European Journal of Soil Science

European Journal of Soil Science, June 2009, **60,** 453–464 doi: 10.1111/j.1365-2389.2009.01121.x

Improved analysis and modelling of soil diffuse reflectance spectra using wavelets

R. A. Viscarra Rossel[a] & R. M. Lark[b]
[a]*CSIRO Land & Water, Bruce E. Butler Laboratory, GPO Box 1666, Canberra ACT 2601, Australia, and* [b]*Rothamsted Research, Harpenden, Hertfordshire, Al5 2JQ, UK*

Summary

Diffuse reflectance spectroscopy using visible (vis), near-infrared (NIR) and mid-infrared (mid-IR) energy can be a powerful tool to assess and monitor soil quality and function. Mathematical pre-processing techniques and multivariate calibrations are commonly used to develop spectroscopic models to predict soil properties. These models contain many predictor variables that are collinear and redundant by nature. Partial least squares regression (PLSR) is often used for their analysis. Wavelets can be used to smooth signals and to reduce large data sets to parsimonious representations for more efficient data storage, computation and transmission. Our aim was to investigate their potential for the analyses of soil diffuse reflectance spectra. Specifically we wished to: (i) show how wavelets can be used to represent the multi-scale nature of soil diffuse reflectance spectra, (ii) produce parsimonious representations of the spectra using selected wavelet coefficients and (iii) improve the regression analysis for prediction of soil organic carbon (SOC) and clay content. We decomposed soil vis-NIR and mid-IR spectra using the discrete wavelet transform (DWT) using a Daubechies's wavelet with two vanishing moments. A multiresolution analysis (MRA) revealed their multi-scale nature. The MRA identified local features in the spectra that contain information on soil composition. We illustrated a technique for the selection of wavelet coefficients, which were used to produce parsimonious multivariate calibrations for SOC and clay content. Both vis-NIR and mid-IR data were reduced to less than 7% of their original size. The selected coefficients were also back-transformed. Multivariate calibrations were performed by PLSR, multiple linear regression (MLR) and MLR with quadratic polynomials (MLR-QP) using the spectra, all wavelet coefficients, the selected coefficients and their back transformations. Calibrations by MLR-QP using the selected wavelet coefficients produced the best predictions of SOC and clay content. MLR-QP accounted for any non-linearity in the data. Transforming soil spectra into the wavelet domain and producing a smaller representation of the data improved the efficiency of the calibrations. The models were computed with reduced, parsimonious data sets using simpler regressions.

Introduction

There is growing interest in the use of diffuse reflectance spectroscopy at visible to near-infrared (vis-NIR) and mid-infrared (mid-IR) wavelengths to characterize soils quickly and cheaply (Viscarra Rossel *et al*., 2006a). Reflectance spectra of the soil have been used to predict multiple soil properties, because the fundamental molecular vibrations of soil components, organic and mineral, determine their mid-IR reflectance properties. The overtones and combinations of these are detected in the NIR and electronic excitations determine absorption of radiation in the visible part of the spectrum. An appropriate method for multivariate calibration should therefore allow predictive quantitative relationships between diffuse reflectance in these parts of the spectra and important soil properties to be developed from a reference data set. However, this process is not without difficulties, because of interferences resulting from the overlapping spectral responses of soil constituents, which are varied and interrelated, and sources of error including instrumental noise and drift, light-scatter and path-length variations that occur during measurements. For this reason various spectral pre-processing algorithms have been developed, such as the Savitzky-Golay smoothing (Savitzky & Golay, 1964), multiplicative signal correction (Geladi & Kowalski, 1986), baseline correction (Barnes *et al*., 1989) and derivatives.

The resulting spectra still present challenges. The information that they contain is in their shape, the peaks and edges that represent the interactions of the soil material with electromagnetic

Correspondence: R. A. Viscarra Rossel. E-mail: raphael.viscarra-rossel@csiro.au
Received 23 April 2008; revised version accepted 16 December 2008

radiation; the data consist of reflectance values in adjacent narrow spectral intervals. To extract the information from the data is not simple, not least because the reflectance in successive spectral intervals is strongly correlated over multiple soil samples. Therefore, data compression techniques such as partial least squares regression (PLSR) (Wold *et al.*, 1983) are often used for modelling and prediction.

The fundamental challenge for the processing of diffuse reflectance spectra is therefore how to extract predictive information, which consists largely of localized features of the spectrum, from noisy and strongly correlated data. In our paper, we propose that the wavelet transform is a natural technique for tackling this problem. The wavelet transform is an integral transform, that is to say it proceeds by computing a set of coefficients that can be used to combine a set of mathematical building blocks, the basis functions, to reconstitute the original data. Lark & Webster (1999) presented a detailed account of the wavelet transform for soil scientists. The particular value of the wavelet transform comes from the fact that: its basis functions represent distinct scales of variation (in the current case a fine-scale component of a spectrum is a fluctuation over a short interval of wavelength, and a coarse-scale component is a broader peak); a single basis function is localized (i.e. the coefficient corresponds to features in a delimited part of the spectrum); and, for many wavelets, adjacent coefficients at a particular scale are typically very weakly correlated for a wide range of signals (Silverman, 1999).

To return to our problem, the soil information in a diffuse reflectance spectrum may consist of a local peak, which is presented as a set of correlated reflectance values for successive wavelength intervals, embedded in noise. Under a wavelet transform this peak may correspond to a wavelet coefficient at some scale, the noise may largely appear at other scales, and the coefficient of interest is expected to be decorrelated with respect to adjacent coefficients. This clearly has potential to improve our quantitative analysis of spectra for prediction. Indeed, wavelet transforms have been used in chemometrics for spectral smoothing and data reduction (e.g. Trygg & Wold, 1998; Teppola & Minkkinen, 2000; Gributs & Burns, 2006). The aim of our study was to investigate their potential for the analysis of diffuse reflectance spectra of soil. Specifically, we wished to: (i) show how wavelets can be used to represent the multi-scale nature of soil diffuse reflectance spectra; (ii) produce parsimonious representations of the spectra using selected wavelet coefficients; and (iii) improve the regression analysis for prediction of soil organic carbon (OC) and clay content.

Materials and methods

The legacy soil samples used in this research originated from Queensland, New South Wales, South Australia and Western Australia. Soils were sampled from the 0 to 10 cm, 10 to 20 cm, 30 to 60 cm and 70 to 80 cm layers and were collected by various people for other research and for commercial agronomic purposes. They were not sampled specifically for this work. Approximately 50% of the samples were surface soils (0–10 cm), 20% were from 10 to 20 cm and the remaining samples originated from deeper soil layers down to 80 cm. Subsamples of these soils were used to develop vis-NIR and mid-IR spectral libraries, but not all of the samples were scanned with both instruments and not all samples in each library correspond. The subsamples used were air-dried and ground to a particle size ≤ 2 mm before submitting a part of each to SOC analysis using the dichromate oxidation method (Rayment & Higginson, 1992) and particle size analysis using the hydrometer method (Gee & Bauder, 1986). The remaining soil was used for the spectroscopic measurements.

vis-NIR soil spectroscopy

The diffuse reflectance spectra of 1139 soil samples were measured using the Agrispec® vis-NIR spectrometer (Analytical Spectral Devices, Boulder, Colorado, USA) with a spectral range of 350–2500 nm (28571–4000 cm^{-1}) and spectral resolution of 3 nm at 700 nm and 10 nm at 1400 nm and 2100 nm, corresponding to its three detectors: one 512-element Si photodiode for the 350–1000 nm range and two separate thermoelectrically cooled graded index InGaAs photodiodes for the 1000–1800 nm and 1800–2500 nm ranges. The soils were scanned using a contact probe (Analytical Spectral Devices, Boulder, Colorado, USA) and a Spectralon® panel (Labsphere, North Sutton, NH, USA) was used for white referencing once every 10 measurements. Each spectrum was made up of 1076 wavelengths and thus the vis-NIR data matrix consisted of 1139 samples and 1076 predictor variables.

mid-IR spectroscopy

The mid-IR diffuse reflectance spectra of 842 soil samples were recorded using a TENSOR 37® Fourier Transform Infrared (FT-IR) spectrometer from Bruker Optics (Billerica, MA, USA) with a Diffuse Reflectance Infrared Fourier Transform (DRIFT) attachment. Spectra were recorded in the range 3992–397 cm^{-1} (2505–25189 nm) with a sampling resolution of 8 cm^{-1} and collecting 64 scans per minute. A KBr white reference background spectrum was recorded at the start of a scanning session and once every hour thereafter during each session. Each spectrum was made up of 933 frequencies and hence the mid-IR data matrix consisted of 842 samples and 933 predictor variables.

Both vis-NIR and mid-IR spectra were recorded as per cent reflectance (R) but were later transformed to log1/R.

Wavelet transform

We do not attempt a detailed account of the wavelet transform here, and refer the reader to the introduction by Lark & Webster

(1999) and the literature cited therein. Here we give an outline of the methods and describe our particular analyses. Figure 1 shows Daubechies's extremal phase wavelet with two vanishing moments (Daubechies, 1988) and the corresponding scaling function.

Note that on Figure 1 the wavelet function fluctuates but its variations damp to zero as x is increased or decreased. Consider a signal (e.g. a spectrum) that is expressed as a function of location (e.g. wavelength). A wavelet coefficient can be obtained by multiplying the wavelet function by the signal, and integrating the result with respect to location. From the form of the wavelet, it can be seen that the resulting coefficient will reflect the variation of the signal within the interval where the wavelet takes non-zero values, but not outside this interval. The wavelet may be translated (i.e. centred at a new location) to generate a wavelet coefficient that describes variation there. It is also possible to expand (dilate) the wavelet coefficient so that the interval over which it responds to variation in a signal (the support) is increased. In the discrete wavelet transform (DWT), the wavelet function is dilated successively by doubling the support, corresponding to scales 2_i, $2^2{}_i$, $2^3{}_i$, … where i is the basic interval between observations, and at any scale, $2^j{}_i$, the wavelet is translated in steps of length $2^j{}_i$. Under these conditions, when the wavelet has certain mathematical properties, the translates and dilates are orthogonal to each other, and so wavelet coefficients for successive translations at a particular scale are decorrelated for a wide range of signals (Silverman, 1999). In practice the wavelet transform is implemented in the pyramid algorithm (see Press *et al.*, 1992), which is computationally efficient. The pyramid algorithm obtains wavelet coefficients for the translations of successive dilations of the wavelet up to a coarsest scale (typically $2^{m-1}{}_i$ when there are 2^m data) by successively applying a high-pass filter obtained from the wavelet function, which extracts the wavelet coefficients, and a low pass filter obtained from the scaling function, which extracts a smooth version of the original signal (Figure 1). The procedure is illustrated in Figure 2.

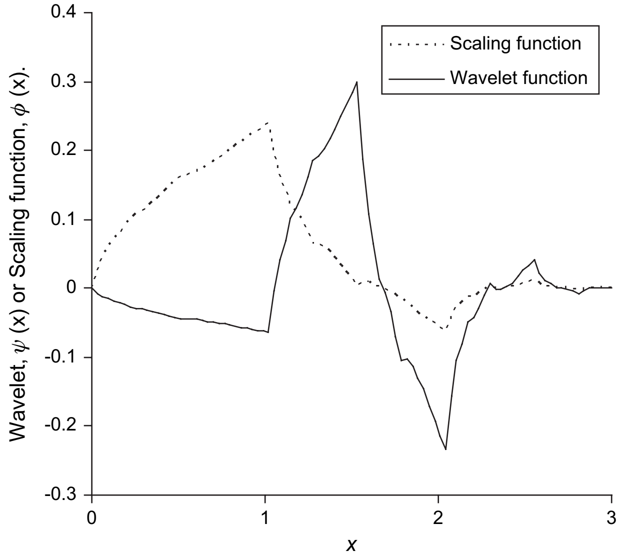

Figure 1 Daubechies wavelet function, $\psi(x)$, with two vanishing moments and its scaling function, $\varphi(x)$.

The wavelet transform may be inverted to reconstruct the original signal perfectly from the wavelet coefficients. If all wavelet coefficients were set to zero, apart from those that correspond to one particular dilation of the wavelet, and hence one scale, and then the inverse transform applied, an additive component of the original data specific to this scale is obtained. Such components, the 'detail components', d, can be obtained for different scales up to some coarsest scale at which a 'smooth or approximation component', a, is obtained, which is a smooth version of the data. The sum of all these components is the original data, and this decomposition is known as a multiresolution analysis (MRA) (Mallat, 1989). A MRA (Figure 2) shows the scale-specific structure of the data, and its components may be localized (because of the finite width of the wavelet's support) with much variation at some locations and little at others.

In our study we used the pyramid algorithm, and the adapted wavelet functions proposed by Cohen *et al.* (1993) for use on the finite interval, to deal with end effects. This is discussed in more detail by Lark & Webster (1999): note that if there are 2^m data this modified pyramid algorithm can be used to compute coefficients for up to $m-2$ dilations of the wavelet. Because the pyramid algorithm requires that there are 2^m data, where m is some integer, we first padded each data set by symmetrical reflection (Percival & Walden, 2000) to extend it to the nearest integer power of two in length. However, the coefficients corresponding to these padded values were excluded from further analysis. In addition to computing these wavelet coefficients, we also computed the detail and smooth components of the MRA.

Selection of wavelet coefficients

We wanted to retain only those coefficients that produced the most parsimonious representation of the data so as to use them in multivariate calibrations for each soil property. For each data set (vis-NIR and mid-IR), we calculated the variance of the wavelet coefficients and ordered them from largest to smallest variance, regardless of scale. Our hypothesis was that wavelet coefficients should be good at separating the information from the noise contained in the spectra. Ordering by variance should order the wavelet coefficients in decreasing information content because the variation between values of a particular coefficient is expected to reflect variations in the composition of the corresponding soils. If only the first few coefficients were retained, information would be lost and if too many were retained, noise would be added. Therefore, selection of wavelet coefficients to retain involved finding the optimum number of coefficients from the ordered set.

To determine which and how many wavelet coefficients to use for prediction, we performed a generalized validation technique

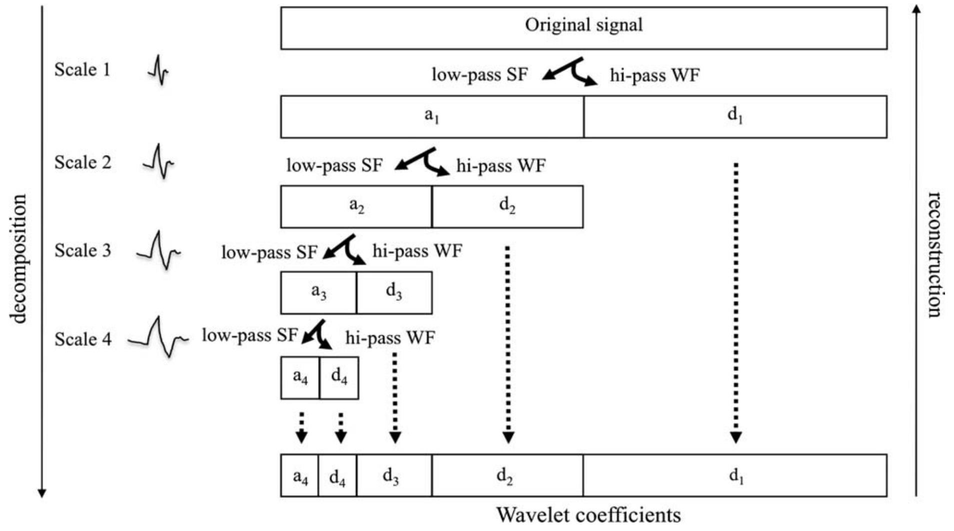

Figure 2 Implementation of the pyramid algorithm for a multiresolution analysis (MRA). At each scale, the algorithm applies a high-pass filter obtained from the wavelet function (WF) and a low-pass filter obtained from the scaling function (SF). The high-pass filter extracts the wavelet coefficients, also referred to as the detailed (d) components of the wavelet decomposition. The low-pass filter extracts the smooth component, which is described by the approximation (a) components to the data. The algorithm allows for a perfect reconstruction of the wavelet coefficients to the original signal.

as follows: we divided samples in each data set at random into training and validation sets of size n_t and n_v for the vis-NIR data set and m_t and m_v for the mid-IR data set. Using the respective training data, we then computed multiple linear regressions with quadratic polynomials (MLRQP) to predict SOC and clay content separately, adding the coefficients as ordered by variance one at a time and testing each regression on the corresponding validation data sets. In this way, we could determine the number of wavelet coefficients that would give the best predictions and the wavelet scales to which they belonged. Prediction accuracy was determined by the adjusted coefficient of determination (R^2_{adj}) and the root-mean squared error (RMSE) of prediction:

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

where $\hat{y}_i$ is the predicted value, y_i is the observed value and N is the number of data.

Back-transforming wavelet coefficients to spectral domain

After selecting which and how many wavelet coefficients to retain, we set all other coefficients to zero and the reduced set for each sample was back transformed into the original spectral domain using the inverse wavelet transformation algorithm. The reconstructed spectra were effectively 'denoised'.

Multivariate calibrations

The vis-NIR and mid-IR spectra were combined with their corresponding measurements of SOC and clay content. Outlier detection was conducted using the Mahalonobis distance statistic (De Maesschalck *et al*., 2000) on the scores of the first five PLS factors. We removed five outliers from the clay-vis-NIR data, seventeen outliers from the SOC-vis-NIR data, three outliers from the clay-mid-IR data and six outliers from the SOC-mid-IR data. To derive training and test data sets, the data for each soil variable were sorted from lowest to highest values and every third row was held out to test models developed using the remaining training data. In this way, the calibrations for each of the soil properties were representative of the entire population and the models were independently validated.

With soil diffuse reflectance spectra it is difficult to find selective wavelengths for the chemical constituents in a sample. Take for instance when calibrating for SOC, no single or few wavelengths in the mid-IR or vis-NIR provide sufficient information. Thus, it is common practice to use multivariate calibrations. Multivariate calibrations of the spectra and the wavelet coefficients were performed for each soil property using: (i) PLSR for a single response variable (Viscarra Rossel, 2008), (ii) multiple linear regressions (MLR), (iii) MLR with quadratic polynomials (MLR-QP), and (iv) the scores of the PLS model regressed using MLR-QP (PLSScores-MLR-QP). The quadratic polynomials were used to account for any nonlinear response in the data.

PLSR is a technique that can be used to relate a response variable to many predictor variables that are strongly collinear; in our case, for example, relating SOC or clay content to the 1076 and 933 collinear vis-NIR and mid-IR frequencies, respectively.

The DWT coefficients are strongly decorrelated, so they could be used directly as predictors in a MLR without serious numerical problems. However, to ensure numerical stability, the least squares regression coefficients $\hat{\beta}$, were estimated by the QR decomposition (Lawson & Hanson, 1974).

Predictions of SOC and clay content using the multivariate calibrations were made on the independent test data. These were assessed using the $R^2_{\mathrm{adj.}}$ and the RMSE of prediction.

Results and discussion

The soils in each of the spectral libraries were diverse and represented by various Australian Soil Classification (ASC) orders, including Vertosols, Ferrosols, Kurosols, Chromosols, Dermosols, Sodosols and a smaller number of Podosols, Rudosols, Tenosols and Calcarosols (Isbell, 2002). Their approximate WRB-FAO classification is: Vertisols, Ferralsols, Planosols, Luvisols, Ferric Calcisols, Solonetz, Podzols, Leptosols and Calcisols (FAO, 1998). In both libraries, there was a large variation in SOC and clay content (Table 1) as the samples originated from different depths in the profiles.

The average spectra of the vis-NIR and the mid-IR libraries are shown in Figure 3(a,b).

The mid-IR spectrum (Figure 3b) contains many more absorption features and hence much more information on soil organic and mineral composition than the vis-NIR (Figure 3a). The vis-NIR portion of the spectrum shows characteristic absorption bands near 1430 nm, 1930 nm and 2220 nm (Figure 3a). The band near 1430 nm may be attributed to the first overtone of the hydroxyl (O–H) stretching vibration of minerals or water. The band near 1930 nm is representative of the combination H–O–H bend and O–H stretch vibrations of free molecular water and water contained in structures of 2:1 minerals like montmorillonite. The combination bands near 2220 nm and 2350 nm are diagnostic for clay mineral identification (Clark *et al.*, 1990) and are characteristic of kaolinite and other aluminosilicates, such as illite and montmorillonite, as well as carbonate (Viscarra Rossel *et al.*, 2006b). The mid-IR spectrum (Figure 3b) shows absorption bands in the region between 3800 and 3600 cm^{-1}, which may be attributed to O–H stretching vibrations of clay minerals. The broad band near 3400 cm^{-1} may be attributed to O–H stretching vibrations of water molecules. The faint absorption bands at 2930 cm^{-1} and 2850 cm^{-1} are particularly useful for the detection of organic matter in soils and may be attributed to alkyl material. Mid-IR spectra contain a number of other absorption bands that are consistent with the presence of organic matter (e.g. absorptions by carboxylic acids near 1725 cm^{-1}, proteins near 1640 cm^{-1} and 1530 cm^{-1}, aliphatic compounds near 1465 cm^{-1}, 1445 cm^{-1} and 1350 cm^{-1}, phenolics near 1275 cm^{-1} and carbohydrates near 1050 cm^{-1}). Other bands around these frequencies may be attributed to clay minerals and include the alumino-silicate lattice vibrations near 1020 cm^{-1} and the Al–OH deformation vibrations at 920 cm^{-1}. Quartz displays a number of characteristic absorption bands, which include the group of three bands at 2000 cm^{-1}, 1870 cm^{-1} and 1790 cm^{-1} (Figure 3b). These are overtones and combination bands of fundamental vibrations at 1080 cm^{-1}, 800 cm^{-1} and 700 cm^{-1}, respectively. The region between 2000 cm^{-1} and 1600 cm^{-1} also has overtones and combination bands for other silicate structures, although these are usually masked by those of quartz.

Table 1 Statistical description of soil organic carbon (SOC) and clay content in the vis-NIR and mid-IR libraries. The number of samples without outliers is given by *n*

		Mean	SD	Median	Minimum	Maximum
	n	%	%	%	%	%
vis-NIR library						
OC	1122	2.61	2.32	1.53	0.01	13.90
Clay	1134	33.1	18.1	26.2	1.8	77.8
mid-IR library						
OC	836	2.70	2.31	1.84	0.032	11.88
Clay	839	38.1	18.5	38.0	5.0	86.1

Multiresolution analysis (MRA)

A MRA of the average vis-NIR and mid-IR spectra (Figure 4) shows both the detail components, d_m, of the wavelet transform at each scale and the approximation component, a_m, at the coarsest scale.

The MRA in Figure 4 shows features of the spectra at different scales. It shows that the high frequency elements of the spectra occur at the finest scales $\lambda = 2$, 4, 8 and 16. In both vis-NIR and mid-IR spectra, these appear near the edges of the signals and at specific absorption features, for example those near 1430 nm, 1930 nm and 2220 nm (Figure 4a), and those near 3700 cm^{-1}, 2850 cm^{-1} and wave numbers < 2000 cm^{-1} (Figure 4b). From Figure 4a, the discontinuities of the vis-NIR signal because of changing detectors at 1000 nm and 1800 nm are also evident at these scales. At $\lambda = 32$ the most prominent feature is near 1930 nm, while at $\lambda = 64$ absorption features occur near 740 nm, 860 nm, 1430 nm, 1640 nm, 1760 nm, 1930 nm, 2220 nm and 2350 nm. At $\lambda = 128$, prominent absorption features occur near 600 nm, 860 nm, 1120 nm, 1460 nm, and 2000 nm, while at $\lambda = 256$ these are broader and occur near 860 nm, 1400 nm, 1890 nm and 2390 nm (Figure 4a). From Figure 4b, the most prominent features at scale parameter 32 include those near 2000 cm^{-1}, 1870 cm^{-1}, 920 cm^{-1}, 700 cm^{-1}, 600 cm^{-1} and 450 cm^{-1}, which are characteristic absorptions for quartz and alumino-silicates, and near 1640 cm^{-1}, 1530 cm^{-1}, 1465 cm^{-1}, 1350 cm^{-1} and 1050 cm^{-1}, which are characteristic of organic material in soil. At $\lambda = 64$ absorption features appear near 3600 cm^{-1}, 3400 cm^{-1} and 2515 cm^{-1}, which may be attributed to either calcite or dolomite present in some of the subsoil samples, near 1640 cm^{-1}, 1350 cm^{-1}, 1275 cm^{-1}, 1050 cm^{-1} and 660 cm^{-1}. At $\lambda = 128$ there are broad absorptions near 3600 cm^{-1}, 3250 cm^{-1}, 2930 cm^{-1}, 2850 cm^{-1}, 2505 cm^{-1}, 2000 cm^{-1}, 1640 cm^{-1}, 1275 cm^{-1} and 800 cm^{-1} and at scale parameter 256 there are prominent features near 3500 cm^{-1} and 1530 cm^{-1}. The MRA shows that a soil diffuse reflectance

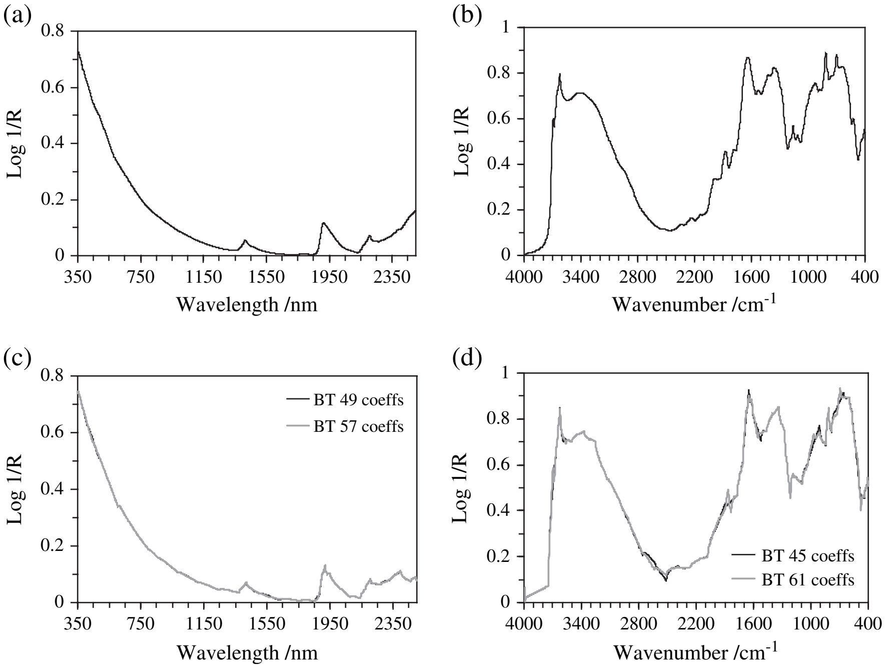

Figure 3 Average (a) vis-NIR and (b) mid-IR spectra showing most characteristic absorptions and average reconstructed (c) vis-NIR and (d) mid-IR spectra from back transformed (BT) wavelet coefficients.

spectrum has local features that contain information about the soil and that these are readily extracted into coefficients corresponding to elements of the wavelet basis. This suggests that the wavelet transform is one way to isolate the informative features of diffuse reflectance spectra.

Selection of wavelet coefficients

The ordered wavelet coefficient variances derived from the vis-NIR and mid-IR data are shown in Figure 5(a,b), while their contributions by wavelet scale are shown in Figure 5(c,d).

For both the vis-NIR and mid-IR data, the wavelet coefficients of coarser scales ($\lambda \geq 32$) accounted for the largest variances. Generally, the average contribution of wavelet coefficients to their variance decreased with scale. The exceptions were at $\lambda = 128$ for the vis-NIR and $\lambda = 64$ for the mid-IR data (Figure 5c,d, respectively), which coincided with the prominence of absorption peaks at these scales (Figure 4).

The results of the generalized validation technique (Figure 6) were used to determine which and how many of these coefficients to retain. As successive wavelet coefficients are added to the model there is first a reduction in the RMSE to a minimum and then a gradual increase. This is consistent with our hypothesis. The original diffuse reflectance spectra contain both information and noise. The wavelet transform can extract these components into separate coefficients in so far as the information corresponds to features that are distinct both with respect to location on the spectrum and the scale of generalization at which they are most apparent. Thus as the first set of coefficients are added information from the spectrum is added to the predictor and the RMSE is reduced. However, as we continue to add coefficients, increasingly those are added that correspond to noise and have no predictive power, and the RMSE rises again. The distinct shape of these graphs with a pronounced reduction and then increase also suggests that ordering the coefficients with respect to their variance over the library is an appropriate way to select the predictors.

Figure 6a shows that multivariate calibrations of the first 49 vis-NIR and 61 mid-IR wavelet coefficients produced the smallest RMSEs when used to predict the SOC content of the validation samples. For clay content, the best calibrations were found by retaining the first 57 vis-NIR and the first 45 mid-IR wavelet coefficients (Figure 6b). In all cases, using additional wavelet coefficients in the regressions only enlarged the RMSEs. Although our hypothesis that ordering the wavelet coefficients by variance should order them by decreasing information content is supported by our results, further evidence is needed to support it. Further improvements in the selection of coefficient to include in the regression may be possible. For example, ordering coefficients by variance could be used as an initial selection procedure to prime the variable selection approach of Lark *et al.* (2007).

The proportions of retained wavelet coefficients by scale are shown in Figure 7.

For each data set, all of the scaling coefficients and all coefficients at scale parameters $\lambda \geq 128$ were retained (Figure 7), as the coefficients at these scales contain the lower frequency

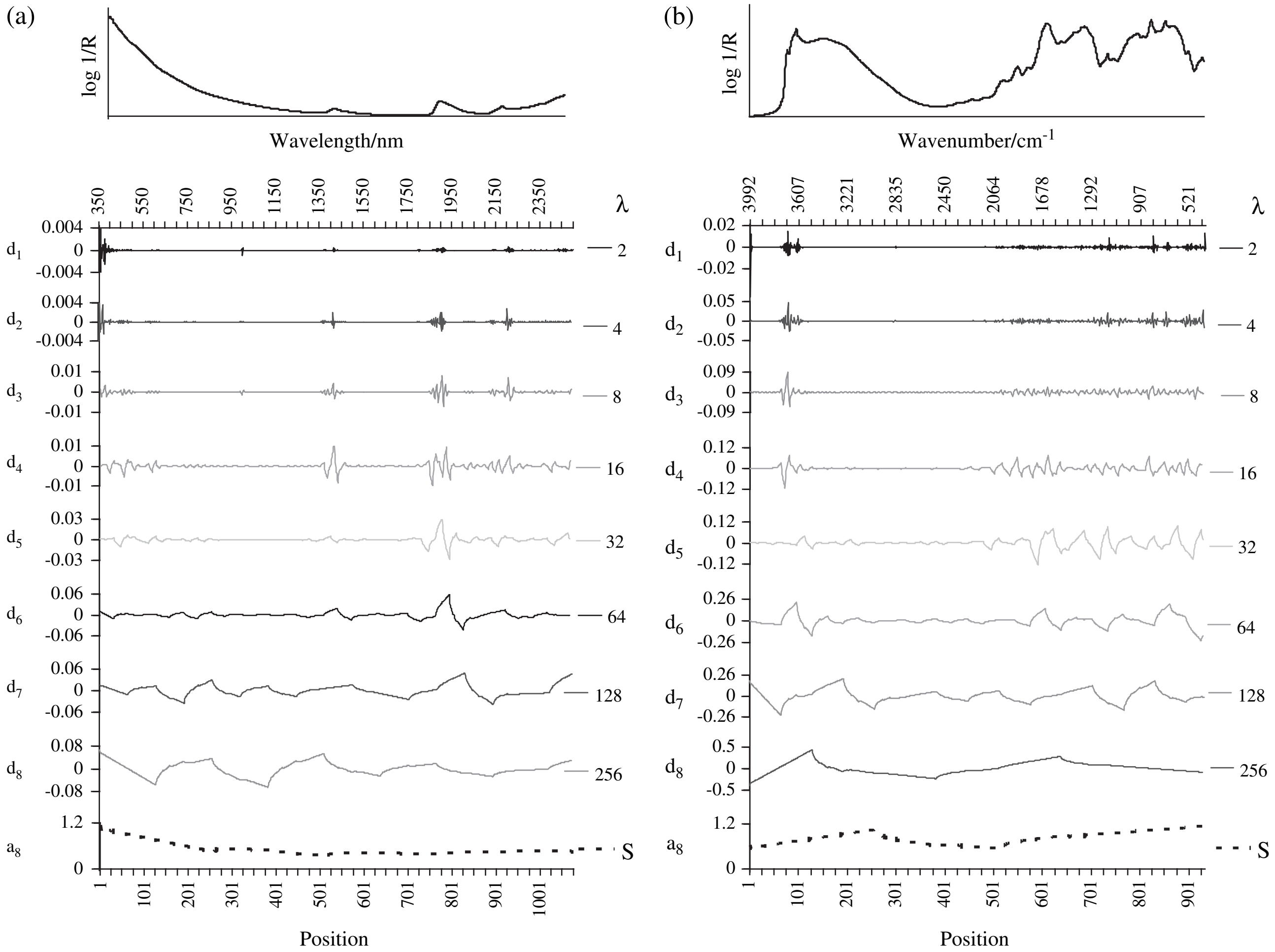

Figure 4 Multiresolution analysis (MRA): approximation (a_m) and detail (d_m) components at each scale, λ, of a (a) vis-NIR and (b) mid-IR soil spectrum. Each component is centred about its own zero.

systematic information in the spectra that are useful in the regressions. The proportion of wavelet coefficients that were retained decreased with scale (Figure 7).

Back-transformed wavelet coefficients

Average back-transformed vis-NIR and mid-IR spectra from the selected wavelet coefficients are shown in Figure 3(c,d). Our aim here was not to 'denoise' the spectra, as the original spectra were relatively smooth (Figure 3a,b), but to select the most relevant wavelet coefficients for the regression analysis. For this reason the back-transformed spectra appear somewhat degraded, displaying some of the characteristic shape of the wavelet function (Figure 3c,d).

Multivariate calibrations

Predictions of SOC by PLSR using the vis-NIR and mid-IR spectra produced RMSE values of 1.08% and 0.54%, and predictions of clay content with RMSE values of 8.17% and 6.35%, respectively (Table 2). These predictions are shown in Figures 8(a,b) and 9(a,b).

PLSR predictions using all wavelet coefficients (DWT-PLSR) were not significantly different to those using the original spectra (Table 2). Predictions of SOC by PLSR using only the selected wavelet coefficients did not improve predictions of SOC and predictions of clay content were poorer than those using all wavelet coefficients (Table 2). Predictions of SOC by PLSR using 49 vis-NIR selected wavelet coefficients (DWT-VAR-PLSR) produced a RMSE of 1.08% and using 61 mid-IR wavelet coefficients produced a RMSE of 0.55%. Predictions of clay content using 57 vis-NIR selected coefficients produced a RMSE of 8.36% and using 45 mid-IR coefficients produced a RMSE of 7.06%. To account for the nonlinear response in the data we regressed the PLS scores of the vis-NIR and mid-IR PLSR models using MLR-QP (PLSScores-MLR-QP); however, these predictions were biased and had the largest RMSE values for both soil properties (Table 2).

Predictions of SOC and clay content by MLR using the selected wavelet coefficients (DWT-VAR-MLR) were generally,

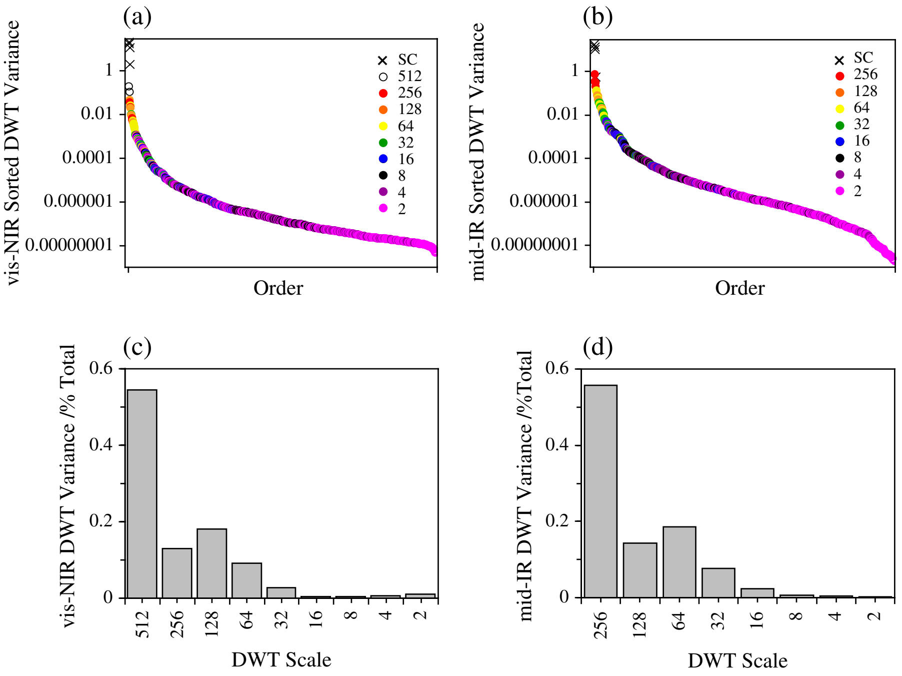

Figure 5 Ordered wavelet coefficients variance from the transformation of the (a) vis-NIR and (b) mid-IR spectra. The per cent contribution to the wavelet variance by each wavelet scale, for (c) the vis-NIR and (d) mid-IR data.

but not significantly, better than those derived by PLSR (Table 2). Multivariate calibrations by MLR-QP using the selected vis-NIR and mid-IR wavelet coefficients (DWT-VAR-MLR-QP) produced the best predictions of SOC and clay content as the MLR-QP accounted for the slightly nonlinear data (Table 2). Predictions of SOC by DWT-VAR-MLR-QP using 49 vis-NIR wavelet coefficients produced a RMSE of 0.86% and using 61 mid-IR coefficients produced a RMSE of 0.52%. DWT-VAR-MLR-QP predictions of clay content using 57 vis-NIR coefficients produced a RMSE of 7.06% and using 45 mid-IR coefficients produced a RMSE of 5.77%. These predictions are shown in Figures 8(c,d) and 9(c,d), respectively. Multivariate calibrations by PLSR using the back transformed spectra (BTDWT-VAR-PLSR) improved predictions of clay content but not of SOC (Table 2). Predictions of clay content using the back transformed vis-NIR and mid-IR spectra produced RMSE values of 7.30% and 6.11%, respectively.

Multivariate calibrations of spectroscopic data usually involve the use of large spectral libraries that need to be re-calibrated for different soil analyses. Transforming soil spectra into the wavelet domain and producing a smaller representation of the data can improve the efficiency of these calibrations. As was

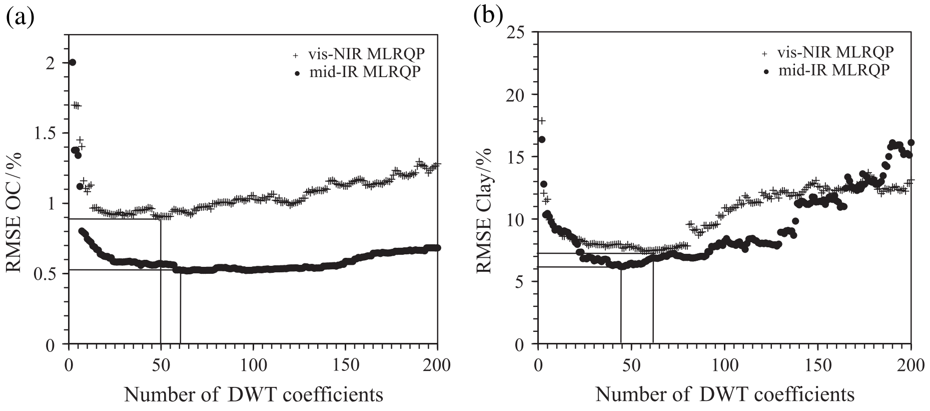

Figure 6 Root mean-square errors (RMSE) of predictions for (a) soil organic carbon (SOC) and (b) clay content show the best number of wavelet coefficients to use in the regressions. Multiple linear regression with quadratic polynomials (MLR-QP) of the wavelet coefficients ordered by their variance performed sequentially by adding one coefficient at a time. Data shown on the graphs are for the first 200 wavelet coefficients only.

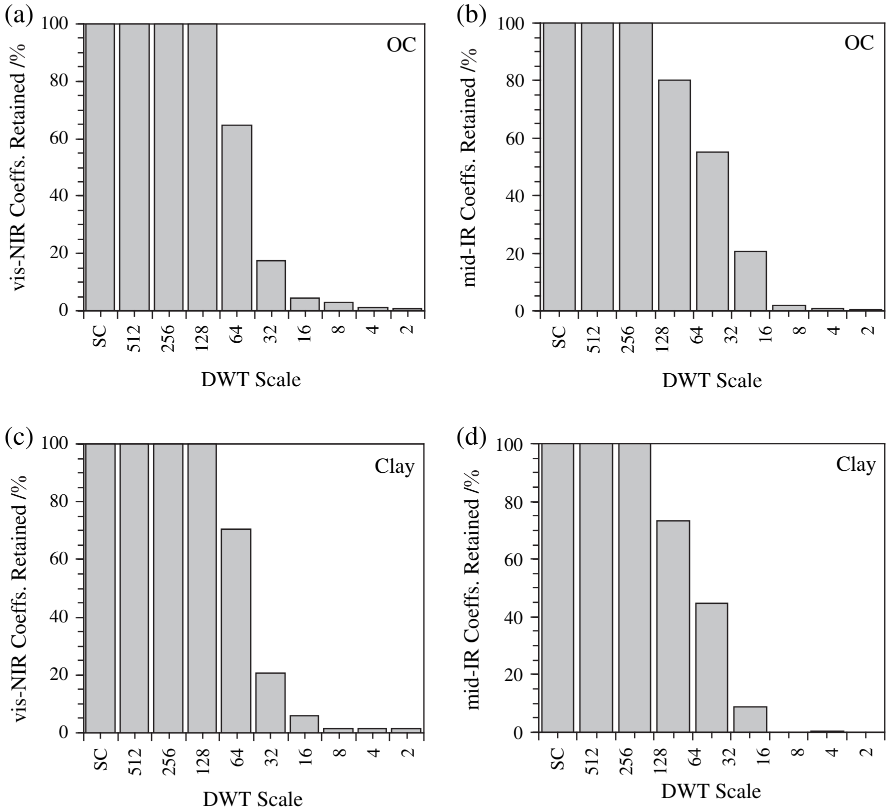

Figure 7 Proportion of selected wavelet coefficients by scale. Proportion of (a) 49 vis-NIR wavelet coefficients and (b) 61 mid-IR wavelet coefficients used for the calibrations to soil organic carbon (SOC). Proportion of (c) 57 vis-NIR wavelet coefficients and (d) 45 mid-IR wavelet coefficients used for the calibrations to clay content.

shown here, the models can be computed with greatly reduced, parsimonious data sets using simpler regression techniques that are simpler and faster to implement than full spectrum PLSR. Furthermore, the DWT is computationally efficient. Once the informative coefficients for a particular spectrum to predict a particular property have been identified, then their extraction could be automated, and the user need not complete a full wavelet analysis in order to compute predictions from spectra of new soil samples.

Data reduction using wavelets relies on the premise that the spectra can be quite accurately represented by a smaller number of wavelet coefficients; that is to say, like many natural phenomena, they have a sparse representation. Our results show that rather than using all of the highly collinear predictor variables in the spectral domain (i.e. 1076 vis-NIR and 933 mid-IR frequencies) we can use a much smaller number of orthogonal wavelet coefficients to derive the multivariate calibrations. In this case, both vis-NIR and mid-IR data were reduced to less than 7% of their original size (Table 3).

Our results showed how wavelets might be used to improve the analysis of soil diffuse reflectance spectra for the prediction of soil properties. We demonstrated this with a legacy soil data set. The use of legacy data for the development of soil spectral libraries is important (Viscarra Rossel *et al.*, 2008); they are a valuable resource. Nonetheless, it would be interesting to repeat our analyses on a data set with more comprehensive metadata, collected specifically for spectroscopic calibrations.

Conclusions

- A multiresolution analysis (MRA) of soil diffuse reflectance spectra can be used to identify different spectral features that occur at different scales. That is, a spectrum has local features that contain information about soil composition and wavelets present a good method to extract them.
- Using the wavelet coefficients variance to select significant coefficients, the wavelet-transformed spectra were reduced to less than 7% of their original size. The data sets were more parsimonious and the multivariate calibrations more accurate.
- Multivariate calibrations of the selected wavelet coefficients by MLR-QP (i.e. DWT-VAR-MLR-QP) were more straightforward and faster than PLSR. Furthermore, MLR-QP was able to account for the slight nonlinearities of the data.
- Predictions of SOC by DWT-VAR-MLR-QP using 49 selected vis-NIR wavelet coefficients produced a RMSE of 0.86% and using 61 mid-IR wavelet coefficients produced

Table 2 Predictions of soil organic carbon (SOC) and clay content using vis-NIR and mid-IR data. NF is the number of factors used in the PLSR models; NV refers to the number of predictor variables used in the models. DWT-VAR refers to the selection of wavelet coefficients by the wavelet variance and MLR-QP refers to modelling by multiple linear regressions with quadratic polynomials. Best predictions are shown in italics

	Soil OC vis-NIR				Soil OC mid-IR			
				RMSE				RMSE
	NF	NV	R^2	%	NF	NV	R^2	%
Spectral domain PLSR	11	1076	0.79	1.08	16	933	0.94	0.54
DWT-PLSR	9	1076	0.80	1.04	14	933	0.95	0.54
DWT-VAR-PLSR	12	49	0.79	1.08	16	61	0.94	0.55
PLS scores-MLR-QP	11	1076	0.74	1.29	16	933	0.94	0.78
DWT-VAR-MLR		49	0.79	1.07		*61*	*0.95*	*0.52*
DWT-VAR-MLR-QP		*49*	*0.86*	*0.86*		*61*	*0.95*	*0.52*
BTDWT-VAR-PLSR	9	1076	0.80	1.06	14	933	0.94	0.55
	Soil clay vis-NIR				Soil clay mid-IR			
				RMSE				RMSE
	NF	NV	R^2	%	NF	NV	R^2	%
Spectral domain PLSR	8	1076	0.82	8.17	20	933	0.88	6.35
DWT-PLSR	8	1076	0.82	8.16	19	933	0.88	6.38
DWT-VAR-PLSR	8	57	0.81	8.36	9	45	0.86	7.06
PLS scores-MLR-QP	8	1076	0.84	9.10	20	933	0.89	8.49
DWT-VAR-MLR		57	0.84	7.54		45	0.89	6.21
DWT-VAR-MLR-QP		*57*	*0.86*	*7.06*		*45*	*0.90*	*5.77*
BTDWT-VAR-PLSR	9	1076	0.86	7.30	15	933	0.89	6.11

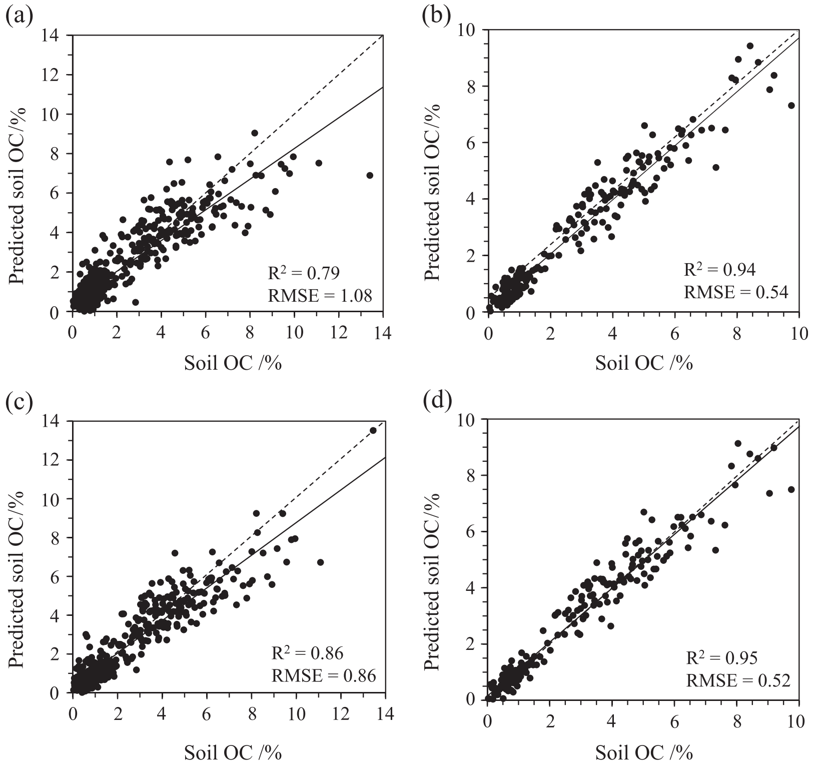

Figure 8 PLSR predictions of soil organic carbon (SOC) using the (a) vis-NIR and (b) mid-IR spectra, and DWT-VAR-MLR-QP predictions of OC using the selected wavelet coefficients derived from the (c) vis-NIR and (d) mid-IR spectra, respectively.

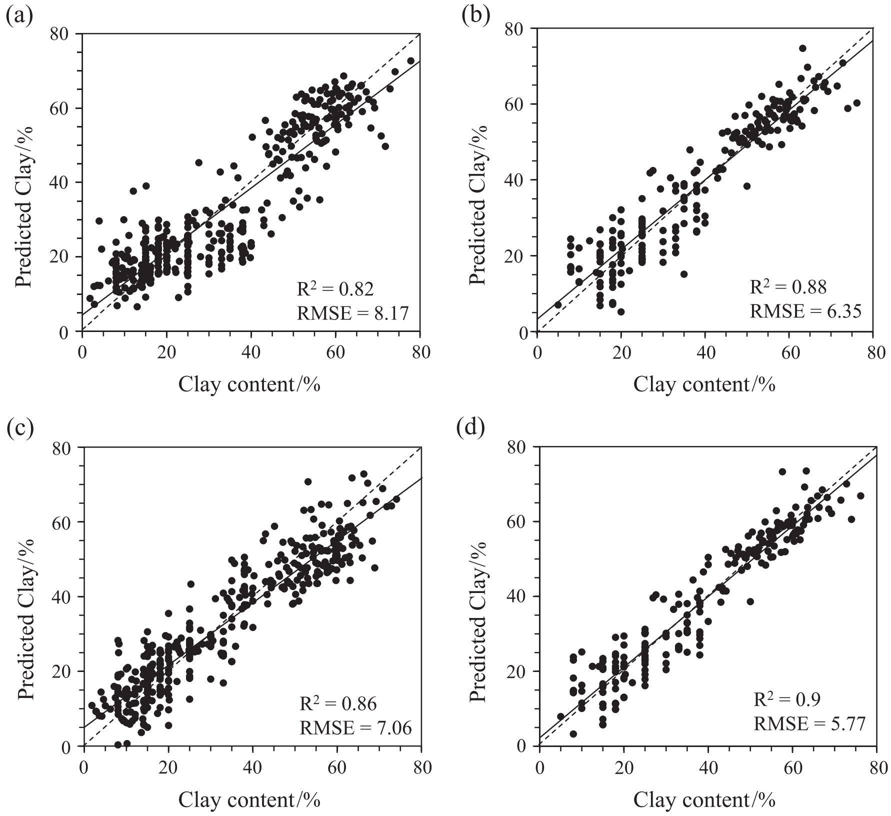

Figure 9 PLSR predictions of clay content using the (a) vis-NIR and (b) mid-IR spectra, and DWT-VAR-MLR-QP predictions of clay using the selected wavelet coefficients derived from the (c) vis-NIR and (d) mid-IR spectra, respectively.

a RMSE of 0.52%. Predictions of clay content by DWT-VAR-MLR-QP using 57 selected vis-NIR coefficients produced a RMSE of 7.06% and using 45 mid-IR coefficients produced a RMSE of 5.77%.

- By transforming soil spectra into the wavelet domain and producing a sparse representation of the data, the overall efficiency of these calibrations was improved as the models were computed with greatly reduced data sets using simpler regressions.

Table 3 Number of predictor variables (spectral bands or wavelet coefficients) and amount of data reduction achieved by the discrete wavelet transform (DWT)

	vis-NIR spectra	vis-NIR DWT	mid-IR spectra	mid-IR DWT
Organic C				
Number of predictor variables	1076	49	933	61
Reduction % original size	—	5	—	7
Clay content				
Number of predictor variables	1076	57	933	45
Reduction % original size	—	5	—	5

Acknowledgements

Dr Viscarra Rossel wishes to thank the University of Sydney, the Cotton Catchment and Communities CRC (CCC CRC) and the Grains Research and Development Corporation (GRDC) for their financial support. Dr Lark's contribution was part of Rothamsted Research's Programme in Mathematical and Computational Biology, sponsored by the U.K. Biotechnology and Biological Sciences Research council (BBSRC)

References

Barnes, R.J., Dhanoa, M.S. & Lister, S.J. 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, **43,** 772–777.

Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G. & Vergo, N. 1990. High spectral resolution reflectance spectroscopy of minerals. *Journal of Geophysical Research*, **95,** 12653–12680.

Cohen, A., Daubechies, I. & Vial, P. 1993. Wavelets on the interval and fast wavelet transforms. *Applied & Computational Harmonic Analysis*, **1,** 54–81.

Daubechies, I. 1988. Orthonormal bases of compactly supported wavelets. *Communications in Pure & Applied Mathematics*, **41,** 909–996.

De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D.L. 2000. The Mahalanobis distance. *Chemometrics & Intelligent Laboratory Systems*, **50,** 1–18.

FAO 1998. *World Reference Base for Soil Resources*. Food and Agriculture Organization of the United Nations, Rome.

Gee, G.W. & Bauder, J.W. 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.

Geladi, P. & Kowalski, B.R. 1986. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, **185,** 1–17.

Gributs, C.E.W. & Burns, D.H. 2006. Parsimonious calibration models for near-infrared spectroscopy using wavelets and scaling functions. *Chemometrics & Intelligent Laboratory Systems*, **83,** 44–53.

Isbell, R.F. 2002. *The Australian Soil Classification*. CSIRO Australia, Collingwood, Victoria.

Lark, R.M. & Webster, R. 1999. Analysis and elucidation of soil variation using wavelets. *European Journal of Soil Science*, **50,** 185–206.

Lark, R.M., Bishop, T.F.A. & Webster, R. 2007. Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties. *Geoderma*, **138,** 65–78.

Lawson, C.L. & Hanson, R.J. 1974. *Solving Least-Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ.

Mallat, S.G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11,** 674–693.

Percival, D.B. & Walden, A.T. 2000. *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. 1992. *Numerical Recipes (Fortran)*, 2nd edn. Cambridge University Press, Cambridge.

Rayment, G.E. & Higginson, F.R. 1992. *Australian Laboratory Handbook of Soil and Water Chemical Methods*. Inkata Press, Melbourne, Australia.

Savitzky, A. & Golay, M.J.E. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, **36,** 1627–1639.

Silverman, B.W. 1999. Wavelets in statistics: beyond the standard assumptions. *Philosophical Transactions of the Royal Society of London, Series A*, **357,** 2459–2473.

Teppola, P. & Minkkinen, P. 2000. Wavelet–PLS regression models for both exploratory data analysis and process monitoring. *Journal of Chemometrics*, **14,** 383–399.

Trygg, J. & Wold, S. 1998. PLS regression on wavelet compressed NIR spectra. *Chemometrics & Intelligent Laboratory Systems*, **42,** 209–220.

Viscarra Rossel, R.A. 2008. ParLeS: software for chemometric analysis of spectroscopic data. *Chemometrics & Intelligent Laboratory Systems*, **90,** 72–83.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J. & Skjemstad, J.O. 2006a. Visible, near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, **131,** 59–75.

Viscarra Rossel, R.A., McGlynn, R.N. & McBratney, A.B. 2006b. Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. *Geoderma*, **137,** 70–82.

Viscarra Rossel, R.A., Jeon, Y.S., Odeh, I.O.A. & McBratney, A.B. 2008. Using a legacy soil sample to develop a mid-IR spectral library. *Australian Journal of Soil Research*, **46,** 1–16.

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.