Geoderma 150 (2009) 309–323

Contents lists available at ScienceDirect

Geoderma

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

Relationships in soil distribution as revealed by a global soil database

J.M. Gray [a,b,*], G.S. Humphreys [b,1], J.A. Deckers [c]

[a] NSW Department of Environment & Climate Change, PO Box 3720, Parramatta (Sydney), NSW 2124, Australia
[b] Department of Physical Geography, Macquarie University (Sydney), NSW 2109, Australia
[c] Catholic University of Leuven, Vital Decosterstraat 102, 3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 14 August 2008
Received in revised form 25 January 2009
Accepted 9 February 2009
Available online 21 March 2009

Keywords:
Quantitative soil survey
Soil prediction
Spatial modelling
Environmental correlation
Pedotransfer function
Soil relationships
Soil genesis

ABSTRACT

The ISRIC WISE Global soil database was used to develop broad quantitative relationships between the environmental factors of climate, parent material and topography and a range of important soil properties (pH, sum of bases, organic carbon, clay content and others). Three different analytical approaches were used in the analysis, involving (i) multiple linear regressions, (ii) fitted decision trees and (iii) categorical analysis with median values.

The strengths of the predictive relationships are generally only moderate at best, with R^2 values in the regression relationships typically in the range 0.2 to 0.4 and a wide spread in inter-quartile values in categorical median results. Testing with 100 random samples revealed only broadly moderate accuracy for the soil property predictions, with root mean square error down to 0.84 for pH predictions and down to 2.6 $cmol_c/kg$ for sum of bases. Despite their low strengths, the predictive relationships can provide useful first approximations of soil character under different environmental conditions and could be applied in broad quantitative soil modelling and mapping programs. They have the potential for widespread application as they should be universally applicable, are based on readily available data and do not require sophisticated quantitative modelling techniques.

The relationships revealed in the study can assist in our understanding of soil formation and soil distribution. Most relationships are in accord with accepted pedological thinking and support the state factor model of soil formation, but some anomalies are observed and deserve further examination. The results reveal the dominant influence of climate and parent material in controlling the distribution of many soil properties, with the influence of topography being less evident, at least at the global scale.

Crown Copyright © 2009 Published by Elsevier B.V. All rights reserved.

1. Introduction

A knowledge and understanding of soil and how it is distributed across the landscape is essential for the effective use, management and conservation of this most vital resource. As soil is an integral component of many physical and ecological processes, soil information is increasingly being required for input into ecological, hydrologic, climatic and other environmental models, particularly due to the ever-rising environmental concerns around the planet. Despite this need, much of the world has only very poor coverage of soil data, with many countries, particularly larger and/or poorer ones, only having data available at very broad scale if at all, and typically of low reliability (McBratney et al., 2003). For example, in Australia, only 50% of the Murray Darling Basin, the country's most important agricultural region, has been mapped at 1:250 000 and only 3% at 1:100 000 or finer; most is at a scale of 1:2 million (Bui and Moran, 2003).

1.1. Need for quantitative relationships

There is a clear need for precise quantitative relationships between soil and key environmental factors to facilitate soil data collection and soil modelling around the globe. Such relationships form the basis of digital soil mapping techniques, which is widely considered to be where the future of soil survey lies (Lagacherie and McBratney, 2007; Hartemink, 2006; Grunwald, 2006a). However, most published information on soil–environment relationships is of a qualitative form, and even these are often not clearly enunciated. For example, it is widely reported that parent material influences soil character, but there is little detail on what the precise influence is, for example how pH changes with increasing mafic character, particularly in quantitative terms. Few universally applicable quantitative relationships have been published and widely accepted to date. Most that have been published relate to specific regions rather than to continental or global scales. Heuvelink (2006) laments that most of our knowledge on soil forming processes is available only in a conceptual or descriptive form and that so far we have not succeeded in building a generic, quantitative, reproducible model that predicts the soil from its controlling factors in a satisfactory way.

* Corresponding author. NSW Department of Environment & Climate Change, PO Box 3720, Parramatta (Sydney), NSW 2124, Australia. Fax: +61 2 9895 7985.
E-mail address: jonathan.gray@environment.nsw.gov.au (J.M. Gray).
[1] Deceased.

0016-7061/$ – see front matter. Crown Copyright © 2009 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2009.02.012

1.2. The state factor model

Soil–environment relationships are generally based on the principles inherent in the now widely accepted state factor model of soil formation and the associated fundamental soil equation. This model, sometimes referred to as the "*clorpt*" model, maintains that the soil or soil property is a function of the factors of climate (*cl*), organism or biota (*o*), relief or topography (*r*), parent material (*p*) and time (*t*) (Jenny, 1941). The equation has been described as "a general statement implying that soils are natural bodies that are distributed in a predictable way in response to a systematic interaction of environmental factors" (Hudson, 1992). McBratney et al. (2003) expanded on this model, proposing a new seven-factor "*scorpan*" model for use in digital soil mapping applications. This has the additional factors of *s* (a soil attribute predictor), and *n* (geographic position predictor) together with the five traditional factors.

1.3. Development of quantitative relationships

The fundamental soil equation provides in theory for a quantitative prediction of soil classes or properties. However, due to the inability to derive effective quantitative indicators of each factor and the complexity of the resulting mathematical relationships, the equation has not been solved in its entirety to date. Partial solving of the equation is possible through the use of individual pedofunctions such as climofunctions or topofunctions where all factors bar the one of interest are held constant (Jenny, 1961). Pedofunctions have been attempted by numerous workers, particularly in the decades following Jenny's 1941 book (see Yaalon, 1975; Bokheim, 1980; Jenny, 1980). The frequent low number of sample points in many of these studies limits their statistical reliability. The majority of the functions were based on relatively localised locations and do not have universal applicability.

There appears to have been few standard pedofunctions published in recent decades. McBratney et al. (2003) note that it is difficult to find studies that explicitly consider the interactions between the different soil forming factors, ie multi-variate functions. They suggest that Webster (1977) came closest with his canonical correlation studies relating soil properties to topography, lithology and time. Rasmussen et al. (2005) established linear relationships for soil carbon and clay contents from an "energy index" based on climate and parent material factors, using the US State Soil Geographic (STATSGO) database.

The advent of digital soil mapping over the past two decades has led to the development of sophisticated multi-variate regression relationships and decision tree models relating soil properties to digital environmental data. These relationships use secondary ancillary data such as from geophysical and satellite image sources to represent the various state factors. Examples include McKenzie and Austin (1993), Skidmore et al. (1997), McKenzie and Ryan (1999), Ryan et al. (2000) and McBratney et al. (2000); others are given in McBratney et al. (2003). However, these relationships tend not to have universal applicability, being only relevant to the study area being surveyed. The sophisticated ancillary data they use may not be readily available for other locations. Thus, in most cases they cannot be applied more generally for soil distribution modelling applications.

Table 1
Topographic classes.

Topographic class	Slope gradient	Landform feature
Sloping sites (S) (sites of depletion)	>5%	Medium and high gradient hill, mountain, ridge, valley and escarpment
Near level lowland sites (L) (sites of accumulation)	<=5%	Plain, valley floor, depression and low-gradient footslope

Table 2
Parent material classes.

Parent material class	Silica range (%)	Average Silica (%)	Ca–Mg–Fe oxides Range (%)	Examples
Extremely siliceous (ES)	>85	90	<3	Quartz sands, quartzite, chert, jasper (generally >90% quartz)
Highly siliceous (HS)	68–85	76.5	3–7	Granite, rhyolite, adamellite, quartz sandstone, quarts siltstone, siliceous tuff (~30–60% quartz)
Intermediate (Int)	52–68	60	7–20	Granodiorite, dacite, trachyte, syenite, monzonite, diorite, andesite, greywacke, lithic sandstone, argillaceous materials (mudstone, shale, slate, phyllite), clay, loess (~0–30% quartz)
Mafic (Maf)	45–52	48.5	20–30	Gabbro, dolerite, basalt, mafic tuff (generally nil quartz)
Calcareous (Cal)	Not applicable	Not applicable	Not applicable	Limestone, dolomite, calcareous shale, calcareous sands (generally low quartz)

Approximate compositions derived from Best (1982) and Duff (1993).

1.4. Prediction of soil properties at two hypothetical sites

An illustration of the current lack of universal and readily applicable quantitative soil–environmental relationships is provided by attempting to predict the precise nature of soils likely to occur in two hypothetical sites. Site A is on moderately sloping terrain (slope 10%) with highly siliceous parent material (eg, granite) and wet climate (1800 mm pa rainfall) while Site B is on near level terrain with mafic parent material (eg, basalt) in a dry climate (400 mm pa rainfall). What soils would we expect at these two sites? There is currently very little published information to assist in this prediction process, even to provide broad qualitative predictions let alone detailed quantitative predictions. These two hypothetical sites are revisited later in this paper.

1.5. Study aims

There appears to be a need for pragmatic, easy-to-apply relationships for predicting soil character under different environmental conditions. Such relationships would assist in soil data collection by providing useful first approximations of soil properties occurring in different locations. They might encourage a more widespread application of quantitative predictive techniques by the soil survey community, perhaps in association with conventional mapping techniques. The relationships may be incorporated into quantitative models of other scientific disciplines, for example in ecological, climatic or hydrological studies. More generally, they may provide valuable insights into soil formation and soil distribution.

This study used the International Soil Reference and Information Centre (ISRIC) WISE Global database in an exploratory analysis to reveal broad quantitative soil–environment relationships. The potential for such "data mining" to develop important environmental relationships is well recognised. Henderson et al. (2005) and Bui et al. (2006) adopted a similar approach to model the distribution of various soil properties

Table 3
Climate classes.

Climate class	Rainfall/evapotranspiration (*R*/EV)	Annual rainfall (*R*) (mm pa)
A. Extremely dry	<0.125	<250
B. Very dry	0.125–0.25	250–500
C. Dry	0.25–0.5	500–750
D. Moderately dry	0.5–0.75	750–1000
E. Moderately wet	0.75–1.0	1000–1250
F. Wet	1.0–1.25	1250–1500
G. Very wet	1.25–1.5	1500–2000
H. Extremely wet	>1.5	>2000

Table 4
Multiple linear regression relationships for key soil properties.

Property	Climate index	Horizon	Regression equation[a]	N	R^2	F model	p value
pH	R/EV	A	$pH = 8.69 - 1.22(R/EV) - 0.025(silica\%) - 0.0072(slope\%)$	490	0.40	106.0	<0.0000
		B	$pH = 8.57 - 1.07(R/EV) - 0.025(silica\%) - 0.0077(slope\%)$	604	0.33	99.9	<0.0000
	R	A	$pH = 8.86 - 0.0008(R) - 0.030(silica\%) - 0.013(slope\%)$	491	0.35	89.3	<0.0000
		B	$pH = 8.59 - 0.0007(R) - 0.029(silica\%) - 0.009(slope\%)$	606	0.29	82.2	<0.0000
Sum of bases (ln $cmol_c$/kg)	R/EV	A	$\ln(SB) = 5.57 - 0.81(R/EV) - 0.051(silica\%)$	781	0.30	168.8	<0.0000
		B	$\ln(SB) = 6.26 - 1.22(R/EV) - 0.057(silica\%) - 0.013(slope\%)$	549	0.33	88.3	<0.0000
	R	A	$\ln(SB) = 5.73 - 0.0007(R) - 0.052(silica\%)$	783	0.31	176.5	<0.0000
		B	$\ln(SB) = 6.46 - 0.0009(R) - 0.061(silica\%) - 0.010(slope\%)$	550	0.34	93.3	<0.0000
Organic C (ln %)	R/EV	A	$\ln(OrgC) = 0.74 + 0.65(R/EV) - 0.014(silica\%) + 0.009(slope\%)$	485	0.21	42.1	<0.0000
		B	$\ln(OrgC) = -0.23 + 0.40(R/EV) - 0.012(silica\%) + 0.009(slope\%)$	591	0.12	26.5	<0.0000
	R	A	$\ln(OrgC) = 0.82 + 0.0003(R) - 0.011(silica\%) + 0.012(slope\%)$	486	0.12	22.9	<0.0000
		B	$\ln(OrgC) = 0.082 + 0.0002(R) - 0.011(silica\%) + 0.011(slope\%)$	592	0.07	15.7	<0.0000
Clay (%)	R/EV	A	$Clay = 76.17 + 4.00(R/EV) - 0.80(silica\%)$	856	0.29	174.6	<0.0000
		B	$Clay = 87.59 + 4.91(R/EV) - 0.86(silica\%)$	908	0.29	183.0	<0.0000
	R	A	$Clay = 74.11 + 0.0042(R) - 0.79(silica\%)$	857	0.30	187.2	<0.0000
		B	$Clay = 82.28 + 0.0064(R) - 0.83(silica\%)$	909	0.33	220.3	<0.0000

[a]Variables excluded where not significant (p value > 0.10).
See Appendices A and B for additional soil properties and statistics.

across the agricultural zone of Australia using a large Australian database, the Australian Soil Resource Information System (ASRIS, Johnson et al., 2003; McKenzie et al., 2005). To date the approach does not appear to have been used to develop soil–environment relationships at the global scale.

Specific aims of the study were to:

- develop broad universal quantitative relationships between environmental factors (climate, parent material and topography) and key soil properties, using three different analytical approaches.
- demonstrate the use of these relationships as pragmatic tools that facilitate a broad quantitative prediction of soil properties under different environmental conditions, for use in quantitative soil mapping and modelling exercises at a local to global scale.
- examine the soil–environment relationships revealed in relation to current accepted thinking on soil formation and distribution.

A semi-quantitative analysis of World Reference Base (WRB) soil type relationships with environmental factors was also undertaken using the ISRIC database. The results may aid in the prediction of WRB

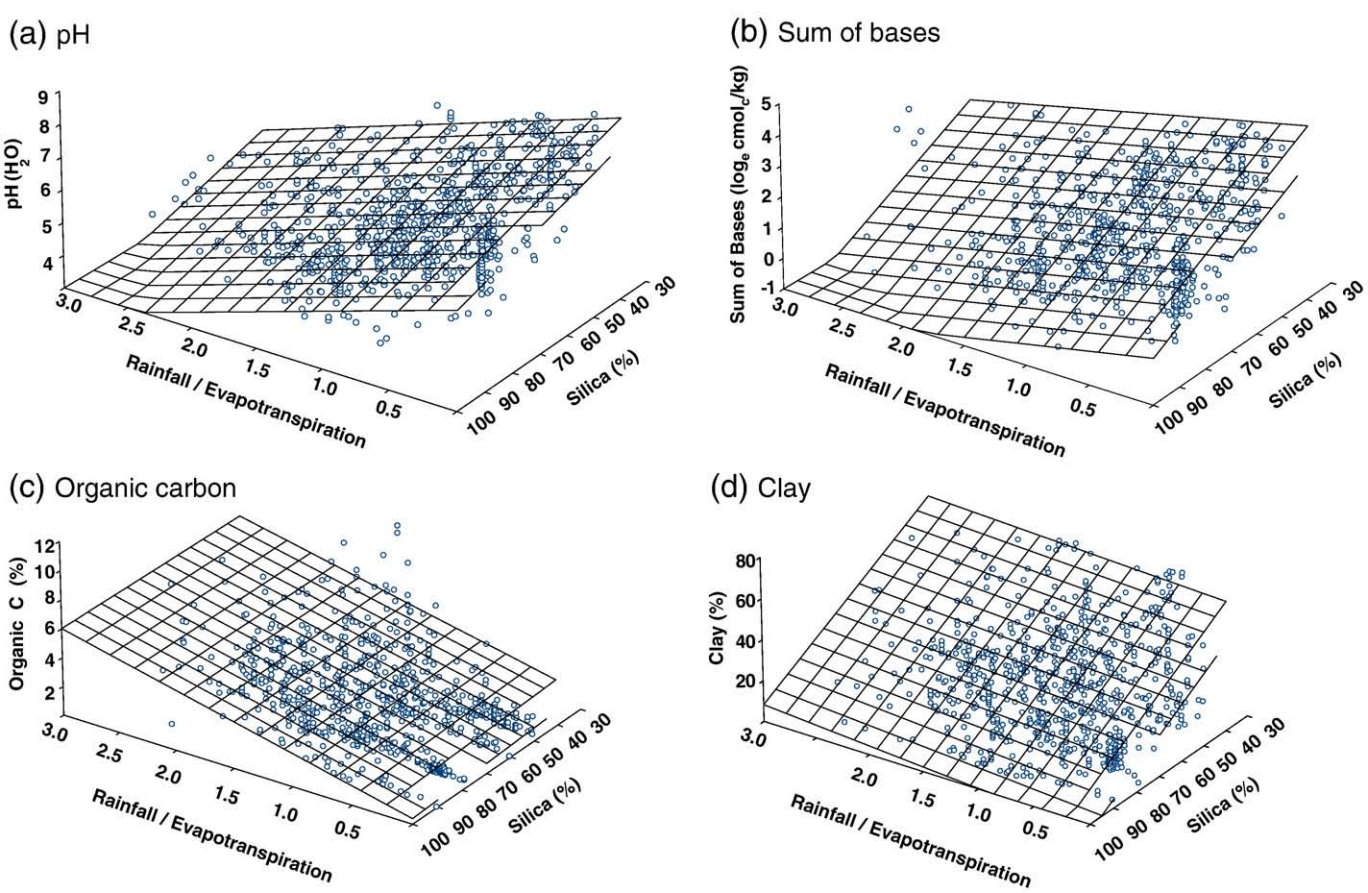

Fig. 1. Regression planes of key soil properties v R/EV and silica%.

soil types under different conditions and will be presented in an upcoming paper (Gray et al., in preparation). This study builds on work presented in Gray (2008), Gray et al. (2007), Gray and Humphreys (2004) and Gray and Murphy (2002).

2. Methods

2.1. The ISRIC WISE Global database

This dataset was examined to extract relationships that could be used to model soil behaviour at the global level. It contains 4382 geo-referenced soil profiles from 123 countries (Batjes, 2002). It was established by ISRIC using the framework of the World Inventory of Soil Emission Potential (WISE) project (Batjes and Bridges, 1994). The dataset includes, for most soil profiles, a comprehensive listing of site conditions, soil type (including FAO 1988 legend) and soil physical and chemical data. The dataset, in Microsoft Access 2000 format, was downloaded from the ISRIC website. Rainfall and potential evapotranspiration data were derived from a world climatic dataset (with 5 arc min grid) contained in FAO/IIASA (2002) and linked to soil profiles using their spatial coordinates.

2.2. Extraction of soil and environmental variables

Multiple queries were created to extract key soil properties and the associated environmental variables relating to topography, parent

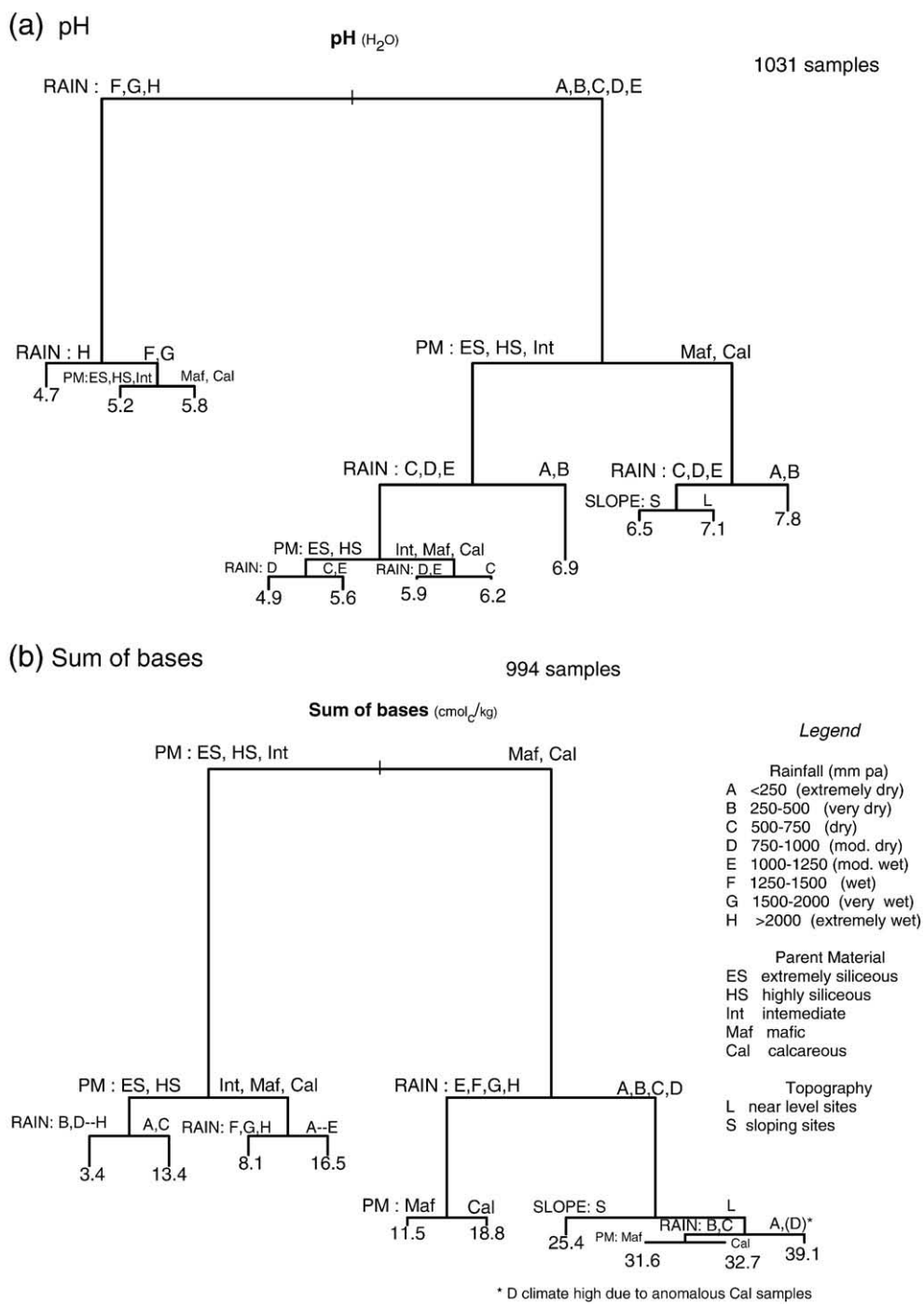

Fig. 2. Fitted decision trees for four key soil properties.

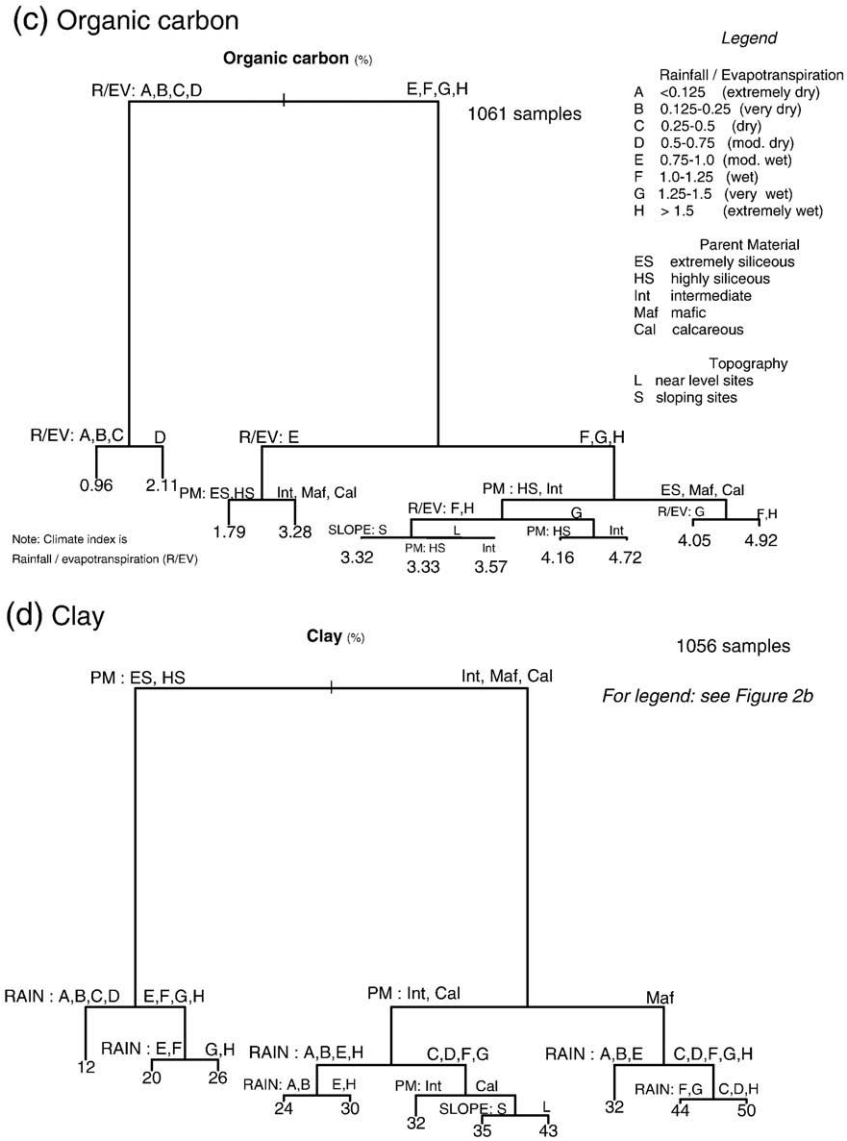

Fig. 2 (*continued*).

material and climate from the database. These variables were divided into classes as outlined below. The variables of age and biotic activity could not be included in the query process due to their lacking meaningful indicators.

Topography — the topographic indicator used was slope percent in conjunction with landform features, as provided for each soil profile in the database. Just two broad classes were established: (i) Sloping sites and (ii) Near level sites as shown in Table 1. These classes, with the 5% slope gradient cut off point, are intended to broadly reflect sites of depletion or accumulation respectively, where soil particles and chemical material is either lost or gained due to erosional and leaching processes. Crest and plateau sites less than 5% slope were excluded.

Parent material — this was categorised on the basis of broad chemical composition, considered the single most important feature of parent material in influencing soil properties. Of the five classes selected, four are based on broad silica (SiO_2) content and one on calcareous ($CaCO_3$) character, as shown in Table 2. Other classes such as ultramafic and organic materials were omitted due to low sample numbers in the database. Table 2 shows the declining base Ca–Mg–Fe oxides with increasing silica contents in parent materials. Classification into these five classes was based on the parent material name and description as given in the database. Assumptions had to be made on many materials, eg, unqualified "loess" and "clay" were considered to be intermediate and unqualified "sand" was considered to be extremely siliceous.

Climate — two straightforward climatic indices were used in the project: annual rainfall (R) and annual rainfall / annual potential evapotranspiration (R / EV). The former (R) has the advantage of being easily understood and accessible at the site or district level, however the latter (R / EV) may be expected to give a better indication of wetness of the environment, as it effectively includes a temperature variable. Eight broad climatic classes based on R and R / EV were established as shown in Table 3. These range from extremely dry to extremely wet climates and attempt to best represent global moisture regimes.

The combination of 2 topographic, 5 parent material and 8 climatic classes resulted in the creation of 80 different environmental regimes, which provided a basis for presenting categorical analysis results.

Only 1646 profiles within the database matched the final query requirements, comprised of 1108 in near level sites and 538 in sloping sites. The main reason for the relatively low return rate was incomplete site data for many profiles, such as insufficient topographic data, or inadequately defined parent material, for example "fluvial" or "aeolian", terms that give little indication of lithological composition. Some environmental regimes were represented only poorly or not at all, particularly those with extremely siliceous parent materials and many arid-sloping regimes.

A range of potentially important soil properties were extracted and/or derived for a single topsoil layer (uppermost A horizon) and subsoil layer (uppermost B horizon) from each profile. Key properties examined in detail included pH, sum of bases, organic carbon and clay content, these being considered to be fundamental indicators of the character and agricultural potential of a soil. Other properties briefly examined included cation exchange capacity (CEC), base saturation, exchangeable sodium percent (ESP), electrical conductivity (ECe) and sand and silt contents.

2.3. Statistical analysis

Data extracted from the database were further organised using Microsoft Excel software. Exploratory statistical analysis was then undertaken using S-Plus (2003) software. Although several approaches were trialled, only results from three are presented in this paper: (i) multi-variate linear regression involving the three environmental variables giving regression relationships for each soil property; (ii) fitted decision trees for each soil property and (iii) a categorical median approach giving median values for each soil property in 80 different environmental regimes. Standard statistical techniques were applied to test the performance of the relationships and the relative contributions of each variable. Graphical plots were prepared for all relationships. Predictions of soil properties derived using the three approaches were tested for accuracy against actual values from 100 random samples drawn from the ISRIC database prior to the development of the relationships.

3. Results

3.1. Soil property–environment relationships

The relationships of the various soil properties with the variables of climate, parent material and topography were examined using the three different analytical approaches. Results are presented in this section, together with key statistical parameters.

3.1.1. Multiple linear regression approach

Regression relationships between the various soil properties and the three environmental variables were established. Results for pH, sum of bases, organic carbon and clay content are presented in Table 4. Results for other properties including CEC, base saturation, ESP, ECe, sand and silt contents are presented in Appendix A. Additional statistics associated with the four key properties including standard errors, *t* values, and ANOVA *F* values are presented in Appendix B. Three-dimensional plots with regression planes were prepared for each key property, as presented in Fig. 1, with rainfall / evapotranspiration (*R* / EV) on the *x* axes and silica content on the *y* axes. These display key trends in soil behaviour with varying climate and parent material.

Coefficient of determination (R^2) values are generally in the low to moderate range for most properties. For pH they are moderate (~0.30–0.40), indicating a moderate correlation and that these regression relationships are at least moderately effective in accounting for the variation of this property in the database. R^2 values are low to moderate for sum of bases and clay (~0.30–0.35) and only low for organic carbon (~0.10–0.25), indicating that these relationships are of only low to moderate effectiveness in accounting for the variation of these properties. Note that for sum of bases and organic carbon, a natural log (ln) transformer is used in the regression to address the positive skew in the data. ANOVA *F* values of the regression models are generally high, in the order of 100 or more, except for organic carbon which are only moderate, and the *p* values are all extremely low (<0.0000) confirming that significant relationships do exist between the properties and the environmental variables.

The similarity in R^2 values between relationships using the two different climatic indices (*R* / EV and *R*) suggests that the two indices have similar effectiveness, however the *R* index appears less effective in the organic carbon relationships.

3.1.2. Decision tree approach

Decision trees were fitted to the data for each of the four key soil properties. For the purposes of simplicity and ease of presentation, the sensitivity of the trees was set so as to finish with approximately 10–12 leaves at the base. The models, as presented in Fig. 2a–d, use rainfall classes as the climatic index for pH, sum of bases and clay content, but rainfall / evapotranspiration (*R* / EV) classes for the organic carbon model as this gave significantly more reliable results. By way of example, it can be seen that hypothetical site A (climate G, parent material HS and topography S) may be expected to have a pH of 5.2 and sum of bases 3.4 $cmol_c$/kg.

The models all reveal a generally consistent change in soil properties as the environmental factors vary, eg, increasing pH with increasing dryness and increasing mafic character. However, some minor anomalies occur, which may be due to low sample numbers or other factors as discussed in Section 4.1. It is apparent that the slope class does not have a strong influence in the models. If a greater number of leaves had been allowed for in the sensitivity settings there would be less grouping of results from different environmental regimes and thus potentially more reliable predictive ability.

3.1.3. Categorical median value approach

This approach involved division of the 1646 profiles into different environmental categories based on the two topographic classes, five parent material classes and the eight climatic classes. This gave rise to 80 different topographic-parent material-climatic regimes. For each regime, the median values were determined for the four key soil properties, as presented in Appendix C.1–C.4 along with sample numbers (using rainfall climate index). The median values may be presented in box plots as shown for the four key properties in the five near level regimes in Fig. 3 (using *R* / EV climate index). These plots also reveal the broad trends in the behaviour of these properties with changing conditions.

The inter-quartile ranges are frequently high, particularly for organic carbon, reflecting the wide variability in results within each regime. Low sample numbers in many regimes reduce the effectiveness of the results. Mean values in each regime, plus associated statistics (standard deviation, standard error, etc) were also derived but were found to be slightly less effective than the median values, presumably due to the effect of occasional significant outliers.

3.2. Soil property behaviour

Broad quantitative trends in the behaviour of the key soil properties with varying environmental conditions were revealed by the analysis, with all three analytical approaches giving similar results.

By way of example, the regression relationships (presented in Table 4) reveal the following trends in behaviour for topsoil pH:

- a pronounced decrease in pH as climate becomes wetter, with an average 1.2 unit decrease per *R* / EV unit increase or 0.8 unit decrease per 1000 mm increase in *R*.
- a moderate decrease in pH as parent material becomes more siliceous, with an average 0.3 unit decrease with each 10% increase in silica content.
- a slight decrease in pH with increasing slope, with an average 0.1 unit decrease for each 10% increase in slope.

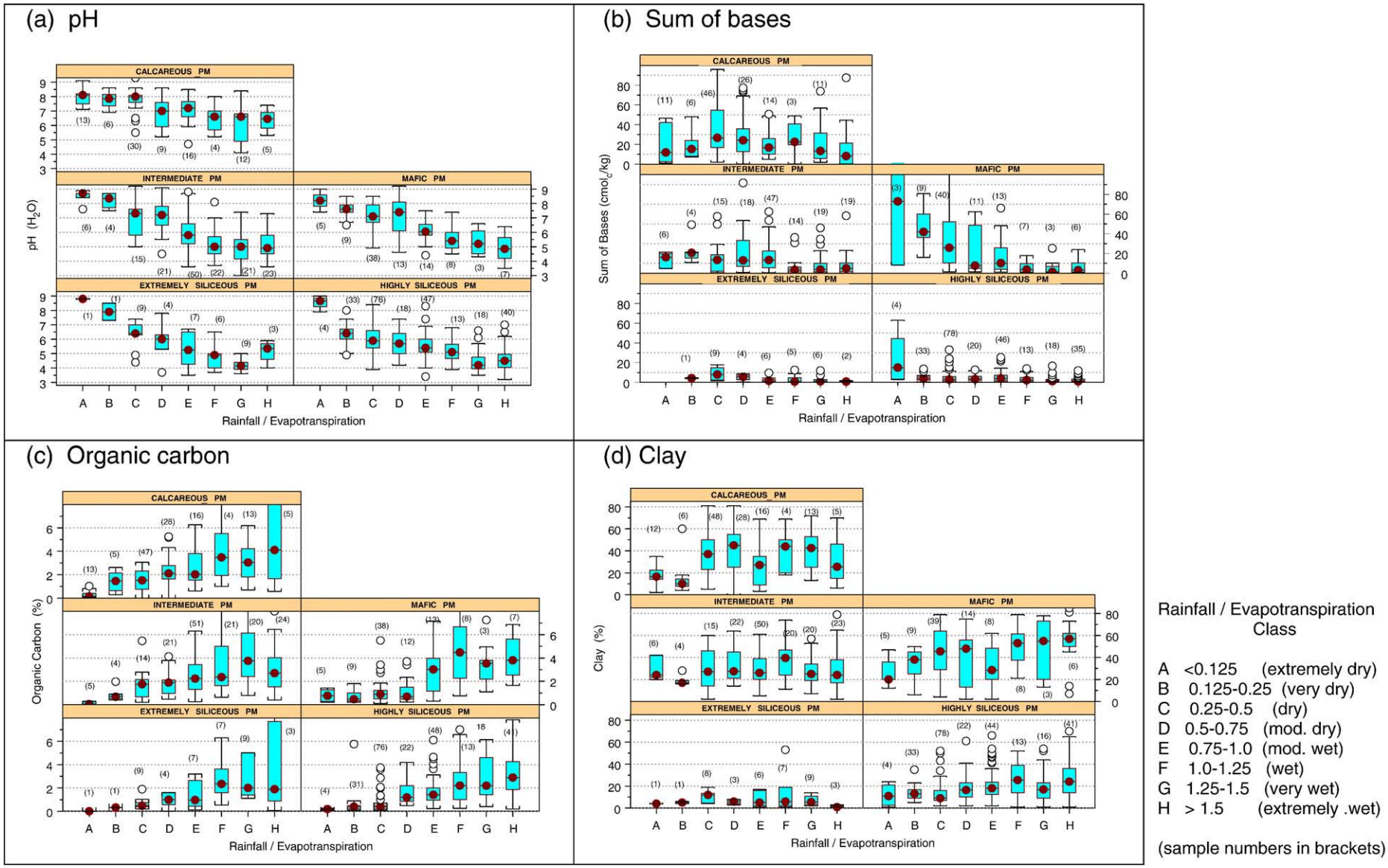

Fig. 3. Box-plots of key topsoil properties (near level regimes).

- the rate of decrease of pH as climate becomes wetter is generally greater with increasing mafic character (decreasing silica %) of parent material (although the ES class does not follow this trend). The trend is even more pronounced for sum of bases.

Examination of other regression relationships, from Table 4 and Appendix A, allows similar quantitative behavioural trends to be determined for other soil properties. The ANOVA results (with R / EV, as presented in Appendices A and B) provide indications of the relative importance of each factor for the different properties, for example:

- For pH, the F and t values indicate that climate exerts the greatest influence (t values −14.5), then parent material (−6.5) with slope gradient exerting the least influence (−2.0).
- For sum of bases, F and t values indicate that parent material has the greatest influence on this property (t values −14.6) closely followed by climate (−10.7) with the influence of slope being insignificant.
- For organic carbon, F and t values indicate that climate has the strongest influence on this property (t value 9.3) with parent material then slope having lesser influence (t values −4.3 and 3.1 respectively).
- For clay content, F and t values indicate the dominant influence of parent material (t value −18.4), the lesser influence of climate (t value 5.3) and the insignificant influence of slope %.

3.3. Prediction of soil properties

Three different approaches to the analysis of the data have been presented, representing different avenues for the prediction of soil properties. For the regression approach, the R^2 values are generally only moderate at best, suggesting that it is only of moderate reliability as a universal predictive tool. The fitted decision tree approach (as presented) resulted in considerable grouping and loss of distinction between different regimes, with several minor anomalies apparent. For the categorical median value approach, relatively large inter-quartile ranges again suggest only moderate reliability. It is apparent that all three approaches could generally only be relied upon for first approximations of likely soil properties under different environmental conditions.

The use of these approaches as predictive tools is demonstrated by attempting predictions for the four key soil properties in the two hypothetical sites A and B referred to in the Introduction, as shown in Table 5. Predictions from each approach vary slightly but are generally broadly equivalent.

3.4. Testing

The accuracy of predictions made from the three analytical approaches was tested. The process involved the determination of root mean square error (RMSE) and mean error (ME) on predictions made on a set of 100 random independent samples, drawn from the database prior to analysis. The difference in predictions made from the three analytical approaches and two climate indices was also tested.

Table 5
Comparison of predictions from the three approaches for hypothetical sites A and B.

Site	Soil property (upper A horizon)	Predictive approach[a]		
		Multi-variate regression	Decision tree	Categorical median
Site A (sloping, highly siliceous parent material, very wet)	pH	4.9	5.2	4.8
	Sum of bases[b]	1.8	3.4	1.5
	Organic C %	1.94	2.94	2.88
	Clay %	22	26	30
Site B (level, mafic parent material, very dry)	pH	7.1	7.8	7.5
	Sum of bases[b]	19.8	31.6	38.4
	Organic C %	1.43	0.80	0.44
	Clay %	38	32	37

[a] Using rainfall climate index.
[b] $cmol_c$/kg.

Table 6
Accuracy of predictions for 100 random samples (topsoils).

Approach	Climate index	pH		Sum of bases[a]		Organic C (%)		Clay (%)	
		RMSE	ME	RMSE	ME	RMSE	ME	RMSE	ME
1. Multiple linear regression	R / EV	0.91	−0.33	2.80	0.86	1.45	0.26	14.7	−3.6
	R	0.91	−0.30	2.80	0.96	1.48	0.28	14.4	−3.4
2. Decision tree	R / EV	0.92	−0.09	3.10	0.65	1.50	−0.25	14.8	−2.8
	R	0.94	−0.12	3.40	0.54	1.75	−0.47	14.7	−2.9
3. Categorical median	R / EV	0.88	−0.12	2.70	0.82	1.57	0.16	16.9	−2.6
	R	0.84	−0.13	2.60	0.95	1.79	0.01	15.0	−1.5

[a] $cmol_c$/kg.

The predicted values for pH, sum of bases, organic carbon and clay content derived from each of the three analytical approaches, for both climate indices, were compared against the actual property values for each of the 100 random samples. The predictions used the climate, parent material and topographical data associated with each sample. Results are presented in Table 6.

The results generally reveal only moderate at best predictive performance of the different approaches. The lowest RMSE recorded for pH was 0.84 pH units, for sum of bases it was 2.6 $cmol_c$/kg, for organic carbon, 1.45% and for clay content, 14.4%. Note that the RMSE for sum of bases was initially calculated in $\log_e$ units and then converted back into normal units. The mean errors suggest that the predictions tended to slightly overestimate values relative to actual sample values for pH and clay contents and slightly underestimate for sum of bases, with no clear trend for organic carbon. In all tests, the five most extreme outliers of the 100 samples were removed, however results are still weakened by other outliers.

All three analytical approaches gave broadly similar levels of accuracy, although the categorical median approach generally gave the lowest RMSE values for pH and sum of bases, while the multiple linear regression approach gave the lowest values for organic carbon and clay content. The actual predictions made from each approach and climate index were also generally very similar. The RMSE values derived from comparisons of these predictions are generally low, with an overall mean value of 0.40 for pH, 1.85 $cmol_c$/kg for sum of bases, 0.87% for organic carbon and 5.2% for clay content. There is a broad tendency for the categorical approaches to give a broader range of predicted values than the regression and decision tree approaches. The latter approaches appear to be associated with a smoothing effect and removal of extremes.

There is generally close agreement between predictions derived from the use of either annual rainfall (R) or rainfall / evapotranspiration (R / EV) as the climatic variable. The two indices have similar overall accuracy levels in the prediction of the test samples as shown by RMSE values in Table 6, although one may be more effective than the other for individual properties and approaches.

4. Discussion

4.1. Variability in the relationships

The regression relationships appear to be generally only of low to moderate effectiveness in accounting for the variation of the soil properties within the ISRIC database. Coefficient of determination (R^2) values typically falls in the range 0.2–0.4. Of the four key properties, the strongest relationships were associated with pH, followed by sum of bases and clay content, then organic carbon with the weakest relationships. The categorical median analysis also revealed only moderate strengths of relationships. The test of accuracy of predictions of the four properties with 100 random samples revealed generally moderate to high RMSE values, indicating only low to moderate predictive reliability.

The relatively low to moderate effectiveness of the three predictive approaches reflects the inherent complexity of the soil property–environmental factor relationships. A high variability in soil properties is demonstrated even within samples in supposedly similar

environmental conditions, at least in terms of topography, parent material and climate. A number of factors may contribute to the lack of strong relationships revealed in the analysis. These include:

- The database does not adequately represent all environmental regimes; it contains an abundance of observations in certain regimes but a paucity of observations in others. Although the original database contained over 4300 profiles, the number of usable profiles with all necessary quantitative data was far less than this. Low sample numbers decrease the reliability of the statistical analyses.
- Variation in other environmental factors apart from climate, parent material and topography is undoubtedly a major cause of the spread in observed soil property values and resulting weak relationships. The factors of age (of soils and/or landscape) and biota (including vegetation, micro-fauna and land management issues) are key elements in controlling soil behaviour but neither of these was directly incorporated into the analysis.
- The simplification of climate into annual rainfall or annual rainfall / annual evapotranspiration does not adequately address the role of varying seasonal conditions (eg, distinctive wet or dry winters) in controlling overall water balance and many soil formation processes. The important influence of micro-climate due to aspect and other topographic features is likewise not considered.
- The classification of parent material into classes, with assumed specific silica contents, is a major simplification. For example, shales may regularly vary from 50–75% silica and accordingly give rise to soils with a range of properties, therefore ascribing a value of 60% silica to all these materials will often be inaccurate. The actual parent material may vary considerably from that suggested from its name in the database, for example, due to being based on a broad geological map rather than actual field observation or to the influence of externally derived materials such as aeolian dust.
- The primary use of slope percent, together with some consideration of landform element, may not adequately represent site behaviour in terms of being sites of material depletion or accumulation. For example, sites with 4% slope may in some cases be subject to accumulation but in others to depletion. Hillcrest and plateau elements were effectively omitted from the analyses.
- The polygenetic nature of many soils must be considered. Where soils are old enough to have been influenced by previous climates, they may carry the imprint of those earlier conditions, and not entirely reflect current climatic conditions.
- A portion of the soil variability giving rise to the weak relationships may be attributable to "intrinsic" factors rather than traditionally accepted "extrinsic" factors of soil formation, as suggested by Phillips (1993a,b, 2001) in his non-linear dynamic system model of soil formation.

4.2. Pragmatic soil prediction tools

The soil–environment relationships derived from the database, although not strong, may serve as general predictive tools to provide useful first approximations of soil properties under different environmental conditions. These may have application in broad quantitative predictive soil modelling and mapping exercises ranging from single site to global scale studies where no other more reliable data is available. They could be applied in a purely manual process to provide simple soil property class maps or in a fully automated process to produce continuous soil property maps on a raster grid, depicting a continuous variation in soil properties across a landscape.

The potential of the predictive approaches to derive first approximations of soil conditions over specific sites is illustrated with the two hypothetical sites A and B. Table 7 gives a detailed listing of soil properties expected over both sites, using results from the categorical median approach. It can be seen that, broadly, the soils from Site A would be expected to be highly acid, low fertility soils, characterised by sandy clay subsoils while those from Site B would be expected to be neutral to alkaline, high fertility, uniformly clayey soils. The table also includes a listing of World Reference Base (WRB) soil types predicted to occur at these sites. This is based on a semi-quantitative analysis of WRB soil type–environment relationships revealed by the ISRIC database, with full results to be presented in an upcoming paper (Gray et al., in preparation).

The relatively straightforward nature of the predictive approaches means that they have the potential for widespread application. They are based on primary environmental data; climate (annual rainfall or rainfall / evapotranspiration), parent material (assumed silica %) and topography (slope % and/or landscape position) that should be widely available around the world. They do not use the more sophisticated but less widely available environmental indices such as derived from geophysical techniques or satellite imagery (see McBratney et al., 2000, 2003). There is no requirement for sophisticated computer technologies or quantitative modelling skills. They can therefore be readily and cheaply applied by all users, including those without access to more sophisticated datasets and technologies.

Levels of certainty associated with predictions could be applied. Results for each predicted soil property, could include the associated residual standard error, where regression relationships are used (from Appendices A and B) or the associated root mean square error values derived from the 100 sample test (from Table 6), eg, pH = 5.2 (associated RMSE = 0.84). Alternatively, results could be presented in terms of particular confidence intervals displayed by the ISRIC database (not presented here), eg, ISRIC derived relationships suggest a 95% probability that the sum of bases in Site A falls between 1.0 and 3.5.

As the relationships have been based on a worldwide dataset, rather than from a particular location or region, they have a claim to be globally applicable. They should be most suitable for the production of broad scale maps at the global or continental scale. Their reliability would decrease with finer scales of map production, but there is no inherent reason why the predictive relationships would not apply at finer levels. They should still give useful first approximations of soil conditions at most regional and even local map scales. Scull et al. (2005), Henderson et al. (2005) and Bui et al. (2006) demonstrated how relationships derived from large soil databases, using data mining techniques, can be used to produce digital predictive soil maps over regional and continental scales.

The results suggest that the two climate indices used (annual rainfall (R) and annual rainfall / annual potential evapotranspiration (R / EV))

Table 7
Predicted soil properties for hypothetical sites A and B.

	Site A		Site B	
	Wet climate (1800 mm pa rain, class G); highly siliceous parent material (granite); sloping site (10% slope)		Dry climate (400 mm pa rain, class B); mafic parent material (basalt), near level site (2% slope)	
Soil property[a]	A horizon (uppermost)	B horizon (uppermost)	A horizon (uppermost)	B horizon (uppermost)
pH	4.8 (28)[b]	4.7 (25)	7.5 (14)	8.0 (16)
CEC ($cmol_c$/kg)	12.6 (28)	7.4 (24)	30.7 (15)	31.0 (17)
Base saturation (%)	12.0 (23)	6.5 (20)	100 (13)	100 (17)
Sum of bases ($cmol_c$/kg)	1.5 (28)	0.4 (20)	36.4 (15)	41.9 (17)
Organic C (%)	2.88 (27)	0.55 (24)	0.44 (15)	0.40 (17)
P available (mgP_2O_5/kg)	2.4 (5)	1.40 (3)	13.7 (7)	4.6 (9)
ECe (DS/m)	0.07 (5)	0.02 (7)	0.20 (7)	0.20 (8)
ESP (%)	0.6 (5)	3.1 (5)	2.1 (15)	4.2 (17)
Clay (%	30 (28)	36 (25)	37 (15)	49 (17)
Silt (%)	18 (28)	16 (25)	20 (15)	21 (17)
Sand (%)	50 (28)	47 (25)	46 (15)	23 (17)
WRB soil type[c]	Cambisols, Ferralsols, Acrisols or possibly Alisols		Luvisols, Vertisols, possibly Chernozems or Calcisols	
WRB qualifier[c]	Umbric, haplic or ferralic		Eutric, chromic or rendzic	

[a] Soil properties derived from median of values for each environmental regime using annual rainfall as climate index (partly presented in Appendix Ca–d).
[b] Numbers in brackets indicate number of samples with lab data in the environmental regime.
[c] World Reference Base soil type and qualifier derived from additional analysis of ISRIC database, to be presented in Gray et al. (in preparation).

have almost equal effectiveness in the predictive relationships, although one may be more effective than the other for specific predictive approaches and properties. For example, when using the regression approach, coefficient of determination values (Table 4) suggests that *R* may be more effective for sum of bases and clay content while *R* / EV may be more effective for pH and organic carbon. The *R* index has important advantages in terms of its inherent simplicity, ease of understanding and ready availability at the local level, which may compensate for a small loss in reliability in many applications. The use of more advanced indices such as the Thornthwaite moisture index or others using monthly or seasonal data may enhance the effectiveness of the relationships.

Any of the three predictive approaches presented could be applied, the choice depending on the nature of the program being undertaken. The categorical predictive approaches for providing median soil properties may be the most readily applied for non-sophisticated programs, as no mathematical processing is required. The regression approach may have the greater potential application in more sophisticated programs as it can accommodate specific continuous data for each variable.

Despite the clear utility and importance of quantitative approaches to soil data collection, they have not yet achieved systematic and widespread application in soil survey around the world. The qualitative application of the state factor model through conventional mapping techniques continues to be the guiding paradigm in national soil survey programs (Grunwald, 2006a). Even with all the technical advances in recent years there appears to have been little movement away from conventional mapping approaches by government agencies in the USA (Scull et al., 2003) and in Australia, at least at the state government level. The inherent and perceived complexities of the quantitative approaches, and the need for sophisticated technologies may be factors that discourage the more widespread application of these approaches in many organisations. The simple and pragmatic tools developed in this study may overcome some of these concerns and thereby facilitate an increased use of these approaches.

The quantitative relationships developed in this study may also be important for inclusion into other environmental and earth science quantitative models. This application is becoming increasingly important as the crucial role of soils in many ecological, climatic, hydrologic and other natural processes is being recognised. They can help us to identify the role of soil in these processes. Grunwald (2006b) suggests that quantitative soil relationships can ultimately contribute to unifying soil science with these other natural sciences.

4.3. Understanding soil formation and distribution

The relationships and associated statistical data derived from the database analysis can assist in our theoretical understanding of soil formation and distribution. They shed light on the precise role and relative influence that the environmental factors of climate, parent material and topography have in controlling various key properties in soils (see section 3.2). They can be examined in relation to currently accepted pedological thinking. Quantitative information has been yielded, rather than the qualitative information that has dominated the literature to date.

The behaviour revealed for each environmental variable for each property examined can generally be explained by first principles of soil science. For example, the pH and base cation behaviour are due to the increased leaching of bases with increasingly wet conditions, the greater supply of bases associated with mafic parent materials and the transferral of bases from higher to lower points in the landscape. The behaviours are generally in accord with standard thinking on the role of each of the environmental factors. However, several anomalies and unexpected findings are observed, including:

- The surprisingly weak influence of topography in controlling pH, base cation behaviour and soil texture (ie, clay, sand and silt contents) — standard thinking would suggest a much greater influence of this factor, with clays and bases being leached and transported from higher to lower points in the landscape (eg, Dan et al., 1968; Birkeland 1999; Murphy 2006). This may be at least partly a reflection of the broad global scale of the study and the low density of sites, as variations due to topography may be more evident at a fine scale, with several sites on single catenas.
- The negligible change in CEC with climate (see Appendix A) — a general decrease with wetter climate would have been expected as high activity clays such as smectites transform to low activity clays such as kaolin or sesqui-oxides like gibbsite (Tardy et al., 1973; Birkeland 1999).
- The relationship between base saturation and silica content (see Appendix A) is more significant than might be expected, as the degree of leaching, in which base saturation is thought to reflect, would be expected to be independent of parent material type — a possible explanation is that more mafic parent materials are more effective in replenishing the soil with additional bases following their depletion by leaching.
- The weak influence of parent material on organic carbon — one might have expected higher values associated with the more fertile soils derived from mafic materials.
- The lack of a significant relationship of available phosphorous with the environmental variables (R^2 values were all less than 0.05) — a moderate relationship with parent material would be expected, with increasing values with increasing mafic content as suggested by geochemical compositions of parent materials. Phosphorous might also be expected to decline with increasing wetness as for base elements.

The results support the state factor model of soil formation, demonstrating how soil distribution is controlled by the combined influence of several environmental factors. They particularly confirm the dominant roles played by climate and parent material in determining the distribution of many soil properties. The influence of topography is also apparent, but appears to exert the least influence of the three factors on the properties examined in this global scale study. The unexpected findings may be a reflection of weaknesses in the databases and analytical processes of the study, or they may be a true reflection of actual pedogenic processes. The findings parallel those of Bui et al. (2006), in their establishment of quantitative relationships derived from ASRIS for the digital mapping of Australia's agricultural zone. They found climate and lithology to be the dominant variables at the broad continental scale and that terrain only became significant at much finer scales.

The results suggest that present day climate has a clear influence on soils properties, even though the soils may have been subject to different climates in the past. This conclusion was also drawn by Bui et al. (2006) who noted that its marked imprint on the old and polygenetic Australian soils challenged the view that present climate has little relevance to the character of these soils (eg, Taylor, 1983). Recent data from Wilkinson and Humphreys (2005) reveal relatively high rates of soil production, meaning that many Australian soils may be younger than previously thought, and thus more markedly subject to the influences of current climatic conditions. This may also be true at the global scale.

The ANOVA results (Appendices A and B) show that the influence of the various factors varies with different properties. For example, pH and organic carbon are most influenced by climate, while clay content is most influenced by parent material and sum of bases almost equally by both those factors. The influence of topography may be more marked in other soil properties not examined in this study; for example, soil thickness would almost certainly be strongly influenced by topographic position.

Although there are many potential sources of variation in the soil data, as discussed in the section 4.1, the observed weak relationships support the notion that soil variability may be attributable to “intrinsic” factors rather than traditionally accepted “extrinsic” factors

of soil formation, as postulated by Phillips (1993a,b, 2001) in his non-linear dynamic system model of soil formation. Intrinsic factors such as deterministic chaos, dynamical instability and divergent self-organisation may be crucial processes. The model recognises that considerable soil variability may occur in apparently homogeneous landscape units and independent of any observable variations in environmental controls.

A better understanding and knowledge of how the state factors influence soil distribution will improve our ability to model and map soils, particularly by conventional methods. Our understanding of the factors and processes contributing to soil formation and distribution has to date been largely based on qualitative relationships and descriptive concepts (eg, Heuvelink, 2005). Our knowledge of soil formation has been suggested to be "far behind reality" (Dobos, 2006). The important role that quantitative soil studies can play in quantifying these relationships and improving our understanding and knowledge of them is clear. Continuing analysis of soil databases as carried out in this study would undoubtedly yield further fine detail and subtle relationships in soil formation and distribution.

5. Conclusion

This study has used the ISRIC WISE Global database to develop broad quantitative relationships between the environmental factors of climate, parent material and topography and a range of key soil properties. The predictive relationships may assist in the collection and modelling of soil data as required to meet the growing demands for this information in many earth, environmental and other scientific modelling programs. More specifically, the relatively simple and pragmatic relationships may:

- provide useful first approximations of soil character under different environmental conditions that could be applied in broad quantitative soil modelling and mapping programs.
- have the potential for widespread application as they should be universally applicable, are based on readily available data and do not require sophisticated quantitative modelling techniques. They may thus facilitate an increased application of digital techniques in addition to conventional techniques in soil survey organisations and in the broader soil science community.
- assist in our understanding of soil formation and soil distribution. Most relationships revealed are in accord with accepted pedological thinking and support the state factor model of soil formation, but some anomalies are observed and deserve further examination.

The project has demonstrated the potential value of large world soil databases for the understanding, modelling and mapping of soil distribution around the globe. Further analyses of the ISRIC and other large soil databases with more sophisticated data mining techniques would undoubtedly help to refine the predictive tools and shed further light on factors influencing soil distribution.

Acknowledgements

Dr Paul Hesse, Macquarie University, assisted in the completion of this research project following the sad passing of the second author. Prof Alex McBratney and Dr Tom Bishop, University of Sydney, provided important technical advice and assistance.

Appendix A

Multiple linear regression relationships for other soil properties.

Property	Climate index	Horizon	Regression equation[a]	N	R^2	F stat.	Residual SE	p value	t values of variables: Climate	Silica	Slope
Cation exchange capacity (ln $cmol_c$/kg)	R/EV	A	$\ln(CEC) = 4.71 - 0.034(silica\%) - 0.011(slope\%)$	486	0.23	71.7	0.8	<0.0000	–	−11.3	3.9
		B	$\ln(CEC) = 4.81 - 0.19(R/EV) - 0.36(silica\%) - 0.009(slope\%)$	603	0.22	55.5	0.9	<0.0000	−3.0	−12.0	3.0
	R	A	$\ln(CEC) = 4.71 - 0.034(silica\%) - 0.011(slope\%)$	486	0.23	71.7	0.8	<0.0000	–	−11.3	3.9
		B	$\ln(CEC) = 5.04 - 0.0003(R) - 0.037(silica\%) - 0.011(slope\%)$	604	0.26	69.6	0.8	<0.0000	−6.3	−12.7	3.7
Base saturation (%)	R/EV	A	$BS = 143.34 - 28.75(R/EV) - 0.89(silica\%) - 0.30(slope\%)$	429	0.35	75.0	29.2	<0.0000	−10.7	−7.4	−2.7
		B	$BS = 136.88 - 32.03(R/EV) - 0.79(silica\%) - 0.39(slope\%)$	544	0.35	98.4	30.0	<0.0000	−13.8	−7.1	−3.4
	R	A	$BS = 147.06 - 0.019(R) - 1.01(silica\%) - 0.40(slope\%)$	428	0.31	65.0	28.9	<0.0000	−9.5	−8.3	−3.5
		B	$BS = 138.66 - 0.022(R) - 0.89(silica\%) - 0.39(slope\%)$	544	0.34	92.2	30.3	<0.0000	−13.1	−7.9	−3.4
Exchangeable sodium percent (ln %)	R/EV	A	$\ln(ESP) = 2.66 - 6.88(R/EV)$[b]	165	0.28	63.3	1.07	<0.0000	−8.0	–	–
		B	$\ln(ESP) = 2.77 - 6.52(R/EV)$[b]	175	0.25	56.9	1.07	<0.0000	−7.5	–	–
	R	A	$\ln(ESP) = 2.03 - 0.0039(R) + 0.010(silica\%)$[c]	193	0.32	43.9	0.97	<0.0000	−8.8	1.7	–
		B	$\ln(ESP) = 2.59 - 0.0034(R)$[c]	261	0.22	70.0	1.10	<0.0000	−8.4	–	–
Electrical conductivity (extract) (ln DS/m)	R/EV	A	$\ln(ECe) = 1.55 - 12.21(R/EV)$[d]	62	0.36	33.0	1.45	<0.0000	−5.7	–	–
		B	$\ln(ECe) = 1.89 - 13.39(R/EV)$[d]	61	0.45	48.7	1.40	<0.0000	−7.0	–	–
	R	A	$\ln(ECe) = 2.68 - 0.012(R)$[e]	47	0.51	46.0	1.49	<0.0000	−6.8	–	–
		B	$\ln(ECe) = 2.43 - 0.011(R)$[e]	39	0.49	35.4	1.42	<0.0000	−5.9	–	–
Sand (%)	R/EV	A	$Sand = -28.3 - 11.49(R/EV) + 1.31(silica\%)$	862	0.39	271.4	21.6	<0.0000	−9.2	22.0	–
		B	$Sand = -35.0 - 11.53(R/EV) + 1.31(silica\%)$	910	0.41	314.8	20.9	<0.0000	−10.1	22.9	–
	R	A	$Sand = -29.63 - 0.0064(R) + 1.28\ (silica\%)$	863	0.36	240.4	22.2	<0.0000	−6.4	21.0	–
		B	$Sand = -35.08 - 0.0075(R) + 1.29\ (silica\%)$	911	0.39	293.4	21.2	<0.0000	−8.5	22.1	–
Silt (%)	R/EV	A	$Silt = 51.53 + 7.80(R/EV) - 0.51(silica\%)$	860	0.18	95.9	15.9	<0.0000	8.5	−11.5	–
		B	$Silt = 47.28 + 6.67(R/EV) - 0.46(silica\%)$	911	0.18	96.5	15.0	<0.0000	8.1	−11.2	–
	R	A	$Silt = 54.93 + 0.0024(R) - 0.49\ (silica\%)$	861	0.12	61.2	16.6	<0.0000	3.1	−10.6	–
		B	$Silt = 52.8 + 0.0011(R) - 0.47\ (silica\%)$	912	0.12	61.9	15.6	<0.0000	1.7	−10.9	–

[a]Variables excluded where not significant (p value >0.10).
[b]Based on $R/EV < 0.4$.
[c]Based on $R < 750$ mm pa.
[d]Based on $R/EV < 0.3$.
[e]Based on $R < 400$ mm pa.
Note: available phosphorous (P) and nitrogen (N) tested but no significant relationships apparent.

Appendix B

Multiple linear regression additional statistics.

Property	Climate index	Horizon	Variable	Value	SE[a]	*t* value	ANOVA *F*	*p* value	Residual SE	*N*
pH	*R* / EV	A	Intercept	8.69	0.27	32.3		0.0000	1.01	490
			R / EV	−1.22	0.085	−14.5	272.9	0.0000		
			Silica	−0.025	0.0039	−6.5	42.0	0.0000		
			Slope	0.0072	0.0036	−2.0	44.0	0.0468		
		B	Intercept	8.57	0.25	35.0		0.0000	0.98	604
			R / EV	−1.07	0.07	−14.4	241.6	0.0000		
			Silica	−0.025	0.0034	−7.4	53.0	0.0000		
			Slope	−0.0077	0.0034	−2.3	5.1	0.0238		
	R	A	Intercept	8.86	0.28	31.4		0.0000	1.04	491
			R	−0.0008	0.0001	−12.9	198.1	0.0000		
			Silica	−0.03	0.0039	−7.7	57.8	0.0000		
			Slope	−0.013	0.0037	−3.4	11.8	0.0006		
		B	Intercept	8.59	0.25	33.8		0.0000	1.01	606
			R	−0.0007	0.0001	12.6	172.3	0.0000		
			Silica	−0.029	0.0035	−8.3	67.9	0.0000		
			Slope	−0.0089	0.0035	−2.5	6.3	0.0120		
Sum of bases (ln $cmol_c$/kg)	*R* / EV	A	Intercept	5.57	0.25	22.6		0.0000	1.21	781
			R / EV	−0.81	0.075	−10.7	183.4	0.0000		
			Silica	−0.051	0.0035	−14.6	312.7	0.0000		
		B	Intercept	6.26	0.38	16.6		0.0000	1.45	549
			R / EV	−1.22	0.11	−10.7	143.2	0.0000		
			Silica	−0.058	0.005	−10.9	116.4	0.0000		
			Slope	−0.013	0.005	−2.3	5.3	0.0230		
	R	A	Intercept	5.73	0.25	22.9		0.0000	1.22	783
			R	−0.0007	0.0001	−11.8	131.5	0.0000		
			Silica	−0.052	0.0035	−14.9	221.6	0.0000		
		B	Intercept	6.46	0.38	17.1		0.0000	1.44	550
			R	−0.0009	0.0001	−11.3	139.7	0.0000		
			Silica	−0.062	0.0053	−11.7	136.5	0.0000		
			Slope	−0.011	0.0054	−1.9	3.7	0.0535		
Organic C (ln, %)	*R* / EV	A	Intercept	0.74	0.22	3.4		0.0009	0.82	485
			R / EV	0.65	0.07	9.3	85.0	0.0000		
			Silica	−0.014	0.0032	−4.3	27.7	0.0000		
			Slope	0.009	0.003	3.1	33.5	0.0023		
		B	Intercept	−0.23	0.2	−1.1		0.2600	0.8	591
			R / EV	0.40	0.06	6.6	50.0	0.0000		
			Silica	−0.0117	0.0028	−4.2	18.4	0.0000		
			Slope	0.0092	0.0028	3.3	11.1	0.0009		
	R	A	Intercept	0.82	0.24	3.5		0.0060	0.86	486
			R	0.0003	0.001	5.8	39.8	0.0000		
			Silica	−0.011	0.0033	−3.3	11.6	0.0007		
			Slope	0.013	0.0031	4.1	17.2	0.0000		
		B	Intercept	−0.082	0.21	−0.4		0.6951	0.82	592
			R	0.0002	0.0000	3.5	18.3	0.0005		
			Silica	−0.011	0.0029	−3.7	14.1	0.0003		
			Slope	0.011	0.0029	3.8	14.7	0.0001		
Clay (%)	*R* / EV	A	Intercept	76.17	3.04	25.1		0.0000	15.6	856
			R / EV	4	0.9	4.4	11.9	0.0006		
			Silica	−0.8	0.043	18.4	343.9	0.0000		
		B	Intercept	87.59	3.36	26.0		0.0000	16.9	908
			R / EV	4.91	0.93	5.3	29.1	0.0000		
			Silica	−0.86	0.047	−18.4	336.8	0.0000		
	R	A	Intercept	74.12	3.04	24.3		0.0000	15.4	857
			R	0.0042	0.0007	6.1	34.6	0.0000		
			Silica	−0.79	0.043	−18.4	339.8	0.0000		
		B	Intercept	82.28	3.3	24.9		0.0000	16.42	909
			R	0.0064	0.0007	9.3	108.7	0.0000		
			Silica	−0.93	0.045	−18.2	332.0	0.0000		

[a]Standard error.

Appendix C.1

Median values of pH (H_2O) in 80 environmental regimes.

Topo. class	Parent material	Horizon	Rainfall class and mm pa															
			A		B		C		D		E		F		G		H	
			<250		250–500		500–750		750–1000		1000–1250		1250–1500		1500–2000		>2000	
			Median	*N*	Median	*N*	Median	*N*	Median	*N*	Median	*N*	Median	*N*	Median	*N*	Median	*N*
Level	ES	A	8.7	2	5.8	3	6.3	10	4.2	11	4.6	2	5.1	3	4.4	5	5.0	4
		B	7.8	4	6.1	4	5.4	10	4.8	13	–	1	5.6	6	5.3	4	5.2	3
	HS	A	8.5	6	6.5	56	5.7	53	5.4	21	5.7	27	5.3	37	4.5	15	4.5	36
		B	8.1	5	6.1	79	5.9	55	5.6	33	5.4	28	5.3	36	4.5	16	4.7	42

Appendix C.1 (*continued*)

Topo. class	Parent material	Horizon	Rainfall class and mm pa															
			A		B		C		D		E		F		G		H	
			<250		250–500		500–750		750–1000		1000–1250		1250–1500		1500–2000		>2000	
			Median	*N*	Median	*N*	Median	*N*	Median	*N*	Median	*N*	Median	*N*	Median	*N*	Median	*N*
Level	Int	A	8.7	8	7.5	16	6.8	36	5.6	22	5.9	29	5.6	20	4.8	7	4.8	24
		B	8.4	8	7.6	11	5.8	31	6.0	30	6.2	35	5.7	26	5.0	7	4.8	27
	Maf	A	8.3	5	7.5	14	6.9	27	8.1	13	6.8	14	5.8	9	5.3	9	4.4	6
		B	8.5	4	8.0	16	7.1	16	7.4	8	6.2	11	6.4	7	5.2	8	4.9	8
	Cal	A	8.2	14	8.0	23	7.6	17	7.2	19	6.4	7	7.6	3	5.8	9	5.8	9
		B	8.2	8	8.1	22	7.7	21	7.8	38	7.8	20	6.0	8	5.4	11	5.4	11
Sloping	ES	A	–	1	–	0	3.8	2	4.4	2	6.0	3	5.8	6	4.9	4	–	0
		B	–	0	–	0	4.5	2	–	1	–	1	5.5	5	–	1	–	0
	HS	A	–	0	5.2	6	4.6	7	4.5	13	5.6	26	5.2	26	4.8	28	4.2	35
		B	–	0	4.8	5	5.5	9	4.6	17	5.7	22	5.0	20	4.7	25	4.7	36
	Int	A	–	1	8.3	3	6.0	14	6.1	11	4.9	4	5.1	16	5.4	16	5.1	34
		B	–	1	7.7	2	5.2	11	5.5	10	4.5	5	4.9	10	5.4	12	4.8	41
	Maf	A	–	1	7.6	6	7.1	11	–	1	6.3	5	5.5	14	5.6	10	5.3	13
		B	–	1	7.7	6	7.6	12	–	1	6.1	3	5.8	13	5.6	9	5.2	16
	Cal	A	–	0	8.0	9	7.4	16	6.7	9	6.5	11	–	1	6.8	14	–	1
		B	–	0	–	1	7.2	16	7.4	14	5.7	11	6.2	3	6.4	23	–	1

Appendix C.2

Median values of sum of bases in 80 environmental regimes.

Topo. class	Parent material	Horizon	Rainfall class and mm pa															
			A		B		C		D		E		F		G		H	
			<250		250–500		500–750		750–1000		1000–1250		1250–1500		1500–2000		>2000	
			Median ($cmol_c$/kg)	*N*	Median ($cmol_c$/kg)	*N*	Median ($cmol_c$/kg)	*N*	Median ($cmol_c$/kg)	*N*	Median ($cmol_c$/kg)	*N*	Median ($cmol_c$/kg)	*N*	Median ($cmol_c$/kg)	*N*	Median ($cmol_c$/kg)	*N*
Level	ES	A	–	1	1.6	3	8.7	9	2.8	10	1.3	2	0.4	3	0.8	4	0.8	3
		B	0.1	3	4.6	4	1.7	6	0.4	10	–	0	2.0	5	0.6	3	–	0
	HS	A	44.6	6	4.1	56	2.6	52	3.8	25	3.1	25	2.9	37	1.2	16	0.7	33
		B	8.0	5	4.9	77	3.7	55	3.0	35	1.5	26	2.6	34	0.7	17	0.4	33
	Int	A	17.0	8	16.7	16	15.3	29	5.6	22	11.0	21	9.6	20	2.8	6	2.4	24
		B	19.5	7	15.8	11	13.3	26	10.1	33	16.1	28	15.1	25	1.5	5	1.0	23
	Maf	A	72.9	3	36.4	15	25.5	28	27.7	12	10.2	13	17.9	9	1.3	8	1.9	6
		B	35.1	3	41.9	17	33.9	18	10.2	9	17.9	10	28.1	7	0.8	7	0.3	7
	Cal	A	11.4	12	28.1	22	23.5	31	28.2	33	12.1	7	15.9	3	11.3	10	24.7	3
		B	15.1	8	41.6	22	23.2	31	23.2	45	10.7	20	8.0	8	8.5	12	8.6	3
Sloping	ES	A	–	1	–	0	0.2	2	10.4	2	3.1	2	1.8	6	0.2	4	–	0
		B	–	0	–	0	–	0	–	1	–	1	0.9	5	–	0	–	0
	HS	A	–	0	2.0	5	1.1	7	1.4	13	2.2	14	1.9	25	1.5	28	1.1	31
		B	–	0	1.2	5	1.4	9	1.2	17	1.1	13	0.8	20	0.4	20	0.6	33
	Int	A	–	1	30.0	2	2.5	13	18.3	10	1.7	4	3.2	16	3.7	15	5.3	27
		B	–	1	46.9	2	4.6	12	5.7	10	0.8	4	1.4	18	7.8	12	0.9	31
	Maf	A	–	0	16.1	4	56.4	11	6.3	3	8.7	4	2.5	14	4.3	9	4.0	13
		B	–	0	41.9	2	62.2	12	2.9	3	23.5	2	4.0	12	1.1	8	2.0	14
	Cal	A	–	0	21.9	9	23.6	16	22.8	8	16.5	8	–	1	15.9	14	–	1
		B	–	0	–	1	23.2	15	22.1	12	12.7	5	–	1	10.6	24	–	1

Appendix C.3

Median values of organic carbon in 80 environmental regimes.

Topo. class	Parent material	Horizon	Rainfall class and mm pa															
			A		B		C		D		E		F		G		H	
			<250		250–500		500–750		750–1000		1000–1250		1250–1500		1500–2000		>2000	
			Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*
Level	ES	A	–	1	0.20	3	0.75	10	3.85	12	2.14	2	1.02	3	1.39	5	2.67	4
		B	0.10	3	0.27	4	1.20	7	0.36	13	–	1	0.31	6	0.29	4	0.11	3
	HS	A	0.20	6	0.40	55	0.30	50	1.74	26	1.24	29	1.07	37	1.67	16	2.47	36
		B	0.25	4	0.30	77	0.20	50	0.48	37	0.43	28	0.40	36	0.43	16	0.73	44
	Int	A	0.29	7	1.00	14	1.97	34	3.15	23	1.92	29	1.70	21	2.01	8	3.10	24
		B	0.18	5	0.40	9	0.70	30	0.50	27	0.60	35	0.88	26	0.68	8	0.70	28
	Maf	A	0.47	5	0.44	15	1.05	26	0.54	13	0.84	13	3.51	8	2.20	9	2.62	6
		B	0.22	4	0.40	17	0.63	18	0.81	10	0.80	11	1.00	7	0.75	7	0.95	9
	Cal	A	0.16	14	0.94	22	2.14	33	1.92	38	2.82	8	1.94	3	3.05	10	5.70	3
		B	0.14	8	0.82	22	0.91	32	0.90	49	0.91	21	0.78	8	1.10	11	1.18	3
Sloping	ES	A	–	1	–	0	2.64	2	–	0	0.91	3	0.70	6	1.59	4	–	0
		B	–	0	–	0	1.44	2	–	1	–	1	0.50	6	–	1	–	0
	HS	A	–	0	1.58	6	1.00	7	2.10	14	2.90	24	2.17	27	2.88	27	2.41	33
		B	–	0	0.30	4	0.30	8	0.80	17	0.56	22	1.00	18	0.55	24	0.61	37

(*continued on next page*)

Appendix C.3 (*continued*)

Topo. class	Parent material	Horizon	Rainfall class and mm pa															
			A		B		C		D		E		F		G		H	
			<250		250–500		500–750		750–1000		1000–1250		1250–1500		1500–2000		>2000	
			Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*
Sloping	Int	A	–	1	0.96	3	2.00	14	3.49	11	3.31	4	2.38	16	3.00	16	2.89	34
		B	–	1	0.44	2	0.61	10	1.03	8	0.53	5	0.53	18	0.87	12	0.80	41
	Maf	A	–	1	1.12	6	0.96	11	1.49	3	2.34	5	4.42	14	3.33	10	4.38	12
		B	–	1	0.82	5	0.86	11	2.84	2	0.87	3	1.51	12	0.90	9	1.27	15
	Cal	A	–	0	1.30	8	2.38	17	2.46	10	3.66	11	–	1	2.93	15	–	1
		B	–	0	–	1	0.83	17	0.56	15	0.90	11	0.70	3	0.80	24	–	1

Appendix C.4
Median values of clay content in 80 environmental regimes.

Topo. class	Parent material	Horizon	Rainfall class and mm pa															
			A		B		C		D		E		F		G		H	
			<250		250–500		500–750		750–1000		1000–1250		1250–1500		1500–2000		>2000	
			Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*	Median (%)	*N*
Level	ES	A	5.0	2	8.5	2	8.5	8	4.0	12	4.5	2	5.0	3	11.0	5	1.0	4
		B	4.5	4	10.0	4	4.0	9	3.5	14	–	1	8.5	6	10.0	4	1.0	3
	HS	A	11.0	4	13.0	57	9.0	50	14.0	22	16.0	29	17.0	37	25.0	16	26.5	36
		B	11.0	5	19.0	79	15.0	54	26.5	36	28.0	29	32.0	36	39.0	17	38.0	42
	Int	A	22.0	8	29.5	16	24.5	34	22.5	22	26.5	28	29.0	21	37.5	8	21.0	23
		B	28.0	8	37.0	11	36.0	31	33.0	33	37.0	34	49.0	26	48.5	8	32.0	28
	Maf	A	20.0	5	37.0	15	39.0	27	58.0	13	46.0	14	56.0	9	38.0	9	62.0	5
		B	43.5	4	49.0	17	53.5	18	59.5	10	45.0	10	52.5	8	59.5	9	65.0	7
	Cal	A	15.0	13	24.5	24	39.0	33	46.5	38	28.5	8	49.0	3	50.5	10	56.0	3
		B	20.0	8	31.0	23	46.5	32	38.0	50	39.0	21	55.5	8	64.0	11	55.0	3
Sloping	ES	A	–	1	–	0	2.5	2	7.5	2	6.5	2	13.5	6	53.0	4	–	0
		B	–	0	–	0	–	1	–	1	–	1	12.0	6	–	1	–	0
	HS	A	–	0	13.5	6	6.0	6	9.5	14	18.0	26	22.0	25	30.0	28	22.5	34
		B	–	0	7.0	5	10.0	8	21.0	20	26.5	22	33.0	19	36.0	25	38.0	37
	Int	A	–	1	19.0	3	23.0	14	31.0	11	36.5	4	41.0	16	34.0	16	27.0	33
		B	–	1	29.0	2	28.0	12	34.0	11	31.0	5	46.0	17	40.0	12	39.0	40
	Maf	A	–	0	18.5	6	59.0	11	13.0	3	13.0	3	52.0	13	45.5	10	59.0	11
		B	–	1	17.0	6	64.0	12	23.0	3	56.0	3	57.0	13	59.0	9	67.0	15
	Cal	A	–	0	24.0	9	34.0	16	25.5	10	29.0	11	–	1	42.0	15	–	1
		B	–	0	–	1	43.0	17	35.0	15	14.0	10	47.0	3	51.5	24	–	1

References

Batjes, N.H., 2002. A homogenised soil profile dataset for global and regional environmental research (WISE, version 1.1), Report 2002/01. International Soil Reference and Information Centre, Wageningen.

Batjes, N.H., Bridges, E.M., 1994. Potential emissions of radiatively active gases from soil to atmosphere with special reference to methane: development of a global database (WISE). Journal of Geophysical Research 99 (D8), 16479–16489.

Best, M.G., 1982. Igneous and Metamorphic Petrology. W.H. Freeman & Co., New York.

Birkeland, P.W., 1999. Soils and Geomorphology, third ed. Oxford University Press, New York.

Bokheim, J.G., 1980. Solution and use of chronofunctions in studying soil development. Geoderma 24, 71–85.

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., Henderson, B.L., Viergever, K., 2006. Knowledge discovery from models of soil properties developed through data mining. Ecological Modelling 191, 431–436.

Dan, J., Yaalon, D.H., Koyumdjisky, H., 1968. Catenary soil relationships in Israel, I. The Netanya catena on coastal dunes of the Sharon. Geoderma 2, 95–120.

Dobos, E., 2006. Future of soil science: a vision from Europe. In: Hartemink, A.E. (Ed.), The Future of Soil Science. International Union of Soil Sciences (IUSS), Wageningen, pp. 34–36.

Duff, D., 1993. Holmes' Principles of Physical Geology. Chapman & Hall.

FAO/IIASA, 2002. Global agro-ecological assessment for agriculture in the twenty-first century. Food and Agriculture Organisation of the United Nations and International Institute for Applied Systems Analysis. FAO Land and Water Digital Media Series, vol. 21.

Gray, J.M., 2008. Relationships in Global Soil Distribution as Revealed by a World Soil Database. Master of Philosophy Thesis, Macquarie University, Sydney (unpublished).

Gray, J.M., Murphy, B.W., 2002. Parent Material and World Soil Distribution, poster paper 2215. Proceedings of 17th World Congress of Soil Science, Bangkok, August 2002.

Gray, J.M., Humphreys, G.S., 2004. Patterns of global soil distribution as revealed by two major soil databases. Proceedings of SuperSoil 2004, Australian and NZ Societies of Soil Science Conference, University of Sydney, Sydney, Australia, December 5–9 2004.

Gray, J.M., Humphreys, G.S., Deckers, J.A., 2007. Use of a large world soil database for modelling global soil distribution. Eurasian Soil Science 40 (9), 928–933.

Gray, J.M., Humphreys, G.S. and Deckers, J.A., in preparation. Patterns in World Reference Base soil distribution as revealed by a global soil database.

Grunwald, S., 2006a. What do we really know about the space-time continuum of soil-landscapes? In: Grunwald, S. (Ed.), Environmental Soil-Landscape Modelling — Geographic Information Technologies and Pedometrics. Taylor & Francis, pp. 3–36.

Grunwald, S., 2006b. Future of soil science. In: Hartemink, A.E. (Ed.), The Future of Soil Science. International Union of Soil Sciences (IUSS), Wageningen, pp. 51–53.

Hartemink, A.E. (Ed.), 2006. The Future of Soil Science. International Union of Soil Sciences (IUSS), Wageningen.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383–398.

Heuvelink, G.B.M., 2006. Foreword. In: Grunwald, S. (Ed.), 2005, Environmental Soil-Landscape Modelling — Geographic Information Technologies and Pedometrics. Taylor & Francis.

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

Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill Book Company Inc., New York.

Jenny, H., 1961. Derivation of state factor equations of soils and ecosystems. Soil Science Society of America Proceeding 25, 385–388.

Jenny, H., 1980. The Soil Resource. Springer-Verlag, New York.

Johnson, R.M., Barry, S.J., Bleys, E., Bui, E.N., Moran, C.J., Simon, D.A.P., Carlile, P., McKenzie, N.J., Henderson, B.L., Chapman, G., Imhoff, M., Maschmedt, D., Howe, D., Grose, C., Shoknecht, N., Powell, B., Grundy, M., 2003. ASRIS: the database. Australian Journal Soil Research 41, 1021–1036.

Lagacherie, P., McBratney, A.B., 2007. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping. Elsevier, Amsterdam, pp. 3–22.

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

McBratney, A.B., Mendonca Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

McKenzie, N.J., Austin, M.P., 1993. A quantitative Australian approach to medium 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., Jacquier, D.W., Maschmedt, D.J., Griffin, E.A., Brough, D.M., 2005. The Australian Soil Resource Information System: Technical Specifications. CSIRO, Canberra.
Murphy, B.W., 2006. The nature of soil. In: Charman, P.E.V., Murphy, B.W. (Eds.), Soils — Their Properties and Management. Oxford University Press, pp. 1–13.
Phillips, J.D., 1993a. Progressive and regressive pedogenesis and complex soil evolution. Quaternary Research 40, 169–176.
Phillips, J.D., 1993b. Stability implications of the state factor model of soils as a non-linear dynamical system. Geoderma 58, 1–15.
Phillips, J.D., 2001. The relative importance of intrinsic and extrinsic factors in pedodiversity. Annals of the Association of American Geographers 94, 609–621.
Rasmussen, C., Southard, R.J., Horwath, W.R., 2005. Modelling energy inputs to predict pedogenic environments using regional environmental databases. Soil Science Society of America 69, 1266–1274.
Ryan, P.J., McKenzie, N.J., O'Connell, D.O., 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.
Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Progress in Physical Geography 27, 171–197.
Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecological modelling 181, 1–15.
Skidmore, A.K., Varekamp, C., Wilson, L., Knowles, E., Delaney, J., 1997. Remote sensing of soils in a eucalypt forest environment. International Journal Remote Sensing 18, 39–56.
Tardy, Y., Bocquier, G., Paquet, H., Millot, G., 1973. Formation of clay from granite and its distribution in relation to climate and topography. Geoderma 10, 271–284.
Taylor, J.K., 1983. The Australian environment. Soils: An Australian Viewpoint. Division of Soils. CSIRO/Academic Press, Melbourne/London, pp. 1–2.
Webster, R., 1977. Canonical correlation in pedology: how useful? Journal of Soil Science 2, 196–221.
Wilkinson, M.T., Humphreys, G.S., 2005. Exploring pedogenesis via nuclide-based soil production rates and OSL-based bioturbation rates. Australian Journal of Soil Research 43, 767–779.
Yaalon, D.H., 1975. Conceptual models in pedogenesis: can soil-forming functions be solved? Geoderma 14, 189–205.