CSIRO PUBLISHING

www.publish.csiro.au/journals/ajsr *Australian Journal of Soil Research*, 2008, **46**, 1–16

Using a legacy soil sample to develop a mid-IR spectral library

R. A. Viscarra Rossel[A,B,C], *Y. S. Jeon*[B], *I. O. A. Odeh*[B], *and A. B. McBratney*[A,B]

[A]Australian Centre for Precision Agriculture, Faculty of Agriculture, Food & Natural Resources, The University of Sydney, NSW 2006, Australia.
[B]Faculty of Agriculture, Food & Natural Resources, The University of Sydney, NSW 2006, Australia.
[C]Corresponding author. Email: r.viscarra-rossel@usyd.edu.au

Abstract. This paper describes the development of a diffuse reflectance spectral library from a legacy soil sample. When developing a soil spectral library, it is important to consider the number of samples that are needed to adequately describe the soil variability in the region in which the library is to be used; the manner in which the soil is sampled, handled, prepared, stored, and scanned; and the reference analytical procedures used. As with any type of modelling, the dictum is 'garbage in = garbage out' and hopefully the converse 'quality in = quality out'. The aims of this paper are to: (*i*) develop a soil mid infrared (mid-IR) diffuse reflectance spectral library for cotton-growing regions of eastern Australia from a legacy soil sample, (*ii*) derive soil spectral calibrations for the prediction of soil properties with uncertainty, and (*iii*) assess the accuracy of the predictions and populate the legacy soil database with good quality information. A scheme for the construction and use of this spectral library is presented. A total of 1878 soil samples from different layers were scanned. They originated from the Upper Namoi, Namoi, and Gwydir Valley catchments of north-western New South Wales (NSW) and the McIntyre region of southern Queensland (Qld). A conditioned Latin hypercube sampling (cLHS) scheme was used to sample the spectral data space and select 213 representative samples for laboratory soil analyses. Using these data, partial least-squares regression (PLSR) was used to construct the calibration models, which were validated internally using cross validation and externally using an independent test dataset. Models for organic C (OC), cation exchange capacity (CEC), clay content, exchangeable Ca, total N (TN), total C (TC), gravimetric moisture content θ_g, total sand and exchangeable Mg were robust and produced accurate results ($R^2_{\text{adj.}} > 0.75$ for both cross and test set validations). The root mean squared error (RMSE) of mid-IR-PLSR predictions was compared to those from (blind) duplicate laboratory measurements. Mid-IR-PLSR produced lower RMSE values for soil OC, clay content, and θ_g. Finally, bootstrap aggregation-PLSR (bagging-PLSR) was used to predict soil properties with uncertainty for the entire library, thus repopulating the legacy soil database with good quality soil information.

Additional keywords: mid-IR diffuse reflectance spectroscopy, spectral library, partial least squares regression, bagging-PLSR, legacy soil data.

Introduction

Diffuse reflectance spectroscopy (DRS) has been used in soil science research since the 1950s and '60s (e.g. Brooks 1952; Bowers and Hanks 1965). However, it is only in approximately the last 20 years, most likely coinciding with the establishment of chemometrics and multivariate statistical techniques in analytical chemistry, that its usefulness and importance have been realised. The majority of this work has involved using the visible and near infrared (vis-NIR: 400–2500 nm) regions of the electromagnetic spectrum (e.g. Dalal and Henry 1986; Stenberg *et al*. 1995; Reeves *et al*. 1999; Chang *et al*. 2001; Dunn *et al*. 2002) but increasingly also the mid infrared (mid-IR: 2500–25 000 nm) (Janik and Skjemstad 1995; Masserschmidt *et al*. 1999; Madari *et al*. 2006).

Both vis-NIR and mid-IR techniques are rapid, accurate, and more economical than conventional methods of soil analysis (e.g. Shepherd and Walsh 2002; Viscarra Rossel *et al*. 2006*a*), they do not use environmentally harmful chemicals, require fewer pre-treatments, are non destructive, and when combined with multivariate calibrations, a single spectrum can provide estimates of several soil properties. The techniques are highly sensitive to both organic and inorganic soil composition, making them potentially useful and powerful tools for the assessment and monitoring of soil, its quality and function. The mid-IR contains much more information on soil mineral and organic composition than the vis-NIR (e.g. Janik and Skjemstad 1995), and its multivariate calibrations across a wide range of soil types are more robust (Viscarra Rossel *et al*. 2006*a*). The reason is that the fundamental molecular vibrations of soil components occur in the mid-IR, while only their overtones and combinations are detected in the NIR. Hence soil NIR spectra display fewer and much broader absorption features compared to mid-IR spectra (Fig. 1).

Mostly, the techniques have been used for rapid, inexpensive, and non destructive soil analyses (e.g. Janik *et al*. 1998; Reeves *et al*. 1999; Dunn *et al*. 2002; Brown *et al*. 2006; Viscarra Rossel *et al*. 2006*a*). However, other applications are emerging, for example, using DRS as a tool for soil surveying (e.g. Demattê

10.1071/SR07099 0004-9573/08/010001

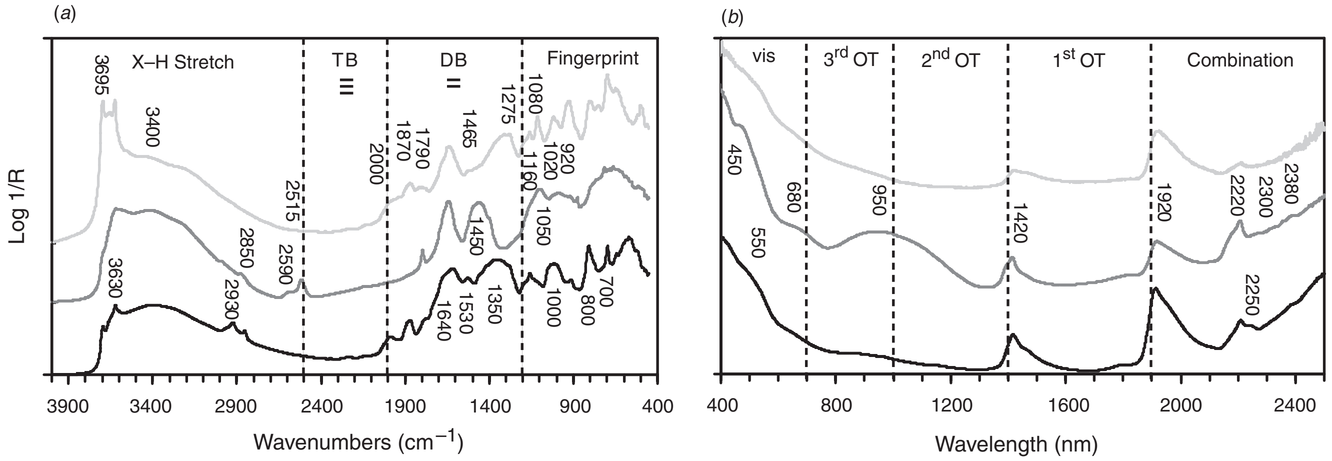

Fig. 1. Soil diffuse reflectance spectra in (*a*) the mid infrared 4000–400 cm^{-1} (or 2500–25 000 nm) showing approximate occurrence of the fingerprint, double bond (DB), triple bond (TB), and X–H stretch regions, and (*b*) the visible and near infrared 400–2500 nm (or 25 000–4000 cm^{-1}) showing approximately where the combination, first, second, and third overtone (OT) vibrations occur as well as the visible (vis) range.

et al. 2004), mineralogical quantification (e.g. Viscarra Rossel *et al.* 2006*b*), the assessment of soil quality and condition (e.g. Vågen *et al.* 2006), soil monitoring (e.g. Cohen *et al.* 2005), soil inference (McBratney *et al.* 2006), digital soil mapping (Viscarra Rossel and McBratney 2008), remote sensing (e.g. Ben-Dor *et al.* 2006), precision agriculture (e.g. Stenberg *et al.* 2002; Wetterlind *et al.* 2007), and proximal soil sensing (e.g. Shibusawa *et al.* 2001; Mouazen *et al.* 2007).

A requirement for using soil DRS to assess soil is the creation of a soil spectral database or library and there is widespread interest in their development. However, only a few that are geographically diverse and specific to soil properties have been described in the literature (e.g. Dunn *et al.* 2002; Shepherd and Walsh 2002; Brown *et al.* 2006). Three important requirements for the development of a soil spectral library are: (*i*) it should contain as many samples as are needed to adequately describe the soil variability in the region in which the library is to be used; (*ii*) the samples should be carefully subsampled, handled, prepared, stored, and scanned (everything that has happened to the sample up to the time of scanning will be embodied in the sample and recorded in the spectra); and (*iii*) the reference soil analytical data used in the calibrations should be acquired using

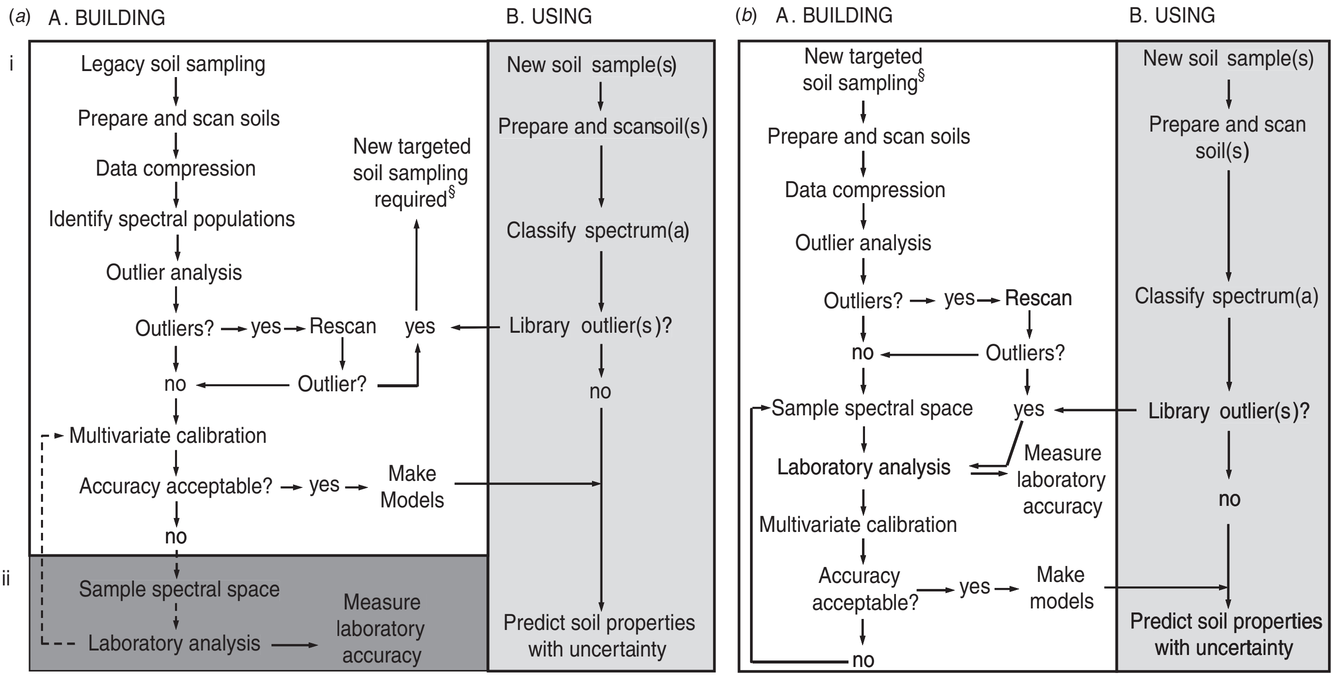

Fig. 2. Developing a soil diffuse reflectance spectral library from (*a*) a legacy soil sampling, and (*b*) a new targeted soil sampling. Construction of the library is shown in (A) and its use by (B).

reliable and accredited analytical procedures. As with any type of modelling, the dictum here is 'garbage in = garbage out' and hopefully the converse 'quality in = quality out'. With relation to (*i*) above, if the library is being developed for a particular region from scratch, then the soil sampling strategy used will be critical (e.g. de Gruijter *et al*. 2006). If the library is being developed from a legacy soil sample, then the recommendation is to scan all of the samples, then correlate the spectra with the relevant soil properties and use these to assess the quality of the legacy soil data.

The aims of this paper are to: (*i*) develop a soil diffuse reflectance spectral library for cotton-growing regions of eastern Australia from a legacy soil sample, (*ii*) derive soil spectral calibrations for the prediction of soil properties with uncertainty, and (*iii*) assess the accuracy of the predictions and populate the legacy soil sample with good quality information.

Schemes for the development of a soil spectral library

Figure 2*a* shows the scheme that was used to develop the soil mid-IR diffuse reflectance spectral library from the legacy soil sample. The scheme is divided into 2 sections: (A) for building the library and (B) for using it.

From Fig. 2*a* (A-i), soil from the legacy sample is prepared and scanned using the diffuse reflectance spectrometer. The spectra are compressed using principal components analysis (PCA) to identify structure, patterns, and possible clustering. Outlier analysis is performed on the multivariate data, and if outliers are present, the samples are rescanned for verification. If there are no outliers, the spectra are combined with the legacy soil data and multivariate calibrations derived and validated using cross validation and independent test set validations (e.g. Martens and Næs 1989). If the accuracies of the calibrations are within acceptable limits, then from Fig. 2*a* (B), the spectral library may be used to predict soil properties from only the diffuse reflectance spectra of new soil samples that belong to the same spectral population as the soils in the library. Conversely, if you expect the accuracies of the calibrations to be good but they are not, then it may be that the soil analytical data from the legacy sample are not reliable. In this case, the spectral data space is sampled and laboratory analysis performed on the selected samples [Fig. 2*a* (A-ii)]. These new soil data are then combined with their corresponding spectra and multivariate calibrations performed. If the accuracies of the calibrations are acceptable, then from Fig. 2(B), they may be used to predict the soil properties for the remaining legacy soil sample. In this way, the legacy data may be repopulated with good quality information. These calibrations may also be used with new soil samples that belong to the same spectral population as those in the library. If samples are not represented by the spectral library, i.e. they are outliers [Fig. 2*a* (A-i) and (B)], these may be removed and a new targeted soil sampling mission may be required to populate the library with samples with similar characteristics. This sampling should aim to characterise the soil variability of the targeted area, and the scheme shown in

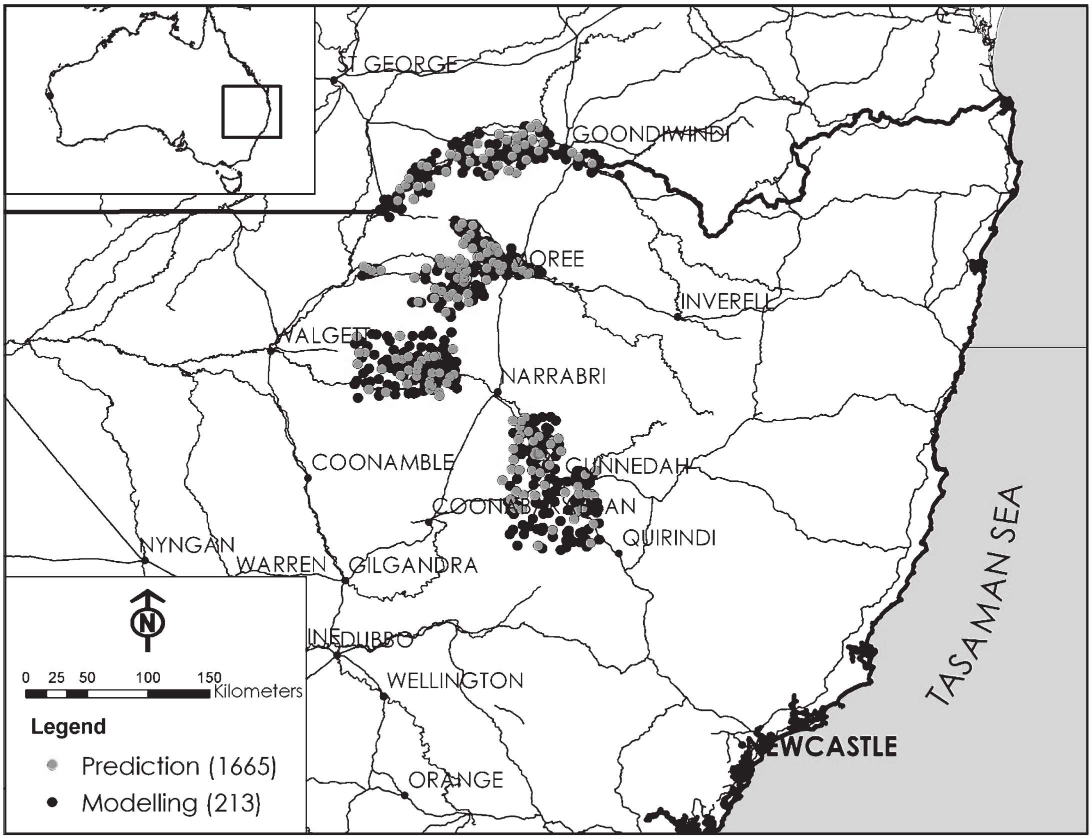

Fig. 3. Geography of soil samples in the legacy soil sampling. Black points show the samples selected for laboratory analysis.

Fig. 2*b*, which is similar to that by Shepherd and Walsh (2002), may be used. In this case, if the new spectra classify poorly because they are not well represented by the library, they may be added to the library after reference laboratory analysis. Hence, the development of soil spectral libraries should be a continual process. Note that in both instances (Fig. 2*a*, *b*), when reference soil samples are analysed, acquiring a measure of laboratory accuracy is suggested.

Methods

Legacy soil sample

The legacy soil sample originated from four of the major cotton-growing regions of eastern Australia over a period of six years by Odeh *et al*. (2004). The soil was sampled from different layers, including the 0–0.10, 0.10–0.20, 0.30–0.40, 0.60–0.70, and 0.70–0.80 m. They were stored in sealed plastic containers as ground samples with a size fraction of $\leq$2 mm. The laboratory analysis that was available for these samples included: soil pH; electrical conductivity (EC); cation exchange capacity (CEC); exchangeable cations K, Na, Mg, and Ca; clay, sand and silt contents; and organic carbon (OC). These analyses were measured in different laboratories and using various techniques, outlined in Odeh *et al*. (2004). A total of 1878 samples were used: 619 from the Upper Namoi (UPN), 352 samples from the Lower Namoi (NAM), and 573 samples from the Gwydir valley (GYD) in New South Wales, and 334 samples from the McIntyre valley (MTY) in southern Queensland (Fig. 3).

Building the mid-IR spectral library

Sample preparation and spectroscopic analysis

From each of the 1878 samples, approximately 5 g was subsampled for spectroscopic measurements. Soil samples were prepared and scanned in batches of approximately 50 samples. Sample preparation involved oven drying the samples for 18 h at 40°C to remove moisture effects, then 1 g of each soil was ground in an agate mortar and pestle for 30 s and then passed through a $\leq$200-µm sieve to ensure constant particle size. These soil powders were placed in aluminium sample cups (5 mm deep and 10 mm in diameter) (Fig. 4*a*) and levelled before introducing them to the spectrometer.

Soil mid-IR diffuse reflectance spectra were collected using a Tensor 37$^{\circledR}$ Fourier Transform Infrared (FT-IR) spectrometer from Bruker Optics (Massachusetts, USA) with a DRIFT attachment (Fig. 4*b*). Spectra were recorded from 3993 to 397 cm^{-1} with 8 cm^{-1} resolution and 64 scans per s, resulting in 933 wavenumbers. A KBr white reference background spectrum was recorded at the start of a scanning session and once every hour thereafter during each session. Spectra were recorded as percent reflectance in single files, which were subsequently merged into a single file and combined with the soil property data for the chemometric analysis.

PCA and outlier analyses

The spectra were mean centred and scaled before compression using PCA. The scores of the first 5 principal components (PCs) were used to visualise the structure, patterns, and possible clustering of the spectra in multivariate space. The Mahalanobis distance (De Maesschalck *et al*. 2000) on the first 4 PCs was used for outlier detection. Spectral outliers were rescanned for verification.

Multivariate calibrations using partial least-squares regression (PLSR)

Multivariate calibrations were performed on mean centred data using the partial least-squares regression 1 (PLSR) algorithm (Wold *et al*. 1983). To derive training and test datasets for each soil variable, the data were sorted from lowest to highest values and 2 out of every 5 consecutive rows were selected to test models developed using the remaining data. In this way, the calibrations for each of the soil properties were representative of

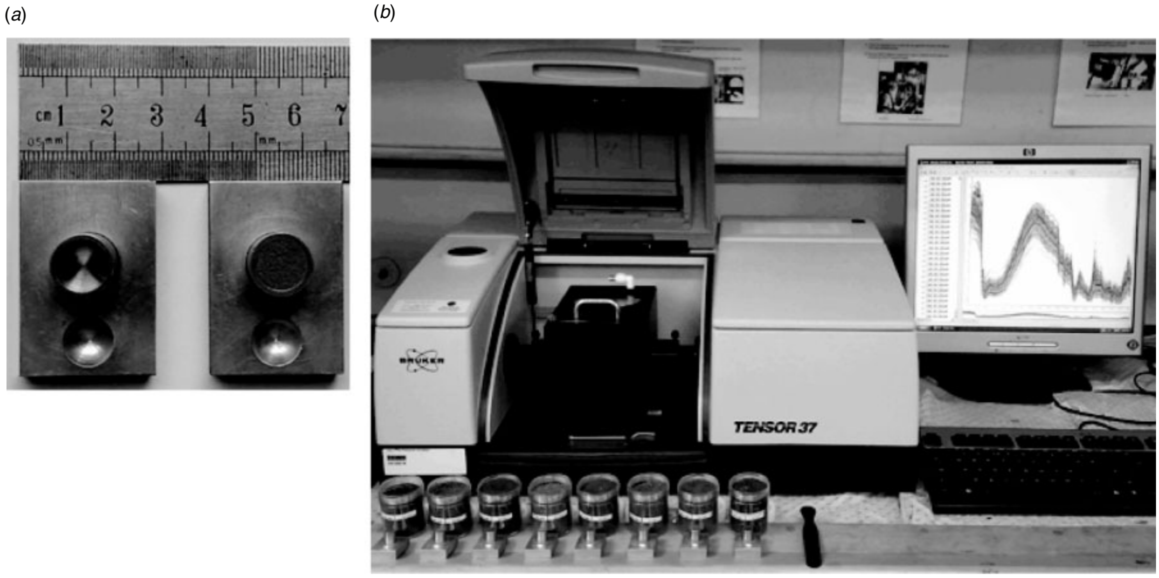

Fig. 4. Mid-IR spectroscopy of soil samples. (*a*) Sample holders and (*b*) Tensor 37$^{\circledR}$ Fourier Transform Infrared (FT-IR) spectrometer.

the entire population and were independently tested. Leave-one-out cross validation was used to select the optimal number of PLSR factors to use in the calibration models. Several spectral preprocessing techniques were tested to improve the robustness of the calibrations. When useful, the techniques used were the standard normal variate transform (SNV) (Barnes *et al.* 1989), the Savitski-Golay filter with either a first or second derivative (Savitzky and Golay 1964). Predictions on the independent test data were assessed using the root mean-square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

the mean error (ME):

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$$

and the standard deviation of the error (SDE):

$$\mathrm{SDE} = \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i - \mathrm{ME})^2}{N-1}$$

where $\hat{y}_i$ is the predicted value, y_i is the observed value and N is the number of data. These were used to measure the accuracy, bias, and precision of the models, respectively. The adjusted coefficient of determination ($R^2{}_{\mathrm{adj.}}$) and the ratio of percentage deviation (RPD) (Williams 1987) were also measured. The RPD is the ratio of the standard deviation of the reference laboratory data to the RMSE of the test set prediction. It is the factor by which the prediction accuracy has been increased compared with using the mean of the original data.

Identification of important mid-IR frequencies for the PLSR predictions were made by using both the PLS regression coefficient, **b** (Haaland and Thomas 1988) and the variable importance for projection (VIP) (Wold *et al.* 2001). The VIP was calculated by:

$$\mathrm{VIP}_k(a) = K\sum_{a} w_{ak}^2\left(\frac{SSY_a}{SSY_t}\right)$$

where $\mathrm{VIP}_k(a)$ is the importance of the *k*th predictor variable based on a model with *a* factors, w_{ak} is the corresponding loading weight of the *k*th variable in the *a*th PLSR factor, SSY_a is the explained sum of squares of **y** by a PLSR model with *a* factors, SSY_t is the total sum of squares of **y**, and *K* is the total number of predictor variables (i.e. in this instance 933 mid-IR frequencies).

The important mid-IR frequencies for each of the PLSR models were identified by setting thresholds for both VIP and the PLS regression coefficients, **b**. The thresholds for the VIP were set to 1 (following recommendations by Chong and Jun 2005) and thresholds for **b** were based on their standard deviations. A mid-IR frequency was deemed important for prediction if both of these conditions were met.

Selection of samples for laboratory analysis

The accuracies of the multivariate calibrations using the original legacy data were low and unsuitable for prediction. Therefore, a representative set of soil samples was selected for laboratory analysis. The selection was based purely on the spectra. The spectral data space was sampled by first deriving the PCA model and submitting the first 10 PCs, which accounted for 99% of the variation in the spectra, to a conditioned Latin Hypercube Sampling (cLHS) (Minasny and McBratney 2006). The cLHS selected a set of representative samples that covered the hypercube of the first 10 PCs. The number of samples was constrained to 213, which was a function of the available funds for laboratory analyses. The soils were analysed for soil OC by dichromate oxidation, total C (TC) and total N (TN) by dry combustion, total P (TP) using the boiling sulfuric acid method, soil EC and pH measured in 1 : 5 soil : water ($\mathrm{pH_w}$) as well as pH in 1 : 5 soil-0.01 M $\mathrm{CaCl_2}$ ($\mathrm{pH_{Ca}}$), exchangeable K, Na, Mg, Ca, and CEC using the combined $\mathrm{BaCl_2/NH_4Cl}$ procedure, Bray-P, and $\mathrm{NO_3}$-N and $\mathrm{NH_4}$ extracted by 2 M KCl and measured by flow injection analysis. These methods are described in Rayment and Higginson (1992). A measure of the exchangeable sodium percentage (ESP = (Na/CEC) × 100) was also derived. The laboratory was ISO accredited and certified by the Australasian Soil and Plant Analysis Council (ASPAC). Gravimetric moisture content (θ_g) and clay, silt, and sand contents were analysed in our research laboratory. The particle size analysis was performed

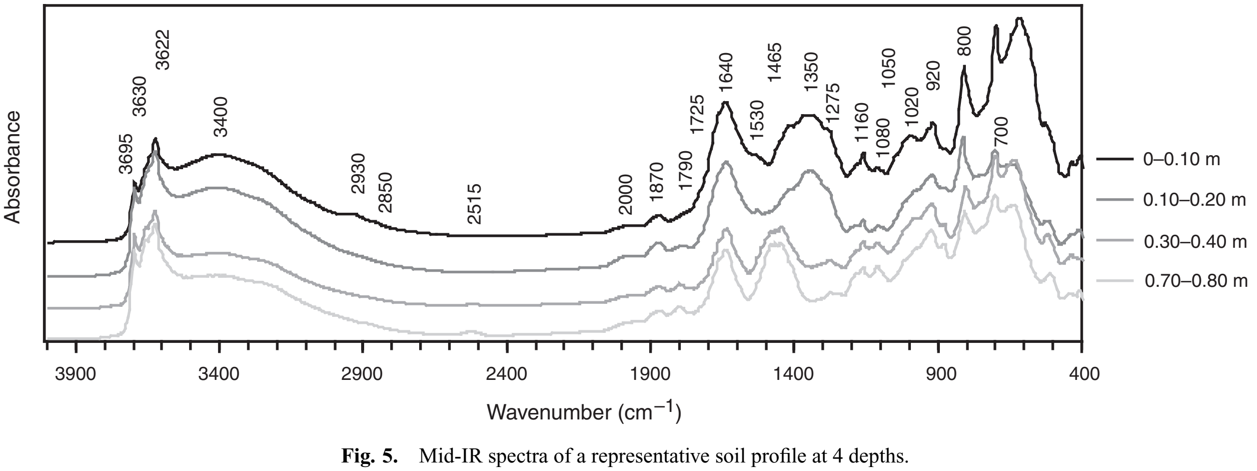

Fig. 5. Mid-IR spectra of a representative soil profile at 4 depths.

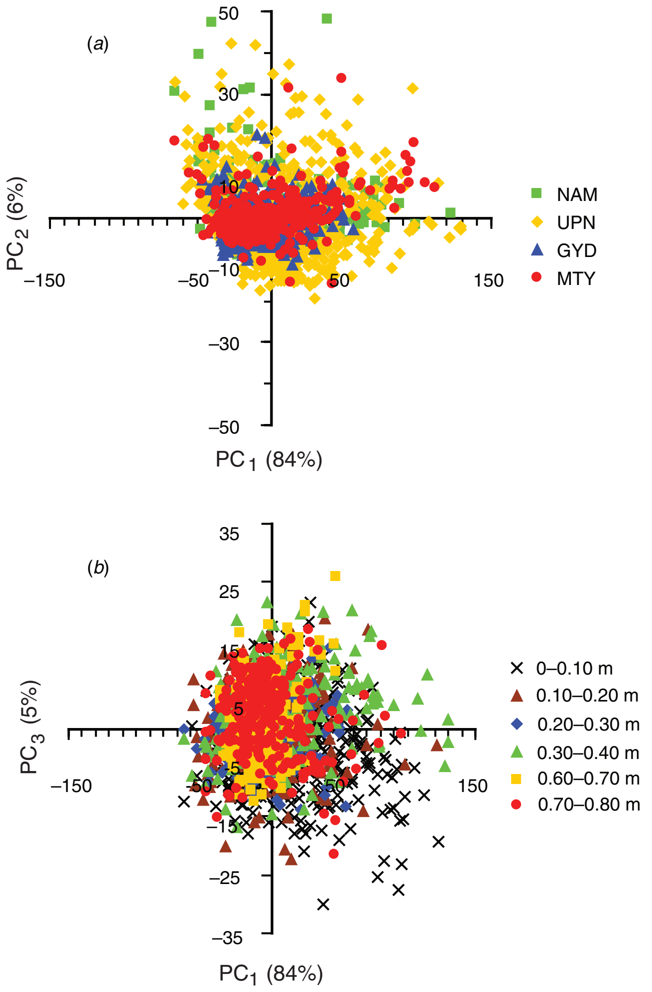

Fig. 6. Principal component scores showing the overall structure of the data. (*a*) Differences between the catchments reflect subtle differences in soil variability and composition, and (*b*) inherent differences between the layers may be primarily due to differences in their clay and organic matter contents.

using the hydrometer method described by Gee and Bauder (1986). Ten samples were presented to the laboratories incognito as blind duplicates to assess laboratory accuracy. This was done using a least-squares procedure and using their relative RMSE, SDE, and ME for statistical assessment. A comparison was then made between the statistics of the spectroscopic predictions and those from the laboratory measurements.

Using the spectral library to populate the legacy soil database

PLSR models were constructed using all 213 data. Then, using the remaining 1629 spectra in the library, bootstrap aggregation-PLSR (or bagging-PLSR) (Viscarra Rossel 2007) was used to make the soil property predictions with uncertainty, which was measured by 95% confidence intervals. Predictions were grouped by region and soil layer and were plotted together with their confidence intervals.

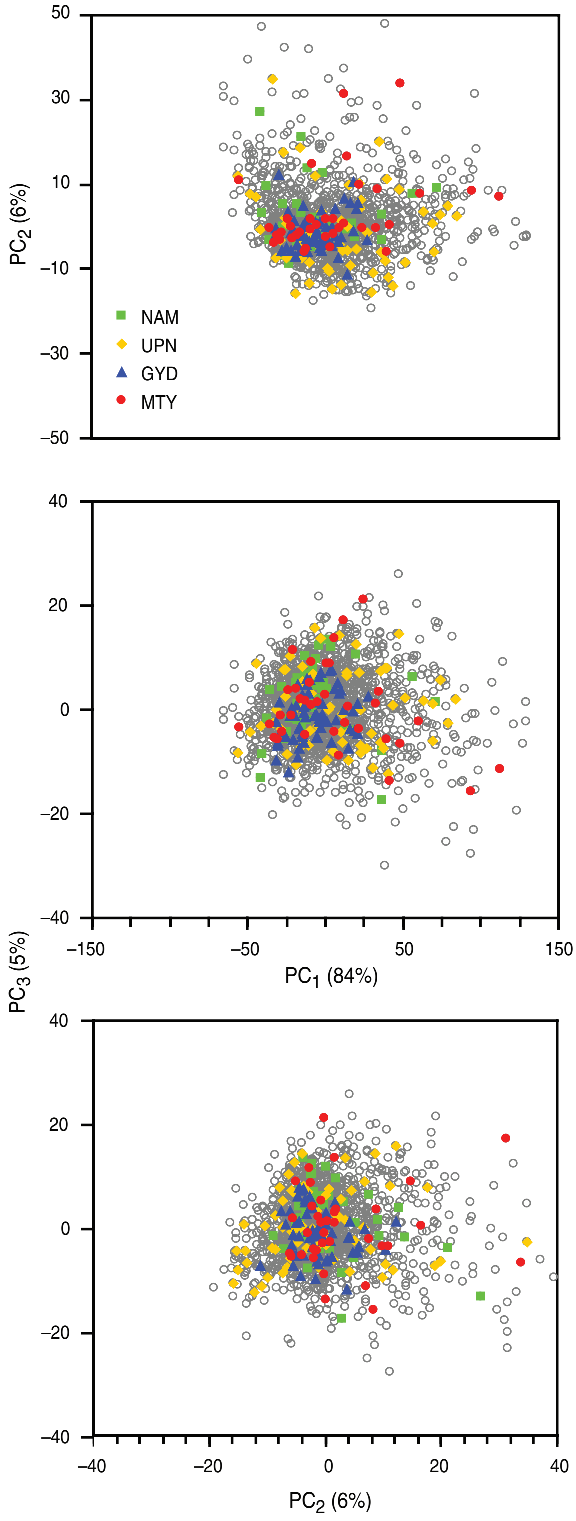

Fig. 7. Distribution in multivariate space of the 213 samples selected for laboratory analyses.

All of the chemometric analysis was performed using the software ParLeS version 3.1 (Viscarra Rossel 2008).

Results

Building the spectral library

Most of the legacy samples were classified according to the Australian Soil Classification as Vertosols, but there were also Kandosols, Dermosols, Sodosols, Tenosols, Rudosols, and Chromosols (Isbell 2002). Mostly, these soils were used for both dryland and irrigated cotton as well as for dryland wheat and other grain production. A smaller number of soils were also collected from native vegetation (Odeh *et al.* 2004). Generally, the mineral composition of the samples was made up of abundant amounts of montmorillonite (smectite) and illite producing a high CEC, but also contained kaolinite and smaller amounts of mica and interstratified minerals (Singh and Heffernan 2002). On average, the soils contained a moderate amount of clay, and had neutral pH and low amounts of organic mater. The spectra of a representative soil profile at four depths shows absorption features that are consistent with these characteristics (Fig. 5).

Table 1. Descriptive statistics of soil data
Laboratory analyses ($n = 213$)

	Units	Mean	s.d.	Median	Min.	Max.
Organic C	%	0.84	0.54	0.71	0.15	3.85
CEC	mmol_c/kg	334.2	101.3	352.1	5.5	599.4
Clay	%	53.19	13.83	55.33	8.03	86.05
Total C	%	1.15	0.53	1.07	0.21	3.87
Total N	%	0.076	0.039	0.06	0.01	0.24
Exch. Ca	mmol_c/kg	187.7	62.8	192.8	1.9	313.5
pH_{Ca}		7.24	0.73	7.5	4.5	8.5
Exch. Mg	mmol_c/kg	110.9	49.8	111.7	2.8	326.6
Total sand	%	21.44	13.91	18.42	1.39	74.61
pH_{w}		8.14	0.8	8.3	5.8	9.4
Silt	%	18.88	6.57	18.33	2.52	41.73
θ_g	g/g	0.056	0.018	0.057	0	0.131
Total P	mg/kg	232.93	165.63	188	0	1043
Exch. Na	mmol_c/kg	27.8	27.3	18.4	0.3	146.2
NH_4	mg/kg	3.69	3.09	3	1	19
ESP	%	8.34	9.21	5.53	0.22	66.8
Exch. K	mmol_c/kg	7.73	3.52	7.2	0	26
EC	dS/m	0.21	0.22	0.16	0.01	2.09
P	mg/kg	24.34	31.72	15	2	310
NO_3-N	mg/kg	12.28	15.81	6.5	1	92

Table 2. Assessment statistics for the (internal) cross validation and the (external) test set validation results of the PLSR models
Assessment statistics of the duplicate laboratory measurements shown for comparison

	Preprocessing	NF	Cross validation ($n = 153$)		Test set validation ($n = 60$)					Duplicate laboratory measurements ($n = 10$)		
			$R^2_{\text{adj.}}$	RMSE	$R^2_{\text{adj.}}$	RMSE	ME	SDE	RPD	RMSE	ME	SDE
OC %	None	13	0.92	0.14	0.91	0.15	0.03	0.14	3.22	0.16	0.09	0.14
CEC mmol_c/kg	None	9	0.91	27.4	0.89	32.5	6.3	32.1	2.86	28.12	21.20	19.50
Clay %	None	14	0.89	4.4	0.88	4.5	0.2	4.5	2.86	4.60	0.80	4.80
TC %	None	8	0.86	0.014	0.85	0.21	–0.002	0.015	2.46	0.10	0.03	0.10
Ca mmol_c/kg	SG	19	0.85	21.2	0.84	22.7	–1.91	22.8	2.51	7.71	13.10	8.00
TN %	SNV; SG; 1st der.	12	0.85	0.014	0.84	0.015	–0.002	0.015	2.50	0.010	0.010	0.010
θ_g g/g	None	9	0.79	0.008	0.78	0.008	–0.001	0.008	2.12	0.012	–0.004	0.012
Tot. sand %	None	16	0.78	6.24	0.77	6.25	–0.10	6.08	2.08	2.21	0.50	2.30
Mg mmol_c/kg	SG	17	0.78	22.6	0.75	23.0	2.23	23.1	2.03	16.74	11.60	12.80
pH_{Ca}	MSC; SG; 1st der.	9	0.73	0.38	0.7	0.41	–0.05	0.41	1.68	0.18	0.09	0.18
pH_{w}	None	6	0.7	0.43	0.69	0.46	0.08	0.46	1.64	0.23	0.15	0.18
K mmol_c/kg	SG; 1st Der.	18	0.6	2.3	0.59	2.3	–0.03	2.3	1.48	1.27	1.20	0.50
TP mg/kg	None	12	0.55	106.1	0.54	111.1	–7	111.8	1.45	38.80	29.00	29.80
Silt %	SNV	15	0.58	4.2	0.54	4.7	–0.3	4.7	1.39	4.47	0.50	4.70
ESP %	MSC; SG; 2nd der.	19	0.68	5	0.43	5.7	0.9	5.7	1.27	2.00	1.60	1.30
NH_4 mg/kg	SNV; SG; 2nd der.	7	0.5	2	0.4	2.1	–0.3	2	1.29	1.52	0.70	1.42
Na mmol_c/kg	None	11	0.43	20.1	0.39	20.3	–5	19.8	1.23	9.49	7.10	6.60
P mg/kg	SNV; SG	14	0.41	17	0.34	20.7	2.4	20.7	1.21	4.56	–2.80	3.80
EC dS/m	SNV; SG; 2nd der.	12	0.35	0.19	0.34	0.19	0.026	0.19	1.18	0.04	0.02	0.03
NO_3-N mg/kg	None	4	0.23	15.3	0.092	13.2	2.1	13.1	1.02	5.93	4.00	4.60

The absorption bands in the region between 3800 and 3600 cm^{-1} may be attributed to hydroxyl stretching vibrations of kaolinite and 2 : 1 layer alumino-silicates like montmorillonite and illite. The latter only shows poorly defined bands in this region and are often masked by kaolinite, amorphous silica, and other clay minerals. Other bands that may be attributed to clay minerals include the alumino-silicate lattice vibrations near 1020 cm^{-1} and the Al-OH deformation vibrations at 920 cm^{-1} (Nguyen *et al.* 1991). The broad band near 3400 cm^{-1} may be attributed to hydroxyl stretching vibrations of water molecules

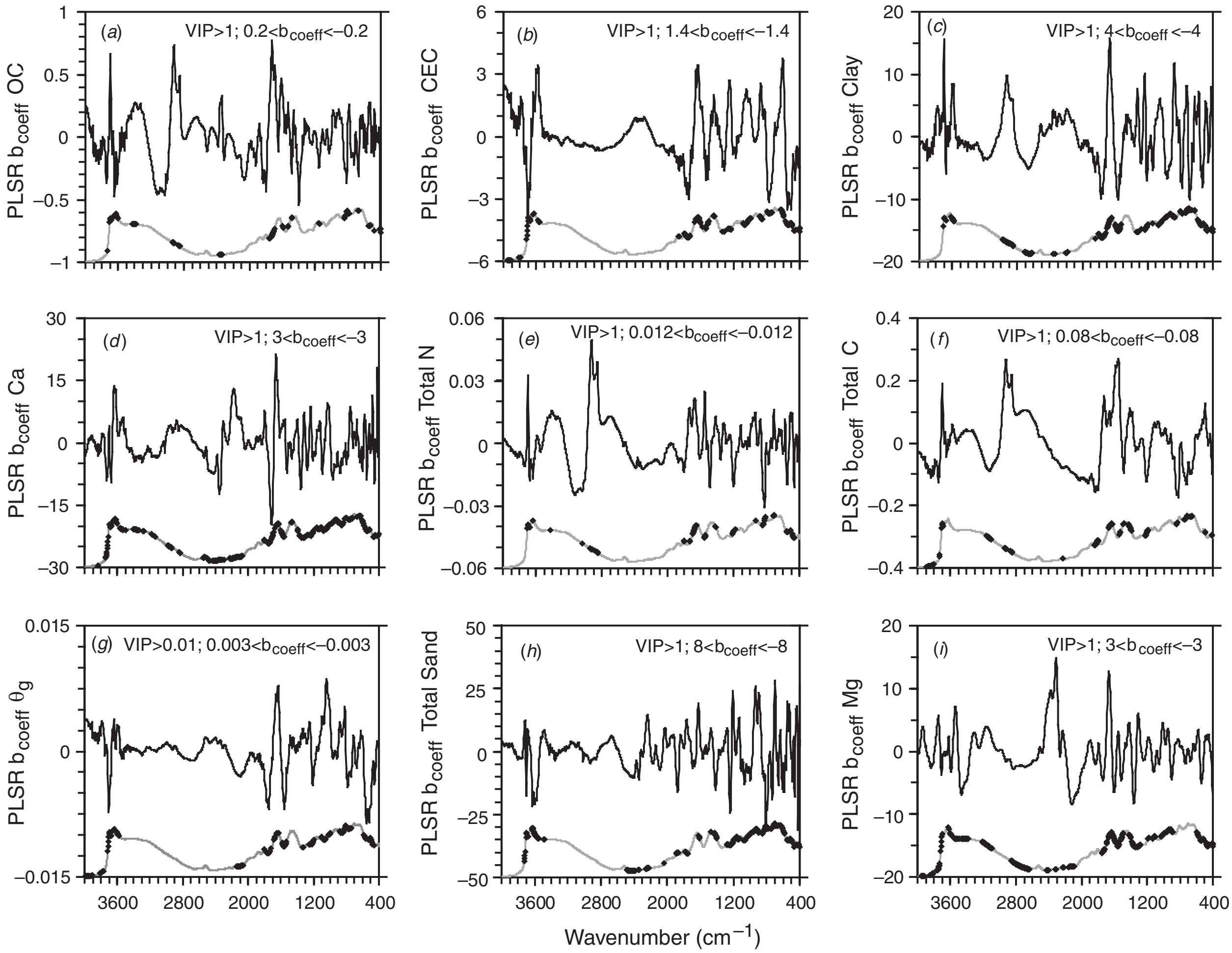

Fig. 8. PLS regression coefficient, **b**, spectra and important frequencies identified by the combined use of **b** and the variable importance for projection (VIP) for (*a*) organic C, (*b*) CEC, (*c*) clay content, (*d*) exchangeable Ca, (*e*) total N, (*f*) total C, (*g*) gravimetric moisture θ_g, (*h*) sand content, and (*i*) exchangeable Mg. For each soil property, the important frequencies in each of the models are highlighted on a sample spectrum. The thresholds used for the VIP and for the PLS regression coefficient, **b** are also shown on each graph.

Table 3. Correlation matrix for the PLS regression coefficients, b, of models with cross validation $R^2_{\text{adj.}} > 0.75$

	OC	CEC	Clay	Ca	TN	TC	θ_g	Tot. sand	Mg
OC	1.00								
CEC	–0.03	1.00							
Clay	0.07	0.38	1.00						
Ca	–0.23	0.14	0.31	1.00					
TN	0.68	–0.03	0.21	0.11	1.00				
TC	0.63	0.00	0.06	0.06	0.71	1.00			
θ_g	–0.07	0.78	0.46	0.09	–0.02	–0.07	1.00		
Tot. sand	0.09	–0.20	–0.39	–0.16	0.01	–0.05	–0.19	1.00	
Mg	–0.05	0.27	0.22	0.00	–0.15	–0.08	0.29	–0.18	1.00

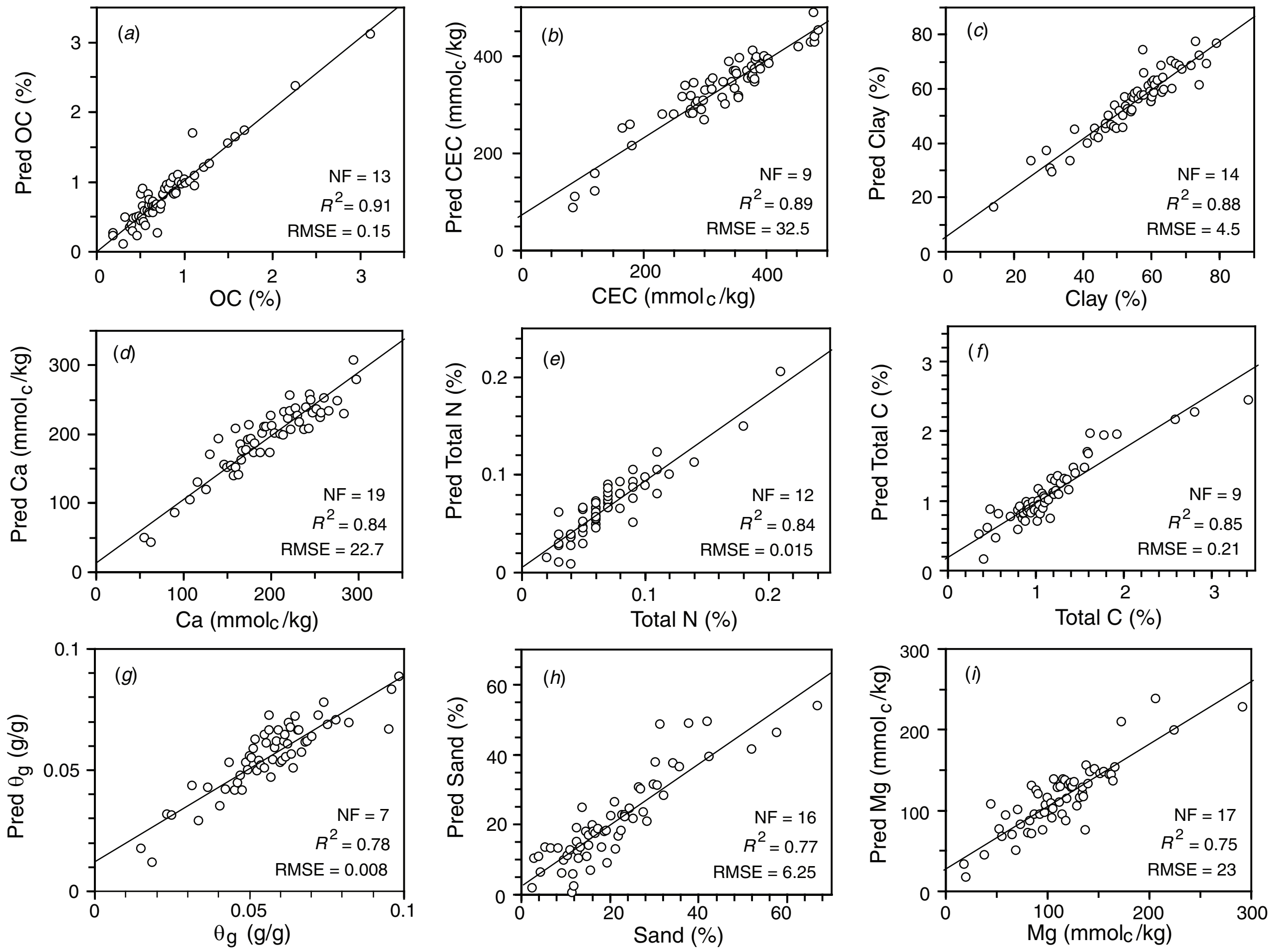

Fig. 9. Observed *v.* predicted (external) test set validations for (*a*) organic C, (*b*) CEC, (*c*) clay content, (*d*) exchangeable Ca, (*e*) total N, (*f*) total C, (*g*) gravimetric moisture θ_g, (*h*) sand content, and (*i*) exchangeable Mg.

that are in the structure of 2 : 1 minerals. The absorption bands at 2930 and 2850 cm^{-1} are particularly useful for the detection of organic matter in soils and may be attributed to alkyl material, as this region is free of overlaps or masking by other more intense vibrations. The spectra also showed absorption bands due to carboxylic acids (1725 cm^{-1}), proteins (1640 and 1530 cm^{-1}), aliphatic compounds (1465, 1445, and 1350 cm^{-1}), phenolics (1275 cm^{-1}), and carbohydrates (near 1100–1050 cm^{-1}) in soil organic matter (Skjemstad and Dalal 1987). This profile contained carbonates in the 0.30–0.40 and 0.70–0.80 m layers (Fig. 5), as shown by the absorption band at 2515 cm^{-1}, which may be attributed to either calcite or dolomite. Quartz displayed several characteristic absorption bands (Fig. 5), which include the group of 3 bands at 2000, 1870, and 1790 cm^{-1}. These are the overtones and combination bands of the fundamental vibrations at 1080, 800, and 700 cm^{-1} (Nguyen *et al.* 1991). The region between 2000 and 1600 cm^{-1} also has overtones and combination bands for other silicate structures, although these are usually masked by those of quartz.

While the intensities and distribution of peaks in mid-IR spectra may be used to qualitatively describe various soil components, the peaks are often masked by overlaps

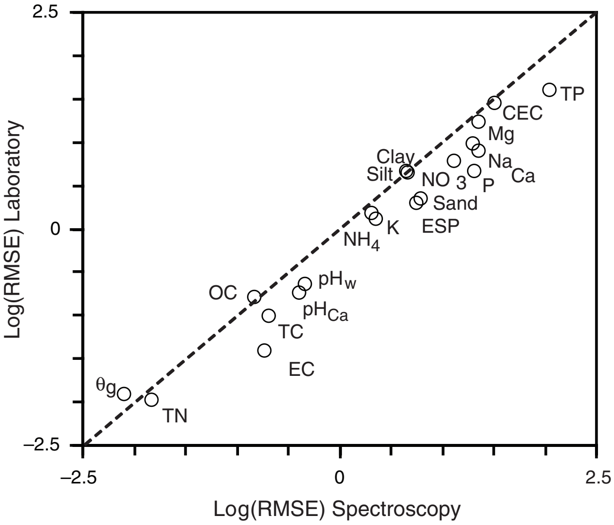

Fig. 10. Comparison of the relative accuracies between the duplicate laboratory measurements *v.* the spectroscopic predictions. Points above the 1 : 1 line show instances where spectroscopic predictions produced lower RMSEs than the laboratory measurements. Points below the line show where the laboratory measurements produced lower RMSEs.

Table 4. Mean (plain text) and standard errors (italics) for the bagging-PLSR predictions of soil properties in each catchment at every depth

Depth (m)	OC %	CEC $mmol_c/kg$	Clay %	TC %	Ca $mmol_c/kg$	TN %	$θ_g$ g/g	Tot. sand %	Mg $mmol_c/kg$	pH_{Ca}	pH_w	K $mmol_c/kg$	TP mg/kg	Silt %	ESP %	NH_4 mg kg	Na $mmol_c/kg$	EC dS/m	P mg/kg	NO_3-N mg/kg
Gwydir																				
0–0.10	0.96	353.22	54.82	1.32	216.55	0.09	0.06	19.00	112.32	7.18	7.91	10.20	193.65	19.67	5.24	4.45	17.24	0.19	20.68	16.45
	0.05	*8.79*	*1.28*	*0.08*	*8.59*	*0.01*	*0.003*	*2.19*	*7.36*	*0.11*	*0.08*	*1.05*	*35.90*	*1.47*	*2.46*	*0.54*	*5.64*	*0.07*	*9.91*	*2.32*
0.10–0.20	0.85	361.01	56.24	1.21	223.21	0.08	0.06	18.70	110.13	7.19	8.17	8.74	195.70	18.29	4.18	4.02	20.69	0.15	20.72	13.48
	0.05	*7.88*	*1.25*	*0.07*	*8.43*	*0.01*	*0.003*	*2.15*	*7.43*	*0.10*	*0.08*	*1.13*	*31.96*	*1.31*	*2.57*	*0.46*	*5.97*	*0.07*	*8.03*	*1.93*
0.20–0.30	0.75	372.94	57.69	1.09	229.84	0.07	0.06	18.54	114.29	7.35	8.32	7.34	188.34	18.32	5.64	3.42	26.77	0.16	19.01	11.36
	0.05	*8.70*	*1.33*	*0.08*	*9.00*	*0.01*	*0.003*	*2.34*	*7.38*	*0.11*	*0.08*	*1.18*	*35.95*	*1.38*	*2.74*	*0.49*	*6.50*	*0.08*	*9.77*	*1.84*
0.60–0.70	0.51	393.78	59.68	0.92	216.43	0.05	0.06	15.22	131.91	7.67	8.59	6.64	162.35	19.46	9.94	2.26	33.44	0.27	11.62	13.58
	0.05	*8.10*	*1.23*	*0.06*	*9.51*	*0.01*	*0.003*	*2.30*	*7.88*	*0.11*	*0.07*	*1.05*	*34.07*	*1.43*	*2.50*	*0.49*	*6.34*	*0.08*	*8.52*	*1.84*
McIntyre																				
0–0.10	0.94	301.58	52.52	1.14	188.82	0.07	0.05	23.01	87.52	7.19	7.99	9.08	161.65	16.59	4.77	4.64	18.95	0.16	19.29	13.43
	0.08	*12.45*	*1.80*	*0.10*	*13.11*	*0.01*	*0.004*	*3.14*	*10.78*	*0.14*	*0.11*	*1.66*	*46.22*	*1.81*	*3.57*	*0.64*	*7.87*	*0.10*	*11.67*	*3.30*
0.20–0.30	0.74	325.19	56.07	0.98	187.78	0.06	0.05	21.11	103.48	7.34	8.27	7.59	123.93	16.19	8.68	3.32	32.11	0.22	23.42	8.35
	0.06	*10.03*	*1.41*	*0.09*	*10.73*	*0.01*	*0.004*	*2.50*	*8.43*	*0.13*	*0.10*	*1.19*	*40.22*	*1.71*	*2.97*	*0.59*	*7.25*	*0.09*	*15.68*	*1.95*
0.60–0.70	0.44	333.21	58.24	0.76	169.87	0.04	0.06	17.95	111.39	7.53	8.39	5.70	76.30	16.67	12.61	1.91	38.18	0.32	14.83	10.61
	0.07	*11.95*	*1.55*	*0.10*	*12.20*	*0.01*	*0.004*	*2.57*	*9.21*	*0.13*	*0.09*	*1.41*	*42.52*	*1.60*	*3.12*	*0.61*	*9.67*	*0.10*	*11.49*	*3.29*
Namoi																				
0–0.10	1.04	336.27	52.94	1.29	195.92	0.09	0.06	22.44	101.91	7.21	8.17	10.24	265.72	17.41	4.35	5.34	18.68	0.13	28.98	17.15
	0.07	*14.03*	*1.73*	*0.11*	*14.89*	*0.01*	*0.004*	*2.76*	*10.03*	*0.13*	*0.11*	*1.65*	*41.15*	*1.77*	*3.20*	*0.64*	*8.37*	*0.08*	*11.18*	*3.18*
0.30–0.40	0.54	349.15	54.43	0.84	176.14	0.05	0.06	21.73	121.22	7.49	8.55	7.21	228.40	16.88	12.34	2.5	42.39	0.26	22.10	7.82
	0.06	*9.63*	*1.31*	*0.08*	*10.81*	*0.01*	*0.003*	*2.41*	*7.89*	*0.13*	*0.09*	*1.42*	*37.09*	*1.66*	*3.30*	*0.52*	*7.58*	*0.09*	*9.61*	*1.84*
0.70–0.80	0.35	344.69	53.79	0.69	159.77	0.04	0.06	21.54	121.86	7.56	8.61	5.91	220.27	18.75	17.53	1.52	51.74	0.38	18.50	6.56
	0.07	*9.49*	*1.39*	*0.09*	*10.85*	*0.01*	*0.003*	*2.54*	*8.37*	*0.12*	*0.09*	*1.35*	*38.45*	*1.67*	*3.24*	*0.51*	*8.28*	*0.09*	*9.66*	*2.03*
Upper Namoi																				
0–0.10	1.68	297.27	45.69	1.82	178.55	0.12	0.05	29.66	105.34	6.46	7.44	8.96	371.50	18.70	1.33	6.46	74.70	0.04	33.55	17.40
	0.11	*14.28*	*2.22*	*0.13*	*16.72*	*0.01*	*0.005*	*3.75*	*13.02*	*0.20*	*0.12*	*2.10*	*63.99*	*2.48*	*4.61*	*0.93*	*10.29*	*0.12*	*16.94*	*3.93*
0.10–0.20	1.09	314.16	47.94	1.37	177.25	0.08	0.05	28.18	115.35	6.88	7.81	6.52	332.94	18.32	3.28	3.73	16.10	0.07	27.59	12.42
	0.10	*13.86*	*2.09*	*0.11*	*15.35*	*0.01*	*0.006*	*3.56*	*12.67*	*0.19*	*0.11*	*1.91*	*62.41*	*2.40*	*4.55*	*0.83*	*9.45*	*0.12*	*17.80*	*3.26*
0.30–0.40	0.76	338.67	54.15	1.10	179.19	0.06	0.06	24.28	130.38	7.32	8.28	5.25	303.16	16.15	7.21	2.05	28.28	0.19	14.93	8.09
	0.11	*15.41*	*2.28*	*0.13*	*17.93*	*0.01*	*0.006*	*4.20*	*14.40*	*0.20*	*0.13*	*2.37*	*68.94*	*2.59*	*4.89*	*0.84*	*11.17*	*0.14*	*16.57*	*3.32*
0.70–0.80	0.45	359.06	55.82	0.84	163.68	0.04	0.06	23.80	140.38	7.83	8.84	3.79	353.02	15.53	12.75	0.66	47.77	0.30	5.96	4.09
	0.13	*16.38*	*2.66*	*0.13*	*20.60*	*0.01*	*0.005*	*5.20*	*16.90*	*0.24*	*0.15*	*2.43*	*77.39*	*3.11*	*5.77*	*0.90*	*13.81*	*0.15*	*18.81*	*3.38*

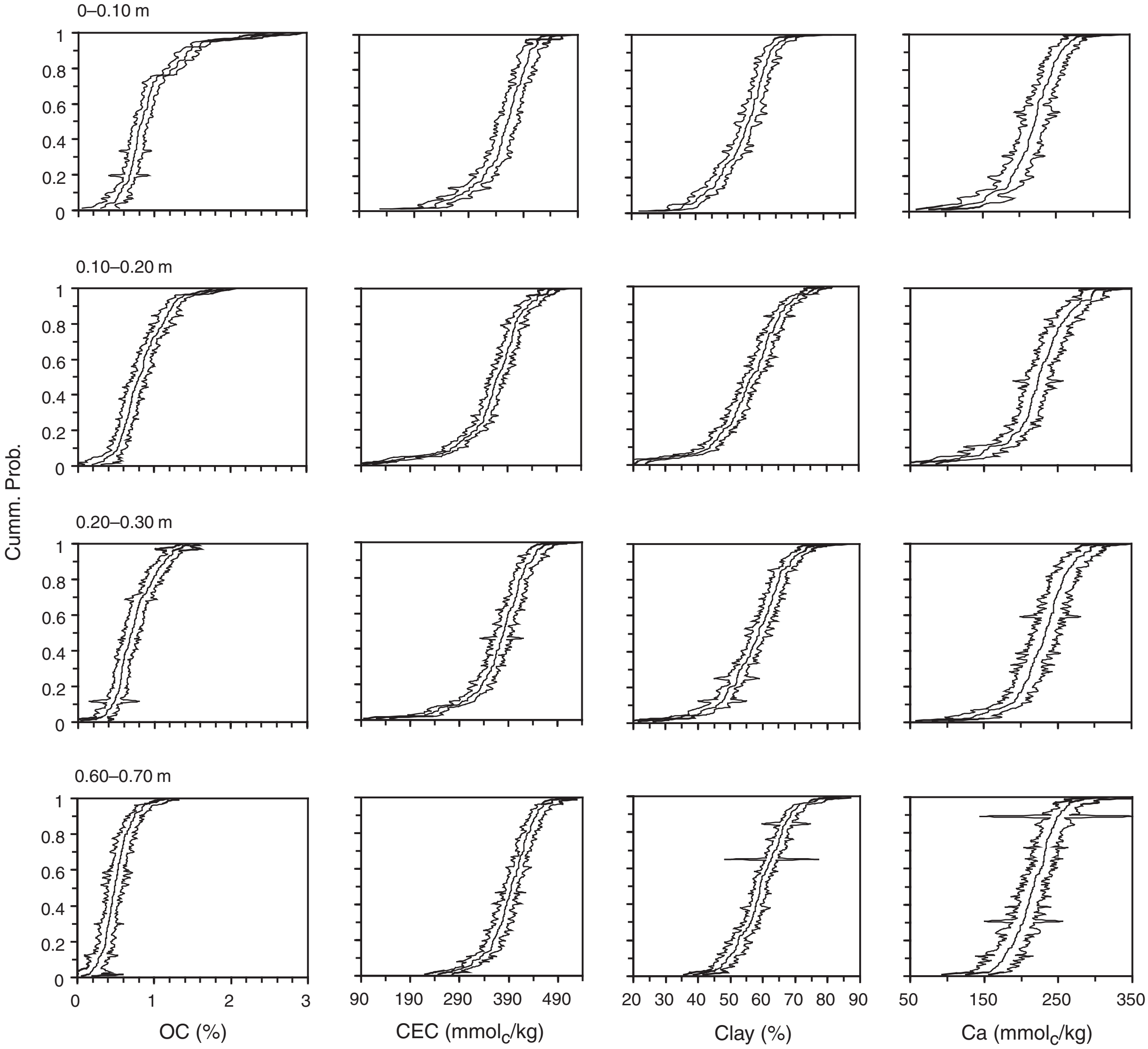

Fig. 11. CDF of bagging-PLSR predictions of soil organic C (OC), CEC, clay content, and exchangeable Ca with uncertainty (measured by 95% confidence intervals) for soil in the Gwydir catchment.

from fundamentals, combinations, and overtones of other soil components, making this type of analysis very difficult. Low concentrations of spectrally active components in soil make this even harder. It is mainly for these reasons that the use of chemometrics and multivariate calibrations are essential to obtain quantitative information from soil spectra.

Principal component analysis

The first 5 principal components (PC) accounted for 98% of the variation in the spectra, and 130 samples were identified as outliers using the Mahalonobis statistic. After rescanning, only 36 spectra were confirmed to be outliers and were left out of the analysis for further investigation. Scores plots showing the overall structure of the data are shown in Fig. 6.

Figure 6*a* shows that there were only small differences between the soil spectra of the catchments, reflecting subtle differences in both their variability and composition. For example, soils from the McIntyre and the Gwydir catchments are spectrally more similar and less variable than those from the (lower) Namoi and upper Namoi valleys. Similarly, Fig. 6*b* shows that there is more variation in the spectra of the topsoil compared with the subsoil, reflecting the inherent differences between the layers, which may be primarily due to differences in their clay and organic matter contents and associated properties.

Selection of soil samples for laboratory analysis

The 1842 (1878—36 outliers) mid-IR spectra were combined with the corresponding soil data from the legacy database. Multivariate calibrations were developed for soil OC, clay content, CEC, and pH$_w$. When these models were tested, the results were much poorer than we expected with $R^2_{adj.} < 0.6$ and RMSE values of 0.8% OC, 20% clay content, 150 mmol$_c$/kg CEC, and 1.1 pH$_w$ units. We believe that the reason for these poor results were various sources of error. Some of

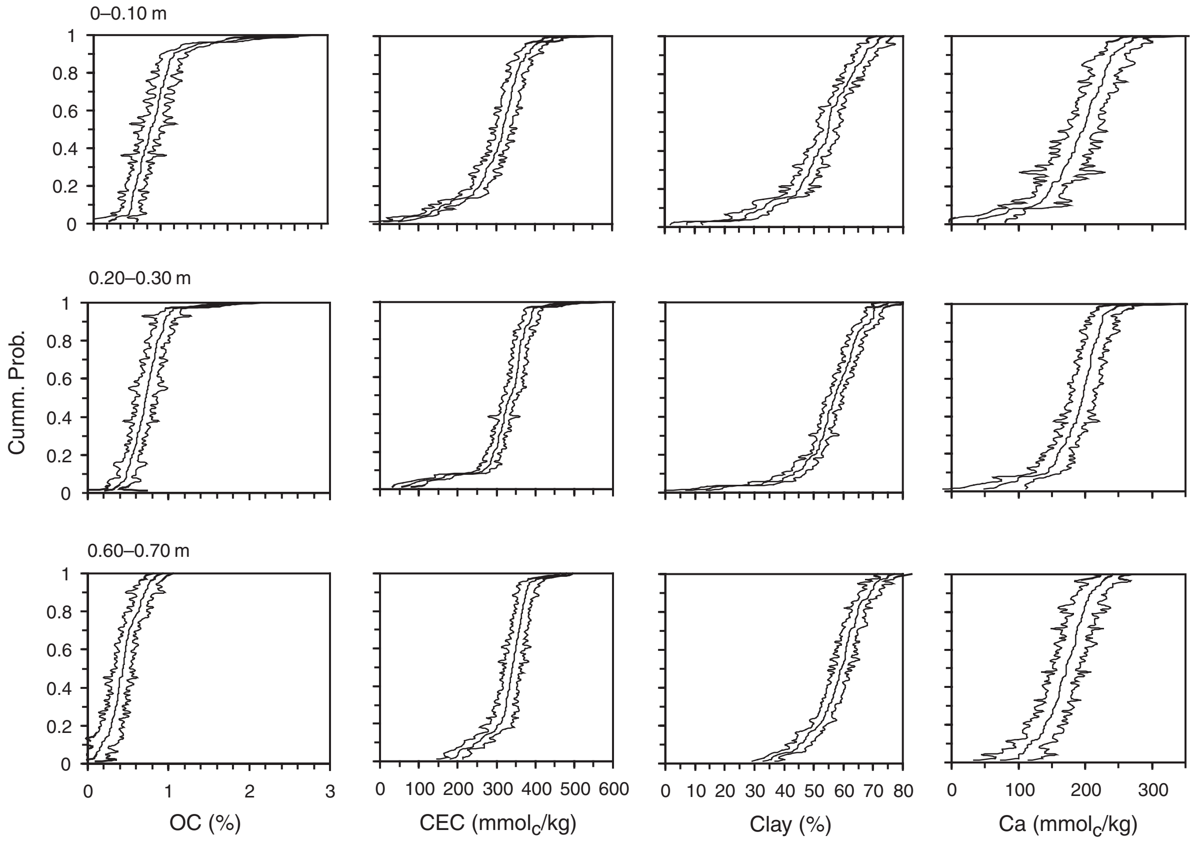

Fig. 12. CDF of bagging-PLSR predictions of soil organic C (OC), CEC, clay content, and exchangeable Ca with uncertainty (measured by 95% confidence intervals) for soil in the McIntyre catchment.

these may include: (*i*) changes in soil properties between the times of laboratory analyses and spectral measurements, (*ii*) differences in the subsamples used for laboratory analyses and spectral measurements, (*iii*) the imprecision (particularly the low reproducibility) of the analytical techniques used, and (*iv*) the different analytical techniques that may have been used over several years to analyse the same soil property. Thus, we needed to select several samples to reanalyse so that they may form the basis for our spectral library. The number of samples that are needed in a spectral library for accurate spectroscopic prediction of soil properties will depend on the extent of soil variability in the region in which the library is to be used, the spectral sampling technique, and the budget. The distribution in multivariate space of the 213 samples selected for laboratory analyses using the cLHS procedure is shown in Fig. 7.

The soil samples selected were representative of the entire spectral population as they were well distributed in PC space (Fig. 7). The samples included 71 soils from the Gwydir, 55 from the upper Namoi, 42 from the Namoi region of NSW, and 71 soils from the McIntyre region of southern Queensland. The statistical description of the laboratory data is given in Table 1.

The properties of the 213 soil samples had wide ranging distributions because the samples originated from top- and subsoils of various soil types occurring in different regions and exposed to various land uses. Soil pH_{Ca} ranged from 4.5 to 8.5 units, OC content from 0.15 to 3.85%, clay content from 8 to 86%, and CEC from 5.5 to 599 mmol$_c$/kg. On average, the soils had a neutral pH, contained only 0.8% organic carbon, 53% clay, and had a CEC of 334 mmol$_c$/kg.

Multivariate calibration using partial least-squares regression (PLSR)

Table 2 shows the cross validation results and the number of PLSR factors and spectral preprocessing applied to the models for the predictions of the soil properties. Figure 8 shows the PLS regression coefficient, **b**, spectra, and the important mid-IR frequencies selected by the combined use of **b** and the VIP, for soil properties with a cross validation $R^2_{\text{adj.}} > 0.75$.

The important frequencies for predictions of soil OC include those near 3400 cm^{-1} which may be attributed to OH and NH stretching vibrations; those near 2930 and 2850 cm^{-1} may be attributed to alkyl$-CH_2$ asymmetric and symmetric stretches; near 1725 cm^{-1} attributed to carboxylic acid and ketones; between 1600 and 1400 cm^{-1} attributed to amides, aromatics, aliphatic acids, and alkyl groups of organic materials in soil; and those near 1100 cm^{-1} attributed to carbohydrates and sugars (Fig. 8*a*). Similar frequencies were identified as important for predictions of total C (Fig. 8*f*) and total N (Fig. 8*e*), as these properties are correlated. The important frequencies for the CEC model (Fig. 8*b*) are similar to those that are diagnostic for clay (Fig. 8*c*). These include those near 3700–3500 cm^{-1}, which may be attributed to Al$-$OH stretching vibrations and

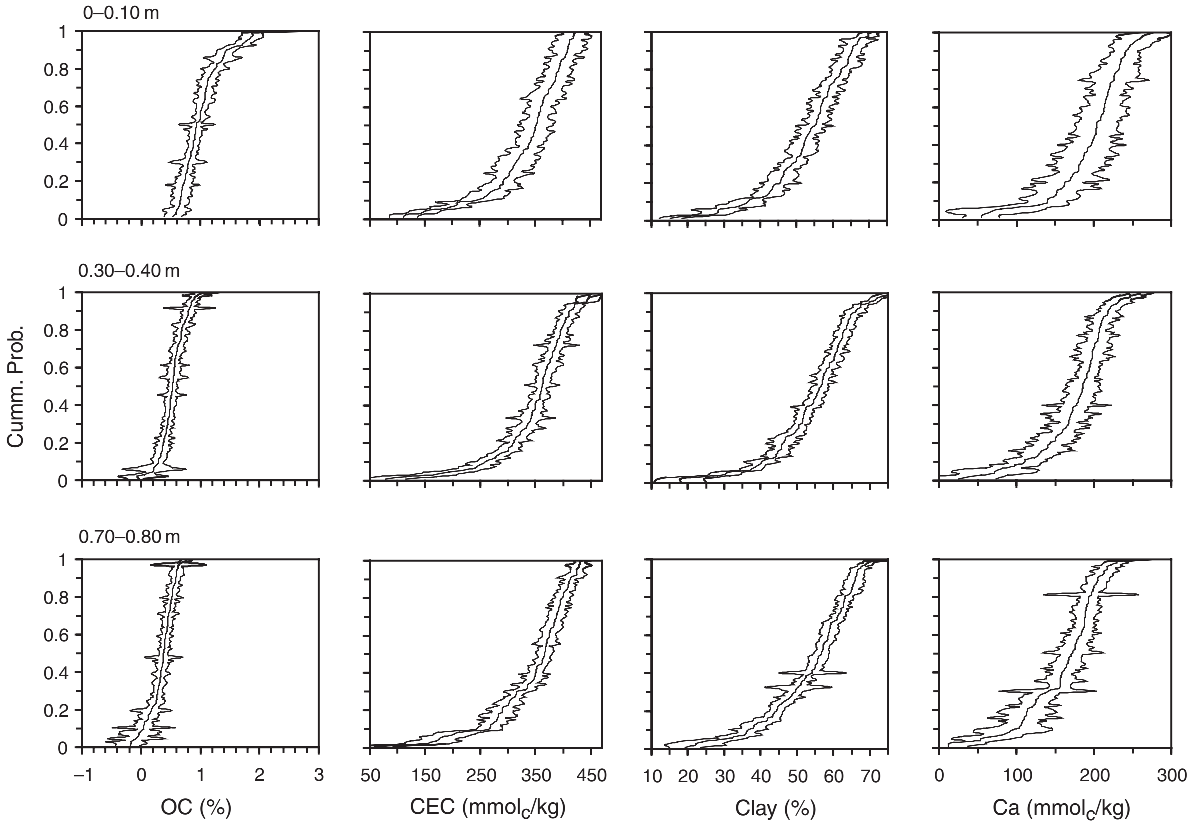

Fig. 13. CDF of bagging-PLSR predictions of soil organic C (OC), CEC, clay content, and exchangeable Ca with uncertainty (measured by 95% confidence intervals) for soil in the Namoi catchment.

others in the fingerprint region, particularly those near 1020 and 920 cm^{-1}, which may represent the Al–OH lattice vibrations and deformation vibrations, respectively, in clay minerals. The important frequencies for predictions of total sand (Fig. 8*h*) are those between 2100 and 1600 cm^{-1}, which may be attributed to the weak overtones and combination bands for quartz and other silicate structures, as well as those in the fingerprint region between 1200 and 400 cm^{-1}, particularly those near 780–800 and 700 cm^{-1} (Nguyen *et al.* 1991). Notice that the inverse relationship between clay content and total sand content is displayed in their **b** spectra. For example, while the **b** spectra of total sand shows negative coefficients near 3700–3500 cm^{-1} and a strong needle-like positive peak at 700 cm^{-1} (Fig. 8*h*), the opposite is true for the coefficients of clay (Fig. 8*c*). The important frequencies for the θ_g model (Fig. 8*g*) are similar to those for clay, CEC, and total sand. Important frequencies for predictions of exchangeable Ca and Mg (Fig. 8*d*, *i*) include those that are diagnostic of calcite and dolomite. For example, the important frequencies for exchangeable Ca (Fig. 8*d*) are those near 3000, 2900, 2515, and 1800 cm^{-1}, all of which are diagnostic peaks for calcite (Nguyen *et al.* 1991). The important frequencies for exchangeable Mg (Fig. 8*i*) are those near 1440, 1470, and 875 cm^{-1} (Fig. 8*i*), which are characteristic of carbonate species and may be due to presence of $MgCO_3$ and dolomite. It should also be noted that other important frequencies for the predictions of Ca and Mg are the same as those for clay and organic matter, indicating their associations to these soil properties. Although the identification of carbonate peaks in the absorbance spectra (Fig. 5) was somewhat difficult because of the masking and overlap caused by other soil components, they are evident in the **b** spectra of Ca and Mg (Fig. 8*d*, *i*). Table 3 shows the correlation between the PLS regression coefficients, **b**, for these soil properties.

The assessment statistics for the PLSR models validated (externally) using independent test data are shown in Table 2. Very good calibrations (RPD > 2) were obtained, in decreasing order, for soil OC, CEC, clay content, total carbon, exchangeable Ca, total nitrogen, θ_g, total sand, and exchangeable Mg (Table 2). Modest calibrations (2 > RPD > 1.5) were obtained for pH_{Ca} and pH_w. Calibrations for exchangeable K, total P, silt content, ESP, NH_4, exchangeable Na, P, EC, and NO_3-N were poor (RPD < 1.5), with EC and NO_3-N being the worst, respectively (Table 2). These results are similar to those reported in other spectroscopic studies of Australian soils (e.g. Janik *et al.* 1998; Viscarra Rossel *et al.* 2006*a*). Figure 9 shows the test set validations for models with an RPD > 2.

Predictions of soil OC, total N, and total C produced RMSE values of 0.15, 0.015, and 0.21% (Fig. 9*a*, *e*, *f*), respectively; CEC and exchangeable Ca and Mg produced RMSE values of 32.5, 22.7, and 23.0 mmol$_c$/kg (Fig. 9*b*, *d*, *i*), respectively; and predictions of clay content, θ_g, and total sand content produced RMSE values of 4.4%, 0.008 g/g, and 6.3%

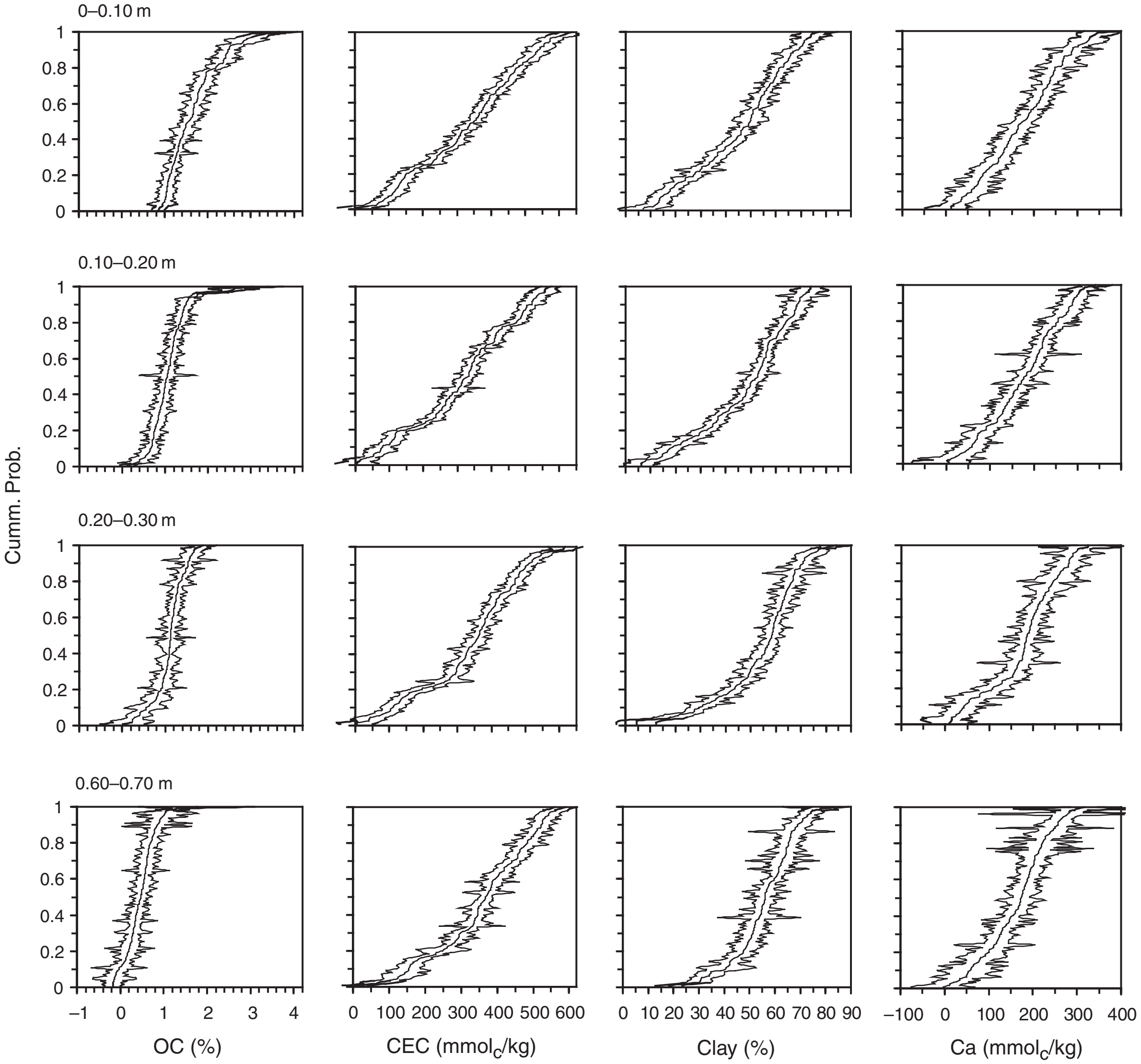

Fig. 14. CDF of bagging-PLSR predictions of soil organic C (OC), CEC, clay content, and exchangeable Ca with uncertainty (measured by 95% confidence intervals) for soil in the upper Namoi catchment.

(Fig. 9*c*, *g*, *h*), respectively. These results can be compared in Table 2 also.

Comparing spectroscopic v. laboratory analysis

Assessment statistics for the spectroscopic predictions and the laboratory measurements are given in Table 2. The relative accuracies of the techniques are compared in Fig. 10.

In Fig. 10, points above the 1 : 1 line show instances where the spectroscopic predictions produced lower RMSEs than the duplicate laboratory measurements, while points plotting below the 1 : 1 line show where the laboratory measurements produced lower RMSEs. Points nearer to the 1 : 1 line indicate greater similarity between the spectroscopic and analytical results. Spectroscopic predictions of soil OC, clay content, and θ_g had lower RMSEs than those from the laboratory replicates (Fig. 10). The RMSE values of the laboratory analyses for CEC, exchangeable Mg, TN, pH_{Ca}, and NH_4 were lower but similar to the spectroscopic predictions (Fig. 10). For all of the other soil properties, the laboratory analyses produced significantly lower RMSEs (Table 2).

Predictions of soil properties with uncertainty

Bagging-PLSR was used to predict soil properties with uncertainty for the entire library, i.e. using the remaining 1629 spectra. Means and standard errors for these predictions, for each region and each soil layer are given in Table 4.

From Table 4, on average in each region, soil OC, total C, and total N were low and decreased with depth while CEC, exchangeable Ca, Mg, Na, and soil pH increased with depth. Clay and sand contents were inversely related and, together with silt, were relatively uniform with increasing depth. N, P, and K decreased with depth, while EC and ESP increased down the profile (Table 4). To illustrate the results, bagging-PLSR predictions of soil OC, CEC, clay content, and exchangeable Ca, and their uncertainty (measured by 95% confidence intervals), for each region and each soil layer are shown in Figs 11–14.

The OC content of the soil in the four catchments was generally low and it decreased with depth (Figs 11–14). Predictions of OC were highest for the 0−0.10 m layer of the upper Namoi, where OC ranged from 0.85 to 4.08%. In the Namoi, 2 soils samples in the 0.30–0.40 m layer and 9 samples in the 0.60–0.70 m layer predicted negative values (Fig. 13). Predictions of OC in the 0.60–0.70 m layer ranged from –0.18 to 0.77%. In the 0.70–0.80 m layer of the upper Namoi, 14 samples predicted negative values such that soil OC ranged from –0.17 to 1.75% (Fig. 14). On average, prediction uncertainty for the upper Namoi was 0.22%, for the McIntyre it was 0.14%, for the Namoi it was 0.13% and for the Gwydir it was 0.1% (Table 4).

Cation exchange capacity and clay content in the four catchments tended to increase slightly with depth (Figs 11–14). In the upper Namoi, predictions of CEC and clay content in all 4 layers had wide-ranging bimodal distributions (Fig. 14). In the 0–0.10 m layer of the upper Namoi samples CEC ranged from 4.5 to 636 mmol$_c$/kg and the range of clay content was 6.9–77.4%. The average CEC and clay content of the larger population of soils, presumably being mostly Vertosols, was 408 mmol$_c$/kg and 53%, respectively. The CEC and clay content of the remaining samples was 125 mmol$_c$/kg and 17%, respectively. Predictions of CEC and clay content for the Gwydir, McIntyre and Namoi samples also had wide ranging distributions, where a small proportion of samples had much lower values (Figs 11–14). This is indicative of the presence of different soil types in the catchments (see above). On average, prediction uncertainty for CEC in the Upper Namoi and in the McIntyre was 30 mmol$_c$/kg, in the Namoi it was 22 mmol$_c$/kg and in the Gwydir it was 17 mmol$_c$/kg (Table 4). Prediction uncertainty for clay content in the upper Namoi was 2.3%, in the McIntyre it was 1.6%, in the Namoi it was 1.5% and in the Gwydir it was 1.3%. In each of the catchments, predictions of exchangeable Ca and their uncertainty followed similar trends to those of clay content and CEC (Figs 11–14).

Conclusions

We developed a soil mid-IR spectral library for cotton-growing soils of eastern Australia from a legacy soil sample. The conditioned Latin hypercube sampling scheme used to sample the spectral data space was effective as it selected a representative set of 213 samples for laboratory analyses. These samples formed the basis for the spectroscopic predictions of the soil properties. Partial least-squares regression produced interpretable models that were robust and accurate for predictions of soil OC, CEC, clay content, exchangeable Ca, total N, total C, θ_g, total sand content, and exchangeable Mg. The RMSEs of the spectroscopic predictions were slightly lower than those from duplicate laboratory analysis for soil OC, clay content, and θ_g. Hence for this range of soils the more efficient mid-IR-PLSR technique may replace the more conventional analytical techniques for these soil properties. Bagging-PLSR produced predictions with uncertainty. These were used to repopulate the legacy soil database with good quality information.

Acknowledgments

We wish to acknowledge the Cotton Catchment and Communities CRC (CCC CRC) and the Grains Research and Development Corporation (GRDC) for their financial support.

References

Barnes RJ, Dhanoa MS, Lister SJ (1989) Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy* **43**, 772–777. doi: 10.1366/0003702894202201

Ben-Dor E, Levin N, Singer A, Karnieli A, Braun O, Kidron GJ (2006) Quantitative mapping of the soil rubification process on sand dunes using an airborne hyperspectral sensor. *Geoderma* **131**, 1–21. doi: 10.1016/j.geoderma.2005.02.011

Bowers SA, Hanks RJ (1965) Reflection of radiant energy from soils. *Soil Science* **100**, 130–138.

Brooks FA (1952) Atmospheric radiation and its reflection from the ground. *Journal of Meteorology* **9**, 41–52.

Brown DJ, Shepherd KD, Walsh MG, Dewayne Mays M, Reinsch TG (2006) Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* **132**, 273–290. doi: 10.1016/j.geoderma.2005.04.025

Chang C-W, Laird DA, Mausbach MJ, Hurburgh CR Jr (2001) Near-infrared reflectance spectroscopy-principal components regression analysis of soil properties. *Soil Science Society of America Journal* **65**, 480–490.

Chong IG, Jun CH (2005) Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* **78**, 103–112. doi: 10.1016/j.chemolab.2004.12.011

Cohen MJ, Shepherd KD, Walsh MG (2005) Empirical reformulation of the universal soil loss equation for erosion risk assessment in a tropical watershed. *Geoderma* **124**, 235–252. doi: 10.1016/j.geoderma.2004.05.003

Dalal RC, Henry RJ (1986) Simultaneous determination of moisture, organic carbon and total nitrogen by near infrared reflectance spectrophotometry. *Soil Science Society of America Journal* **50**, 120–123.

De Maesschalck R, Jouan-Rimbaud D, Massart DL (2000) The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* **50**, 1–18. doi: 10.1016/S0169-7439(99)00047-7

Demattê JAM, Campos RC, Alves MC, Fiorio PR, Nanni MR (2004) Visible–NIR reflectance: a new approach on soil evaluation. *Geoderma* **121**, 95–112. doi: 10.1016/j.geoderma.2003.09.012

Dunn BW, Beecher HG, Batten GD, Ciavarella S (2002) The potential of near-infrared reflectance spectroscopy for soil analysis – a case study from the Riverine Plain of south-eastern Australia. *Australian Journal of Experimental Agriculture* **42**, 607–614. doi: 10.1071/EA01172

Gee GW, Bauder JW (1986) Particle size analysis. In 'Methods of soil analysis. Part I'. 2nd edn, Agronomy Monograph 9. (Ed. A Klute) pp. 383–411. (ASA and SSSA: Madison, WI)

de Gruijter J, Brus DJ, Bierkens MFP, Knotters M (2006) 'Sampling for natural resource monitoring.' (Springer Verlag)

Haaland DM, Thomas EV (1988) Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry* **60**, 1193–1202. doi: 10.1021/ac00162a020

Isbell RF (2002) 'The Australian Soil Classification.' (CSIRO Publishing: Collingwood, Vic.)

Janik LJ, Merry RH, Skjemstad JO (1998) Can mid infra-red diffuse reflectance analysis replace soil extractions? *Australian Journal of Experimental Agriculture* **38**, 681–696. doi: 10.1071/EA97144

Janik LJ, Skjemstad JO (1995) Characterisation and analysis of soils using mid-infrared partial least squares. II. Correlations with some laboratory data. *Australian Journal of Soil Research* **33**, 637–650. doi: 10.1071/SR9950637

Madari BE, Reeves JB III, Machado PLOA, Guimarães CM, Torres E, McCarty GW (2006) Mid- and near-infrared spectroscopic assessment of soil compositional parameters and structural indices in two Ferralsols. *Geoderma* **136**, 245–259. doi: 10.1016/j.geoderma.2006.03.026

Martens H, Næs T (1989) 'Multivariate calibration.' (John Wiley & Sons: Chichester, UK)

Masserschmidt I, Cuelbas CJ, Poppi RJ, De Andrade JC, De Abreu CA, Davanzo CU (1999) Determination of organic matter in soils by FTIR/diffuse reflectance and multivariate calibration. *Journal of Chemometrics* **13**, 265–273. doi: 10.1002/(SICI)1099-128X(199905/08)13:3/4<265::AID-CEM552>3.0.CO;2-E

McBratney AB, Minasny B, Viscarra Rossel RA (2006) Spectral soil analysis and inference systems: a powerful combination for solving the soil data crisis. *Geoderma* **136**, 272–278. doi: 10.1016/j.geoderma.2006.03.051

Minasny B, McBratney AB (2006) A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences* **32**, 1378–1388. doi: 10.1016/j.cageo.2005.12.009

Mouazen AM, Maleki MR, De Baerdemaeker J, Ramon H (2007) On-line measurement of some selected soil properties using a VIS–NIR sensor. *Soil and Tillage Research* **93**, 13–27. doi: 10.1016/j.still.2006.03.009

Nguyen TT, Janik LJ, Raupach M (1991) Diffuse reflectance infrared fourier transform (DRIFT) spectroscopy in soil studies. *Australian Journal of Soil Research* **29**, 49–67. doi: 10.1071/SR9910049

Odeh IOA, Stephen CR, Triantafilis J, McBratney AB, Taylor J (2004) A simple soil database management assistant for Australian cotton industry. In 'Proceedings of Australian and New Zealand Soils Conference – SuperSoil 2004'. The University of Sydney, NSW. CD-ROM. (ASSSI: Sydney)

Rayment GE, Higginson FR (1992) 'Australian laboratory handbook of soil and water chemical methods.' (Inkata Press: Melbourne)

Reeves JB III, McCarty GW, Meisinger JJ (1999) Near infrared reflectance spectroscopy for the analysis of agricultural soils. *Journal of Near Infrared Spectroscopy* **7**, 179–193.

Savitzky A, Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **36**, 1627–1639. doi: 10.1021/ac60214a047

Shepherd KD, Walsh MG (2002) Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society of America Journal* **66**, 988–998.

Shibusawa S, Imade Anom SW, Sato S, Sasao A, Hirako S (2001) Soil mapping using the real-time soil spectrophotometer. In 'ECPA 2001. Third European Conference on Precision Agriculture. Vol. 1'. (Eds G Grenier, S Blackmore) pp. 497–508. (Agro: Montpellier)

Singh B, Heffernan S (2002) Layer charge characteristics of smectites from Vertosols (Vertisols) of New South Wales. *Australian Journal of Soil Research* **40**, 1159–1170. doi: 10.1071/SR02017

Skjemstad JO, Dalal RC (1987) Spectroscopic and chemical differences in organic matter of two Vertisols subjected to long periods of cultivation. *Australian Journal of Soil Research* **25**, 323–335. doi: 10.1071/SR9870323

Stenberg B, Jonnson A, Borjesson T (2002) Near infrared technology for soil analysis with implications for precision agriculture. In 'Near Infrared Spectroscopy: Proceedings of the 10th International Conference'. (Eds A Davies, R Cho) pp. 279–284. (NIR Publications: Chichester, UK)

Stenberg B, Nordkvist E, Salomonsson L (1995) Use of near infrared reflectance spectra of soils for objective selection of samples. *Soil Science* **159**, 109–114. doi: 10.1097/00010694-199502000-00005

Vågen TG, Shepherd KD, Walsh MG (2006) Sensing landscape level change in soil fertility following deforestation and conversion in the highlands of Madagascar using Vis-NIR spectroscopy. *Geoderma* **133**, 281–294. doi: 10.1016/j.geoderma.2005.07.014

Viscarra Rossel RA (2007) Robust modelling of soil diffuse reflectance spectra by 'bagging-'PLSR'. *Journal of Near Infrared Spectroscopy* **15**, 39–47. doi: 10.1255/jnirs.694

Viscarra Rossel RA (2008) ParLeS: Software for chemometric analysis of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems* **90**, 72–83. doi: 10.1016/j.chemolab.2007.06.006

Viscarra Rossel RA, McBratney AB (2008) Diffuse reflectance spectroscopy as a tool for digital soil mapping. In 'Digital soil mapping with limited data'. Developments in Soil Science series. (Eds AE Hartemink, AB McBratney, L Mendonça-Santos) (Elsevier Science: Amsterdam)

Viscarra Rossel RA, McGlynn RN, McBratney AB (2006*b*) Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. *Geoderma* **137**, 70–82. doi: 10.1016/j.geoderma.2006.07.004

Viscarra Rossel RA, Walvoort DJJ, McBratney AB, Janik LJ, Skjemstad JO (2006*a*) Visible, near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **131**, 59–75. doi: 10.1016/j.geoderma.2005.03.007

Wetterlind J, Stenberg B, Soderstrom M (2007) Farmsoil mapping using NIR-technique for increased sample point density. In 'Precision Agriculture '07'. (Ed. JV Stafford) pp. 265–270. (Wageningen Academic Publishers: AE Wageningen, The Netherlands)

Williams PC (1987) Variables affecting near-infrared reflectance spectroscopic analysis. In 'Near-infrared technology in the agricultural and food industries'. (Eds P Williams, K Norris) pp. 143–166. (American Association of Cereal Chemists: St Paul, MN)

Wold S, Martens H, Wold H (1983) The multivariate calibration method in chemistry solved by the PLS method. In 'Proceedings of the Conference on Matrix Pencils, Lecture Notes in Mathematics'. (Eds A Ruhe, B Kagstrom) pp. 286–293. (Springer-Verlag: Heidelberg)

Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**, 109–130. doi: 10.1016/S0169-7439(01)00155-1

Manuscript received 13 July 2007, accepted 21 November 2007