soil properties have been observed.

Carré and Girard (2002) used a continuous method for horizon and profile classification called OSACA. This was based mainly on field soil morphological attributes. Their method is unique in that it models the taxonomic distance to each of the class centroids at each observation site. Because these distances are continuous variables multiple linear regression on environmental variables was used as the 'supervised classification' step. One regression equation was developed for each class and the distances predicted at each site on their prediction raster.

3.4. Spatial considerations

The older corpt approach has no intrinsic or formal spatial component other than the functions are predicted in a spatial context, i.e., spatial position is not taken into consideration. This seems unwise for a mapping application. Spatialisation can be introduced by considering spatial components of the environmental and soil variables (Section 3.4.1) and by perpending the spatial correlation structure of the residuals (Section 3.4.1), as was briefly discussed in Section 3.3.1.

3.4.1. Decomposition of Q factors into spatial components

All of the methods described above find or induce relationships between soil *S* and the predictor variables *Q*. Potentially, each of the seven scorpan factors (with the possible exception of *n?*) can be described by a series of mapped spatial variables. Each of these variables can be decomposed into separate spatial components and mapped separately. Two ways to do this is by factorial kriging analysis (Bourgault, 1994; Wen and Sinding-Larsen, 1997; Oliver et al., 2000) and wavelets (Garguet-Duport, 1997; Zhu and Yang, 1998; Carvalho et al., 2001). Both methods decompose the separate variables into separate hierarchical spatial components of decreasing spatial resolution. The factorial kriging method assumes stationarity but the wavelet method does not require this. On the other hand, the factorial kriging method finds the scale of the components from the observations, whereas in wavelets the various scales are dictated by the size of

the image, i.e., the scales are increasing powers of 2 pixels. These components could all be derived and used as separate layers in the fitting of *s*. It is more than likely that the short spatial range components (e.g., the nugget component) might not relate to soil and can be removed. In a quite a different application Oliver et al. (2000) found that land use was related to a long-range spatial component in SPOT imagery and not to two shorter-range components. Lark et al. (2003) regressed soil properties on the wavelet components of proximally sensed soil electrical conductivity data.

3.4.2. Structure in e-generalised least squares and geostatistics

It would be naïve to imagine that there is no spatial structure in *e*. If *e* has a spatial structure, then generalised least squares or geostatistics can be applied. Why would *e* have a spatial structure? The answers could be:

- scorpan is incorrect.
- Attributes used to describe scorpan are inadequate.
- Interactions are misspecified.
- Form of *f*() is misspecified.
- Something intrinsic—such as spatial diffusion, interaction or inhibition processes.

Variograms of the fitted parts of the soil spatial prediction functions for the various factors will be instructive in elaborating these possibilities.

3.4.2.1. Generalised least squares. In generalised least squares (GLS) (Cressie, 1993): $\mathbf{s} = \mathbf{Qb} + \mathbf{e}$; errors $\mathbf{e}$ belong to multivariate normal distribution with mean 0 and covariance matrix $\mathbf{V}$: $\mathbf{N}(0, \mathbf{V})$. For spatial data, it can be further simplified assuming the error is homogenous with variance σ^2, thusm $\mathbf{V}$ can be replaced by $\sigma^2\mathbf{C}$, where $\mathbf{C}$ is the correlation matrix of the errors (Lark, 2000). For spatial data, the correlation matrix of the residuals can be computed from the semivariogram:

$$C_{ij} = 1 - \frac{\hat{\gamma}^*(d_{ij})}{\sigma^2}$$

where $\hat{\gamma}^*$ is a semivariogram function (which will produce a positive definite correlation matrix), d_{ij} is the Pythagorean distance between the *i*th and *j*th points. The log-likelihood criterion function for GLS then is:

$$L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2\sigma^2}(\mathbf{s}-\mathbf{Qb})^T\mathbf{C}^{-1}(\mathbf{s}-\mathbf{Qb})$$

The estimates of the parameters need to be done iteratively using a nonlinear optimisation technique (Pinheiro and Bates, 2000) minimising:

$$R = (\mathbf{s}-\mathbf{Qb})^T\mathbf{C}^{-1}(\mathbf{s}-\mathbf{Qb})$$

The procedures are:

- estimate parameter vector $\mathbf{b}$: $\hat{\mathbf{b}}=[\mathbf{Q}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Q}]^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{s}$.
- Calculate the residuals $\mathbf{e}$.
- Calculate the semivariance of the residuals.
- Fit a semivariogram model to the data.
- Calculate the correlation matrix of the residuals $\mathbf{C}$.
- Calculate *R*.
- Repeat until *R* is minimised.

GLS has been used in soil science literature. For example, Samra et al. (1991) predicted tree growth from soil sodicity parameters with spatially correlated errors. Other examples include Aiken et al. (1991), Opsomer et al. (1999) and Vold et al. (1999). Lark (2000) provided the theory and example of using GLS for mapping soil organic matter content. Pachepsky et al. (2001) used generalised least squares model with correlated error for prediction of water retention from terrain attributes, where *e* were modelled by the semivariance. The drawback with this method is the heavy computation time when large number of data is involved, as the calculation time of semivariance and inverse of the correlation matrix $\mathbf{C}$ will increase (approximately cubically with sample size). Nevertheless, Opsomer et al. (1999) showed some mathematical manipulation to avoid the inversion of the whole matrix. Pace and Barry (1997) developed spatial autoregressive model, which utilised the sparse matrix technique to allow for quick computation for large spatial data. Hengl et al. (2003a) utilised GLS combined with regression kriging for spatial prediction of soil properties in Croatia.

3.4.2.2. Geostatistics—scorpan kriging. Here, we recognise that the spatial "trend" can be described by $f(s,c,o,r,p,a,n)$ and the residuals e modelled by variograms and some form of kriging. The final prediction is the sum of $f()$ and e.

scorpan-universal kriging. Universal kriging allows incorporation of both deterministic and stochastic components in kriging:

$$S(\mathbf{x}) = \sum_{j=0}^{p} b_j q_j(\mathbf{x}) + e(\mathbf{x}).$$

The first term represents the nonstationary trend, which is modelled as a set of linear functions of the environmental variables $\mathbf{Q}$ with parameter vector $\boldsymbol{b}$, and the second term is the stochastic component modelled by variogram. Universal kriging can be solved by modifying the kriging system. However, the trend function is only limited to linear functions, and when the number of variables p is large, the matrix inversion to solve the system can consume heavy computation time.

Alternatively, the trend function can be modelled separately, where kriging is combined with regression (Ahmed and DeMarsily, 1987; Knotters et al., 1995). This method involves regression of the soil attributes as a function of predictor variables. This is followed by kriging of the regressed values, where the variance of the predicted (from the regression model) is used as the uncertainty of the modified kriging system. This is also known as kriging with uncertain data (Ahmed and DeMarsily, 1987). Odeh et al. (1994, 1995) defined regression kriging where model $f()$ is used to describe the relationship between predictors and soil attributes:

$$S(\mathbf{x}) = f(\mathbf{Q}, \mathbf{x}) + e'(\mathbf{x})$$

where $f(\mathbf{Q}, \mathbf{x})$ is a function describing the structural component of S as a function of $\mathbf{Q}$ at $\mathbf{x}$, $e'(\mathbf{x})$ is the locally varying, spatially dependent residuals from $f(\mathbf{Q}, \mathbf{x})$. In regression kriging, the soil property S at unvisited site is first predicted by $f()$, and followed by kriging of the residuals of the model.

As discussed at the end of Section 3.3.10, Carré and Girard (2002) used a continuous method for horizon and profile classification followed by multiple linear regression on environmental variables. One regression equation was developed for the taxonomic distance to each class centroid and the taxonomic distances predicted at each site on their prediction raster. The residual taxonomic distances to each class centroid were then spatially predicted onto the raster using ordinary kriging and added to the taxonomic distances from regression analysis. The summed taxonomic distances were then displayed and manipulated to make class maps. In this way, regression kriging can be used to make continuous or discrete soil class maps taking into account the spatial correlation structure of the residuals from the fitted classes at each data point.

scorpan–simple kriging. Because $f()$ is modelled under the assumption that e has zero mean, simple kriging can be applied to the residuals of the model. Simple kriging allows prediction of the spatially correlated residuals with known mean where the weights of the kriging equation do not need to sum to unity (Webster and Oliver, 1990).

scorpan–compositional kriging. So far, kriging has been used mainly to predict soil attributes, for prediction of soil classes incorporating predictor variables $\mathbf{Q}$, a form of compositional kriging with external trend is proposed:

$$\Pr[S_c(\mathbf{x})] = f(\mathbf{Q}, \mathbf{x}) + e'_c(\mathbf{x})$$

where $\Pr[S_c(\mathbf{x})]$ is the probability of the soil at $\mathbf{x}$ belongs to soil class c. The probability of the soil classes $c = 1, \ldots K$ must sum to 1 and the residuals of the probability must sum to 0:

$$\sum_{c=1}^{K} \Pr[S_c] = 1$$

$$\sum_{c=1}^{K} e'_c = 0$$

Solution of this method will involve prediction of soil classes using a form of $f()$, such as logistic regression, and compositional kriging of their residuals (Walvoort and De Gruijter, 2001).

3.4.2.3. Geostatistics—co-kriging and coregionalisation analysis. Another method is co-kriging. Any of the q layers can be a covariate in co-kriging. Indeed, many people have used this. The major problems

with co-kriging have been twofold. First, the parameters of all the $[(q+1)q]/2$ variograms and cross-variograms have to be estimated and the parameters have to obey a strict inequality (Wackernagel, 1987). Secondly, and more importantly, the co-kriging model really assumes linear relationships between the predictor and predicted variables. The use of categorical predictors and predicted variables is also difficult. Odeh et al.'s (1995) experience was that co-kriging did not perform as well as $S=f(r)+e$ (regression kriging) and was more cumbersome to use. We prefer the scorpan model, but co-kriging should not be dismissed and the difficulties and restrictions will be overcome. It can be argued that co-kriging is a kind of generalised linear model. Co-kriging can be used for soil attributes and compositional (co)kriging (de Gruijter et al., 1997; Walvoort and De Gruijter, 2001) can be used for indicators or probabilities of discrete soil classes or memberships of continuous ones.

Coregionalisation analysis. Even if co-kriging is not done, coregionalisation analysis (e.g., Lark and Papritz, 2003) is a very instructive way of studying the linear spatial relationships between soil and the predictor variables Q. This will indicate the spatial scales over which we might expect linear relationships to hold.

3.4.3. Other spatial methods

Bayesian maximum entropy (BME) was introduced by Christakos (1990, 2000). This approach allows the incorporation of a wide variety of hard and soft data in a spatial estimation context. The data sources may come in various forms, such as intervals of values, probability density functions (pdf) or physical laws (Christakos, 2000). Bogaert and D'or (2002) used BME algorithm and a Monte Carlo procedure (BME/MC) to generate map of particle-size distributions from a limited number of accurate measurements and a spatially exhaustive soil map. Compared with ordinary kriging (OK), this approach has the advantage of using soft information on a sound theoretical basis.

Fractal interpolation has arisen as an effort to preserve the spatial variability of original data when transferred across scales. Bindlish and Barros (1996) used a fractal interpolation method to map digital elevation data at different spatial resolutions. With a fractional Brownian surface as the interpolating basis function, they found that the fractal interpolation approach preserved well the spatial structure and the vertical scale of the data. Kim and Barros (2002) presented a downscaling model which includes spatially and temporally varying scaling functions, and the scaling functions are linear combinations of the spatial distributions of ancillary data. They demonstrated it with downscaling soil moisture fields from 10 to 1 km resolution using remote-sensing data.

3.5. Previous studies

Although the scorpan model has not previously been formalised, various authors have fitted parts of it. Here, we summarise the work of a large number of studies in Tables 3 and 4. We believe the tables cover a large proportion of the relevant studies. At least 70 studies have been completed over the last decade or so. About 35% of the studies are from Australia, 25% from the USA and 10% from France.

We can see that soil attributes have been estimated more often (70% of studies) than soil classes (30% of studies). The number of observations ranges from 30 to 2448 with a median value of 180, although several studies rely on digitised soil maps as the principal data source. The extent varies from a minimum of 0.007 km^2 to a maximum of 1,058,000 km^2 with a median of 30 km^2. The data density expressed as the number of observations per square kilometre varies from 0.00009357 to 1080 with a median of 5. The pixel size of the digital maps produced range from 2 m to 1 km with a median of around 20 m.

The key predictor factors are r (80% of studies) followed by s (35%), o and p (both 25%), n (20%) and c (5%) whereas a does not seem to have been used as a factor. A single factor is used in 40% of studies, two factors are used in 40% of studies, around 10% of studies considered three of the seven factors and 2% considered four factors. No studies considered five or more factors. The most common combination was r and s. Most studies used a DEM as the main source of ancillary data, followed by remotely sensed imagery and preexisting soil coverages.

A wide variety of methods for modelling $f()$ have been attempted. Generalised linear models mainly in the form of multiple regression has been the most common analysis tool to model $f()$, followed by co-

kriging. The use of regression trees and neural networks has so far not been widespread.

4. Sources of data—the seven scorpan factors

There are seven factors or sets of variables in the scorpan model which makes it different from Jenny's model. The aim is to obtain information on all of these. It will be a matter of convenience (access to data sources) and scientific contention which variables are used to represent the factors. Indeed, this is an area that has not been well enough studied. The creation of these digital maps of the input environmental variables representing the six factors in the scorpan model is seen as an integral part of the digital soil resource assessment approach and a very valuable, environmentally useful, by-product of the new approach. The layers can be used for other modelling purposes. Much of the earth science and ecological research of the last twenty years has been contributing towards the creation of these layers.

These digital surfaces themselves will be created using some kind of surface modelling procedure regression kriging or Laplacian smoothing splines or TINS. For D3 surveys, they should probably be produced on a 100 m raster with a block size of say 100×100 m.

Fig. 1 highlights the useful parts of the electromagnetic spectrum for obtaining information on soil and environmental variables through remote and proximal sensing.

Some technologies such as satellite remote sensing offer the possibility of providing indirect information on a number of the scorpan factors. For example, emission and subsequent detection of radio frequen-

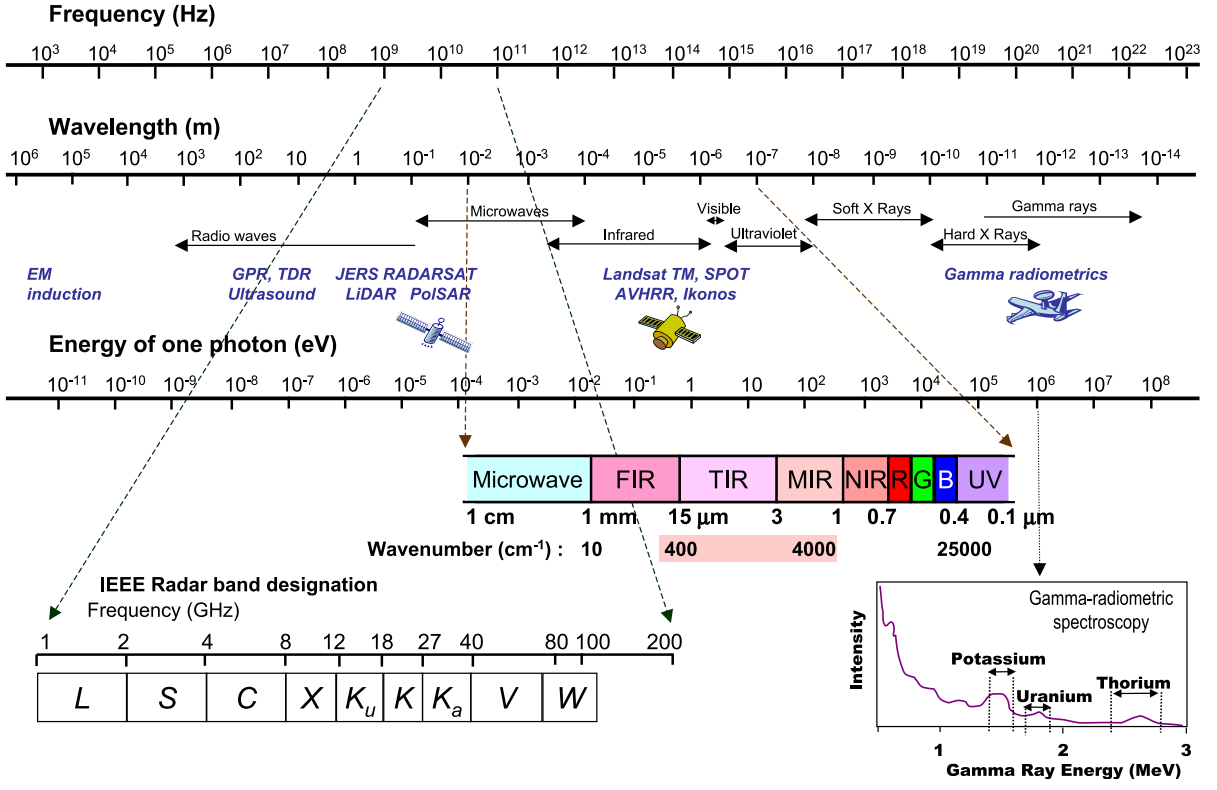

Fig. 1. The electromagnetic spectrum, highlighting the useful parts for obtaining information on soil and environmental variables through remote and proximal sensing. The boundaries for the infrared spectrum (NIR, MIR and FIR) are not consistent and vary between the chemical and remote-sensing literatures. The terms in this figure are based on remote-sensing literature, while the other literature defines the wavelength as: NIR: 0.7 to (2.5–5) μm, MIR: (2.5–5) to (25–40) μm and FIR: (25–40) to 1000 μm.

cies (radar) is promising for surface cover (*o*), biomass (*o*), surface roughness (*r* or *s*) and surface soil moisture (*s*) estimation. Radar is further enhanced by interferometric and polarimetric methods: InSAR interferometric synthetic aperture radar, and PolSAR polarimetric synthetic aperature radar. Radar operates over a number of frequency ranges: in order of decreasing frequency the so-called X, C, L and P bands. The C and L bands have been used in satellite platforms, e.g., C on ERS1 and 2 and L on JERS. The P band has so far only been available on airborne platforms. Hoekman and Quiñones (1999) and Quiñones and Hoekman (2001) have demonstrated the use of P band and PolSAR for mapping landcover and biomass in Colombian rainforests. LiDAR (light detection and ranging) uses pulses of laser light to illuminate the soil surface can be used to map surface roughness or topography (Bunkin and Bunkin, 2000; Schmugge et al., 2002).

4.1. s

Remote and proximal active and passive sensing gives detailed information on the soil themselves—these reflections or emissions or transmissions are intrinsic properties of the soil material and profile they may indicate other soil attributes like texture or mineralogy. This factor is likely to becoming increasingly important as technology advances.

4.1.1. Surface multi and hyperreflectance

Hyperspectral sensors are those which measure a large amount of bands with spectral resolutions less than 20 nm (Palacios-Orueta and Ustin, 1998). In terms of mapping soil information, hyperspectral sensors have been found to be useful in mapping mineralogical features such as iron oxides (King et al., 1995) carbonates and sulphates (Crowley, 1993). The amount of information generated by hyperspectral sensors poses computational problems in terms of extracting useful information in an efficient manner (see Section 3.3.1—principal component regression and partial least squares).

In a general way, the use of digital remotely sensing imagery for mapping soil has been problematic because vegetation cover obscures much of the soil response making it necessary to search for indirect evidence that may be visible at the surface (Campbell, 1987). Thus, remote sensing cannot be applied alone to soil studies (Lee et al., 1988). The use of proxies (such as topography, vegetation, drainage patterns) and field observations are important approaches for inferences about soil. In fact, mapping forest soil directly from remotely sensed data is difficult because of the complexity of environmental factors contributing to the spectral reflectance measured by a sensor. Nevertheless, forest soil was correctly mapped only where they were correlated with species or when vegetation is sparse or absent as a result of cultivation or drought (Post et al., 1994). The principal problems to soil delineation from such imagery, in addition to vegetation cover, are:

- Soil moisture content can interfere with the spectral reflectance (specially in the infrared, thermal and microwave regions (Obukhov and Orlov, 1964);
- Atmospheric effects (Cipra et al., 1980);
- Physical soil characteristics (Huette, 1988) and,
- Observation conditions (e.g., intensity and direction of illumination).

However, some research has been done in adjusting some of these drawbacks, by creating some indices with the main purpose of removing the effect of soil spectral influence (Huette, 1988) or even adjusting the images for vegetation interference and in combination with other information and developing prediction models for improving soil mapping (Odeh and McBratney, 2000). In this way, Dobos et al. (2001) used satellite data complemented with DEM data, in order to correct the distortions caused by topographic variations of the landscape and provide additional data for soil–landscape modelling.

The presence of vegetation cover attenuates electromagnetic radiation at most wavelengths (Skidmore et al., 1997), and the reflectance at the soil surface does not always reflect soil variation at depth (Agbu et al., 1990). While infrared and visible sensors only measure surface characteristics, radar and gamma radiometry can provide spectral information beyond the vegetative cover and the soil surface.

4.1.2. Radar attenuation

Information can be obtained by radar especially if there is a light-textured soil (low dielectric constant) over a heavier textured horizon (high dielectric) or a

water table (very high dielectric) and radar sensors can penetrate through soil to a depth that is equal to 10–25% of their wavelength (Lascano et al., 1998). The longest wavelength radar sensor available from space platforms is 23.5 cm (L band) on the JERS-1 satellite.

So far, we have only considered empirical methods where the multivariate prediction methods have been used to incorporate the remotely sensed imagery into prediction models. In the case of synthetic aperture radar (SAR), the back scattering signal is largely dependent on the dielectric properties of the media that is reflecting the signal, in the case of soil this is the volumetric moisture content (Moran et al., 1997). Therefore, physical models based on the theory of the diffraction of electromagnetic waves have been developed, an example being the Integral Equation Model (IEM) (Fung et al., 1992). The model calculates a backscattering coefficient that is based on:

radar sensor configuration, e.g., observation frequency, polarisation, incidence angle,
surface characteristics, e.g., roughness and dielectric properties.

The model was adapted successfully to predicting soil moisture from bare soil surfaces by Altese et al. (1996). The presence of vegetation complicates matters, and it is currently too difficult to create models describing the interaction between soil–vegetation layers and microwaves in real world applications. Therefore, in the presence of vegetation, empirical approaches are required (e.g., Dobson and Ulaby, 1986; Wood et al., 1993).

4.1.3. Electrical conductivity

Soil bulk electrical conductivity (or its reciprocal soil electric resistivity) reflects a combination of soil mineralogy, salts, moisture and texture, hence, it is a good compound measure of soil. Two commonly used kinds of instruments are: electromagnetic induction (EMI) and electrical conductivity/resistivity based on rolling electrodes (ECRE). The most widely used instruments for soil studies are the EMI devices from Geonics in Canada and two types of ECRE devices, a US design (Lund et al., 1999) and a French one (Tabbagh et al., 2000). These instruments have been used extensively in precision agriculture for mapping soil types and properties (such as Bishop and McBratney, 2001; Sudduth et al., 2001; Anderson-Cook et al., 2002). Such proximal sensing offers the possibility of producing high resolution maps of soil properties (D1 surveys of Table 1). Regression equations have been developed to predict moisture content, topsoil thickness, and clay content. EMI instruments can be placed in airborne platforms for catchment and regional mapping.

4.1.4. Gamma radiometrics

Gamma-Ray Spectrometry (GRS) provides a direct measurement of natural gamma radiation from the top 30–45 cm of the soil (Bierwith, 1996). A gamma-ray spectrometer is designed to detect the gamma rays associated with radioactive elements, and to accurately sort the detected gamma rays by the respective energies (Grasty et al., 1991). Airborne radiometrics survey measures the radiation naturally emitted from the earth surface, using gamma-emitters like ^{40}K and daughter radionuclides of ^{238}U and ^{232}Th. K is a major constituent of most rocks and is the predominant alteration element in most mineral deposits. Uranium and Thorium are present in trace amounts, as mobile and immobile elements, respectively. As the concentration of these radioelements varies between different rock types, we can use the information provided by a gamma-ray spectrometer to map rocks. Airbone methods (air gamma-ray spectrometer, AGRS) provide valuable, systematic coverage of large areas, by providing information about the distribution of K, U and Th that is directly interpretable in terms of surface geology. Nevertheless, AGRS is a surface technique only—interpretation requires an understanding of the nature of the surficial materials and their relationship to bedrock geology.

Although this technique has been employed for geological and mineral resource mapping for over 20 years, it has just become an interesting tool in soil science for detecting spatial variation of soil-forming materials (the *p* factor). It can also be considered as a direct, albeit compound, measure of the mineralogical and textural composition of the soil itself (*s*). Gamma-radiation data are usually available are provided in three channels corresponding to spectral windows for K, U and Th radiation. The apparent K concentration is likely to be most easily interpreted by pedologists. The value of gamma-radiometric data is increasing with the knowledge of their

relation with soil-forming materials and when considered jointly with other information such as terrain models or aerial photography (Cook et al., 1996b) has become an important source of data for digital soil mapping. This technique has also been applied to estimate variation in surface soil moisture content (Carrol, 1981).

4.1.5. Preexisting soil class or property maps or expert knowledge

An existing soil map for parts of an area can be used to build a prediction model, or an experienced surveyor's expertise can be used to make simple rules that can be applied to a DEM, etc. These soil layers should be used as part of the information to build the new model—they should not really be the model itself. This is arguable, there is further discussion in Section 5.1.2.

4.2. c

Climate could be represented by mean annual temperature (T) and mean annual rainfall (P) and perhaps some measure of potential evapotranspiration (E). In the old literature P/E was used to separate pedalfers from pedocals. A $P/E \ll 1$ would imply semiarid conditions and precipitation of carbonates, etc. at depth.

Attempt to use soil moisture and temperature regimes for classification of soil has also been made (Donatelli et al., 2002). They made a classification using dynamic simulation of physical processes which describe the system soil-grass as affected by weather. The daily step simulation of a reference grass offered an insight into the soil moisture and temperature regimes of different soils.

Climate surfaces can be produced from meteorological stations interpolated by Laplacian smoothing splines (Hutchinson, 1998a,b). This has been implemented in a program called ANUCLIM. The climate variables used are monthly mean values for minimum temperature, maximum temperature, precipitation, solar radiation, evaporation and others. The climate surfaces can be used to generate secondary information, e.g., bioclimatic parameters such as mean temperature of warmest period, precipitation of driest quarter, etc., which are useful in determining the climatic envelope for plant and animal species.

Published work suggests that remote-sensing analysis can be used for estimating and mapping air temperature, soil moisture and atmospheric humidity at regional to global scales. Air temperature can be inferred from normalised difference vegetation index (NDVI) data from the NOAA Advanced Very High Resolution Radiometer (AVHRR). Several studies have shown that the surface albedo can be estimated using remote sensing data (i.e., Brest and Goward, 1987), and that net radiation can be calculated with sufficient accuracy (Boegh et al., 2002; Kustas and Norman, 1996).

4.2.1. Temperature

Surface temperature can be derived from remote sensing such as: AVHRR, Geostationary Orbiting Earth Satellite (GOES) (Diak et al., 1998) and TIROS operational vertical sounder (TOVS) (Susskind et al., 1997). The TOVS has two sensors: the high-resolution infrared sounder and the brightness temperatures of the microwave sounding unit. These data can provide estimates daily air temperature, humidity profiles and surface temperature (Susskind et al., 1997).

Goetz et al. (1995) compared surface temperature derived from a multispectral radiometer (MMR) mounted on a helicopter (resolution ~ 5 m pixel), a C-130-mounted thematic mapper simulator (TMS) (~ 20 m pixel) and the Landsat 5 thematic mapper (120 m pixel). Differences between atmospherically corrected radiative temperatures and near-surface measurements ranged from less than 1 °C to more than 8 °C. Corrected temperatures from helicopter MMR and TMS were in general agreement with near-surface infrared radiative, thermometer measurements collected from automated meteorological stations while the Landsat 5 TM systematically overestimated surface temperature.

4.2.2. Precipitation

Spatially distributed precipitation estimates can be derived from rainfall gauge measurements (interpolated using splines or other techniques) or by remote sensing. Records of gauge measurements of monthly precipitation are available throughout the entire twentieth century, while satellite estimates can provide monthly to hourly resolution since 1974. A review has been given by New et al. (2001).

4.2.3. Evapotranspiration

Estimation of evaporation is mainly derived from the energy-balance equation:

$$R_n - G = H + \lambda E,$$

where R_n is net radiation, G is soil heat flux, H is the sensible heat flux and λE is the latent heat flux. Using remote-sensing techniques, we could infer the magnitude of the fluxes of the heat component from surface and air temperatures as derived from radiometric temperature.

Li and Lyons (2002) estimated regional evapotranspiration in central Australia, using limited routine meteorological data and the AVHRR data. Their model attempts to minimise the difference between model-predicted surface temperature and satellite-derived temperature to adjust the estimated soil moisture. They suggested that radiometric surface temperature can be used to adjust simple water-balance estimates of soil moisture providing a simple and effective means of estimating large-scale evapotranspiration in remote arid regions.

Boegh et al. (2002) used Landsat TM data to estimate a composite evaluation of atmospheric resistance, surface resistance and evapotranspiration. The input parameters were: surface temperature, net radiation, soil heat flux, air temperature, and air humidity. The application of the technique in a remote-sensing monitoring context was demonstrated for a Danish agricultural landscape containing crops at different stages of development.

4.2.4. Water-balance components

The familiar water-balance formulation is:

$$P + I = \Delta S + E + T + R + D,$$

where P = precipitation, I = irrigation, ΔS = change in soil moisture, E = evaporation, T = transpiration, R = surface runoff, D = deep drainage.

Precipitation and evapotranspiration can be estimated from remote-sensing data. The most important component relating to soil itself is soil moisture, and much research has been focused on estimating spatially distributed soil moisture (see also Section 4.1.2). Jackson et al. (1996) gave an overview of remote-sensing techniques for estimating soil moisture. Many studies have successfully demonstrated the use of infrared, passive and active microwave sensors to estimate soil moisture (Hoeben and Troch, 2000). Microwave remote sensing of soil moisture is based on the soil's dielectric properties. The large difference between the dielectric properties of dry soil and moisture enables good calibration. The analysis is based on a model that simulates radar backscattering given known surface characteristics such as moisture and roughness. Passive microwave sensors have the advantage of less dependence on soil surface roughness. The main disadvantage with spaceborne sensors is that they produce low-resolution images. This problem is overcome in active microwave sensing through the use of synthetic aperture radar (SAR) sensors (10–100 m). This has been used to monitor spatial and temporal soil moisture at catchment scale (10–1000 km^2) both in vegetated and nonvegetated areas (Lin et al., 1994; Su et al., 1997). Mancini et al. (1999) evaluated the use of multifrequency radar observations in the laboratory for estimating soil moisture.

4.3. o

The main soil forming or altering organisms are vegetation or humans, although other organisms can have an appreciable soil-modifying effect locally (Hole, 1981). In pristine or newly developed environments, the 'natural' vegetation class should represent some kind of equilibrium relation with soil type. Human effects can be seen through land use changes whereupon humans choose different soil types for various purposes. Therefore, land use and land cover are useful indicators of soil properties and class. More recently, estimates of biomass for both crops and more natural vegetation have been obtained; once again these can reflect soil differences. Estimates of vegetation type, land use and land cover and biomass have all been obtained from visible and infrared reflectance by remote sensing and have been enhanced more recently by microwave imagery (Clevers and van Leeuwen, 1996).

4.3.1. Vegetation maps

Due to the global interest concerning the management and conservation of native forest, the development of rapid, cost-effective methods for forest mapping is becoming a challenge. The development

of digital remote sensing is a promising technology to help in reducing time and costs in mapping vegetation. For example, Townshend et al. (1991) used remote sensing and GIS technology to characterise land cover and the production of thematic maps for very large areas. Zhu and Evans (1994) have used AVHRR combined with regression analysis of multitemporal and multisource data, in order to predict forest types and percent of forest cover in the USA on a regional scale. They concluded that multitemporal AVHRR data can be used to produce fairly detailed forest cover maps, since sufficient ancillary data are available for identification of spectral classes. Other studies have been made combining remote sensing GIS, statistical analysis, DEM and ancillary information to map vegetation (Dymond et al., 1992; Hoersch et al., 2002; Lees and Ritman, 1991; Michaelsen et al., 1994; Moore et al., 1991). DeFries et al. (2000) proposed applying a linear mixture model to l-km AVHRR data to estimate proportional cover for three important vegetation characteristics: life form (percent woody vegetation, percent herbaceous vegetation and percent bare ground), leaf type (percent needle leaf and percent broadleaf), and leaf duration (percent evergreen and percent deciduous).

Owens et al. (1999) estimate the current and pre-European mineral soil carbon (C) content of a forested landscape by utilising current forest stand information and pre-European settlement forest data. The forest stands and vegetation patches of the current and pre-European settlement land covers were assigned to one of the three soil C classes based on the type of vegetation present. Using organic matter data from soil surveys of the area, a range of mineral soil C values was determined for each soil mapping unit and vegetation combination.

Verboom and Pate (2003) showed that radiometric data can be indicative of plant distribution. They showed that highly weathered low K soils co-concentrated U and Th and were vegetated mainly by cluster root-bearing *Proteaceae* and *Casuarinaceae*. In granitic soils, ratios of U to Th were higher and cluster root bearing taxa much less prominent, except where ferricrete gravels were concentrated. Draping of radiometric imagery over a digital elevation model showed spiral waveforms of high and low U and Th signal which were largely independent of topography but demarcated different oligotrophic communities.

4.3.2. Land cover and land use classification

Land cover classification is one of the principal motivations and successes of satellite remote sensing. This classification is obtained by supervised classification from some ground-control points. The interest for digital soil mappers is to detect areas of bare soil, or of particular crops representing where humans have picked out soil with particular qualities. Chen et al. (1999) have developed a 1-km landcover dataset of China using AVHRR data (suitable for D4 mapping) (Table 1). Presumably similar land use coverage could be obtained regionally and nationally from Landsat and SPOT imagery.

4.3.3. Biomass and yield maps

Besides landcover itself, improved soil discrimination might be afforded by estimates of biomass variation within particular land uses. This estimation has been developed using visible and near-infrared vegetation indices, such as the Normalised Difference Vegetation Index (NDVI) for natural vegetation and for crops (Lobell et al., 2003). Improved estimation can be obtained by hyperspectral imagery in the visible-NIR range (Gupta et al., 2001) or through the use of microwave imagery (Clevers and van Leeuwen, 1996). The use of yield monitors on harvesting machines also provides a source of spatial biomass information (Stafford et al., 1996). Bishop and McBratney (2001) used yield-monitored wheat yield to aid in the prediction of soil clay content. Yield-monitored data are currently useful for D1 mapping (Table 3).

4.4. r

This is now mainly derived from digital elevation models. Sources of elevation data can be from digitising contour and streamline data, point measurements of elevation from traditional land surveys or from vehicle-mounted high-resolution GPS receivers, or remotely sensed elevation data. The first step to use this information is parametisation of the surface model or the numerical description of continuous surface form (Pike, 1988; Wood, 1996). Parametisation is to quantitatively measure properties

of a landscape that can be used to describe form (Wood, 1996). These parameters are aimed to characterise the geomorphometry of the surface or for landform classification.

Different attributes can be parameterised from a DEM, such as altitude, slope, aspect, different curvatures, upslope area, compound topographic index, etc. Topography has been recognised as one of the soil-forming factors (Jenny, 1941). Aandahl (1948) is perhaps the first scientist to quantitatively relate landscape attributes to soil properties. He derived the distribution of N based on slope length. Troeh (1964) fit a cylindrical parabola to contour lines to derive slope and curvatures. The landform parameters were derived to correlate to soil drainage classes. Many would argue that digital terrain modelling is the most useful and quantitatively developed factor for predicting soil attributes and soil classes (McKenzie et al., 2000).

4.4.1. Primary terrain attributes

The digital elevation model is the basis for calculation of surface attributes, which include slope, aspect, curvature and upslope contributing area. Primary attributes have been used successfully in numerous studies (see Table 4) to predict different soil attributes and classes. We can separate the primary attributes into parameters that were derived locally (using the local neighbourhood points) and derived from the DEM of the whole area (regional) using some specific rules. Shary et al. (2002) divided further the local and regional into scale-specific and scale-invariant. Evans (1972, 1998) provide an overview of primary terrain attributes in relation to their geomorphologic meaning.

A quadratic trend local surface is usually fitted to the local neighbours, such as Evans (1980):

$$z = ax^2 + by^2 + cxy + dx + ey + f.$$

Other methods have also been proposed (Zeverbergen and Thorne, 1987; Shary, 1995). The standard method involves calculating the parameters of a central cell and its eight neighbourhood in a moving 3×3 cell window. The purpose of this fitting is that it enables the easy calculation of the first and second derivative of the surface, and these values can be used to calculate slope, aspect and various curvatures. Shary et al. (2002) defined 12 types of curvature that potentially can be used for landform classification and spatial prediction.

Calculation of the first and second derivatives using a local window is scale dependent. The derived parameters are only relevant to the resolution of the DEM and the neighbourhood cells used for calculation. Wood (1996) proposed a multiple-scale parametisation by generalising the calculation for different window sizes.

For regional parameters, upslope contributing area is one of the most important ones. This parameter (also called drainage or catchment area) is the area above a certain length of contour that contributes flow across the contour. There are different algorithms for estimating this quantity, such as the single flow-direction, randomised single-flow direction (Rho8) and DEMON Stream tube (Gallant and Wilson, 2000). Dobos et al. (2001) proposed a potential drainage density (PDD) designed to highlight relative terrain differences even on a relatively level land surface.

4.4.2. Secondary terrain attributes

Secondary terrain attributes are computed from the primary attributes. These have been described in detail by Wilson and Gallant (2000), Moran and Bui (2002). These attributes usually combine two or more primary attributes to characterise the spatial variability of specific processes in the landscape. The most widely used is Compound Topographic Index (CTI) or also called wetness index:

$$\mathrm{CTI} = \ln\left(\frac{A_s}{\tan\beta}\right)$$

where A_s is the upslope area and β is the slope. Wilson and Gallant (2000) also provide routines for the calculation of erosion, solar radiation and dynamic wetness indices.

4.4.3. Terrain or landscape classification

Traditional landform classification is based on qualitative description from surface shape. Automated classification of the landform from quantitative digital terrain models is the ultimate desire. Pennock et al. (1987) define seven-unit landforms based on an analysis of local surface shape. Wood (1996) was able to differentiate six landform features (peak, ridge,

Table 4
Sources of the scorpan factors for predicting soil classes and/or attributes in previous quantitative studies

Authors	Predicting factors					Predicted factor
	s	*c*	*o*	*r*	*p*	
Troeh (1964)				Slope, curvatures		Soil drainage classes
Walker et al. (1968)				Elevation from contour map		Soil properties
Pavlik and Hole (1977)	Soil survey data			DEM		Soil classes
Legros and Bonneric (1979)	Soil survey data			DEM	Geology map, lithology	Soil classes
Pennock et al. (1987)				DEM		Soil properties
Lee et al. (1988)	Landsat TM			DEM: altitude, slope, aspect		Soil classes
Frazier and Cheng (1989)			Landsat TM			Soil properties
Bhatti et al. (1991)	Lab. analysis Landsat TM					Soil properties
Odeh et al. (1991, 1992, 1994, 1995)	Soil morphology, physical and chemical properties			DEM		Soil classes
Bell et al. (1992)				DEM	Bedrock and superficial geology	Soil drainage classes
McKenzie and Austin (1993)				DEM	Soil survey, air photo	Soil properties
Moore et al. (1993)				Slope, Wetness Index		Soil properties
Dymond and Luckman (1994)				DEM	Regolith map	Soil series
Arrouays et al. (1995)	Soil analysis	Climate data		Relief data		Soil properties
Gessler et al. (1995)				Plan curvature, CTI		Soil properties
Knotters et al. (1995)	ECa					Soil horizon thickness
Cook et al. (1996a)	Soil survey data					Soil properties
Cook et al. (1996b)					Airborne gamma radiometric data	Soil parent material classes
Skidmore at al. (1996)	Aerial photograph			DEM		Soil classes
Zheng et al. (1996)				DEM		Soil properties
Cialella et al. (1997)			AVIRIS	DEM		
Dobermann and Oberthur (1997)	Soil physical and chemical properties					Soil fertility classes
Lagacherie and Holmes (1997)				DEM	Geology map	Soil units

Table 4 (*continued*)

Authors	Predicting factors					Predicted factor
	s	*c*	*o*	*r*	*p*	
Skidmore et al. (1997)						
Thompson et al. (1997, 2001)				Slope, profile curvature, elevation above local depression		Soil Hydromorphic Index
Voltz et al. (1997)	Soil map					θ wilting point
Zhu and Band (1994), Zhu et al. (1996, 1997, 2001)			Landsat TM (canopy coverage)	DEM	Geological map	Soil series and soil properties
Zhu et al. (1997)				DEM	Geological map	Soil properties
Bui et al. (1999)				DEM	Geological map	Soil classes
Lilburne et al. (1998)				DEM		Soil properties
McKenzie and Ryan (1999)		PI estimated from DEM		Slope, specific catchment area, CTI, flow direction	Geological map, gamma radiation	
Oberthur et al. (1999)	Soil map, aerial photo, farmer's knowledge					Soil texture
Sinowski and Auerswald (1999)	Soil survey			Elevation, slope, upslope catchment area		Soil properties
Thomas et al. (1999)				DEM	Geological map	Soil classes
Bell et al. (2000)	Soil map		Aerial photograph	DEM		Soil organic C content
Bourennane et al. (2000)				DEM		Soil horizon thickness
Chaplot et al. (2000)				Elevation above streambank, slope, specific catchment area, CTI		Hydromorphic Index
Dobos et al. (2000)				Elevation, slope, PDD		Soil classes
Lagacherie and Voltz (2000)				DEM		θ wilting point
McIntosh et al. (2000)				Relief map		
Ryan et al. (2000)		Rainfall, temperature, net radiation, PI	Landsat TM	DEM	Airborne gamma radiometric	Soil properties
Odeh and McBratney (2000)			AVHRR			Clay content
McBratney et al. (2000)	ECa		Crop yield	DEM		Clay content, CEC

(continued on next page)

Table 4 (*continued*)

Authors	Predicting factors					Predicted factor
	s	*c*	*o*	*r*	*p*	
Zhu (2000)			Landsat TM (Canopy Coverage Index)	DEM	Geological map	Soil series
Bishop and McBratney (2001)	Soil ECa, aerial photograph, Landsat TM		Crop yield	DEM		CEC
Bui and Moran (2001)	Soil map, Landsat TM			DEM	Lithology map	Soil classes
Boruvka et al. (2002)	Soil properties map: pH, CEC, OC, texture					
Campling et al. (2002)			Landsat TM	DEM: slope, aspect, profile, tangential and plan curvatures, CTI, stream power index, slope-aspect index		Soil drainage classes
Carré and Girard (2002)	Soil map		SPOT	DEM	Geological map	Soil classes
Chaplot and Walter (2002)				DEM: upslope catchment area, CTI		Hydromorphic Index
Florinsky et al. (2002)				DEM		
Florinsky and Eilers (2002)				DEM: horizontal and vertical curvatures		Soil OC content
Hengl et al. (2002)		Climatic map	AVHRR	DEM		Soil pH and OM content
Mayr et al. (2003)				DEM	Statigraphic geology map	Soil classes
Park and Vlek (2002)	Soil map		Vegetation map	DEM		Soil properties
Peng et al. (2003)	Soil map, Landsat TM, IKONOS, DOQ			DEM		Soil drainage class
Scull et al. (2003a)	Landsat TM	DEM		DEM		Surface gravel content
Sommer et al. (2003)	EM induction survey		Airborne multispectral scanner	DEM, curvature and upslope catchment area		Soil classes
Hengl et al. (2003a)				DEM		Thickness of topsoil, soil pH and OM content

pass, plane, channel and pit) from locally derived parameters (slope; cross-sectional, longitudinal, minimum and maximum curvatures).

MacMillan et al. (2000) produced a landform classification based on quantitative digital variables. They proposed a conceptual design for a multilevel, hierarchical system of automated landform classification. The model incorporated hydrological and geomorphic criteria to define and delineate spatial entities at multi scales. MacMillan et al. recognised that spatial entities delineated solely on the basis of geomorphic shape are insufficient to meet many inventory and modelling needs, as they lack the information required to establish linkages, interactions and flows between spatial entities. Similarly, landform units defined solely on the basis of hydrological criteria are incomplete in that they do not differentiate areas of different surface morphology or relative landscape context. Using 37 different terrain attributes, they are able to classify 15 landform elements by employing a fuzzy logic.

Other approaches on terrain classification include fuzzy *k*-means analysis (Burrough et al., 2000; Ventura and Irvin, 2000) or based on the relative elevation within a search radius (Blaszcynski, 1997; Fels and Matson, 1996).

4.5. p

Parent material information can be obtained from digitised geological maps—maps that focus on lithology and not so much on stratigraphy will probably be more useful for soil prediction. Some kind of quantitative information about surface mineralogy and texture (related to parent material) can be obtained by gamma radiometrics (see the previous discussion in Section 4.1.4). Geomorphological and weathering models (Dickson et al., 1996) have been used to identify the distribution of soil-forming materials.

Additionally, the natural fields of the earth, gravitational, electrical (Andriani et al., 2001), magnetic (Galdeano et al., 2001) and electromagnetic (Beard, 2000) can be used to provide information on underlying geological structure. Taylor and Eggleton (2001) discuss regolith mapping. Regolith should be predictive of soil, as soil is the upper part of it. The regolith can be thought of as representing either *s* or *p*. Indeed, regolith itself may more useful than the soil profile for hydrological environmental modelling. Regolith maps are produced in Australia from a combination of multi- and hyperspectral and airborne geophysical data and expert knowledge: reflecting the scorpan approach.

4.6. a

a represents age or elapsed time. This will give limits on how long pedogenesis has been occurring and should differentiate soil classes and properties. One useful estimate of *a* is the age of the ground surface, which may be very old indeed (Twidale, 1985). Alternatively, *a* can be represented by the age of the material in which soil has developed, suggesting that the scorpan approach will not deal well with polycyclic soils. It is theoretically likely that soil development will follow some logarithmic or square–root time relationship, suggesting more need to differentiate between younger materials than older ones. Schaetzl et al. (1994) discuss the form of soil chronofunctions.

Geomorphologists and stratigraphers can presumably draw maps of *a* independent of soil maps. In fact such "gues(s)timated" maps along with an estimate of uncertainty could be used to represent this factor. There are methods for soil and material dating of course, e.g., ^{14}C, $\delta^{18}O$, thermoluminescence (inter alia, Matt and Johnson (1996)) and $^{40}Ar/^{39}Ar$ (inter alia, Van Niekerk et al. (1999)). None of these are, as far as we are aware, capable of scanning and producing full coverages in a true remote-sensing fashion. Ground electromagnetic methods have been used for stratigraphic mapping, e.g., Sinha (1990). *a* remains difficult to characterise well. Indeed, it seems that more than any factor expert knowledge is still needed to derive *a*. Considerable advances in technology and knowledge are needed.

4.7. n

As was discussed in Section 2, soil can be predicted from spatial coordinates alone. Obtaining these is now much easier due to the advent of GPS with 5-m accuracy receivers costing less than US$1000. This may indeed reflect some other environmental variable such as climate, and because of this it can be argued that *n* is not really a factor, but simply putting the

coordinates is a simple way to ensure that spatial trends not included in the other environmental variables are not missed. Therefore, n could also be described by some linear or nonlinear (nonaffine) transformation of the original spatial coordinates, e.g., a new coordinate could be the closest distance of each location to the coast (Webster, 1977, p. 201), or distance uphill from the nearest discharge area (Bui and Moran, 2000). This factor is potentially a valuable yet cheap source of environmental information, and should never be disregarded.

5. Discussion

Having presented a review of what has been done by others, and having suggested and reviewed a generic predictive model and potentially useful environmental data layers, we now put this into a framework for soil mapping based on scorpan and soil spatial prediction functions (SSPF) and spatially auotcorrelated errors (e). The scorpan-SSPFe approach is now outlined with some brief discussion of each of the steps. It is a proposal. Uses, problems, and other implications of the scorpan-SSPFe approach are discussed subsequently.

5.1. Summary of the scorpan-SSPFe method

The scorpan-SSPFe method essentially involves the following steps.

5.1.1. Define soil attribute(s) of interest and decide resolution ρ and block size β

These are the design specifications for the survey. Define soil attribute(s) of interest, i.e., a soil property or set of soil properties and/or a set of soil classes, usually from some predefined classification system. The resolution may be defined by the resolution of the environmental variables, e.g., 30 m Landsat pixels, but should be a design specification from the intended use of the information. Referring back to Table 1, we believe the methodology discussed here is most appropriate for D3 surveys, therefore, pixel or block size ρ, is equal to pixel spacing β, and will be in the range 20–200 m. The linking of ρ and β is a simplification, and is a point that requires further study (see Bishop et al., 2001 for further discussion). At this stage, the uncertainty limits that can be tolerated may be also be defined.

5.1.2. Assemble data layers to represent Q

Assemble the data layers with consideration of the number of layers describing each factor and any prior evidence as to the importance of each factor. This was discussed in detail in Section 4. At this stage, we do not know the relative importance of the data layers. Balance is probably important. At this phase of development, because of the relative availability of DEMs, it would be easy to obtain 15 or 50 terrain attributes (r), e.g., as described in Shary et al. (2002), and rather difficult to represent a or c. An attempt should be made to represent all the factors however.

5.1.3. Spatial decomposition or lagging of data layers

This is suggested as a step because it is felt that predictions might be scale dependent and it is important to find the appropriate spatial associations. This can be achieved either by a wavelet decomposition (e.g., Epinat et al., 2001, applied to airborne NDVI imagery), or geostatistically. The geostatistical approach involves modelling the correlation structure in the imagery by decomposing the variogram into independent spatial components, and then taking each component in turn and kriging it, thereby separating it from the others. Oliver et al. (2000) used this approach on SPOT imagery. Both of these methods will allow the removal of short-range uncorrelated noise components from subsequent sampling, modelling and prediction stages. The spatial decomposition of the environmental variables begs the question of whether the target attribute should also be spatially decomposed. In most cases, there would probably be insufficient observations to do this effectively, but it could be done where this is not the case.

An alternative approach to spatial decomposition is that of spatial lagging, i.e., to fit a model such as,

$$S(x,y) = f(s(x+u,y+v), c(x+u,y+v), \ldots) + g(x+u,y+v)$$

where the soil attribute of interest is "regressed" on the layers representing the scorpan attributes and on spatially lagged ($+u$, $+v$) copies of them, with u and v

variable. This approach seems somewhat more cumbersome than the spatial decomposition approach.

5.1.4. Sampling of assembled data (Q) to obtain sampling sites

In most cases, soil sampling will be required to set up the model (except perhaps when the aim is map updating). We have a lot of prior information on the environmental variables which we can use to guide the sampling. The aim is to construct predictive equations for the soil attributes of interest in terms of the environmental variables. This is a kind of calibration exercise and, therefore, it would seem wise to span the range of values of each variable so that the prediction model will not be required to extrapolate beyond its bounds. One possibility is to use Latin hypercube sampling (McKay et al., 1979), a constrained Monte Carlo sampling scheme. It selects μ different values from each of v variables $q_1, \ldots q_v$ in the following manner. The range of each environmental variable is divided into v nonoverlapping intervals on the basis of equal probability. One value from each interval is selected at random with respect to the probability density in the interval. The v values thus obtained for q_1 are paired in a random manner (equally likely combinations) with the μ values of q_2. These μ pairs are combined in a random manner with the μ values of q_3 to form μ triplets, and so on, until μ v-tuplets are formed. We can search through the data and find the locations that are taxonomically most similar to the combination of values chosen, or find locations that match the intervals in the various variables. In either case, we will then have a set of μ spatial coordinates (locations) at which we can observe the soil attribute(s) of interest. Clearly, μ should be related to the number (v) of environmental attributes q (~ 7), and the number of parameters, φ, and degrees of freedom, $\psi = \mu - \varphi$, in the model to be fitted. Intuitively, we feel that nothing much will be achieved if μ is less than 100. Perhaps, μ should be related to M the number of pixels in the target map—a guess would be $0.0001M < \mu < 0.001M$. If there are a large number of environmental variables, then the sampling could be based on a smaller number of principal components, or canonical variates if there is any prior information on soil classes.

This kind of sampling should produce a reasonably efficient way of sampling the soil and its environment so that the range of conditions are encountered, ensuring a good chance of fitting relationships if they exist. Hengl et al. (2003b) suggest a somewhat related sampling scheme for this purpose.

An alternative procedure is suggested by the work of Lagacherie et al. (2001) that define a reference area (Favrot, 1989) which, through the spatial data layers of the environmental variables, extrapolates well to a larger region. Sample the reference area purposively or systematically (fit the model and extrapolate to the rest of the area). This might give a better chance of fitting local relationships with a given sampling effort, and should be more efficient in field time. The advantage hinges, however, on how well the extrapolation can be done.

5.1.5. GPS field sampling and laboratory analysis to obtain soil class or property data

Step (iii) yields a set of μ spatial coordinates at which the observations of the soil attribute(s) are to be made in the field. These can be located with a GPS receiver, and samples taken for subsequent laboratory analyses, in the usual manner. At a subset of these locations, say 5%, a duplicate observation should be made at a distance (say) half the resolution (ρ) to get an estimate of the short-range field variability in the soil attribute(s). This will help in subsequent spatial modelling. If a specific purpose or numerical classification or allocation to a conventional classification system is required, then the observed soil data can be processed to obtain discrete or continuous class labels for each observation site.

5.1.6. Fit quantitative relationships (observing Ockham's razor)

We can now assemble the soil data for the left side of the scorpan equation, and the environmental data for the right side. We can now fit the model $f()$ representing the μ locations using any of the techniques described in Section 3.3. Ockham's razor (the principle that states that ''Entities should not be multiplied unnecessarily'') should be applied to find the model (with the least number of parameters) that fits best. The residuals (e) of the soil property or class probability or membership at each of μ sites should be estimated also (and kriged using either scorpan-simple or scorpan-compositional kriging as mentioned in Section 3.4.2.2). A further improvement would be

an iterative scheme whereby the $f()$is estimated, the e estimated and then the $f()$ reestimated, etc. as in the following algorithm, written for a soil property, where k_b represents simple block kriging of a block of size $\rho = \beta$.

Iterative scorpan-SSPFe algorithm (ISSA)

Fit $S = f(Q) + e$
Iteration 1
$S = f_1(Q)$
$c_1 = k_{b1}(S - f_1(Q)) = k_{b1}(e_1)$
$n_1 = e_1 - c_1$
Iteration 2
$s - c_1 = f_2(Q)$
$c_2 = k_{b2}(S - f_2(Q)) = k_{b2}(e_2)$
$n_2 = e_2 - c_2$
Iteration i
$S - c_{i-1} = f_i(Q)$
$c_i = k_{bi}(S - f_i(Q)) = k_{bi}(e_i)$
$n_i = e_i - c_i$
Stop when $|f_i - f_{i-1}| < \varepsilon$ and $|c_i - c_{i-1}| < \phi$

This algorithm is similar to the Generalised Estimating Equations (GEE) approach which was conceived by Liang and Zeger (1986) to extend conventional generalised linear models to deal with correlated data. The first applications were longitudinal data correlated in time, e.g., Diggle et al. (1994), but some spatial applications have been reported, e.g., Albert and McShane (1995) and Pebesma et al. (2000). ISSA is slightly more general in that $f()$ need not be linear. If classes are to be mapped, a similar though numerically more difficult algorithm could be written for iterative scorpan-compositional kriging (Section 3.4.2.2).

A single model may not be adequate especially if there are strong pedogeomorphic or geological contrasts within the study area—each subregion may show quite different relationships between soil and environmental variables. In that case, it may be necessary to a scorpan-SSPFe model for each of the pedogeomorphic or geological subregion. Care needs to be taken that the prediction surfaces produced are realistically continuous. This approach demands a higher data burden than a single model for a whole geographic region. Once again, the total number of parameters in the final model needs to be considered.

If we have a lot of local information for both sides of the equation (like an existing soil map, or proximally sensed soil data), we can fit a local model,

$$S[x,y] = f_l([x,y]) + e_l(x,y),$$

where the l subscript on $f_l([x,y])$ and $e_l(x,y)$ refers to models fitted to a local (moving) neighbourhood centred on $[x,y]$ rather than to the whole area to be mapped. It will be rare that m will be large enough to fit this model. This kind of model was used geostatistically, i.e., $S[x,y] = e_l(x,y)$ for soil salinity mapping by Walter et al. (2001).

5.1.7. Predict digital map

We now have a model $f()$ fitted to the μ locations, which we can now apply to the $(M - \mu)$ locations where we have no soil observations, but have environmental observations. Additionally, kriging of the residuals (e) at is also done at the $(M - \mu)$ locations and the results added together. Additionally, the uncertainties of the predictions of $f()$ and e at the M locations should be evaluated. Raster maps can then be made of the soil attribute(s) and their associated uncertainties.

5.1.8. Field sampling and laboratory analysis for corroboration and quality testing

It must not be assumed that the digital information is perfect with minimal information on the quality of the information. Indeed, we cannot expect this kind of map to be more accurate than conventional ones. There are two reasons for this: (i) local variation (at whatever resolution) is a limiting factor, there is a lot of soil variation within 10 m or 100 m or 1 km; and (ii) there is uncertainty in environmental layers and this can propagate errors (Florinsky, 1998; Lagacherie and Holmes, 1997; Heuvelink and Burrough, 2002).

Soil sampling is required to provide an independent estimate of the quality of the map produced (both the map and its uncertainty estimate) should be done (de Gruijter and Marsman, 1985). Of course, we could put some of the original observations to one side. These days, it seems fashionable to put one quarter to one third aside for validation. This fraction does not appear to be based on any empirical evidence. The Marsman and de Gruijter (1986) approach is more efficient in that the sampling is designed specifically

for corroboration. Therefore, we would recommend it as part of the process.

There has been little work on corroboration of digital soil maps, especially of classes. This is an area of research need.

5.1.9. If necessary simplify legend or decrease resolution by returning to (i) or improve map by returning to (v)

If we find that the map does not meet design specifications, i.e., class purity is less than x%, or the confidence intervals for a soil property are too wide over parts of the map, then we can simplify the legend or decrease spatial resolution (i) but these should be design specifications, or more sensibly, target further sampling (v) in areas where predictability appears to be poor, and recalculate the maps.

5.2. Uses

There are at least three potential uses of the scorpan-SSPF approach. The first is the production of digital soil maps as a replacement for the paper-based choropleth soil map of the past. The second is the use of the approach to extrapolate exiting soil maps into unmapped areas. The third is the construction of dynamic soil maps. The first and second may be the most important initially, and the third eventually.

5.2.1. Digital soil maps

The main use of the scorpan-SSPF approach is to replace the polygon-based soil maps of the past with digital maps of soil properties and classes and their associated uncertainties for areas previously mapped, or for new areas. These maps will be stored and manipulated in digital form in a GIS creating the possibility of vast arrays of data for analysis and interpretation.

The first digital soil maps were simply representations of the observations without interpolation or relation to the environment (e.g., Webster et al., 1979). Some authors have worked on better ways to present digital spatial soil information, chronologically (de Gruijter and Bie, 1975; de Gruijter et al., 1997; Grunwald et al., 2001). This is an area requiring considerable research. One goal must be fully operational multiresolution digital soil maps.

While we cannot necessarily expect the maps to more accurate than conventional ones, we can expect to have a quantitative estimate of the uncertainty (Section 5.1); sampling effort should be expended to achieve this. Laba et al. (2002) compared conventional (Congalton and Green, 1999) and fuzzy (Gopal and Woodcock, 1995) methods to assess the accuracy and uncertainty of land cover maps produced at high taxonomic resolution. These methods could be applied to digital soil maps.

Survey commissioners, decision makers and users in general would perhaps be more comfortable with a concept of certainty rather uncertainty. This answers the question, "how well do we know the value at some location?", rather than concentrating on "how badly we know it". A potentially adequate standardised ($0 \rightarrow 1$ or $0\% \rightarrow 100\%$) measure of certainty is $f = 1 - \min(2s/V, 1)$, where s the standard deviation of the estimate. e.g., if we have an estimate V of clay content of 50% and an s of 5% then $f = 0.8$ or $f\% = 80\%$. More sophisticated measures may be required, such as a certainty characteristic—the probability that a statement C is true within a distance d or an increasing neighbourhood A. Clearly, more work is needed on standards for digital soil maps.

5.2.2. Interpolation or extrapolation of existing soil maps

If s for a previously mapped region is put on the right of scorpan equation, the legend is retained largely, and new samples are collected, this might be considered by some to be map updating. There is another possibility; this is where the previous s is put on the left rather than the right side of the scorpan equation (an example is given in Bui et al., 1999). The advantage of this is that no new sampling is required for fitting—although corroboration sampling should be done (see Section 5.1 (viii) above). This would allow quantitative elaboration of the existing (but unknown) models; for classes, this is what Girard (1983) referred to as (the usually unknown) 'chorological rules', and their subsequent extrapolation to new areas. Possible problems include repeating old models which may be wrong, or extrapolation outside the range of associated environmental data sets—Lagacherie and Holmes (1997) and Lagacherie and Voltz (2000) discuss this for landscapes, while McBratney et al. (2002) were concerned about this for pedotransfer functions. It is also impor-

tant to establish for which purpose existing data have been collected. Some soil surveyors use auger observations to confirm their mental model, while other use them to find inclusions or boundaries. Therefore, this approach should be used with a deal of caution.

5.2.3. Environmental change—partially dynamic 'scenario' soil maps

One major criticism of conventional soil maps is that they are essentially static statements. Digital soil maps created with the scorpan-SSPF methodology offer new and necessary possibilities. It is becoming increasingly important for environmental reasons to know not just $S[x,y]$ but $S[x,y,t]$. If we know any of the partial differentials, $\delta s/\delta t$, $\delta c/\delta t$, $\delta o/\delta t$, etc., the first two perhaps being the more important—we can project the existing soil map forward by some time u by calculating most simply say $c+u(\delta c/\delta t)$ for all points and running the new c layer(s) through the prediction function. "Change-detection analysis" (Mücher et al., 2000) is well developed for land use and vegetation change (components of $\delta o/\delta t$) using remotely sensed imagery and/or aerial photographs average and localised values of o can be estimated from rasterised images taken at two or many times (Munyati, 2000). Other derivatives may be obtained from models (e.g., temperature and rainfall changes) or from a few monitoring stations (e.g., soil changes) within the area of interest (Mendonça Santos et al., 1997).

This potential approach has limitations compared with a fully fledged dynamic simulation model, such as lack of feedback and possible extrapolation problems, where for example $c+u(\delta c/\delta t)$ takes us (well) outside the range of the original training data. Nevertheless we still have a relatively quick and easy way to produce first-cut 'scenario' soil maps of both properties and classes.

5.3. General discussion

We now discuss some general points relating to the scorpan-SSPF approach to making digital soil maps.

5.3.1. Pedotransfer functions

Some people might wish to call the soil spatial prediction functions pedotransfer functions (PTFs) but we would caution against that. We believe they should be called pedotransfer functions only when soil attributes (i.e., classes or properties) appear on both sides of the equation $s=f(s)$ (and not when $s=f(o,r,p,a,n)$), and when some principles outlined by McBratney et al. (2002), principally the effort principle, *do not predict something that is easier to measure than the predictor*, are observed. For example, McBratney et al. (2002) do not consider functions that predict soil classes from soil properties to be useful or legitimate PTFs (Table 2), whereas they might be perfectly useful SSPFs. Admittedly, it is debatable whether for example $s=f(r)$ should be considered a PTF, we simply think that this extends the definition too far. There is a possible intersection, or area of overlap, between PTFs and SSPFs, i.e., when $s_1=f(s_2,\ldots)$, they obey the PTF principles, and they are located spatially i.e., they are a function of spatial coordinates. These are *spatial* pedotransfer functions. Fig. 2 is an attempt to illustrate the differences and possible overlap between PTFs and SSPFs. Pachepsky et al. (2001), Rawls and Pachepsky (2002) and Romano and Palladino (2002) almost illustrate examples of spatial or contextual pedotransfer functions, but they are really examples of $s=f(s,r)$, so they are more SSPFs than PTFs.

5.3.2. Spatial aspects of the scorpan-SSPFe and other models

The corpt or environmental correlation approach has no formal spatial component except perhaps for some contextual variables. On the other hand, the

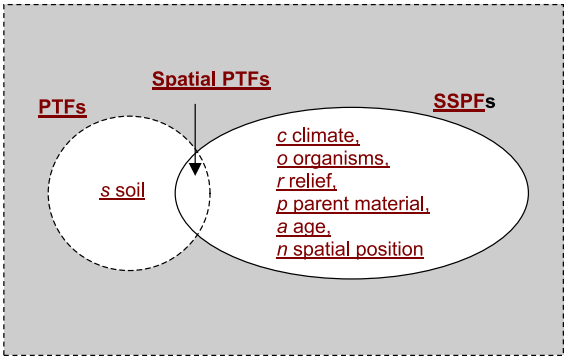

Fig. 2. A Venn diagram showing the relationship between pedotransfer functions (PTFs), soil spatial prediction functions (SSPFs) and their intersection, spatial PTFs.

geostatistical approach is almost purely spatial and has a rather simple model of soil–environment relationships—the linear model of co-regionalisation. For realistic modelling geostatistics needs at least (i) a local linear model of co-regionalisation—but this is difficult to fit automatically, and/or (ii) a nonlinear (or nonaffine) model of coregionalisation.

The scorpan-SSPFe approach, outlined in this paper, has sophisticated soil–environment relationships and incorporates spatial aspects in three ways:

1. *n* is explicitly used as a factor,
2. the environmental variables are spatially decomposed or spatially lagged,
3. the residuals are treated spatially.

This model could also be applied in a local moving neighborhood.

5.3.3. Is this the right approach?

As was said at the beginning of Section 5.1, this is a proposal, not a fait accompli, albeit based on a deal of work and experience worldwide. Whether or not this turns out to be the right approach hinges on a number of factors, not all of them scientific, but we shall deal with those first. Scientifically one could ask a number of questions. Are there other soil-forming factors? Are we missing key variables? For example, have we successfully incorporated hydrological effects? Is the underlying idea of a soil somehow in equilibrium with its environment reasonable enough to be predictive? Or, is the soil too chaotic for prediction by other factors? Will the proposed methodology give similar answers as traditional approaches, and do we want it to? Further experiment and experience will indubitably answer these questions.

There are other socio-economico-political factors that will have a bearing also. The socio-political factors demand recognition of, and solutions to, environmental problems. We believe the approach has the right kind of economics—it is potentially cheaper than traditional approaches and gives the desired kinds of information. Most of the hardware and software tools are in place to put this approach into practice. Clearly, integrated systems have to be devised. Research into aspects is always needed, principally efficient sampling designs and useful certainty estimation. The biggest stumbling block is the teams of personnel with the skills required to complete the task. Education of skilled and knowledgeable personnel for those teams is a key priority.

5.3.4. Dangers—let the user beware!

There are real dangers with this, or any new approach, if it is misused or abused. Here we outline briefly some obvious pitfalls.

1. Data quantity and quality. The first danger is not using enough real soil observations to fit the models, or with using poor quality (missing or noisy) predictor variables. This can to a degree be handled by uncertainty analysis—a large topic (Heuvelink, 1998), which has not been discussed formally in this paper. There is a lower limit below which any fitted models will be meaningless.
2. Overfitting the data. It is easy to overfit models; this could be because of lack of observations but more because of lack of parsimony, especially a problem for tree-based methods. Overfitted models predict poorly. It is imperative to apply Ockham's razor—this will help with evaluating poorly fitting or overfitted models. The use of cross-validation, pruning and boosting methods (Hastie et al., 2001) might also help.
3. Circularity. A third hazard comes from the possible circularity of the model, e.g., a DEM producing climate surfaces producing soil variables as an input to soil class prediction. Once again uncertainty analysis will help.
4. Databases and data mining. During the past decade, soil scientists have created regional, national, continental and worldwide databases. Data mining is a phrase for a class of database applications that look for hidden patterns in such groups of data (Hastie et al., 2001). Unfortunately, the term is sometimes misused to describe software that presents data in new ways. Proper data mining software attempts to discover previously unknown relationships among the data. Data mining is a broad concept from supervised learning (prediction) to unsupervised learning and includes all the methods described in Section 3.3 above-neural networks, classification trees with boosting. There are a large number of commercial software

products available to do this. They incorporate one or often several of the methods described in Section 3.3. This will make evaluation difficult as different soil science groups use different software products for fitting $f()$, therefore, comparative studies will be important to evaluate the best approaches. In addition large national or international databases of legacy soil data will be available (e.g., Bui et al., 2002); they also have potential problems because of their unknown site selection probabilities—which are not usually equal—some of the data from the island of Britain where a 5 km grid survey has been completed, are an exception!

5.3.5. A new paradigm?

Hudson (1992) described soil survey as a paradigm-based science. Paradigm is a much overused and hackneyed word these days but it has a precise philosophical meaning. Much of the following two paragraphs is paraphrased from Rosenberg (2000). It is a term employed by Kuhn (1996) to characterise a scientific tradition, including its theory, apparatus, methodology and scientific philosophy. The soil scientist's task is to apply the paradigm to the solution of problems. Failure to solve problems is the fault of the scientist not the paradigm. Persistent failure makes a problem an anomaly and threatens a revolution which may end the paradigm's hegemony.

What's the difference between the scorpan-SSPFe approach and the conventional Jenny landscape model? Both are models—they are simplified descriptions of regularities governing a natural process, usually mathematical and sometimes derived from a more general or simplified theory. Ontologically, they are similar—they both require soil objects and attributes which are a function of their environment. The conventional paradigm is a qualitative theory. The approach outlined here is a quantitative, partially deterministic, partially probabilistic, empirical theory. So methodologically, they are quite different. The apparatus is also different, here we require digital information, computers, GIS, etc. The Jenny landscape model may be fall under the deductive-nomological model of scientific explanation but because of its somewhat probabilistic nature the scorpan-SSPFe approach may fall under the inductive-statistical model of explanation (Rosenberg, 2000). Therefore, the scorpan-SSPFe approach to soil mapping probably represents an emerging paradigm eventually leading to a complete paradigm shift.

This begs the question, does $f()$ have to be empirical? The Vienna school of logical empiricists would be generally happy with scorpan-SSPFe approach, although perhaps they would have difficulties with its partially probabilistic nature. The lack of a mechanistic theory for predicting soil tugs at the soil scientist's cloak of explanation. Perhaps this is an unnecessary concern, philosophical empiricists believe there is nothing to causation beyond a regular sequence. Any testing of the mechanistic theory will require empirical observation of the real world. The first attempts at a mechanistic approach have begun (Minasny and McBratney, 2001) but it will be a long time before the mechanistic theoretical approach will be competitive in the predictive sense.

6. Conclusions

We have reviewed various approaches to predictive modelling and data acquisition and proposed a methodology for producing digital soil maps.

6.1. Summary of the method

The scorpan-SSPFe method essentially involves the following steps.

1. Define soil attribute(s) of interest and decide resolution ρ and block size β.
2. Assemble data layers to represent Q.
3. Spatial decomposition or lagging of data layers.
4. Sampling of assembled data (Q) to obtain sampling sites.
5. GPS field sampling and laboratory analysis to obtain soil class or property data.
6. Fit quantitative relationships (observing Ockham's razor) including spatially autocorrelated residual errors.
7. Predict digital map.
8. Field sampling and laboratory analysis for corroboration and quality testing.
9. If necessary simplify legend, or decrease resolution by returning to (i) or, improve map by returning to (v).

All of the hardware and software tools, technologies and knowledge, are in place to make this approach operationable. This is clearly an exciting time for soil resource assessment.

6.2. Future work—open questions

Clearly, we need to try out the methodology outlined above and by experience we shall discover the useful forms of $f()$ and the serviceable Q layers are. These are the key open questions. In summary, topics to be further researched include:

1. Environmental covariates for digital soil mapping.
2. Spatial decomposition and/or lagging of soil and environmental data layers.
3. Sampling methods for creating digital soil maps.
4. Quantitative modelling for predicting soil classes and attributes (including generalised linear and additive models, classification and regression trees, neural networks, fuzzy systems, expert knowledge and geostatistics).
5. Quality assessment of digital soil maps.
6. (Re)presentation of digital soil maps.
7. Economics of digital soil mapping.

Nevertheless, we believe the methodology can be used now for real-world applications.

> Even if there is only one possible unified theory, it is just a set of rules and equations. What is it that breathes fire into the equations and makes a universe for them to describe? The usual approach of science of constructing a mathematical model cannot answer the questions of why there should be a universe for the model to describe. Why does the universe go to all the bother of existing? (Stephen W. Hawking, 1998. A Brief History of Time: From the Big Bang to Black Holes, Bantam, NY, p. 174.)

Acknowledgements

We thank Delfim Moreira and Almirante Guilhem for their inspiration, and EMBRAPA Solos and the University of Sydney for making this collaboration possible. We thankfully acknowledge funding from the following projects: FAPERJ E-170.023/2001 (EMBRAPA SEP 12.2002.001)-“Aplicação de técnicas quantitativas digitais para otimizar o mapeamento de solos para fins de planejamento e gestão ambiental”, FAPERJ E-26/171.360/2001 (EMBRAPA SEP 01.2002.202) “Modelagem da magnitude e distribuiç ão espacial do carbono orgânico nos solos do Estado do Rio de Janeiro, usando técnicas quantitativas, SIG e Base de Dados” and “The University of Sydney Sesquicentennial Project (A general approach to making digital soil maps)”. We thank Dr. Philippe Lagacherie, INRA, Montpellier and Dr. Thomas Mayr, National Soil Resource Institute, Silsoe for constructive comments on an earlier draft.

References

Aandahl, A.R., 1948. The characterization of slope positions and their influence on the total N content of a few virgin soils in Western Iowa. Soil Science Society of America Proceedings 13, 449–454.

Agbu, P.A., Fehrenbacher, D.J., Jansen, I.J., 1990. Soil property relationships with SPOT satellite digital data in east central Illinois. Soil Science Society of America Journal 54, 807–812.

Ahmed, S., DeMarsily, G., 1987. Comparison of geostatistical methods for estimating transmisivity using data on transmisivity and specific capacity. Water Resources Research 23, 1717–1737.

Aiken, R.M., Jawson, M.D., Grahammer, K., Polymenopoulos, A.D., 1991. Positional, spatially correlated and random components of variability in carbon-dioxide efflux. Journal of Environmental Quality 20, 301–308.

Albert, P.S., McShane, L.S., 1995. A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data. Biometrics 51, 627–638.

Altese, E., Bolognani, O., Mancini, M., Troeh, P.A., 1996. Retrieving soil moisture over bare soil from ERS-1 synthetic aperture radar data: sensitivity analysis based on theoretical surface scattering model and field data. Water Resources Research 32, 653–661.

Anderson, K.E., Furley, P.A., 1975. An assessment of the relationship between surface properties of chalk soils and slope form using principal component analysis. Journal of Soil Science 26, 130–143.

Anderson-Cook, C.M., Alley, M.M., Roygard, J.K.F., Khosla, R., Noble, R.B., Doolittle, J.A., 2002. Differentiating soil types using electromagnetic conductivity and crop yield maps. Soil Science Society of America Journal 66, 1570–1652.

Andriani, T., Balia, R., Loddo, M., Pecorini, G., Tramacere, A., 2001. Structural features of the Middle Tirso Valley (Central

Sardinia, Italy) from geoelectrical and gravity data. Annali di Geofisica 44, 739–753.
Arrouays, D., Vion, I., Kicin, J.L., 1995. Spatial analysis and modeling of topsoil carbon storage in temperate forest humic loamy soils of France. Soil Science 159, 191–198.
Barshad, I., 1958. Factors affecting clay formation. 6th National Conference on Clays and Clay Mineralogy, pp. 110–132.
Beard, L.P., 2000. Comparison of methods for estimating earth resistivity from airborne electromagnetic measurements. Journal of Applied Geophysics 45, 239–259.
Bell, J.C., Cunningham, R.L., Havens, M.W., 1992. Calibration and validation of a soil–landscape model for predicting soil drainage class. Soil Science Society of America Journal 56, 1860–1866.
Bell, J.C., Cunningham, R.L., Havens, M.W., 1994. Soil drainage probability mapping using a soil–landscape model. Soil Science Society of America Journal 58, 464–470.
Bell, J.C., Grigal, D.F., Bates, P.C., 2000. A soil–terrain model for estimating spatial patterns of soil organic carbon. In: Wilson, J.P., Gallant, J.C. (Eds.), Terrain Analysis—Principles and Applications. Wiley, New York, pp. 295–310.
Bhatti, A.U., Mulla, D.J., Frazier, B.E., 1991. Estimation of soil properties and wheat yields on complex eroded hills using geostatistics and thematic mapper images. Remote Sensing of Environment 37, 181–191.
Bierwith, P.N., 1996. Gamma-radiometrics, a remote sensing tool for understanding soils. Australian Collaborative Land Evaluation Program Newsletter 5, 12–14.
Bindlish, R., Barros, A., 1996. Aggregation of digital terrain data using a modified fractal interpolation scheme. Computers and Geosciences 22, 907–917.
Bishop, T.F.A., McBratney, A.B., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103, 149–160.
Bishop, T.F.A., McBratney, A.B., Whelan, B.M., 2001. Measuring the quality of digital soil maps using information criteria. Geoderma 105, 93–111.
Blaszcynski, J.S., 1997. Landform characterization with geographic information systems. Photogrammetric Engineering and Remote Sensing 63, 183–191.
Boegh, E., Soegaard, H., Thomsen, A., 2002. Evaluating evapotranspiration rates and surface conditions using Landsat TM to estimate atmospheric resistance and surface resistance. Remote Sensing of Environment 79, 329–343.
Bogaert, P., D'Or, D., 2002. Estimating soil properties from thematic soil maps: the Bayesian Maximum Entropy approach. Soil Science Society of America Journal 66, 1492–1500.
Boruvka, L., Kozak, J., Nemecek, J., Penizek, V., 2002. New approaches to the exploitation of former soil survey data. 17th World Congress of Soil Science, Bangkok, Thailand, August 14–21. Paper no. 1692.
Boulaine, J., 1980. Pédologie Appliquée. Collection Sciences Agronomiques. Masson, Paris.
Bourennane, H., King, D., Chery, P., Bruand, A., 1996. Improving the kriging of a soil variable using slope gradient as external drift. European Journal of Soil Science 47, 473–483.
Bourennane, H., King, D., Couturier, A., 2000. Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. Geoderma 97, 255–271.
Bourgault, G., 1994. Robustness of noise filtering by kriging analysis. Mathematical Geology 26, 733–752.
Brest, C.L., Goward, S.N., 1987. Deriving surface albedo measurements from narrow-band satellite data. International Journal of Remote Sensing 8, 351–367.
Breiman, L., 1996. Bagging predictors. Machine Learning 26, 123–140.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression. Tress. Wadsworth, Belmont, CA.
Bui, E.N., 2003. Soil survey as a knowledge system. Geoderma (in press).
Bui, E.N., Moran, C.J., 2000. Regional-scale investigation of the spatial distribution and origin of soluble salts in central north Queensland. Hydrological Processes 14, 237–250.
Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103, 79–94.
Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. Geoderma 111, 21–44.
Bui, E.N., Loughhead, A., Corner, R., 1999. Extracting soil–landscape rules from previous soil surveys. Australian Journal of Soil Research 37, 495–508.
Bui, E.N., Henderson, B., Moran, C.J., Johnston, R., 2002. Continental-scale spatial modelling of soil properties. Proceedings of the 17th World Congress of Soil Science, Bangkok, Thailand, August 14–20, 2002. Symposium no. 52, paper no. 1470, 8 pp.
Bunkin, A.F., Bunkin, F.V., 2000. Lidar sensing of water, ground, and plant surfaces. Atmospheric and Oceanic Optics 13, 54–60.
Burgess, T.M., Webster, R., 1980a. Optimal interpolation and isarithmic mapping of soil properties: I. The semivariogram and punctual kriging. Journal of Soil Science 31, 315–331.
Burgess, T.M., Webster, R., 1980b. Optimal interpolation and isarithmic mapping of soil properties: II. Block kriging. Journal of Soil Science 31, 333–341.
Burrough, P.A., 1993. Soil variability: a late 20th century view. Soils and Fertilizers 56, 529–562.
Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford Univ. Press, Oxford.
Burrough, P.A., van Gaans, P.F.M., McMillan, R.A., 2000. High-resolution landform classification using fuzzy *k*-means. Fuzzy Sets and Systems 113, 37–52.
Campbell, J.B., 1987. Introduction to Remote Sensing. The Guilford Press, New York.
Campling, P., Gobin, A., Feyen, J., 2002. Logistic modeling to spatially predict the probability of soil drainage classes. Soil Science Society of America Journal 66, 1390–1401.
Carré, F., Girard, M.C., 2002. Quantitative mapping of soil types based on regression kriging of taxonomic distances with landform and land cover attributes. Geoderma 110, 241–263.
Carrol, T.R., 1981. Airbone soil moisture measurement using natural terrestrial gamma radiation. Soil Science 132, 358–366.
Carvalho, L.M.T., Fonseca, L.M.G., Murtagh, F., Clevers, J.G.P.W.,

2001. Digital change detection with the aid of multiresolution wavelet analysis. International Journal of Remote Sensing 22, 3871–3876.
Chang, D.H., Islam, S., 2000. Estimation of soil physical properties using remote sensing and artificial neural network. Remote Sensing of Environment 74, 534–544.
Chang, C.-W., Laird, D.W., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy—principal components regression analyses of soil properties. Soil Science Society of America Journal 65, 480–490.
Chaplot, V., Walter, C., 2002. The suitability of quantitative soil–landscape models for predicting soil properties at a regional level. 7th World Congress of Soil Science, Bangkok, Thailand, August 14–21. Paper no. 2331.
Chaplot, V., Walter, C., Curmi, P., 2000. Improving soil hydromorphy prediction according to DEM resolution and available pedological data. Geoderma 97, 405–422.
Chen, X., Tateishi, R., Wang, C., 1999. Development of a 1-km landcover dataset of China using AVHRR data. ISPRS Journal of Photogrammetry and Remote Sensing 54, 305–316.
Christakos, G., 1990. A Bayesian/maximum-entropy view to the spatial estimation problem. Mathematical Geology 22, 763–777.
Christakos, G., 2000. Modern Spatiotemporal Geostatistics. Oxford Univ. Press, New York.
Cialella, A.T., Dubayah, R., Lawrence, W., Levine, E., 1997. Predicting soil drainage class using remotely sensed and digital elevation data. Photogrammetric Engineering and Remote Sensing 63, 171–178.
Cipra, J.E., Franzmeir, D.P., Bauer, M.E., Boyd, R.K., 1980. Comparison of multispectral measurements from some nonvegetated soils using Landsat digital data and a spectroradiometer. Science Society of America Journal 44, 80–84.
Clark, L.A., Pregibon, D., 1992. Tree-based models. In: Chambers, J.M., Hastie, T.J. (Eds.), Statistical Models. S. Wadsworth and Brooks, California, USA, pp. 377–420.
Clevers, J.G.P.W., van Leeuwen, H.J.C., 1996. Combined use of optical and microwave remote sensing data for crop growth monitoring. Remote Sensing of the Environment 56, 42–51.
Congalton, R.G., Green, K., 1999. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices Lewis, New York.
Cook, S.E., Corner, R., Grealish, G.J., Gessler, P.E., Chartres, C.J., 1996a. A rule-based system to map soil properties. Science Society of America Journal 60, 1893–1900.
Cook, S.E., Corner, R., Groves, P.R., Grealish, G.J., 1996b. Use of airborne gamma radiometric data for soil mapping. Australian Journal of Soil Research 34, 183–194.
Corner, R.J., Cook, S.E., Moore, G.A., 1997. Expector: a knowledge based soil attribute mapping method. ACLEP (Australian Collaborative Land Evaluation Project) Newsletter 6, 9–11.
Cox, G.M., Martin, W.M., 1937. Use of a discriminant function for differentiating soils with different azotobacter populations. Iowa Experimental J451, 323–332.
Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley, New York.
Crowley, J.K., 1993. Mapping playa evaporite minerals with AVIRIS data: a 1st report from Death Valley, California. Remote Sensing of the Environment 44, 337–356.
Dale, M.B., McBratney, A.B., Russell, J.S., 1989. On the role of expert systems and numerical taxonomy in soil classification. Journal of Soil Science 40, 223–234.
Davies, B.E., Gamm, S.A., 1969. Trend surface analysis applied to soil reaction values from Kent, England. Geoderma 3, 223–231.
De Bruin, S., Stein, A., 1998. Soil–landscape modeling using fuzzy *c*-means clustering of attribute data derived from a Digital Elevation Model (DEM). Geoderma 83, 17–33.
DeFries, R.S., Hansen, M.C., Townshend, J.R.G., 2000. Global continuous fields of vegetation characteristics: a linear mixture model applied to multi-year 8 km AVHRR data. International Journal of Remote Sensing 21, 1389–1414.
De Gruijter, J.J., Bie, S.W., 1975. A discrete approach to automated mapping of multivariate systems. In: Wilford-Brickwood, J.M., Bertrand, R., van Zuylen, L. (Eds.), Automation in Cartography. Proceedings Technical Working Session, Commission III. International Cartography Association, Enschede, pp. 17–28.
De Gruijter, J.J., Marsman, B.A., 1985. Transect sampling for reliable information on mapping units. In: Nielsen, D.R., Bouma, J. (Eds.), Soil Spatial Variability. Pudoc, Wageningen, pp. 150–163.
De Gruijter, J.J., Walvoort, D.J.J., van Gaans, P.F.M., 1997. Continuous soil maps—a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. Geoderma 77, 169–195.
Delhomme, J.P., 1978. Kriging in the hydrosciences. Advances in Water Resources 1, 251–266.
Diak, G.R., Anderson, M.D., Bland, W.L., Norman, J.M., Mecikalski, J.M., Aune, R.M., 1998. Agricultural management decision aids driven by real-time satellite data. Bulletin of the American Meteorological Society 79, 1345–1355.
Dickson, B.L., Fraser, S.J., Kinsey-Henderson, A., 1996. Interpreting aerial gamma-ray surveys utilising geomorphological and weathering models. Journal of Geochemical Explorations 57, 75–88.
Diggle, P.J., Liang, K.Y., Zeger, L.S., 1994. Analysis of Longitudinal Data. Oxford Univ. Press, Oxford.
Dobermann, A., Oberthur, T., 1997. Fuzzy mapping of soil fertility—a case study on irrigated riceland in the Phlippines. Geoderma 77, 317–339.
Dobos, E., Micheli, E., Baumgardner, M.F., Biehl, L., Helt, T., 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. Geoderma 97, 367–391.
Dobos, E., Montanarella, L., Negre, T., Micheli, E., 2001. A regional scale soil mapping approach using integrated AVHRR and DEM data. International Journal of Applied Earth Observation and Geoinformation 3, 30–41.
Dobson, M.C., Ulaby, F.T., 1986. Active microwave soil moisture research. IEEE Transactions on Geosciences and Remote Sensing 24, 23–36.
Donatelli, M., Stockle, C., Constantini, E.A., Nelson, R., 2002. SOILR: a model to estimate soil moisture and temperature regimes. http://www.inea.it/isci/mdon/research/bottom_model_soil.htm.
Duda, R.O., Hart, P., Barrett, J.G., Gashnig, K., Reboh, R., Slocum,

J., 1978. Development of PROSPECTOR consultation system for mineral exploration. SRI Projects 5821 and 6415, SRI International Artificial Intelligence Center, Menlo Park, CA.

Dymond, J.R., Luckman, P.G., 1994. Direct induction of compact rule-based classifiers for resource mapping. International Journal of Geographical Information Systems 8, 357–367.

Dymond, J.R., Stephens, P.R., Newsome, P.F., Wilde, R.H., 1992. Percent vegetation cover of a degrading rangeland from SPOT. International Journal of Remote Sensing 13, 1999–2007.

Edmonds, W.J., Campbell, J.B., 1984. Spatial estimates of soil temperature. Soil Science 138, 203–208.

Efron, B., Tibshirani, R.J., 1993. An introduction to the bootstrap. Monographs on Statistics and Applied Probability, vol. 57. Chapman & Hall, London, UK.

Epinat, V., Stein, A., de Jong, S.M., Bouma, J., 2001. A wavelet characterization of high-resolution NDVI patterns for precision agriculture. ITC Journal 2, 121–132.

Escoufier, Y., 1970. Echantillonage dans une population de variables aléatoires réellles. Publications Institute of Statistics, University of Paris 19, 1–47.

Evans, I.S., 1972. General geomorphometry, derivatives of altitude, and descriptive statistics. In: Chorley, R.J. (Ed.), Spatial Analysis in Geomorphology. Methuen, London, pp. 17–90.

Evans, I.S., 1980. An integrated system of terrain analysis and slope mapping. Zeitschrift für Geomorphologie Supplementband 36, 274–295.

Evans, I.S., 1998. What do terrain statistics really mean? In: Lane, S.N., Richards, K.S., Chandler, H. (Eds.), Land Monitoring, Modelling and Analysis. Wiley, Chichester, pp. 119–138.

Favrot, J.C., 1989. Une stratégie d'inventaire cartographique à grand échelle: la méthode des secteurs de réference. Science du sol 27, 351–368.

Favrot, J.C., Lagacherie, P., 1993. La cartographie automatisee des sols: une aide a la gestion ecologique des paysages ruraux. Comptes Rendus de L'Academie d'Agriculture de France 79, 61–76.

Fels, J.E., Matson, K.C., 1996. A cognitively-based approach for hydrogeomorphic land classification using digital terrain models. 3rd International Conference on Integrating GIS and Environmental Modeling, Santa Fe, New Mexico.

Feng, C., Michie, D., 1994. Machine learning of rules and trees. In: Michie, D., Spiegelhalter, D.J., Taylor, C.C. (Eds.), Machine Learning, Neural and Statistical Classification. Ellis Horwood, Hertfordshire, pp. 50–83.

Fidêncio, P.H., Ruisanchez, I., Poppi, R.J., 2001. Application of artificial neural networks to the classification of soils from Saõ Paulo state using near-infrared spectroscopy. Analyst 126, 2194–2200.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188.

Florinsky, I.V., 1998. Accuracy of land topographical variables derived from digital elevation models. International Journal of Geographic Information Science 12, 47–61.

Florinsky, I.V., Eilers, R.G., 2002. Prediction of the soil organic carbon content at micro-, meso- and macroscales by digital terrain modelling. 7th World Congress of Soil Science, Bangkok, Thailand, August 14–21. Paper no. 24.

Florinsky, I.V., Eilers, R.G., Manning, G.R., Fuller, L.G., 2002. Prediction of soil properties by digital terrain modelling. Environmental Modelling and Software 17, 295–311.

Frazier, B.E., Cheng, Y., 1989. Remote sensing of soils in the eastern Palouse region with Landsat thematic mapper. Remote Sensing of the Environment 28, 317–325.

Freund, Y., Schapire, R.E., 1996. Game theory, on-line prediction and boosting. Proceedings of the 9th Annual Conference on Computing and Learning Theory, pp. 325–332.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139.

Fridland, V.M., 1972. Pattern of the soil cover. Israel Program for Scientific Translations, Jerusalem.

Friedman, J.H., 1991. Multivariate adaptive regression splines (with discussion). Annals of Statistics 19, 1–67.

Friedman, J.H., 1999. Greedy function approximation: A gradient boosting machine. Technical report, Department of Statistics, Stanford University.

Fung, A.K., Li, Z., Chen, K.S., 1992. Backscattering from a randomly rough dielectric surface. IEEE Transactions on Geosciences and Remote Sensing 30, 356–369.

Furley, P.A., 1968. Soil formation and slope development: 2. The relationship between soil formation and gradient angle in the Oxford area. Zeitschrift für Geomorphologie 12, 25–42.

Galdeano, A., Asfirane, F., Truffert, C., Egal, E., Debeglia, N., 2001. The aeromagnetic map of the French Cadomian belt. Tectonophysics 331, 99–122.

Gallant, J.C., Wilson, J.P., 2000. Primary topographic attributes. In: Wilson, J.P., Gallant, J.C. (Eds.), Terrain Analysis—Principles and Applications. Wiley, New York, pp. 51–86.

Garguet-Duport, B., 1997. WaveMerg: a multiresolution software for merging SPOT panchromatic and SPOT multispectral data. Environmental Modelling and Software 12, 85–92.

Gershenfeld, N., 1999. The Nature of Mathematical Modelling. Cambridge Univ. Press, Cambridge, UK.

Gessler, P.E., Moore, I.D., McKenzie, N.J., Ryan, P.J., 1995. Soil–landscape modelling and spatial prediction of soil attributes. International Journal of Geographical Information Systems 9, 421–432.

Giltrap, D.J., 1977. Mathematical techniques for soil survey design. Doctor of Philosphy thesis, University of Oxford.

Girard, M.C., 1983. Recherche d'une modélisation en vue d'une représentation spatiale de la couverture pédologique. Application á une région des plateaux jurassiques de Bourgogne. Thèse d'Etat, Sols no. 12. INA-PG, 430 pp.

Goetz, S.J., Halthore, R.N., Hall, F.G., Markham, B.L., 1995. Surface temperature retrieval in a temperate grassland with multiresolution sensors. Journal of Geophysical Research 100 (D12), 25397–25410.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimisation and Machine Learning. Addison-Wesley, New York.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89, 1–45.

Gopal, S., Woodcock, C., 1995. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. Photogrammetric Engineering and Remote Sensing 60, 181–188.

Goulard, M., Voltz, M., 1992. Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. Mathematical Geology 24, 269–286.

Grasty, R.L., Mellander, H., Parker, M., 1991. Airborne Gamma-Ray Spectrometer Surveying. International Atomic Energy Agency, Vienna.

Grunwald, S., McSweeney, K., Rooney, D.J., Lowery, B., 2001. Soil layer models created with profile cone penetrometer data. Geoderma 103, 181–201.

Gupta, R.K., Vijayan, D., Prasad, T.S., 2001. New hyperspectral vegetation characterization parameters. Advances in Space Research 28, 201–206.

Hastie, T.J., Pregibon, D., 1992. Generalized linear models. In: Chambers, J.M., Hastie, T.J. (Eds.), Statistical Models. S. Wadsworth and Brooks, California, USA, pp. 195–248.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London, UK.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning: data mining, inference and prediction. Springer Series in Statistics. Springer-Verlag, New York.

Hay, R.L., 1960. Rate of clay formation and mineral alteration in a 4000-years-old volcanic ash soil on St. Vincent, B.W.I. American Journal of Science 258, 354–368.

Henderson, R., Ragg, J.M., 1980. A reappraisal of soil mapping in an area of Southern Scotland: Part II. The usefulness of some morphological properties and of a discriminant analysis in distinguishing between the dominant taxa of four mapping units. Journal of Soil Science 31, 573–580.

Hengl, T., Rossiter, D.G., Husnjak, S., 2002. Mapping soil properties from an existing national soil data set using freely available ancillary data. 17th World Congress of Soil Science, Bangkok, Thailand, August 14–21. Paper no. 1140.

Hengl, T., Heuvelink, G.B.M., Stein, A., 2003a. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma (in press).

Hengl, T., Rossiter, D.G., Stein, A., 2003b. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Australian Journal of Soil Research (in press).

Heuvelink, G.B.M., 1996. Identification of field attribute error under different models of spatial variation. International Journal of Geographical Information Science 10, 921–935.

Heuvelink, G.B.M., 1998. Error Propagation in Environmental Modelling with GIS. Taylor and Francis, London.

Heuvelink, G.B.M., Burrough, P.A., 2002. Developments in statistical approaches to spatial uncertainty and its propagation. International Journal of Geographical Information Science 16, 111–113.

Heuvelink, G.B.M., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma 100, 269–301.

Hewitt, A.E., 1993. Predictive modelling in soil survey. Soils and Fertilizers 56, 305–314.

Hoeben, R., Troch, P.A., 2000. Assimilation of active microwave observation data for soil moisture profile estimation. Water Resources Research 36, 2805–2819.

Hoekman, D.H., Quiñones, M.J., 1999. P-band SAR for tropical forest and landcover change observation. ESA Earth Observation Quarterly 61, 18–22.

Hoersch, B., Braun, G., Schmidt, U., 2002. Relation between landform and vegetation in alpine regions of Wallis, Switzerland. A multiscale remote sensing and GIS approach. Computers, Environment and Urban Systems 26, 113–139.

Hole, F.D., 1981. Effects of animals on soil. Geoderma 25, 75–112.

Hudson, B.D., 1992. The soil survey as paradigm-based science. Soil Science Society of America Journal 56, 836–841.

Huette, A.R., 1988. A soil adjusted vegetation index (SAVI). Remote Sensing of Environment 25, 295–309.

Hutchinson, M.F., 1998a. Interpolation of rainfall data with thin plate smoothing splines: I. Two dimensional smoothing of data with short range correlation. Journal of Geographic Information and Decision Analysis 2, 152–167.

Hutchinson, M.F., 1998b. Interpolation of rainfall data with thin plate smoothing splines: II. Analysis of topographic dependence. Journal of Geographic Information and Decision Analysis 2, 168–185.

Hutchinson, M.F., Gessler, P.E., 1994. Splines—more than just a smooth interpolator. Geoderma 62, 45–67.

Jackson, T.J., Schmugge, J., Engman, E.T., 1996. Remote sensing applications to hydrology: soil moisture. Hydrological Sciences Journal 41, 517–530.

Jang, J.S.R., 1993. ANFIS: adaptive-network-based fuzzy inference systems. IEEE Transactions on Systems, Man, and Cybernetics 23, 665–685.

Jang, J.S.R., 1997. Classification and regression trees. In: Jang, J.-S.R., Sun, C.T., Mizutani, E. (Eds.), Neuro-Fuzzy and soft computing. A computational approach to learning and machine intelligence. Prentice-Hall, Upper Saddle River, pp. 403–422.

Jenny, H., 1941. Factors of Soil Formation, A System of Quantitative Pedology. McGraw-Hill, New York.

Jones, M.J., 1973. The organic matter content of the savanna soils of West Africa. Journal of Soil Science 24, 42–53.

Kim, G., Barros, A.P., 2002. Downscaling of remotely sensed soil moisture with a modified fractal interpolation method using contraction mapping and ancillary data. Remote Sensing of Environment 83, 400–413.

King, T.V.V., Clark, R.N., Ager, C., Swayze, G.A., 1995. Remote mineral mapping using AVIRIS data at Summitville, Colorado and the adjacent San Juan Mountains, Summitville Forum ’95. Special Publication, Geological Survey, Colorado.

King, D., Jamagne, M., Arrouays, D., Bornand, D., Favrot, J.C., Hardy, R., Le Bas, C., Stengel, P., 1999. Inventaire cartographique et surveillance des sols en France. Etat d’avancement et exemples d’utilisation. Étude et Gestion des Sols 6, 215–228.

Kiss, J.J., de Jong, E., Martz, L.W., 1988. The distribution of fallout Cesium-137 in southern Saskatchewan, Canada. Journal of Environmental Quality 17, 445–452.

Knotters, M., Brus, D.J., Oude Voshaar, J.H., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67, 227–246.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics 43, 59–69.

Kravchenko, A.N., Bollero, G.A., Omonode, R.A., Bullock, D.G., 2002. Quantitative mapping of soil drainage classes using topo-

graphical data and soil electrical conductivity. Soil Science Society of America Journal 66, 235–243.
Kuhn, T.S., 1996. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Kustas, W.P., Norman, J.M., 1996. Use of remote sensing for evapotranspiration monitoring over land surfaces. Hydrological Sciences Journal 41, 495–516.
Laba, M., Gregory, S.K., Braden, J., Ogurcak, D., Hill, E., Fegraus, E., Fiore, J., DeGloria, S.D., 2002. Conventional and fuzzy accuracy assessment of the New York gap analysis project land cover map. Remote Sensing of Environment 81, 443–455.
Lagacherie, P., 1992. Formalisation des lois de distribution des sols pour automiser la cartographie pedologique a partir d'un secteur pris comme reference. PhD thesis, Université Montpellier II, France.
Lagacherie, P., Holmes, S., 1997. Addressing geographical data errors in a classification tree soil unit prediction. International Journal of Geographical Information Science 11, 183–198.
Lagacherie, P., Voltz, M., 2000. Predicting soil properties over a region using sample information from a mapped reference area and digital elevation data: a conditional probability approach. Geoderma 97, 187–208.
Lagacherie, P., Robbez-Masson, J.M., Nguyen-The, N., Barthès, J.P., 2001. Mapping of reference area representativity using a mathematical soilscape distance. Geoderma 101, 105–118.
Lane, P.W., 2002. Generalized linear models in soil science. European Journal of Soil Science 53, 241–251.
Lapen, D.R., Topp, G.C., Gregorich, E.G., Hayhoe, H.N., Curnoe, W.E., 2001. Divisive field-scale associations between corn yields, management, and soil information. Soil and Tillage Research 58, 193–206.
Lark, R.M., 2000. Regression analysis with spatially autocorrelated error: simulation studies and application to mapping of soil organic matter. International Journal of Geographical Information Science 14, 247–264.
Lark, R.M., Papritz, A., 2003. Fitting a linear model of coregionalization for soil properties using simulated annealing. Geoderma 115, 245–260.
Lark, R.M., Kaffka, S.R., Corwin, D.L., 2003. Multiresolution analysis of data on electrical conductivity of soil using wavelets. Journal of Hydrology 272, 276–290.
Lascano, R.J., Baumhardt, R.L., Hicks, S.K., Landivar, J.A., 1998. Spatial and temporal distribution of surface water content in a large agricultural field. In: Robert, P.C., Rust, R.H., Larson, W.E. (Eds.), Precision Agriculture. ASA-CSSA-SSSA, Madison, WI, USA, pp. 19–30.
Laslett, G.M., McBratney, A.B., Pahl, P.J., Hutchinson, M.F., 1987. Comparison of several spatial prediction methods for soil pH. Journal of Soil Science 38, 325–341.
Lee, K.S., Lee, G.B., Tyler, E.J., 1988. Determination of soil characteristics from thematic mapper data of a cropped organic–inorganic soil landscape. Soil Science Society of America Journal 52, 1100–1104.
Lees, B.G., Ritman, K., 1991. Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. Environmental Management 15, 823–831.
Legros, J.P., 1975. Occurrence des podzols dans l'Est du Massif Central. Science du Sol 1, 37–49.
Legros, J.P., Bonneric, P., 1979. Modelisation informatique de la repartition des sols dans le Parc Naturel Régional du Pilat. Annales de l'Université de Savoie, Tome 4, Sciences, 63–68.
Li, F., Lyons, T.J., 2002. Remote estimation of regional evapotranspiration. Environmental Modelling and Software 17, 61–75.
Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.
Lilburne, L., Hewitt, A., McIntosh, P., Lynn, I., 1998. GIS-driven models of soil properties in the high country of the south island. 10th Colloquium of the Spatial Information Research Centre, University of Otago, New Zealand, pp. 173–180.
Lin, D.S., Wood, E.F., Troch, P.A., Mancini, M., Jackson, T.J., 1994. Comparison of remote sensed and model simulated soil moisture over a heterogenous watershed. Remote Sensing of the Environment 48, 159–171.
Lobell, D.B., Asner, G.P., Ortiz-Monasterio, J.I., Benning, T.L., 2003. Remote sensing of regional crop production in the Yaqui Valley, Mexico: estimates and uncertainties. Agriculture, Ecosystems and Environment 94, 205–220.
Lösel, G., 2003. Application of heterogeneity indices to coarse-scale soil maps. Abstracts, Pedometrics 2003, International Conference of the IUSS Working Group on Pedometrics, Reading University, Reading, England, September 11–12.
Lund, E.D., Colin, P.E., Christy, D., Drummond, P.E., 1999. Applying soil conductivity technology to precision agriculture. In: Robert, P.C., Rust, R.H., Larson, W.E. (Eds.), Proceedings of the Fourth International Conference on Precision Agriculture. ASA-CSSA-SSSA, Madison, WI, pp. 1089–1100.
MacMillan, R.A., Pettapiece, W.W., Nolan, S.C., Goddard, T.W., 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets and Systems 113, 81–109.
Mancini, M., Hoeben, R., Troch, P.A., 1999. Multifrequency radar observations of bare surface soil moisture content: a laboratory experiment. Water Resources Research 35, 1827–1838.
Marsman, B.A., de Gruijter, J.J., 1986. Quality of soil maps, a comparison of soil survey methods in a study area. Soil Survey papers no. 15. Netherlands Soil Survey Institute, Stiboka, Wageningen, The Netherlands.
Martens, G., Naes, T., 1989. Multivariate Calibration. Wiley, New York.
Matt, P.B., Johnson, W.C., 1996. Thermoluminescence and new ^{14}C age estimates for late Quaternary loesses in southwestern Nebraska. Geomorphology 17, 115–128.
Maulik, U., Bandyopadhyay, S., 2000. Genetic algorithm-based clustering technique. Pattern Recognition 33, 1455–1465.
Mayr, T.R., Evans, L., Mann, A.D., 2003. Comparison of two statistical techniques for soil–landscape analysis. Geoderma (submitted for publication).
McBratney, A.B., Webster, R., 1983. Optimal interpolation and isarithmic mapping of soil properties: V. Co-regionalization and multiple sampling strategy. Journal of Soil Science 34, 137–162.
McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Sha-

tar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97, 293–327.

McBratney, A.B., Minasny, B., Cattle, S., Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109, 41–73.

McCullagh, P., Nelder, J.A., 1983. Generalized Linear Models. Cambridge Univ. Press, Cambridge, UK.

McIntosh, P.D., Lynn, I.H., Johnstone, P.D., 2000. Creating and testing a geometric soil–landscape model in dry steeplands using a very low sampling density. Australian Journal of Soil Research 38, 101–112.

McKay, M.D., Conover, W.J., Beckman, R.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 221, 239–245.

McKenzie, N.J., Austin, M.P., 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57, 329–355.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67–94.

McKenzie, N.J., Gessler, P.E., Ryan, P.J., O'Connel, D., 2000. The role of terrain analysis in soil mapping. In: Wilson, J.P., Gallant, J.C. (Eds.), Terrain Analysis—Principles and Applications. Wiley, New York, pp. 245–265.

Mendonça Santos, M.L., Guenat, C., Thevoz, C., Bureau, F., Vedy, J.C., 1997. Impacts of embanking on the soil–vegetation relationships in a floodplain ecosystem of a pre-alpine river. Global Ecology and Biogeography Letters 6, 339–348.

Michaelsen, J., Schimel, D.S., Friedl, M.A., Davis, F.W., Dubayah, R.C., 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. Journal of Vegetation Science 5, 673–686.

Michalski, R.S., Chilauski, R.L., 1980. Knowledge acquisition by encoding expert rules versus computer induction from examples: a case study involving soybean pathology. International Journal of Man–Machine Studies 12, 63–87.

Milne, G., 1935. Some suggested units of classification and mapping particularly for East African soils. Soil Research 4, 183–198.

Minasny, B., McBratney, A.B., 2001. A rudimentary mechanistic model for soil production and landscape development: II. A two-dimensional model. Geoderma 103, 161–179.

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

Moore, D.M., Lees, B.G., Davey, S.M., 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. Environmental Management 15, 59–71.

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal 57, 443–452.

Moran, C.J., Bui, E., 2002. Spatial data mining for enhanced soil map modelling. International Journal of Geographical Information Science 16, 533–549.

Moran, M.S., Inoue, Y., Barnes, E.M., 1997. Opportunities and limitations for image-based remote sensing in precision crop management. Remote Sensing of the Environment 61, 319–346.

Mücher, C.A., Steinnocher, E.T., Kressler, F.P., Heunks, C., 2000. Land cover characterization and change detection for environmental monitoring of pan-Europe. International Journal of Remote Sensing 21, 1159–1181.

Munyati, C., 2000. Wetland change detection on the Kafue Flats, Zambia, by classification of a multitemporal remote sensing image dataset. International Journal of Remote Sensing 21, 1787–1806.

New, M., Todd, M., Hulme, M., Jones, P., 2001. Precipitation measurements and trends in the twentieth century. International Journal of Climatology 21, 1922–1999.

Noy-Meir, I., 1974. Multivariate analysis of the semiarid vegetation in south-eastern Australia: II. Vegetation catena and environmental gradients. Australian Journal of Botany 22, 115–140.

Nye, P.H., Greenland, D.J., 1960. The Soil under Shifting Cultivation. Commonwealth Bureau of Soils, Harpenden, UK.

Oberthur, T., Goovaerts, P., Dobermann, A., 1999. Mapping soil texture classes using field texturing, particle size distribution and local knowledge by both conventional and geostatistical methods. European Journal of Soil Science 50, 457–479.

Obukhov, A.I., Orlov, D.S., 1964. Spectral reflectivity of the major soil groups and possibility of using diffuse reflection in soil investigations. Soviet Soil Science, 174–184.

Odeh, I.O.A., McBratney, A.B., 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. Geoderma 97, 237–254.

Odeh, I.O.A., Chittleborough, D.J., McBratney, A.B., 1991. Elucidation of soil–landform interrelationships by canonical ordination analysis. Geoderma 49, 1–32.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992. Soil pattern recognition with fuzzy-*c*-means: application to classification and soil–landform interrelationships. Soil Science Society of America Journal 56, 505–516.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63, 197–214.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67, 215–225.

Odeh, I.O.A., McBratney, A.B., Slater, B.K., 1997. Predicting soil properties from ancillary information: non-spatial models compared with geostatistical and combined methods. 5th International Geostatistics Congress, Wollongong, pp. 22–27.

Oliver, M.A., Webster, R., Slocum, K., 2000. Filtering SPOT imagery by kriging analysis. International Journal of Remote Sensing 21, 735–752.

Opsomer, J.D., Ruppert, D., Wand, M.P., Holst, U., Hossjer, O., 1999. Kriging with nonparametric variance function estimation. Biometrics 55, 704–710.

Owens, K.E., Reed, D.D., Londo, A.J., Maclean, A.L., Mroz, G.D., 1999. A landscape level comparison of pre-European settlement and current soil carbon content of a forested landscape in upper Michigan. Forest Ecology and Management 113, 179–189.

Pace, R.K., Barry, R., 1997. Quick computations of regressions with a spatially autoregressive dependent variable. Geographical Analysis 29, 232–247.

Pachepsky, Y.A., Timlin, D.J., Rawls, W.J., 2001. Soil water retention as related to topographic variables. Soil Science Society of America Journal 65, 1787–1795.

Pal, S.K., Bandyopadhyay, S., Murthy, C.A., 1998. Algorithms for generation of class boundaries. IEEE Transactions on System, Man and Cybernetics 28, 816–828.

Pal, S.K., Bandyopadhyay, S., Murthy, C.A., 2001. Genetic classifiers for remotely sensed images: comparison with standard methods. International Journal of Remote Sensing 22, 2545–2569.

Palacios-Orueta, A., Ustin, S.L., 1998. Remote sensing of soil properties in the Santa Monica mountains I. Spectral analysis. Remote Sensing of the Environment 65, 170–183.

Park, S.J., Vlek, L.G., 2002. Prediction of three-dimensional soil spatial variability: a comparison of three environmental correlation techniques. Geoderma 109, 117–140.

Park, S.J., McSweeney, K., Lowery, B., 2001. Identification of the spatial distribution of soils using a process-based terrain characterization. Geoderma 103, 249–272.

Pavlik, H.F., Hole, F.D., 1977. Soilscape analysis of slightly contrasting terrains in southeastern Wisconsin. Soil Science Society of America Journal 41, 407–413.

Pebesma, E.J., Duin, R.N.M., Bio, A.M.F., 2000. Spatial Interpolation of Sea Bird Densities on the Dutch Part of the North Sea. Centre for Geo-Ecological Research, Utrecht.

Pedrycz, W., Waletzky, J., 1997. Fuzzy clustering with partial supervision. IEEE Transactions on Systems, Man, and Cybernetics—Part B 27, 787–795.

Peng, W., Wheeler, D.B., Bell, J.C., Krusemark, M.G., 2003. Delineating patterns of soil drainage class on bare soils using remote sensing analyses. Geoderma 115, 261–279.

Pennock, D.J., Corre, M.D., 2001. Development and application of landform segmentation procedures. Soil and Tillage Research 58, 151–162.

Pennock, D.J., Zebarth, B.J., De Jong, E., 1987. Landform classification and soil distribution in hummocky terrain, Saskatchewan, Canada. Geoderma 40, 297–315.

Phillips, J.D., 2001. The relative importance of intrinsic and extrinsic factors in pedodiversity. Annals of the Association of American Geographers 91, 609–621.

Pike, R.J., 1988. The geometric signature: quantifying landslide terrain types from digital elevation models. Mathematical Geology 20, 491–511.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-Plus. Springer, New York.

Post, D.F., Horvath, E.H., Lucas, W.M., White, S.A., Ehasz, M.J., Batchily, A.K., 1994. Relations between soil color and Landsat reflectance on semiarid rangelands. Soil Science Society of America Journal 58, 1809–1816.

Quinlan, J.R., 1992. Learning with continuous classes. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, pp. 343–348.

Quiñones, M.J., Hoekman, D.H., 2001. Biomass mapping using biophysical forest type characterization of SAR polarimetric images. Proceedings of the Third International Symposium on Retrieval of Bio- and Geophysical Parameters for SAR Data for Land Applications. University of Sheffield, Sheffield, UK.

Rawls, W.J., Pachepsky, Y.A., 2002. Using field topographic descriptors to estimate soil water retention. Soil Science 167, 423–435.

Romano, N., Palladino, M., 2002. Prediction of soil water retention using soil physical data and terrain attributes. Journal of Hydrology 265, 56–75.

Rosenberg, A., 2000. Philosophy of Science: A Contemporary Introduction. Routledge, London.

RuleQuest Research, 2000. Cubist. RuleQuest Research, Sydney, Australia.

Ryan, P.J., McKenzie, N.J., O'Connell, D., Loughhead, A.N., Leppert, P.M., Jacquier, D., Ashton, L., 2000. Integrating forest soils information across scales: spatial prediction of soil properties under Australian forests. Forest Ecology and Management 138, 139–157.

Samra, J.S., Stahel, W.A., Kunsch, H., 1991. Modeling tree growth sensitivity to soil sodicity with spatially correlated observations. Soil Science Society of America Journal 55, 851–856.

Schaetzl, R.J., Barrett, L.R., Winkler, J.A., 1994. Choosing models for soil chronofunctions and fitting them to data. European Journal of Soil Science 45, 219–232.

Schmugge, T.J., Kustas, W.P., Ritchie, J.C., Jackson, T.J., Rango, A., 2002. Remote sensing in hydrology. Advances in Water Resources 25, 1367–1385.

Scull, P., Chadwick, O.A., Franklin, J., Okin, G., 2003a. A comparison of prediction methods to create spatially distributed soil property maps using soil survey data for an alluvial basin in the Mojave Desert California. Geoderma (in press).

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003b. Predictive soil mapping: a review. Progress in Physical Geography 27, 171–197.

Shary, P.A., 1995. Land surface in gravity points classification by a complete system of curvatures. Mathematical Geology 27, 373–390.

Shary, P.A., Sharayab, L.S., Mitusov, A.V., 2002. Fundamental quantitative methods of land surface analysis. Geoderma 107 (1–32), 1–43.

Shatar, T.M., McBratney, A.B., 1999. Empirical modelling of relationships between sorghum yield and soil properties. Precision Agriculture 1, 249–276.

Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Science Society of America Journal 66, 988–998.

Simonett, D.S., 1960. Soil genesis in basalt in North Queensland. Transactions of the 7th International Congress of Soil Science, Madison, Wisconsin, pp. 238–243.

Sinha, A.K., 1990. Stratigraphic mapping of sedimentary formations in southern Ontario by ground electromagnetic methods. Geophysics 55, 1148–1157.

Sinowski, W., Auerswald, K., 1999. Using relief parameters in a discriminant analysis to stratify geological areas with different spatial variability of soil properties. Geoderma 89, 113–128.

Skidmore, A.K., Ryan, P.J., Dawes, W., Short, D., O'Loughlin, E., 1991. Use of an expert system to map forest soils from a geographical information system. International Journal of Geographical Information Science 5, 431–445.

Skidmore, A.K., Watford, F., Luckananurug, P., Ryan, P.J., 1996. An operational GIS expert system for mapping forest soils. Photogrammetric Engineering and Remote Sensing 62, 501–511.

Skidmore, A.K., Varekamp, C., Wilson, L., Knowles, E., Delaney, J., 1997. Remote sensing of soils in a eucalypt forest environment. International Journal of Remote Sensing 18, 39–56.

Soil Survey Staff, 1993. Soil Survey Manual. Handbook No. 18. USDA, Washington, DC.

Sommer, M., Wehrhan, M., Zipprich, M., Castell, Z.W., Weller, U., Castell, W., Ehrich, S., Tandler, B., Selige, T., 2003. Hierarchical data fusion for mapping soil units at field scale. Geoderma 112, 179–196.

Stafford, J.V., Ambler, B., Lark, R.M., Catt, J., 1996. Mapping and interpreting the yield variation in cereal crops. Computers and Electronics in Agriculture 14, 101–119.

Su, Z., Troch, P.A., De Troch, F.P., 1997. Remote sensing of bare surface soil moisture using EMAC/ESAR data. International Journal of Remote Sensing 18, 2105–2124.

Sudduth, K.A., Drummond, S.T., Kitchen, N.R., 2001. Accuracy issues in electromagnetic induction sensing of soil electrical conductivity for precision agriculture. Computers and Electronics in Agriculture 31, 239–264.

Susskind, J., Piraino, P., Rokke, L., Iredell, L., Mehta, A., 1997. Characteristics of the TOVS pathfinder path. A data set. Bulletin of the American Meteorology Society 78, 1449–1472.

Tabbagh, A., Dabas, M., Hesse, A., Panissod, C., 2000. Soil resistivity: a non-invasive tool to map soil structure horizonation. Geoderma 97, 393–404.

Taylor, R., Eggleton, R.A., 2001. Regolith Geology and Geomorphology. Wiley, Chichester, UK.

Thomas, A.L., King, D., Dambrine, E., Couturier, A., Roque, A., 1999. Predicting soil classes with parameters derived from relief geologic materials in a sandstone region of the Vosges mountains (Northeastern France). Geoderma 90, 291–305.

Thompson, J.A., Bell, J.C., Butler, C.A., 1997. Quantitative soil–landscape modeling for estimating the areal extent of hydromorphic soils. Soil Science Society of America Journal 61, 971–980.

Thompson, J.A., Bell, J.C., Butler, C.A., 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil–landscape modeling. Geoderma 100, 67–89.

Townshend, J., Justice, C., Li, W., Gurney, C., McManus, J., 1991. Global land cover classification by remote-sensing—present capabilities and future possibilities. Remote Sensing of Environment 35, 243–255.

Triantafilis, J., Ward, W.T., Odeh, I.O.A., McBratney, A.B., 2001. Creation and interpolation of continuous soil layer classes in the lower Namoi valley. Soil Science Society of America Journal 65, 403–413.

Triantafilis, J., Odeh, I.O.A., Minasny, B., McBratney, A.B., 2003. Elucidation of hydrogeological units using fuzzy *k*-means classification of EM34 data in the lower Namoi Valley, Australia. Environmental Modelling and Software 18, 667–680.

Troeh, F.R., 1964. Landform parameters correlated to soil drainage. Soil Science Society of America Proceedings 28, 808–812.

Twidale, C.R., 1985. Old land surfaces and their implications for models of landscape evolution. Revue de Géomorphologie Dynamique 34, 131–147.

Van Niekerk, H.S., Gutzmer, J., Beukes, N.J., Phillips, D., Kiviets, G.B., 1999. An $^{40}Ar/^{39}Ar$ age of supergene K–Mn oxyhydroxides in a post-Gondwana soil profile on the Highveld of South Africa. South African Journal of Science 95, 450–454.

Vauclin, M., Vieira, S.R., Vachaud, G., Nielsen, D.R., 1983. The use of cokriging with limited field soil observations. Soil Science Society of America Journal 47, 175–184.

Venables, W.N., Ripley, B.D., 1994. Modern Applied Statistics with S-PLUS. Springer-Verlag, New York, USA.

Ventura, S.J., Irvin, B.J., 2000. Automated landform classification methods for soil–landscape studies. In: Wilson, J.P., Gallant, J.C. (Eds.), Terrain Analysis—Principles and Applications. Wiley, New York, pp. 267–294.

Verboom, W.H., Pate, J.S., 2003. Relationships between cluster root-bearing taxa and laterite across landscapes in southwest Western Australia: an approach using airborne radiometric and digital elevation models. Plant and Soil 248, 321–333.

Vold, A., Breland, T.A., Soreng, J.S., 1999. Multiresponse estimation of parameter values in models of soil carbon and nitrogen dynamics. Journal of Agriculture, Biology and Environmental Science 4, 290–309.

Voltz, M., Webster, R., 1990. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. Journal of Soil Science 41, 473–490.

Voltz, M., Lagacherie, P., Louchart, X., 1997. Predicting soil properties over a region using sample information from a mapped reference area. European Journal of Soil Science 48, 19–30.

Wackernagel, H., 1987. Multivariate Geostatistics. Springer, Berlin.

Walker, P.H., Hall, G.F., Protz, R., 1968. Relation between landform parameters and soil properties. Soil Science Society of America Proceedings 32, 101–104.

Walter, C., McBratney, A.B., Douaoui, A., Minasny, B., 2001. Spatial prediction of topsoil salinity in the Chelif Valley, Algeria, using local ordinary kriging with local variograms versus whole-area variogram. Australian Journal of Soil Research 39, 259–272.

Walvoort, D.J.J., de Gruijter, J.J., 2001. Compositional kriging: a spatial interpolation method for compositional data. Mathematical Geology 33, 951–966.

Webster, R., 1977. Canonical correlation in pedology: how useful? Journal of Soil Science 28, 196–221.

Webster, R., Burgess, T.M., 1980. Optimal interpolation and isarithmic mapping of soil properties, III. Changing drift and universal kriging. Journal of Soil Science 31, 505–524.

Webster, R., Burrough, P.A., 1974. Multiple discriminant analysis in soil survey. Journal of Soil Science 25, 120–134.

Webster, R., Oliver, M.A., 1990. Statistical Methods in Soil and Land Resource Survey. Oxford Univ. Press, Oxford.

Webster, R., Harrod, T.R., Staines, S.J., Hogan, D.V., 1979. Grid

sampling and computer mapping of the Ivybridge area, Devon. Technical Monograph no.12, Soil Survey of England and Wales, Harpenden, Herts.

Wen, R., Sinding-Larsen, R., 1997. Image filtering by factorial kriging-sensitivity analysis and application to GLORIA side-scan sonar images. Mathematical Geology 21, 433–468.

Wielemaker, W.G., de Bruin, S., Epema, G.F., Veldkamp, A., 2001. Significance and application of the multi-hierarchical landsystem in soil mapping. Catena 43, 15–34.

Wilson, J.P., Gallant, J.C., 2000. Secondary topographic attributes. In: Wilson, J.P., Gallant, J.C. (Eds.), Terrain Analysis—Principles and Applications. Wiley, New York, pp. 87–132.

Wood, J., 1996. The Geomorphological Characterisation of Digital Elevation Models. PhD thesis, University of Leicester, UK. Available at: http://www.soi.city.ac.uk/~jwo/phd.

Wood, E.F., Lin, D.S., Mancini, M., 1993. Inter-comparisons between active and passive microwave remote sensing, and hydrological modelling for soil moisture. Advances in Space Research 13, 5167–5176.

Wösten, J.H.M., Pachepsky, Y.A., Rawls, W.J., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. Journal of Hydrology 251, 123–150.

Wrigley, N., 1978. Probability surface mapping. In: Davis, J.C., Levi de Lopez, S. (Eds.), Mapas por computadora para el analisis de los recursos naturales; memorias de la reunion internacional. Univ. Nac. Auton. Mex., Inst. Geogr. Mexico City and Univ. Kans., Kans. Geol. Surv., Lawrence, Kansas, Mexico City, Mexico, pp. 39–49.

Yaalon, D.H., 1975. Conceptual models in pedogenesis: can soil-forming functions be solved? Geoderma 14, 189–205.

Zeverbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surface Processes and Landforms 12, 47–56.

Zheng, D., Hunt, E.R., Running, S.W., 1996. Comparison of available soil water capacity estimated from topography and soil series information. Landscape Ecology 11, 3–14.

Zhu, A.X., 1997. A similarity model for representing soil spatial information. Geoderma 77, 217–242.

Zhu, A.X., 2000. Mapping soil landscape as spatial continua: the neural network approach. Water Resources Research 36, 663–677.

Zhu, A.X., Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. Canadian Journal of Remote Sensing 20, 408–418.

Zhu, Z.L., Evans, D.L., 1994. US forest types and predicted percent forest cover from AVHRR data. Photogrammetric Engineering and Remote Sensing 60, 525–533.

Zhu, C., Yang, X., 1998. Study of remote sensing image texture analysis and classification using wavelet. International Journal of Remote Sensing 19, 3197–3203.

Zhu, A.X., Band, L.E., Dutton, B., Nimlos, T.J., 1996. Automated soil inference under fuzzy logic. Ecological Modelling 90, 123–145.

Zhu, A.X., Band, L.E., Vertessy, R., Dutton, B., 1997. Derivation of soil properties using a soil land inference model (SoLIM). Soil Science Society of America Journal 61, 523–533.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal 65, 1463–1472.

Zighed, A., 1985. Methodes et utils pour les processus d'interrogation non-arborescents. PhD thesis, Université de Lyon 1, Lyon, France.

Zighed, D.A., Rakotomalala, R., 2000. Graphes d'Induction-Apprentissage et Data Mining, Hermes.

Zighed, A., Rakotomalala, R., Rabsaeda, S., 1996. A discretization method of continuous attributes in induction graphs. Proceedings of the 30th European Meeting on Cybernetics and Systems Research, Vienna, pp. 997–1002.