Geoderma 158 (2010) 46–54

Contents lists available at ScienceDirect

Geoderma

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

Using data mining to model and interpret soil diffuse reflectance spectra

R.A. Viscarra Rossel [a,*], T. Behrens [b]

[a] CSIRO Land and Water, Bruce E. Butler Laboratory, GPO Box 1666 Canberra ACT 2601, Australia
[b] Institute of Geography, Physical Geography, University of Tübingen, Rümelinstraße 19-23, 72074, Tübingen, Germany

ARTICLE INFO

Article history:
Received 19 May 2009
Received in revised form 4 November 2009
Accepted 29 December 2009
Available online 11 February 2010

Keywords:
Vis–NIR
Diffuse reflectance spectroscopy
Regression
Feature selection
Data mining and knowledge discovery
Wavelets

ABSTRACT

The aims of this paper are: to compare different data mining algorithms for modelling soil visible–near infrared (vis–NIR: 350–2500 nm) diffuse reflectance spectra and to assess the interpretability of the results. We compared multiple linear regression (MLR), partial least squares regression (PLSR), multivariate adaptive regression splines (MARS), support vector machines (SVM), random forests (RF), boosted trees (BT) and artificial neural networks (ANN) to estimate soil organic carbon (SOC), clay content (CC) and pH measured in water (pH). The comparisons were also performed using a selected set of wavelet coefficients from a discrete wavelet transform (DWT). Feature selection techniques to reduce model complexity and to interpret and evaluate the models were tested. The dataset consists of 1104 samples from Australia. Comparisons were made in terms of the root mean square error (RMSE), the corresponding R^2 and the Akaike Information Criterion (AIC). Ten-fold-leave-group out cross validation was used to optimise and validate the models. Predictions of the three soil properties by SVM using all vis–NIR wavelengths produced the smallest RMSE values, followed by MARS and PLSR. RF and especially BT were out-performed by all other approaches. For all techniques, implementing them on a reduced number of wavelet coefficients, between 72 and 137 coefficients, produced better results. Feature selection (FS) using the variable importance for projection (FS_{VIP}) returned 29–31 selected features, while FS_{MARS} returned between 11 and 14 features. DWT–ANN produced the smallest RMSE of all techniques tested followed by FS_{VIP}–ANN and FS_{MARS}–ANN. However, both the FS_{VIP}–ANN and FS_{MARS}–ANN models used a smaller number of features for the predictions than DWT–ANN. This is reflected in their AIC, which suggests that, when both the accuracy and parsimony of the model are taken into consideration, the best SOC model was the FS_{MARS}–ANN, and the best CC and pH models were those from FS_{VIP}–ANN. Analysis of the selected bands shows that: (i) SOC is related to wavelengths indicating C—O, C=O, and N—H compounds, (ii) CC is related to wavelengths indicating minerals, and (iii) pH is related to wavelengths indicating both minerals and organic material. Thus, the results are sensible and can be used for comparison to other soils. A systematic comparison like the one presented here is important as the nature of the target function has a strong influence on the performance of the different algorithms.

1. Introduction

There is widespread interest for using visible–near infrared (vis–NIR) diffuse reflectance spectroscopy for soil analysis and to provide data for digital soil mapping (DSM). The technique is rapid, cost effective, requires minimal sample preparation, can be used in situ (Viscarra Rossel et al., 2009), is non-destructive, no hazardous chemicals are used, and importantly, several soil properties can be measured from a single scan (Viscarra Rossel et al., 2006). This multi-parameter feature of diffuse reflectance spectroscopy implies that one spectrum holds information about various soil constituents and indeed, vis–NIR spectra are sensitive to both organic and inorganic soil composition.

Absorptions in the visible region (400–700 nm) are primarily associated with minerals that contain iron (e.g. haematite, goethite) (e.g. Sherman and Waite, 1985). Soil organic matter can also have broad absorptions in the visible, which are dominated by chromophres and the darkness of humic acid and absorptions in the NIR (700–2500 nm) from the overtones and combination absorptions of O–H, C–H, and N–H (Clark et al., 1990; Clark, 1999). Water has strong absorption features in the vis–NIR spectra of soils, most visibly near 1400 nm and 1900 nm, but there are weaker overtone bands elsewhere. Clay minerals and carbonate absorb in the NIR and absorptions are due to metal–OH bend plus O–H stretch combination and C–O (e.g. Hunt and Salisbury, 1970).

Nonetheless, soil vis–NIR spectra are largely non-specific, quite weak and broad due to overlapping absorptions of soil constituents and their often small concentrations in soil. Therefore, the information needs to be mathematically extracted from the spectra so that they may be correlated with soil properties. Hence, the analysis of soil diffuse reflectance spectra requires the use of chemometric techniques and

* Corresponding author. Tel.: +61 2 6246 5945.
E-mail address: raphael.viscarra-rossel@csiro.au (R.A. Viscarra Rossel).

0016-7061/$ – see front matter.
doi:10.1016/j.geoderma.2009.12.025

multivariate calibration (Martens and Næs, 1989). In these cases, to be useful quantitatively, spectra must be related to a set of known reference samples through a calibration model. The set of reference samples used in the models need to be representative of the range of soils in which the models are to be used. Partial least squares regression (PLSR) is the most common algorithm used to calibrate vis–NIR spectra to soil properties (Wold et al., 1983). However, other approaches have also been used, for example, stepwise multiple linear regression (Dalal and Henry, 1986), principal components regression (PCR) (Chang et al., 2001), artificial neural networks (ANN) (e.g. Daniel et al., 2003), multivariate adaptive regression splines (MARS) (Shepherd and Walsh, 2002), boosted regression trees (Brown et al., 2006), PLSR with bootstrap aggregation (bagging-PLSR) (Viscarra Rossel, 2007), support vector machines SVM and penalised spline signal regression (Stevens et al., 2008).

The aims of this paper are: (i) to compare different state-of-the-art regression algorithms from the broader field of data mining for modelling soil vis–NIR diffuse reflectance spectra and (ii) to provide interpretation of the results by analysing feature selection approaches provided by the best performing algorithms.

Surely, the question arises why another study on different calibration approaches? The problem is, that different studies return different results, which is based on the fact that the nature of the target function has a strong influence of the performance of the different prediction approaches (Friedman, 2001). So here, we compare the performance of different algorithms for prediction of three soil properties on a large spectral library.

2. Material and methods

2.1. The spectral library

The soil samples in the vis–NIR library used in this research were not sampled specifically for this work. They are historical soil samples collected by various people for other research and for commercial agronomic purposes. They originate from Queensland, New South Wales, South Australia, Victoria and Western Australia. Soils were sampled from the 0–10 cm, 10–20 cm, 30–60 cm and 70–80 cm layers. Approximately 50% of all 1104 samples were surface soils (0–10 cm), 20% were from the 10–20 cm and the remaining samples were from deeper layers down to 80 cm. The soils were diverse and represented by various Australian Soil Classification orders, including Vertosols, Ferrosols, Kurosols, Chromosols, Dermosols, Sodosols and a smaller number of Podosols, Rudosols, Kandosols, Tenosols and Calcarosols (Isbell, 2002). Their approximate World Reference Base for Soil Resources classification is: Vertisols, Ferralsols, Planosols, Luvisols, Ferric Calcisols, Solonetz, Podzols, Leptosols and Calcisols (FAO, 1998).

2.2. Sample preparation and laboratory analysis

Samples were air-dried and ground to a particle size $\leq$2 mm before submitting a part of each to soil organic carbon (SOC) analysis using the dichromate oxidation method and pH using a 1:5 soil to H_2O solution ratio (pH), both described in Rayment and Higginson (1992). Clay content (CC) was measured using the hydrometer particle size analyses method (Gee and Bauder, 1986).

2.3. vis–NIR spectroscopy

The diffuse reflectance spectra were measured using the Agrispec® vis–NIR spectrometer (Analytical Spectral Devices, Bolder, Colorado, USA) with a spectral range of 350–2500 nm. The soils were scanned using a contact probe (Analytical Spectral Devices, Bolder, Colorado, USA) and a Spectralon® panel was used for white referencing once every 10 measurements. Each spectrum was made up of 876 wavelengths (or features) and thus, the vis–NIR spectral library consisted of 1104 samples and 876 predictor variables. The vis–NIR spectra were recorded as percent reflectance (R%).

The spectra were compressed using a principal component analysis (PCA), which was used to summarise the data and examine its structure. The PCA scores were submitted to a *k*-means clustering algorithm (Steinhaus, 1956) to classify the different types of spectra in the spectral library. A scree plot of the percent variance explained by cluster was used to select the number of clusters. The continuum-removed spectrum (Clark and Roush, 1984) of each of the resulting classes was used to help interpret the main absorption features in the spectra.

2.4. Algorithms compared

We tested the performance of different techniques for calibrating vis–NIR reflectance spectra to SOC, CC and pH. These included methods such as partial least squares regression (PLSR), multivariate adaptive regression splines (MARS), Support vector machines (SVM), random forests (RF), and boosted trees (BT). For brevity, we provide only a summary of each of these techniques, including the wavelet analysis and feature selection techniques, in the following sections. We cite some key references where the interested reader may find more detailed information.

2.4.1. Partial least squares regression (PLSR)

Wold et al. (1983) developed partial least squares regression (PLSR). It has become a popular technique used in chemometrics that is used for quantitative analysis of diffuse reflectance spectra. PLSR is closely related to principal components regression (PCR). It is used to construct predictive models when there are many predictor variables that are highly collinear. Both PLSR and PCR compress the data prior to performing the regression. However, unlike PCR, the PLSR algorithm integrates the compression and regression steps and it selects successive orthogonal factors that maximize the covariance between predictor and response variables. The number of factors to use in the models is selected by cross validation. By fitting a PLSR model, one hopes to find a few PLSR factors that explain most of the variation in both predictors and responses (Martens and Næs, 1989).

2.4.2. Multivariate adaptive regression splines (MARS)

MARS is a non-parametric regression approach introduced by Friedman (1991). It is a generalization of recursive partitioning regression approaches such as *classification and regression trees* (CART, Breiman et al., 1984), which generate piece-wise linear models instead of piece-wise constant models like CART. By applying linear basis functions between the splits of the partitioned space, MARS overcomes the "fairly severe restriction" (Friedman, 1991) of the approximation functions used in regression trees, which return discontinuous response surfaces (Burr, 1994). Hence, prediction accuracy can be expected to be higher with MARS, at least when the underlying function is continuous (Friedman, 1991). Technically, the generated piece-wise functions are aggregated in terms of an additive model. In 1996 MARS was described as "one of the most important advances in applied statistics in the last 10 years" (Burr, 1994).

2.4.3. Random forests (RF)

One of the most recent improvements in ensemble learning, which is already widely used, is RF (Breiman, 2001; Grimm et al., 2008). RF aggregates multiple predictions based on changes in the training dataset through resampling (Maclin and Opitz, 1997). It consist of multiple classification or regression trees (Breiman, 2001) generated based on a combination of bootstrap aggregation (or bagging) (Breiman, 1996) and the random subspace method (Ho, 1998) applied at each split in the tree. In regression the final prediction is the average of the individual tree outputs, whereas in classification, trees are voted by majority (Breiman, 2001). Generally, the most important advantages of RF, which

makes it a favourable choice for soil predictions, are: suitability for datasets with many predictors (features) and few samples (instances), robustness to noise and irrelevant features, it does not overfit, and almost no fine-tuning of the parameters is necessary to produce good predictions (Díaz-Uriarte and de Andrés, 2006).

2.4.4. Boosted regression trees (BT)

We apply stochastic gradient boosting for decision trees also known as TreeNet or MART (Friedman, 2001, 2002). Similar to MARS, BT produces an additive regression model (Friedman, 2002) or a tensor product based approximation (Friedman, 2001). As boosting is based on multiple predictions based on resampling and weighing it also belongs to the group of ensemble techniques. In stochastic gradient boosting small trees or stumps are sequentially built based on the residuals of the preceding tree(s). An overview of the mathematical and technical details behind stochastic gradient boosting can be found in Ridgeway (2008). Friedman (2001) shows that BT is competitive to MARS.

2.4.5. Support vector machines (SVM)

SVM are a kernel-based learning method from statistical learning theory (Vapnik, 1995). Kernel-based learning methods use an implicit mapping of the input data into a high dimensional feature space defined by a kernel function (Karatzoglou et al., 2008). Using this so-called kernel-trick (Boser et al., 1992) it is possible to derive a linear hyperplane as a decision function for non-linear problems and then apply a back-transformation in the non-linear space. We use radial basis function kernel, which is the typical general-purpose kernel (Karatzoglou et al., 2008). SVM is reported to return comparable results to RF models (Díaz-Uriarte and de Andrés, 2006; Bosch et al., 2007). However, Behrens and Scholten (2006) found SVM to perform poorly compared to other machine learning approaches in digital soil mapping.

2.5. The discrete wavelet transform (DWT)

The performance of the previously described algorithms was also tested using wavelet coefficients in the regressions. We used wavelets to produce a more sparse representation of the spectra so as to use a much smaller, more parsimonious number of wavelet coefficients in the models being compared. We implemented the wavelet transform using the pyramid algorithm and Daubechies wavelets with 2 vanishing moments (Daubechies, 1988). Once the wavelet decomposition was performed, we wanted to retain only those coefficients that would be useful in the models. We followed the methodology presented in Viscarra Rossel and Lark (2009), which involved the selection of the wavelet coefficients based on their variance, regardless of wavelet scale. The selected coefficients were then used as the predictors in the techniques described above.

2.6. Feature selection

The RF, PLSR, and MARS algorithms provide feature selection or ranking procedures in the spectral domain, while the wavelet technique (described above) also selects important features but in the wavelet domain. As RF does not return a selected subset of features, but a relative measure of feature importance for all features, we only tested feature selection approaches based on PLSR, MARS and DWT. To select the important features from the PLSR model we used the variable importance for projection (FS_{VIP}, Wold et al., 2001), which is calculated by:

$$FS_{VIP_k}(a) = K \sum_a w_{ak}^2 \left(\frac{SSY_a}{SSY_t}\right),$$

where $VIP_k(a)$ is the importance of the kth predictor variable based on a model with a factors, w_{ak} is the corresponding loading weight of the kth variable in the ath PLSR factor, SSY_a is the explained sum of squares of the response variable, by a PLSR model with a factors, SSY_t is the total sum of squares of the response variable, and K is the total number of predictor variables. The important features for each of the PLSR models were identified by selecting those that occurred at wavelengths where peak maxima occurred above a VIP threshold of 1 (Chong and Jun, 2005).

For the MARS approach (FS_{MARS}) we use the actual features that are used in the final MARS models.

Analysing feature importance (Liu and Motoda, 1998) is conducted for two main reasons: (i) it aims to reduce the dimensionality of datasets by eliminating redundant and/or noisy features and thus might increase prediction accuracy (Pechenizkiy et al., 2003); (ii) knowledge about important features helps understanding and validating the models. With respect to soil spectroscopy it allows to identify the physical basis for the predictions. We compared and interpreted the selected features from each of these techniques by analysing the occurrence of vis–NIR overtone and combination bands of organic and mineral soil constituents that were calculated from the fundamental mid infrared (mid-IR) frequencies for each of the soil constituents.

2.7. Algorithms implemented on the selected features

We tested the use of (i) multiple linear regressions with quadratic polynomials (MLR_{QP}) and (ii) artificial neural networks (ANN) on the selected wavelet coefficients and on the variables selected using feature selection techniques. MLR was used only on the selected features and not on all of the spectra to prevent problems with collinearity and ill-conditioned regressions. ANNs were used only on the selected features because of computational limitations of running the ANNs on all of the spectra.

2.7.1. Multiple linear regression (MLR)

MLR was also tested on the wavelet coefficients and the selected features resulting from the feature selection approaches. As the selected wavelet coefficients are strongly decorrelated and the feature selection techniques provide only a small number of predictors, they could be used directly in a MLR without serious numerical problems. We used MLR with quadratic polynomials to account for non-linear response in the data. To ensure numerical stability, the least squares regression coefficients, were estimated by the QR decomposition (Lawson and Hanson, 1974).

2.7.2. Artificial neural networks (ANN)

The design and the basic concept of ANNs has been adopted from data processing in biological nervous systems, as there are cells for the reception of information, others for its forwarding and storage, and another group for the outward release of information. The network topology is characterised by the number of layers and units per layer. The first concepts date back to McCulloch and Pitts (1943). In the learning phase the weights between the cells are optimized. An ANN based on the backpropagation (of error) algorithm (Rumelhart et al., 1986) minimizes the learning error in the inverse direction from the output layer towards the input layer (Gallant, 1993). Backpropagation is a method for computing the gradient of the case-wise error function with respect to the network to minimize the overall network error (Sarle, 1995). To prevent overfitting, for each soil property we used only one hidden node and an overfitting penalty (or weight decay), which puts a penalty on the size of the parameter estimates.

2.8. Assessment of the techniques

Validation as well as model fitting was carried out using a 10-fold leave-group-out cross validation approach also known as repeated

random sub-sampling validation (Vapnik, 1995). The accuracy of the cross validation is given by the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2}$$

where X_i is the predicted value and Y_i is the observed value.

To compare the best performing algorithms, the Akaike Information Criterion (AIC) was used to determine the method that most satisfactorily compromised between model accuracy and model parsimony (Akaike, 1973). The AIC was calculated by:

$$AIC = n \ln RMSE + 2p$$

where n is the number samples and p the number of features used in the prediction. The model with the smallest AIC is best. We also compared the results using the coefficient of determination (R^2).

3. Results and discussion

3.1. Soil properties and the spectra

There was a large variation in soil OC and CC and pH (Fig. 1), which reflect the different land uses and the different depths from which the samples originate.

To aid with interpretation, the PCA scores were classified into two groups using the *k*-means algorithm (Fig. 2) and their average reflectance spectra are shown in Fig. 3a. To highlight the absorption features of the spectra in each group, the continuum-removed spectrum of each is shown in Fig. 3b. The spectra in both groups have absorptions due to charge transfers near 430 nm, 485 nm and 650 nm (Fig. 3b), which are indicative of the presence of the iron oxides (Sherman and Waite, 1985). The spectrum in the first group has a very slight absorption near 920 nm (Fig. 3b), indicating the presence of small amounts of goethite (FeOOH). The 650 nm shoulder in the spectrum of the second group (Fig. 3b) may indicate the presence of small amounts of haematite (Fe_2O_3).

The strong absorptions near 1412 nm and 1908 nm in the spectra of both groups indicate the occurrence of water bound in the interlayer lattices (Bishop et al., 1994). The absorption near 2204 nm in group one is due to the absorption of Al–OH, and the small absorption near 2280 nm may be due to Fe–OH as Fe is substituted in the octahedral sheet, e.g. in montmorillonite. In the spectrum of group one, the absorption near 2380 nm, the slight shoulder near 2350 nm, as well as that near 2345 nm may represent illite or mixtures of smectite and illite (Post and Noble, 1993). In the spectrum of group two, the absorption

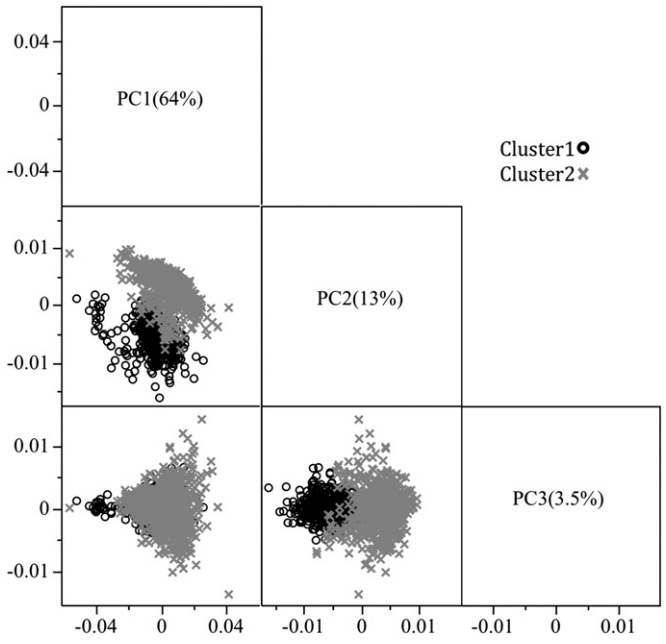

Fig. 2. Scatter plots of the first three principal component (PC) scores showing the two spectral clusters.

near 2165 nm and 2204 nm result from the Al–OH bend plus O–H stretch combination vibrations in poorly ordered kaolin. Slight absorptions near 2380 nm and 2345 nm (Fig. 3b) also indicate the presence of some illite in these group two soils. Therefore, on average, group one soils contain some goethite and are primarily mixed layer smectitic–illitic soils, while group two soils contain more haematite and are primarily poorly ordered kaolinitic soils.

Table 1 shows the correlation between the soil properties and the first two principal component scores, for all of the data as well as for each group.

Overall, SOC was negatively correlated to both CC (correlation coefficient, $\rho = -0.53$) and pH ($\rho = -0.45$), while there was a positive correlation between CC and pH ($\rho = 0.62$). PC1 was better correlated to SOC ($\rho = 0.54$) than to pH ($\rho = -0.26$) or CC ($\rho = -0.2$), while PC2 was better correlated to CC ($\rho = -0.72$) and pH ($\rho = -0.66$) than to SOC ($\rho = 0.5$) (Table 1).

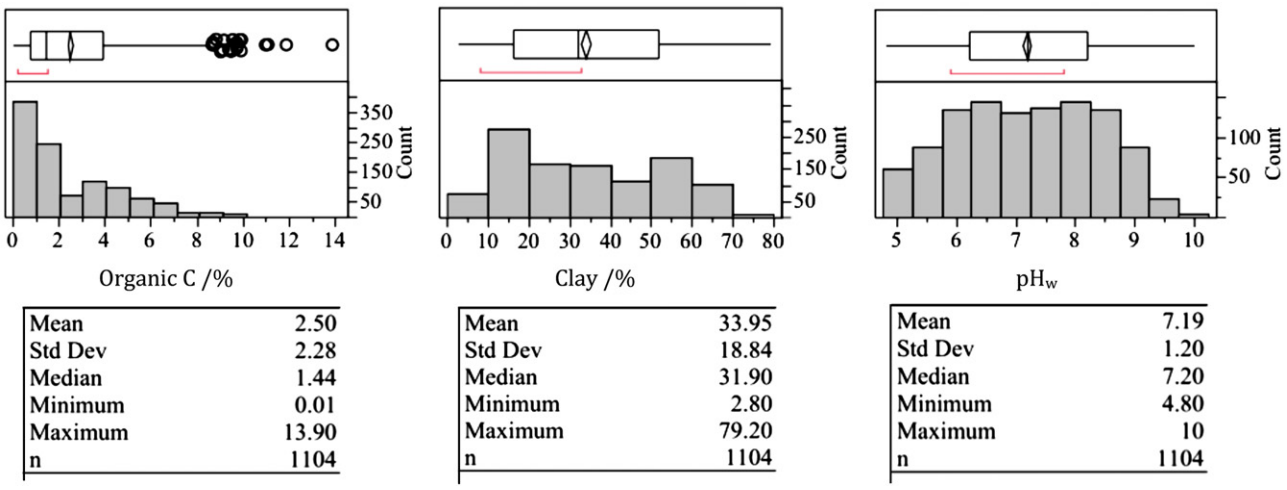

Fig. 1. Histograms, outlier box-plots and descriptive statistics for the three soil properties.

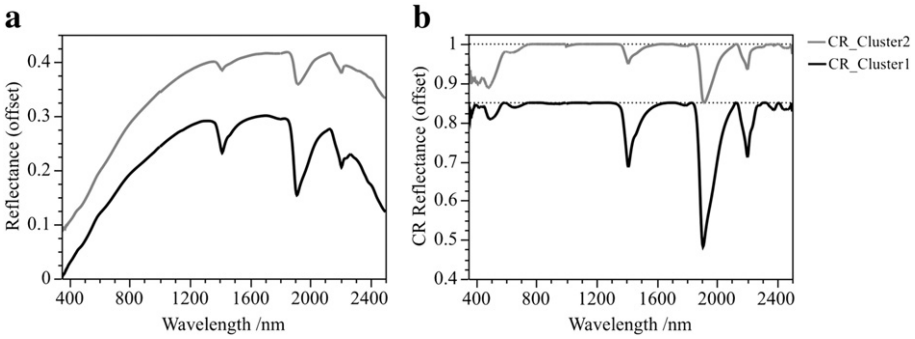

Fig. 3. The vis–NIR spectra of the two spectral clusters showing differences in mineral composition: (a) are the reflectance spectra and (b) the continuum-removed (CR) spectra.

Table 1
Upper triangular correlation matrix between soil properties and the first two principal component scores of the spectra.

All	Organic C	pH_w	Clay	PC1(64%)	PC2(13%)
Organic C	1	−0.45	−0.53	0.54	0.5
pH_w		1	0.62	−0.26	−0.66
Clay			1	−0.20	−0.72
PC1(64%)				1	0
PC2(13%)					1

3.2. Comparison of predictions using different algorithms

3.2.1. Calibration

The comparison of prediction accuracy and model parsimony from the different algorithms are presented in Table 2.

The upper part of Table 2 show results from the five data mining algorithms, the middle section results from the models that use the DWT coefficients, and the lower part results from the neural network models that use the DWT coefficients and specific wavelengths selected using the FS_{VIP} and FS_{MARS} feature selection techniques.

3.2.1.1. Data mining algorithms. The SVM model with all 876 vis–NIR wavelengths produced the smallest prediction RMSEs for the three soil properties, SOC (RMSE = 0.92%), CC (RMSE = 7.63%) and pH (RMSE = 0.61 pH units) (Table 2). However, predictions by PLSR and MARS were close with RMSE values of 0.96% and 1.02% SOC, 7.77% and 7.79% CC, respectively. Predictions of pH produced identical RMSE of 0.63 pH units. The PLSR model for SOC used 11 factors, for CC 14 factors and for pH 18 factors. For all three soil properties, the RF and BT models produced the largest prediction RMSE values and were thus the least accurate (Table 2).

3.2.1.2. Modelling with the wavelet coefficients. When we ordered the wavelet coefficients in decreasing order by their variance, the coefficients of the coarser scales accounted for the largest variances and their contribution decreased with scale. Fig. 4 was used to determine which and how many wavelet coefficients to retain.

For the three soil properties (Fig. 4a, b and c), as successive wavelet coefficients are added we see first a reduction in the RMSE to a minimum and then a gradual increase. The reason for this is that original vis–NIR 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 we add information from the spectrum to the predictor and the RMSE is reduced. However, as we continue to add coefficients, increasingly we add those that correspond to noise and have no predictive power, and the RMSE rises again. Fig. 4a shows that 72 coefficients produced the smallest RMSEs when used to predict SOC content. For CC, it was 132 coefficients (Fig. 4b) and for pH it was 137 coefficients (Fig. 4c). In all cases, using additional wavelet coefficients in the regressions only enlarged the RMSEs. The selected coefficients were then used as the predictor variables in the spectroscopic models using each of the techniques being compared.

Table 2
Comparison of predictions using different algorithms and combinations. PLSR models used 11 factors for organic C, 14 for clay content and 18 for pH_w.

	Organic C				Clay				pH_w			
	p	RMSE/%	R^2	AIC	*p*	RMSE/%	R^2	AIC	*p*	RMSE/%	R^2	AIC
PLSR	876	0.96	0.82		876	7.77	0.83		876	0.63	0.73	
MARS	876	1.02	0.80		876	7.79	0.83		876	0.63	0.72	
SVM	876	0.92	0.84		876	7.63	0.84		876	0.61	0.75	
BT	876	1.49	0.62		876	9.44	0.75		876	0.77	0.62	
RF	876	1.23	0.71		876	8.93	0.77		876	0.73	0.63	
DWT–PLSR	72	1.03	0.81		132	7.82	0.82		137	0.63	0.72	
DWT–MARS	72	0.88	0.85		132	7.59	0.84		137	0.63	0.73	
DWT–SVM	72	0.87	0.86		132	7.35	0.85		137	0.62	0.75	
DWT–BT	72	0.99	0.83		132	8.37	0.81		137	0.66	0.70	
DWT–RF	72	0.93	0.84		132	7.53	0.84		137	0.63	0.72	
DWT–MLR_{QP}	72	0.91	0.84		132	7.73	0.83		137	0.62	0.74	
FS_{VIP}–MLR_{QP}	29	0.98	0.82		31	7.63	0.84		29	0.61	0.74	
FS_{MARS}–MLR_{QP}	14	0.99	0.81		13	7.35	0.85		11	0.60	0.75	
DWT–ANN	72	0.75	0.89	−3078	132	6.42	0.88	−2470	137	0.53	0.81	−2247
FS_{VIP}–ANN	29	0.83	0.87	−3052	31	6.80	0.87	−2608	29	0.56	0.79	−2402
FS_{MARS}–ANN	14	0.82	0.87	−3095	13	7.08	0.86	−2600	11	0.58	0.77	−2399

Note: *p* is the number of features in the model.

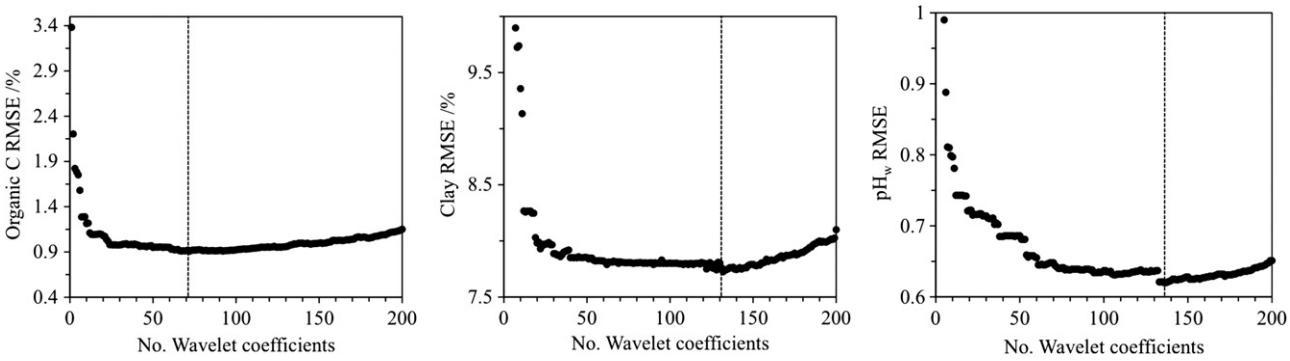

Fig. 4. Root mean-square errors (RMSE) of predictions for soil organic carbon, clay content and pH, showing the best number of wavelet coefficients to use in the regressions. Data shown on the graphs is for the first 200 wavelet coefficients only.

The SVM model using the wavelet coefficients (DWT–SVM) produced the smallest prediction RMSEs for the three soil properties (Table 2). For SOC, the DWT–SVM predictions (RMSE = 0.87%) were only slightly better than those of the MARS model using the DWT (DWT–MARS: RMSE = 0.88%), the DWT–MLR_{QP} (RMSE = 0.91%) and the DWR–RF (RMSE = 0.93%). Predictions of SOC using the DWT–BT and DWT–PLSR were poorest with RMSE values of 0.99% and 1.03%, respectively (Table 2). For CC, the DWT–SVM predictions (RMSE = 7.35%) were again better than those of the DWT–RF (RMSE = 7.53%) and the DWT–MARS model (RMSE = 7.59%). The DWT–MLR_{QP} produced an RMSE of 7.73%, which was better than the DWT–PLSR mode (RMSE = 7.82%). The poorest predictions were obtained from the DWT–BT model with an RMSE of 8.37% (Table 2). Predictions of pH were less variable than those for clay or SOC. Both the DWT–SVM and the DWT–MLR_{QP} techniques produced an RMSE of 0.62 pH units and the DWT–PLSR, DWT–MARS and DWT–BT an RMSE of 0.63 pH units. Like for CC, DWT–BT produced the poorest predictions with an RMSE of 0.66 pH units (Table 2). Compared to the untransformed dataset, modelling with wavelet coefficients resulted in higher prediction accuracies (lower RMSEs). However, the general trend in the performance of the algorithms was similar to the untransformed dataset. The performance of DWT–PLSR was poorer compared to DWT–MARS and DWT–SVM. DWT–RF was now competitive, whereas DWT–BT was not.

3.2.1.3. Selected features with multiple linear regressions (MLR). Interestingly, DWT–MLR_{QP}, FS_{VIP}–MLR_{QP} and FS_{MARS}–MLR_{QP} produced predictions of SOC, clay and pH that were comparable, and in occasions, better than those from more complex algorithms (Table 2). For example, the RMSE of the FS_{MARS}–MLR_{QP} predictions for CC was 7.35%, which is the same as that of the DWT–SVM prediction and for pH predictions were better (Table 2). These results suggest that the quadratic polynomials in the MLRs accounted for some of the non-linearity in the data and that with effective feature selection, good predictions can be achieved using relatively simple algorithms.

3.2.1.4. Selected features with artificial neural networks (ANN). The reason for not using ANN to directly model the vis–NIR spectra with all of its frequencies (in this case 876) was the exceedingly long computation time. However, as ANN is a flexible approach we tested it on the reduced data from the DWT as well as those features that were selected by the VIP (FS_{VIP}–ANN) and the MARS (FS_{MARS}–ANN) techniques.

The ANN model using the wavelet coefficients (DWT–ANN) produced the smallest prediction RMSEs for the three soil properties (Table 2). The RMSE of the DWT–ANN predictions of SOC was 0.75%, which was the best prediction RMSE for SOC of all of the methods tested. The RMSE of the FS_{VIP}–ANN and FS_{MARS}–ANN predictions were similar with values of 0.83% and 0.82%, respectively. Predictions of CC using DWT–ANN produced an RMSE of 6.42%. This was again the best prediction of all of the methods tested (Table 2). The prediction RMSE of the FS_{VIP}–ANN model was 6.8%, which was better than the 7.08% produced by the FS_{MARS}–ANN. For pH, the prediction RMSE of the DWT–ANN model was 0.53 pH units, which was also the best of all other techniques compared. The FS_{VIP}–ANN and the FS_{MARS}–ANN models produced similar RMSE values of 0.56 and 0.58 pH units, respectively (Table 2).

3.2.1.5. Comparing the best models. Using the ANNs on the reduced feature sets returned the best results for all three soil properties (Table 2). The DWT–ANN produced the smallest cross validation RMSE values, suggesting that the procedure used to select the wavelet coefficients was effective. The FS_{VIP}–ANN technique produced a smaller RMSE than FS_{MARS}–ANN for predictions of CC (Table 2). For SOC and pH the FS_{MARS}–ANN model returned only slightly lower

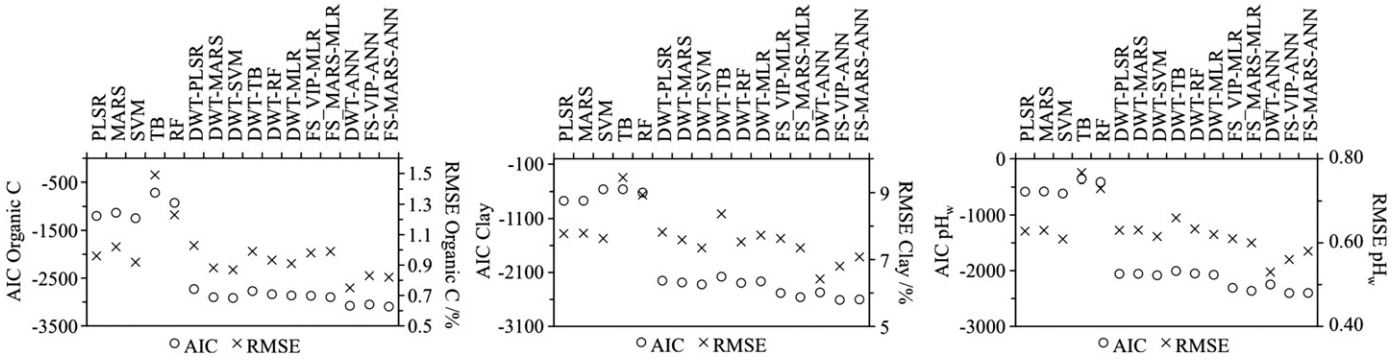

Fig. 5. Root mean squared error (RMSE) and Akaike Information Criterion (AIC) for predictions of soil organic carbon, clay content and pH using different data mining algorithms.

Table 3
Band assignments for fundamental mid-IR absorptions of soil constituents and their overtones and combinations in the vis–NIR. Compilation of fundamental mid-IR absorptions from: Oinuma and Hayashi (1965), White (1971), van der Marel and Beutelspacher (1975), Clark et al. (1990), Srasra et al. (1994), Madejova and Komadel (2005), Nayak and Singh (2007).

Soil constituent	Fundamental (cm^{-1})	vis–NIR wavelength (nm)	vis–NIR mode
Fe oxides			
Goethite		434, 480, 650, 920	Electronic transitions
Haematite		404, 444, 529, 650, 884	Electronic transitions
Water	ν_1 O–H 3278 cm^{-1}	1915	$\nu_2+\nu_3$
	ν_2 H–O–H 1645 cm^{-1}	1455	$2\nu_2+\nu_3$
	ν_3 O–H 3484 cm^{-1}	1380	$\nu_1+\nu_3$
		1135	$\nu_1+\nu_2+\nu_3$
		940	$2\nu_1+\nu_3$
Hydroxyl	ν_1 O–H 3575 cm^{-1}	1400	$2\nu_1$
		930	$3\nu_1$
		700	$4\nu_1$
Clay minerals			
Kaolin doublet	ν_{1a} O–H 3695 cm^{-1}	1395	$2\nu_{1a}$
	ν_{1b} O–H 3620 cm^{-1}	1415	$2\nu_{1b}$
	δ Al–OH 915 cm^{-1}	2160	$\nu_{1a}+\delta$
		2208	$\nu_{1b}+\delta$
Smectite	ν_1 O–H 3620 cm^{-1}	2206	$\nu_1+\delta_a$
	δ_a Al–OH 915 cm^{-1}	2230	$\nu_1+\delta_b$
	δ_b AlFe–OH 885 cm^{-1}		
Illite	ν_1 O–H 3620 cm^{-1}	2206	$\nu_1+\delta$
		2340	
		2450	Poorly defined
Carbonate	ν_3 CO_3^{-2} 1415 cm^{-1}	2336,	$3\nu_3$
Organics			
Aromatics	ν_1 C–H 3030 cm^{-1}	1650	$2\nu_1$
		1100	$3\nu_1$
		825	$4\nu_1$
Amine	δ N–H 1610 cm^{-1}	2060	$\nu_1+\delta$
	ν_1 N–H 3330 cm^{-1}	1500	$2\nu_1$
		1000	$3\nu_1$
		751	$4\nu_1$
Alkyl asymmetric–symmetric doublet	ν_3 C–H 2930 cm^{-1}	1706	$2\nu_3$
	ν_1 C–H 2850 cm^{-1}	1754	$2\nu_1$
		1138	$3\nu_3$
		1170	$3\nu_1$
		853	$4\nu_3$
		877	$4\nu_1$
Carboxylic acids	ν_1 C=O 1725 cm^{-1}	1930	$3\nu_1$
		1449	$4\nu_1$
Amides	ν_1 C=O 1640 cm^{-1}	2033	$3\nu_1$
		1524	$4\nu_1$
Aliphatics	ν_1 C–H 1465 cm^{-1}	2275	$3\nu_1$
		1706	$4\nu_1$
Methyls	ν_1 C–H 1445–1350 cm^{-1}	2307–2469	$3\nu_1$
		1730–1852	$4\nu_1$
Phenolics	ν_1 C–OH 1275 cm^{-1}	1961	$4\nu_1$
Polysaccharides	ν_1 C–O 1170 cm^{-1}	2137	$4\nu_1$
Carbohydrates	ν_1 C–O 1050 cm^{-1}	2381	$4\nu_1$

RMSE values (Table 2). Both the FS_{VIP}–ANN and FS_{MARS}–ANN models used a smaller number of features for the predictions than DWT–ANN. This is reflected in their AIC, which suggests that, when both the accuracy and parsimony of the model are taken into consideration, the best SOC model was the FS_{MARS}–ANN and for CC and pH it was the FS_{VIP}–ANN (Table 2). These results are also summarised in Fig. 5.

The performance of the ANN approaches is positively correlated to the number of features used in the models, indicating that ANNs are able to extract more relevant information when more features are used. Even though ANNs are 'black box' systems the combination of feature selection and non-linear modelling helps to achieve good predictions and interpretability.

3.3. Feature importance

It is possible to identify regions of the vis–NIR spectrum that show absorption due to water, mineral and organic matter in soil. The reason is that their fundamental molecular vibrations occur in the mid-IR, while their overtones and combinations occur in the NIR. In the visible–short wave NIR, the main process by which molecules absorb energy is electronic transitions in atoms from ground to higher energy states. The overtones and combinations in the NIR are much weaker than the fundamental bands in the mid-IR and can be difficult to distinguish with the naked eye. Their location is also often slightly shifted from the exact expected location because real molecules do not behave totally harmonically (Bishop et al., 1994). Therefore, interpretations of soil vis–NIR spectra can be difficult, particularly when the target species is present in only small amounts in the soil or when the soil property being studied is dependent on correlations to the target species. Table 3 presents a summary of important fundamental absorptions in the mid-IR and the occurrence of their overtones and combinations in the vis–NIR, which can be used to help with interpretation.

The absorption features in the visible–short wave NIR (400–1000 nm) are mostly due to the Fe oxides — in soil mainly haematite and goethite, while those in the NIR between 1000 nm and 2500 nm can be due to water, clay minerals and organic matter (Table 3).

Fig. 6 shows the features identified as important by the FS_{VIP} and FS_{MARS} algorithms, plotted on an average soil spectrum. The graphs in Fig. 6 also indicate the wavelengths that correspond to the iron, water, hydroxyl, mineral and organic components identified in Table 3.

For SOC, the FS_{VIP} and FS_{MARS} techniques selected 29 and 14 important wavelengths, respectively. Mostly, these coincided with wavelengths that are related to the iron oxides, water and organics, with a smaller number of wavelengths coinciding with those that are related to minerals (Fig. 6a). For CC, FS_{VIP} and FS_{MARS} selected, 31 and 13 important wavelengths, respectively. A large proportion of these coincided with wavelengths that are related to iron oxides, hydroxyl, water and minerals, and only a smaller number with wavelengths related to organics (Fig. 6b). For pH, the FS_{VIP} and FS_{MARS} techniques selected 29 and 11 important wavelengths, respectively, which coincided with wavelengths related to mostly iron oxides and organics, but also to organics (Fig. 6c).

Table 4 shows the assignment of absorption features for similar wavelengths selected by the FS_{VIP} and FS_{MARS} techniques.

4. Conclusion

Our results show that: (i) in the full spectral domain SVM produced the best predictions (lowest RMSE) for all three soil properties; (ii) transforming the spectra to the wavelet domain and performing the calibrations on the selected wavelet coefficients improved all of the predictions; (iii) combining good feature selection techniques with simple MLR_{QP} can produce comparable predictions to those using more complex algorithms, as long as there are only simple non-linear structures in the data; (iv) ANN on the wavelet coefficients produced the best predictions for all three soil properties, because ANN can model complex non-linear interaction in the data; (v) the use of ANN on a reduced number of wavelengths selected using feature selection techniques results in good calibrations that can be easily interpreted; (vi) both tree ensemble approaches (RF and BT) performed weakly in the spectral and wavelet domains, whereas PLSR and MARS produced competitive predictions to SVM. ANN based on feature subsets selected by VIP and MARS provide both high calibration performance as well as interpretability. In terms of interpretability, we show that (i) SOC is related to wavelengths that represent absorptions due to organic molecules and proteins with C—O, C=O, and N—H bonds; (ii) CC is related to wavelengths that represent absorptions that are due to iron oxides and clay minerals; and (iii) pH is related to wavelength that represents absorptions due to both organic material, iron oxides and clay minerals.

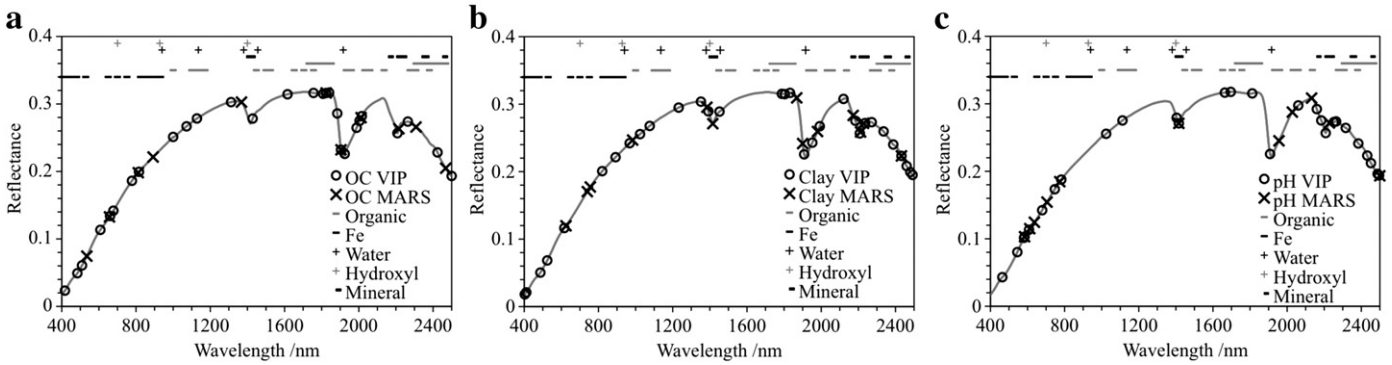

Fig. 6. Variable selection by VIP (o) and MARS (x), plotted on a sample spectrum for (a) organic carbon (OC), (b) clay content and (c) pH. Smaller symbols on the upper portion of the graphics correspond to wavelength positions for overtone and combination bands of organic and inorganic materials identified in Table 3.

Table 4
Absorption band assignments for vis–NIR wavelengths selected by both VIP and MARS algorithms.

Organic C absorptions (nm)			Clay absorptions (nm)			pH_w absorptions (nm)		
λ_{VIP}	λ_{MARS}	Possible assignments	λ_{VIP}	λ_{MARS}	Possible assignments	λ_{VIP}	λ_{MARS}	Possible assignments
660	657	Iron oxides (650)	608	625	Iron oxides (650)	584	584	Haematite (529)
816	812	$4\nu_1$ C–H (825)	988	984	Goethite (920)	608	612	Iron oxides (650)
					$3\nu_1$ O–H (930) $2\nu_1+\nu_3$ water (940)			
1824	1824	$4\nu_1$ C–H (1852)	1392	1384	$2\nu_{1a}$ O–H kaolin (1395)	782	774	$4\nu_1$ N–H (751)
					$\nu_1+\nu_3$ water (1380)			$4\nu_1$ C–H (825)
1904	1904	$\nu_2+\nu_3$ water (1915) $3\nu_1$ C=O (1930)	1908	1900	$\nu_2+\nu_3$ water (1915)	1416	1416	$2\nu_{1b}$ O–H kaolin (1415)
2016	2012	$3\nu_1$ C=O (2033)	2184	2172	$\nu_{1a}+\delta$ Al–OH kaolin (2160)	2496	2500	$3\nu_1$ C–H (2469)
		$\nu_1+\delta$ *N*–H (2060)						
2208	2216	$\nu_{1b}+\delta$ Al–OH kaolin (2208)	2208	2212	$\nu_{1b}+\delta$ Al–OH kaolin (2208)			
			2236	2228	$\nu_1+\delta_b$ AlFe–OH smectite (2230)			
			2432	2432	Poorly defined illite band (2450)			

References

Akaike, H., 1973. Information theory and an extension of maximum likelihood principle. In: Petrov, B.N., Csáki, F. (Eds.), Second International Symposium on Information Theory. Akadémia Kiadó, Budapest, Hungary, pp. 267–281.

Behrens, T., Scholten, T., 2006. A comparison of data mining approaches in predictive soil mapping. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping. Developments in Soil Science 31. Elsevier, p. 658.

Bishop, J.L., Lane, M.D., Dyar, M.D., Brown, A.J., 1994. Reflectance and emission spectroscopy study of four groups of phyllosilicates: smectites, kaolinite–serpentines, chlorites and micas. Clay Minerals 43, 35–54.

Bosch, A., Zisserman, A., Munoz, X., 2007. Image classification using random forests and ferns. IEEE 11th International Conference on Computer Vision, 2007, pp. 1–8.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Haussler, D. (Ed.), 5th Annual ACM Workshop on COLT. ACM Press, Pittsburgh, PA, pp. 144–152.

Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth, Belmont (CA).

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132, 273–290.

Burr, T.L., 1994. Predicting Linear and Nonlinear Time Series with Applications in Nuclear Safeguards and Nonproliferation. Los Alamos National Laboratory, Technical report LA-12766-MS / UC-700. URL: http://www.osti.gov/bridge/servlets/purl/10145200-uTyt76/native/10145200.PDF.

Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh Jr., C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analysis of soil properties. Soil Science Society of America Journal 65, 480–490.

Chong, I.G., Jun, C.H., 2005. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 78, 103–112.

Clark, R.N., 1999. Spectroscopy of rocks and minerals, and principles of spectroscopy. In: Rencz, N. (Ed.), Remote Sensing for the Earth Sciences: Manual of Remote Sensing. John Wiley & Sons, New York, pp. 3–52.

Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. Journal of Geophysical Research 89, 6329–6340.

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.

Dalal, R.C., Henry, R.J., 1986. Simultaneous determination of moisture, organic carbon and total nitrogen by near infrared reflectance spectrophotometry. Soil Science Society of America Journal 50, 120–123.

Daniel, K.W., Tripathi, N.K., Honda, K., 2003. Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). Australian Journal of Soil Research 41, 47–59.

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

Díaz-Uriarte, R., de Andrés, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3.

FAO, 1998. World reference base for soil resources. Food and Agriculture Organization of the United Nations, Rome.

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

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29, 1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. Computational Statistics and Data Analysis, Volume 38, Number 4. Elsevier, pp. 367–378.

Gallant, S.I., 1993. Neural Network Learning and Expert Systems. MIT Press, Cambridge, MA. 365.

Gee, G.W., Bauder, J.W., 1986. Particle-size analysis, methods of soil analysis, part 1. Physical and Mineralogical Methods, 2nd edition: Agronomy Monograph No. 9. ASA-CSSA-SSSA, Madison, Wisconsin, pp. 383–411.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, A., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — digital soil mapping using Random Forests analysis. Geoderma 146 (1–2), 102–113.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8), 832–844.

Hunt, G.R., Salisbury, J.W., 1970. Visible and near-infrared spectra of minerals and rocks. I. Silicate minerals. Modern Geology 1 (4), 283–300.

Isbell, R.F., 2002. The Australian Soil Classification: Australian Soil and Land Survey Handbook. CSIRO Publishing, Melbourne.

Karatzoglou, A., Smola, A., Hornik, K., 2008. kernlab: Kernel-based Machine Learning Lab. (At: http://cran.r-project.org/web/packages/kernlab/index.html. Accessed: 24/03/2009).

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

Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery and Data Mining. Kluver.

Maclin, R. and Opitz, D. 1997. An empirical evaluation of bagging and boosting. In Proceedings of the Fourteenth National Conference on Artificial Intelligence, pp. 546–551 Cambridge, MA. AAAI Press/MIT Press.

Madejova, J., Komadel, P., 2005. Baseline studies of the clay minerals society source clays: infrared methods. Clays and Clay Minerals 49, 410–432.

Martens, H., Næs, T., 1989. Multivariate Calibration. John Wiley & Sons, Chichester. 419 pp.

McCulloch, W., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Biophysics 7, 115–133.

Nayak, P.S., Singh, B.K., 2007. Instrumental characterization of clay by XRF, XRD and FTIR. Bulletin of Material Sciences 30, 235–238.
Oinuma, K., Hayashi, H., 1965. Infrared study of mixed-layer clay minerals. The American Mineralogist 50, 1213–1227.
Pechenizkiy, M., Puuronen, S., Tsymbal, A., 2003. Feature extraction for classification in knowledge discovery systems. Proc. 7th Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems, Berlin, pp. 526–532.
Post, J.L., Noble, P.N., 1993. The near-infrared combination band frequencies of dioctahedral smectites, micas, and illites. Clays and Clay minerals 41, 639–644.
Rayment, G.E., Higginson, F.R., 1992. Australian Laboratory Handbook of Soil and Water Chemical Methods. Inkata Press, Australia.
Ridgeway, G., 2008. gbm: Generalized Boosted Regression Models. (At: http://cran.r-project.org/web/packages/gbm/index.html. Accessed: 24/03/2009).
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Learning Internal Representations by Error Propagation. The MIT Press, Cambridge, MA, pp. 318–362.
Sarle, W., 1995. Stopped training and other remedies for overfitting. 27th Symposium on the Interface Computing Science and Statistics, Pittsburgh, PA.
Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Science Society of America Journal 66, 988–998.
Sherman, D.M., Waite, T.D., 1985. Electronic spectra of Fe^{3+} oxides and oxyhydroxides in the near infrared to ultraviolet. American Mineralogist 70, 1262–1269.
Srasra, E., Bergaya, F., Fripiat, J.J., 1994. Infrared spectroscopy study of tetrahedral and octahedral substitutions in an interstratified illite–smectite clay. Clays and Clay Minerals 42, 237–241.
Steinhaus, H., 1956. "Sur la division des corp materiels en parties" (in French). Bull. Acad. Polon. Sci.query 4 (12), 801–804.
Stevens, A., Van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., Ben-Dor, E., 2008. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. Geoderma 144, 395–404.
van der Marel, H.W., Beutelspacher, H., 1975. Atlas of infrared spectroscopy of clay minerals and their admixtures. Elsevier Scientific Publishing Company, Amsterdam, The Netherlands.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag.
Viscarra Rossel, R.A., 2007. Robust modelling of soil diffuse reflectance spectra by bagging-partial least squares regression. Journal of Near Infrared Spectroscopy 15, 39–47.
Viscarra Rossel, R.A., Lark, R.M., 2009. Improved analysis and modelling of soil diffuse reflectance spectra using wavelets. European Journal of Soil Science 60, 453–464.
Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75.
Viscarra Rossel, R., Cattle, S., Ortega, A., Fouad, Y., 2009. In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy. Geoderma 150, 253–266.
White, W.B., 1971. Infrared characterisation of water and hydroxyl ion in the basic magnesium carbonate minerals. The American Mineralogist 56, 46–53.
Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration method in chemistry solved by the PLS method. In: Ruhe, A., Kagstrom, B. (Eds.), Proc. Conf. Matrix Pencils, Lecture Notes in Mathematics. Springer-Verlag, Heidelberg, pp. 286–293.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58, 109–130.