Computers and Electronics in Agriculture 158 (2019) 201–210

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Optimizing wavelength selection by using informative vectors for parsimonious infrared spectra modelling

Wartini Ng[a,*], Budiman Minasny[a], Brendan P. Malone[a], M.C. Sarathjith[b], Bhabani S. Das[c]

[a] School of Life and Environmental Sciences, Sydney Institute of Agriculture, The University of Sydney, NSW 2006, Australia
[b] Kelappaji College of Agricultural Engineering and Technology, Malappuram, Kerala, India
[c] Agricultural and Food Engineering Department, Indian Institute of Technology Kharagpur, West Bengal, India

ARTICLE INFO

Keywords:
Variable selection
Wavelength selection
Informative vector
Near-infrared spectroscopy
Soil inference

ABSTRACT

Infrared spectroscopy has been widely adopted by various agricultural research. The typical spectra variables contain thousands of wavelengths. These large number of spectra variables often contribute to collinearity, and redundancies rather than relevant information. Variable selection of the predictors is an important step to create a robust calibration model from these spectra data. This paper presents an algorithm for spectra variable selection based on a combination of informative vectors and an ordered predictor selection (OPS) approach with an exponentially decreasing function (EDF) selection. Informative vectors are features derived from statistical principles that can be used to describe the relationship between the dependent variables and the predictors (spectra). The informative vectors analysed include regression coefficient vector (*b*), variable influence on projection *(V)*, residual vector (*S*), net analyte signal vector (*Na*), linear correlation vector (*COR*), biweight mid-correlation vector (*BIC*), mutual information based on adjacency matrix (*AMI*), covariance procedures matrix (*COV*). These eight informative vectors can be joined in pairs and become 22 combination vectors. This approach was tested with near-infrared soil spectra for predicting the properties of pH, clay and sand content, cation exchange capacity (CEC), and total carbon content. This example used the Cubist regression tree and partial least squares regression (PLSR) models for calibration. By utilizing the subset of the spectra (retaining those that are significant based on the absolute values of the informative vectors), the regression models were still able to enhance the prediction capability. Overall, the PLSR model performed better than the Cubist model. The informative vector *b* (and its combinations) and *S* (and its combinations) were found to be the ones that provide the most accurate predictions for this dataset. Although the performance of the subset model does not perform better than the full spectra model, the number of wavelengths variable used in the model is significantly reduced to, on average, 25%.

1. Introduction

Infrared spectroscopy has been widely adopted for characterizing materials in various fields. Near Infrared (NIR) has been utilized to detect salt concentration in wastewater (Inagaki et al., 2010) and heavy metals contaminations in soil (Kemper and Sommer, 2002). In the food industries, NIR spectroscopy has been used to detect the quality of fruit (Kurz et al., 2010), and milk (Wu et al., 2008) products. NIR spectroscopy was demonstrated to be able to identify seedling quality (Shrestha et al., 2017), origins of coffee beans (Marquetti et al., 2016), and tea polyphenol concentration (Li et al., 2015).

In soil science, several soil properties such as soil organic matter content, total nitrogen content, pH, cation exchange capacity (CEC) and soil texture can be effectively estimated using infrared spectroscopy (Chang et al., 2001; Shepherd and Walsh, 2002; Islam et al., 2003; Viscarra Rossel et al., 2005; Stenberg et al., 2010). By measuring the infrared reflectance of grapevine leaves, Pascoa et al. (2016) were able to discriminate between soil types in vineyards.

In general, infrared spectroscopy provides an alternative to conventional analytical methods because it is rapid, less-expensive and non-destructive. Furthermore, with proper calibration, several properties of a material may be characterized from the single spectrum (Viscarra Rossel et al., 2005; Stenberg et al., 2010; Horta et al., 2015). When a sample is illuminated with electromagnetic radiation, it interacts and induces vibrations of chemical bonds in different molecules present in that sample (Viscarra Rossel et al., 2010). Absorptions of

* Corresponding author at: Sydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, Biomedical Building, 1 Central Avenue, Australian Technology Park, Eveleigh, NSW 2015 Australia.
E-mail address: wartini.ng@sydney.edu.au (W. Ng).

https://doi.org/10.1016/j.compag.2019.02.003
Received 3 May 2018; Received in revised form 29 January 2019; Accepted 7 February 2019

such electromagnetic energy in different wavelength regions create a basic stretching and bending of various bonds providing a unique spectral response in the form of reflectance or absorption features across potentially thousands of individual wavelengths. Extraction of these spectral features from measured infrared spectra is difficult because of the complex mixtures and the scattering effects (Horta et al., 2015). Typically, a chemometric approach is employed to retrieve these features to create a spectral algorithm for inferring a specific property of the material. Unfortunately, the presence of high dimensional or ultra-spectral data poses many challenges, mainly because the number of spectral variables is relatively larger than the number of samples; also known as "large p and small n" problems (Mehmood et al., 2012). Moreover, the spectra variables are often highly correlated; highly collinear variables could deplete the model predictions (Li et al., 2009; Vohland et al., 2014). Due to such high dimensional information, statistical models created with these irrelevant variables often cause overfitting - not all variables are equally crucial for the predictive models. Thus, the analysis of spectral data requires efficient feature extraction and multivariate calibration approaches.

In general, both linear and non-linear multivariate techniques such as partial least squares (PLS) regression, principal component regression (PCR) (Chang et al., 2001), stepwise multiple linear regression (Dalal and Henry, 1986), regression trees (Brown et al., 2006), support vector machines (Devos et al., 2009), multivariate adaptive regression splines (Shepherd and Walsh, 2002) and artificial neural networks (ANN) (Daniel et al., 2003) have been used for analysing spectral data. Among these, the PLS regression approach is the most common method of estimating soil properties from a set of soil spectra (Viscarra Rossel and Behrens, 2010).

Recently, several studies show that the variable selection (wavelength selection) may further improve the performance of the regression models (Li et al., 2009; Teofilo et al., 2009; Zou et al., 2010; Vohland et al., 2014; Sarathjith et al., 2016). Variable selection is an iterative process to create subset variables which give the lowest prediction errors. The benefits of variable selection include: (i) improving the predictive ability of the model by removing uninformative variables, (ii) improving the interpretability of the models, (iii) decreasing the computational time needed to analyse the data (Guyon and Elisseeff, 2003; Teofilo et al., 2009; Zou et al., 2010). By identifying the subset variables, variable selection should yield the smallest errors when used to predict samples outside the training dataset.

Variable selection may be done manually (based on expert knowledge) (Zou et al., 2010). Expert users might be able to identify specific regions of the spectra that have poor information quality. However, this manual selection approach might also remove portions of spectra that could be useful in creating a robust model, and there is no guarantee the same section of the data will be removed between datasets (Zou et al., 2010). Variable selection techniques that are based on statistical principles can provide robust models for non-expert users with a limited expert intervention (Zou et al., 2010). A number of variable selection approaches have been examined, such as simulated annealing (Kirkpatrick et al., 1983), genetic algorithm (GA) (Li et al., 1992; Xuemei et al., 2010), uninformative variable elimination (UVE) (Centner et al., 1996), interval PLS (Nørgaard et al., 2000; Zou et al., 2007; Xuemei et al., 2010), backward interval PLS (Zou et al., 2007), forward interval PLS (Zou et al., 2007), successive projections algorithm (SPA) (Araujo et al., 2001), wavelet transform (Viscarra Rossel and Lark, 2009), and competitive adaptive reweighted sampling (CARS) (Li et al., 2009) among the few. Li et al. (2009) successfully employed CARS algorithm in conjunction with regression coefficients in a variable selection process. Vohland et al. (2014) used the CARS approach to create a PLS regression model that only integrates the relevant wavelengths. However, the CARS approach that used the Monte Carlo strategy does not provide a unique solution. Another approach is the use of informative vectors in conjunction with ordered predictor selection (OPS) approach and exponentially decreasing function (EDF) from the CARS algorithm as suggested by Sarathjith et al. (2016).

The purpose of this work is to (i) implement variable selection based on the combination use of *informative vectors* with OPS and EDF approaches which can be used in various regression models (Partial Least Square regression (PLSR) and Cubist regression tree model), (ii) provide the algorithms as codes in R statistical language, and (iii) to identify the important wavelength variables and relate them to the fundamental bands of the relevant.

2. Materials and methods

2.1. Theory

The overall variable selection process can be summarized as follows:

Calculate the *informative vectors* from the spectra (**X**) and the corresponding response variable (**y**).
Sort the *informative vectors* in a descending order using the ordered predictor selection (OPS) approach. The higher the absolute value of the informative vector, the more informative it is.
Create a regression model and evaluate its performance.
Apply the exponentially decreasing function (EDF) to estimate the ratio of the wavelengths to be kept for regression, and only retain informative wavelengths.
Go back to step (iii) to create a model with the selected wavelengths.
This process is repeated until N-th number of iterations is achieved, generating N subset models. However, to prevent the loss of spectra information, in this study, a minimum of 80 retained spectra variables was set.

The whole algorithm is coded in R statistical language and open-source software (R Core Team, 2016), available from the authors. This algorithm is also available in Matlab. The schematic of the process is included in Fig. 1.

In the next section, various informative vectors of the spectra will be described. This is followed by the description of the ordered predictor selection (OPS) approach and the exponentially decreasing function (EDF) procedure to remove uninformative variables.

2.2. Informative vectors

Informative vectors are descriptors of the relationship between the predictors (**X**, spectra variables) and the response variables (**y**). There are various ways to calculate the informative vectors, here we describe eight informative vectors:

2.2.1. Regression coefficient vector (b)

This vector is defined as the change in the response per unit change in the predictor variables. The vector can be estimated using Eq. (1.1)

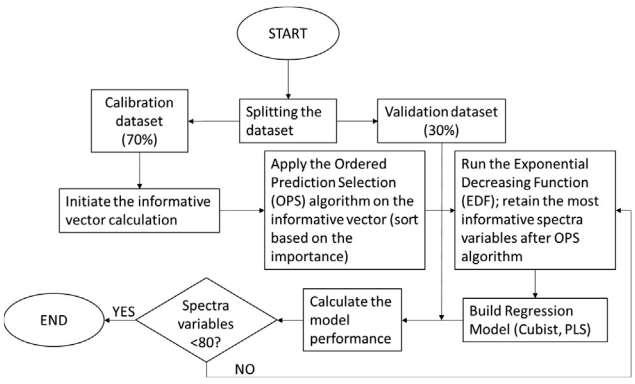

Fig. 1. Schematic of the variable selection process.

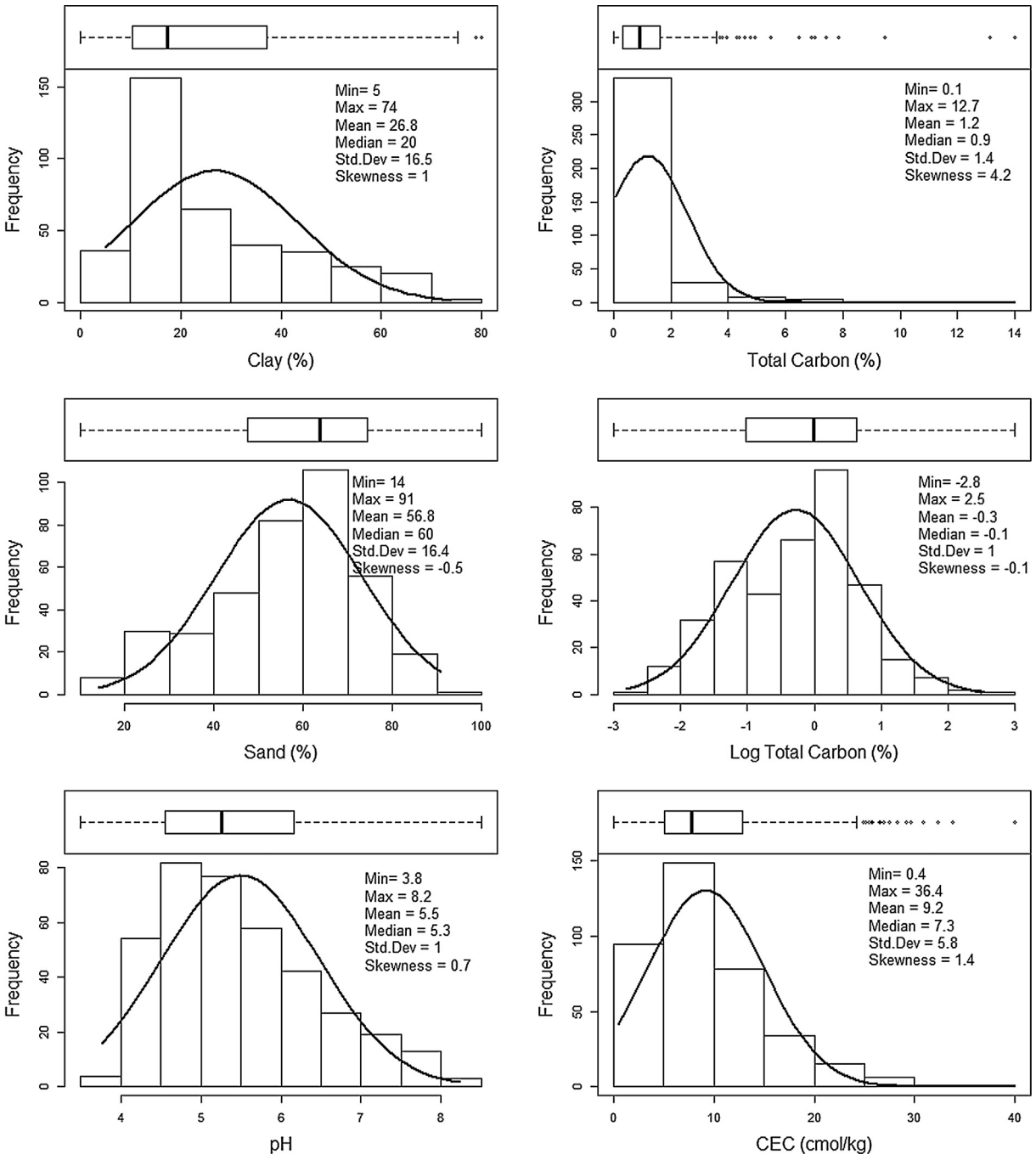

Fig. 2. Boxplots, histograms and descriptive statistics of the soil attributes used in this study. Min.: minimum, Max.: maximum, Std Dev.: standard deviation.

(Teofilo et al., 2009):

$$\mathbf{y} = \mathbf{X}\boldsymbol{b} \rightarrow \mathbf{y} = \mathbf{U_j R_j V_j^t b} \rightarrow \boldsymbol{b} = \mathbf{V_j R_j^{-1} U_j^t y}. \quad (1.1)$$

where **y** is the response variable, **X** is a matrix of spectra data (size I × J), *b* is the regression coefficient vector. **U** (I × J), and **V** (I × J) are matrices with orthonormal columns, which satisfy $\mathbf{U^tU} = \mathbf{V^tV} = \mathbf{I}$, and **R** (J × J) is a bi-diagonal matrix.

2.2.2. Variable influence on projection (V)

This vector was first proposed by Wold et al. (1993). It estimates the importance of **X** variables to explain the variation in **y**. It was calculated as the weighted sum of squares of the PLS weights using Eq. (1.2):

$$V_j = \sqrt{\frac{J \times \sum_{k=1}^{K} (W_{jk}^2 \times SSY_{comp,k})}{SSY_{cum}}} \quad (1.2)$$

where J is the total number of wavelengths, W_{jk} is the loading of the j-th wavelength in k-th factor, $SSY_{comp,k}$ is the explained sum of squares of **y** explained by the PLS regression model with k factor, and SSY_{cum} is the total sum of squares of **y**.

2.2.3. Residual vector (S)

S is a vector comprising residuals from the eliminated information when the original matrix (**X**) is truncated into a reconstructed matrix with h components $\widehat{\mathbf{X}}_\mathbf{h}$ (Teofilo et al., 2009). *S* can be defined as:

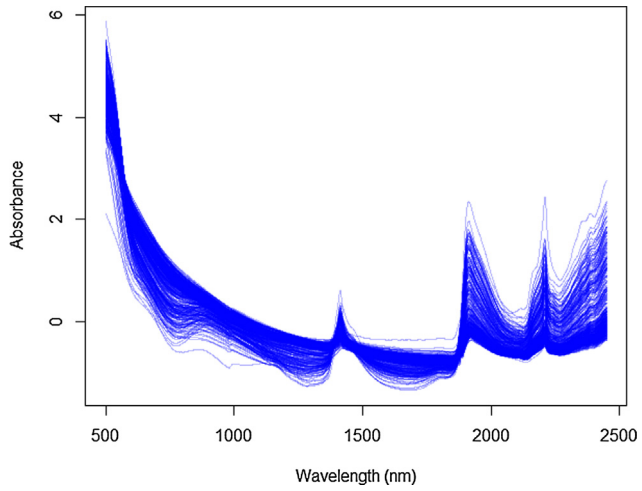

Fig. 3. Vis-NIR absorption spectra of soil samples after pre-processing.

$$\mathbf{E_h} = \mathbf{X} - \widehat{\mathbf{X}}_\mathbf{h}; \mathbf{S_j} = \frac{1}{\mathbf{e_j^t e^j}} \tag{1.3}$$

where $\mathbf{X}$ is the original matrix, $\widehat{\mathbf{X}}_\mathbf{h}$is the truncated matrix, $\mathbf{e_j}$ is the j-th column of $\mathbf{E_h}$, $\mathbf{S_j}$ is the informative vector that calculates the inverse of the sum of square residuals.

2.2.4. Net analyte signal vector (Na)

Net analyte signal is a measure of part of the spectra that is orthogonal to the spectra of the other components (Faber, 1998). Faber suggested the calculation of *Na* vector with the regression vector as:

$$\boldsymbol{Na_j} = (\mathbf{y_j}/\mathbf{b^T b})\mathbf{b} \tag{1.4}$$

where *b* is the regression vector.

2.2.5. Linear correlation coefficient (COR) vector

COR is a vector measuring the linear association between two variables x and y. It was first developed by Pearson in 1895 (Rodgers and Nicewander, 1988). It is calculated as:

$$COR = \frac{\sum (\mathrm{x_i} - \bar{\mathrm{x}}) \sum (\mathrm{y_i} - \bar{\mathrm{y}})}{\sqrt{\sum (\mathrm{x_i} - \bar{\mathrm{x}})^2} \sqrt{\sum (\mathrm{y_i} - \bar{\mathrm{y}})^2}} \tag{1.5}$$

This coefficient ranges from the values of −1 to +1 with zero value indicating the absence of correlation, positive values indicating directly related and negative if inversely related.

2.2.6. Bi-weight mid-correlation (BIC)

This vector was first proposed by Wilcox (2012) to find similarity between two genes. This vector was shown to be more robust to outliers than the Pearson correlation (Song et al., 2012; Wilcox, 2012). *BIC* can be defined as:

$$BIC = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \mathrm{med(x)}) \mathrm{w_i^{(x)}} (\mathrm{y_i} - \mathrm{med(y)}) \mathrm{w_i^{(y)}}}{\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{n}} ((\mathrm{x_j} - \mathrm{med(x)}) \mathrm{w_j^{(x)}})^2} \sqrt{\sum_{\mathrm{k}=1}^{\mathrm{n}} ((\mathrm{y_k} - \mathrm{med(y)}) \mathrm{w_k^{(x)}})^2}} \tag{1.6}$$

where w_i is the weight factor, defined as:

$$w_i^{(x)} = (1 - u_i^2)^2 I(1 - |u_i|) \quad u_i = \frac{x_i - med(x)}{9mad(x)}$$

$$w_i^{(y)} = (1 - v_i^2)^2 I(1 - |v_i|) \quad v_i = \frac{y_i - med(y)}{9mad(y)}$$

Here, *med* is median, and *mad* is the median absolute deviation.

2.2.7. Mutual information based on adjacency matrix (AMI)

AMI measures the non-linear dependency between variables, which provides a better and more general criterion compared to Pearson correlation. The symmetric uncertainty based mutual information adjacency matrix is estimated as (Song et al., 2012):

$$dx = discretize(\mathbf{X});\ no.\ bins = \sqrt{nrow(\mathbf{X})}$$
$$MI(dx, dy) = Entropy(dx) + Entropy(dy) - Entropy(dx, dy)$$
$$A_{ij}^{MI,Symmetric\quad Uncertainty} = \frac{2MI(dx_i, dx_j)}{Entropy(dx_i) + Entropy(dx_j)} \tag{1.7}$$

In this case, **X** (spectra data) is partitioned into equal-width bins with the default number of bins, given by no. of bin $=\sqrt{nrow(\mathbf{X})}$. The mutual Information adjacency matrix can then be calculated based on the discretised vectors and entropy estimation with Miller-Madow method (Miller, 1955). In this study, the universal version mutual information based adjacency matrix was used, which can be calculated as:

$$A_{ij}^{MI,Universal\quad Ver} = \frac{A_{ij}^{MI,Symmetric\quad Uncertainty}}{2 - A_{ij}^{MI,Symmetric\quad Uncertainty}} \tag{1.8}$$

2.2.8. Covariance procedures (COV) vector

It is a vector obtained by the combination of the PLS and standard linear regression methods. This approach concentrates on balancing the fit and prediction. *COV* can be defined as (Reinikainen and Hoskuldsson, 2003; Teofilo et al., 2009):

$$COV = \mathrm{diag}(\mathbf{X^T y y^T X}) \tag{2.11}$$

where **y** is the response variable and **X** is a matrix of spectra data.

In general, all the *informative vectors* mentioned above can be broadly categorized as PLS-dependent (*b, V, S, Na*) and PLS-independent categories (*COR, BIC, AMI, COV*) (Sarathjith et al., 2016). New informative vectors can be created by combining two independent vectors, as well as combining PLS-independent and PLS-dependent vectors. These combination vectors are obtained by multiplying one vector to the other vector after normalization as described by Teofilo et al. (2009). This pairing results in a total of 22 combination vectors, which include *b-V, b-S, b-Na, V-S, V-Na, S-Na, b-COR, b-BIC, b-AMI, b-COV, V-COR, V-BIC, V-AMI, V-COV, S-COR, S-BIC, S-AMI, S-COV, Na-COR, Na-BIC, Na-AMI, Na-COV*. In total, these 30 informative vectors (the singular informative vectors and the combination informative vectors) will be compared.

After we obtain the informative vectors, we need to select the important predictors (wavelengths). This is achieved through the Ordered Predictor Selection (OPS) and Exponentially Decreasing Function (EDF) as described below.

2.3. Ordered predictor selection (OPS)

In essence, this approach sorts the informative vectors based on their corresponding absolute values for each predictor (Teofilo et al., 2009). The higher the absolute values, the more relevant those variables are for the regression models.

2.4. Exponentially decreasing function (EDF)

EDF is a function to remove wavelengths with low absolute values of informative vectors. The ratio of the wavelengths (*r*) to be kept for regression modelling is defined as an exponentially decreasing function:

$$r_i = ae^{-(ki)}$$
$$a = \left(\frac{\mathrm{p}}{2}\right)^{\frac{1}{\mathrm{N}-1}}$$
$$k = \frac{\ln(p/2)}{N-1} \tag{1.9}$$

where $i = 1,2,3,\ldots, N$ represents the number of iteration or run, and p is the total number of wavelengths. Both a and k are constants that satisfy the following conditions: (i) in the first sampling run, $\mathrm{r}_1 = 1$, (ii) in the

Table 1
Regression statistics for the pre-processed full spectra models (Number of Spectra Variables (NSV) = 1951) in comparison to the subset models (NSV < 1951) using various informative vectors. Values of R^2 and RMSE (Root Mean Squared Error) are represented as the mean and standard deviation (in brackets) of 50 iterations.

Soil attribute	*Cubist*							PLS						
	Informative Vector	Informative Vector	NSV	Calibration		Validation		Informative Vector	Informative Vector	NSV	Calibration		Validation	
				R^2	RMSE	R^2	RMSE				R^2	RMSE	R^2	RMSE
Full Model														
CEC (cmol/kg)	–	–	1951	0.88 (0.02)	2.00 (0.20)	0.78 (0.05)	2.73 (0.40)	–	–	1951	0.84 (0.02)	2.34 (0.18)	0.80 (0.05)	2.59 (0.32)
Clay (%)	–	–	1951	0.89 (0.03)	5.43 (0.74)	0.77 (0.05)	7.97 (0.87)	–	–	1951	0.81 (0.02)	7.15 (0.40)	0.77 (0.04)	7.83 (0.81)
pH	–	–	1951	0.82 (0.04)	0.42 (0.04)	0.63 (0.07)	0.60 (0.06)	–	–	1951	0.74 (0.04)	0.50 (0.04)	0.66 (0.07)	0.57 (0.05)
Sand (%)	–	–	1951	0.84 (0.03)	6.49 (0.56)	0.68 (0.05)	9.37 (0.76)	–	–	1951	0.71 (0.02)	8.84 (0.28)	0.69 (0.05)	9.11 (0.67)
Total C (%)	–	–	1951	0.90 (0.02)	0.30 (0.04)	0.77 (0.06)	0.46 (0.07)	–	–	1951	0.85 (0.02)	0.37 (0.03)	0.79 (0.05)	0.43 (0.05)
Overall Performance				0.87		0.73					0.79		0.74	
Best Calibration (within 5%)														
CEC (cmol/kg)	b-AMI	17	479	0.88 (0.02)	2.00 (0.19)	0.76 (0.06)	2.89 (0.45)	b	1	416	0.83 (0.03)	2.41 (0.20)	0.78 (0.05)	2.70 (0.35)
Clay (%)	S-COR	23	551	0.88 (0.03)	5.67 (0.65)	0.76 (0.05)	8.15 (1.00)	AMI	7	156	0.80 (0.02)	7.29 (0.31)	0.77 (0.04)	7.80 (0.70)
pH	b-AMI	17	362	0.82 (0.04)	0.42 (0.04)	0.65 (0.07)	0.58 (0.06)	b-BIC	16	273	0.74 (0.04)	0.50 (0.04)	0.67 (0.06)	0.56 (0.05)
Sand (%)	S-BIC	24	479	0.83 (0.03)	6.72 (0.66)	0.68 (0.05)	9.35 (0.88)	b-BIC	16	206	0.71 (0.02)	8.88 (0.28)	0.69 (0.04)	9.14 (0.63)
Total C (%)	b-AMI	17	730	0.88 (0.04)	0.32 (0.05)	0.75 (0.06)	0.48 (0.06)	b-BIC	16	416	0.84 (0.03)	0.39 (0.04)	0.77 (0.05)	0.46 (0.05)
Overall Performance				0.86		0.72					0.78		0.74	
Best Validation (within 5%)														
CEC (cmol/kg)	b-Na	11	416	0.88 (0.03)	2.04 (0.22)	0.78 (0.05)	2.69 (0.32)	S	3	237	0.81 (0.03)	2.50 (0.20)	0.78 (0.05)	2.70 (0.35)
Clay (%)	S-COV	26	156	0.87 (0.03)	5.91 (0.61)	0.76 (0.05)	8.21 (0.83)	AMI	7	156	0.80 (0.02)	7.29 (0.31)	0.77 (0.04)	7.80 (0.70)
pH	b-AMI	17	179	0.82 (0.04)	0.42 (0.04)	0.65 (0.07)	0.58 (0.06)	b-AMI	17	156	0.74 (0.03)	0.50 (0.03)	0.66 (0.05)	0.57 (0.05)
Sand (%)	S-COV	26	118	0.80 (0.05)	7.32 (0.82)	0.68 (0.05)	9.40 (0.80)	AMI	7	156	0.69 (0.03)	9.13 (0.35)	0.67 (0.05)	9.40 (0.75)
Total C (%)	V-COV	22	135	0.88 (0.03)	0.33 (0.04)	0.76 (0.06)	0.47 (0.06)	S-COV	26	416	0.83 (0.02)	0.40 (0.02)	0.78 (0.05)	0.45 (0.05)
Overall Performance				0.85		0.72					0.77		0.73	
Best Calibration II (within 5%)														
CEC (cmol/kg)	b-BIC	16	551	0.87 (0.03)	2.06 (0.24)	0.77 (0.05)	2.76 (0.32)	b	1	967	0.83 (0.03)	2.41 (0.20)	0.78 (0.05)	2.70 (0.35)
Clay (%)	S-COR	23	551	0.88 (0.03)	5.67 (0.65)	0.76 (0.05)	8.15 (1.00)	b-Na	11	416	0.81 (0.02)	7.14 (0.40)	0.78 (0.04)	7.77 (0.78)
pH	b-AMI	17	416	0.82 (0.04)	0.42 (0.04)	0.65 (0.07)	0.58 (0.06)	b-BIC	16	551	0.74 (0.04)	0.50 (0.04)	0.67 (0.06)	0.56 (0.05)
Sand (%)	S-COR	23	840	0.83 (0.03)	6.76 (0.64)	0.68 (0.05)	9.36 (0.75)	b-BIC	16	551	0.71 (0.02)	8.88 (0.28)	0.69 (0.04)	9.14 (0.63)
Total C (%)	S-COR	23	967	0.87 (0.03)	0.34 (0.04)	0.74 (0.06)	0.49 (0.06)	b-BIC	16	1280	0.84 (0.03)	0.39 (0.04)	0.77 (0.05)	0.46 (0.05)
Overall Performance				0.85		0.72					0.78		0.74	

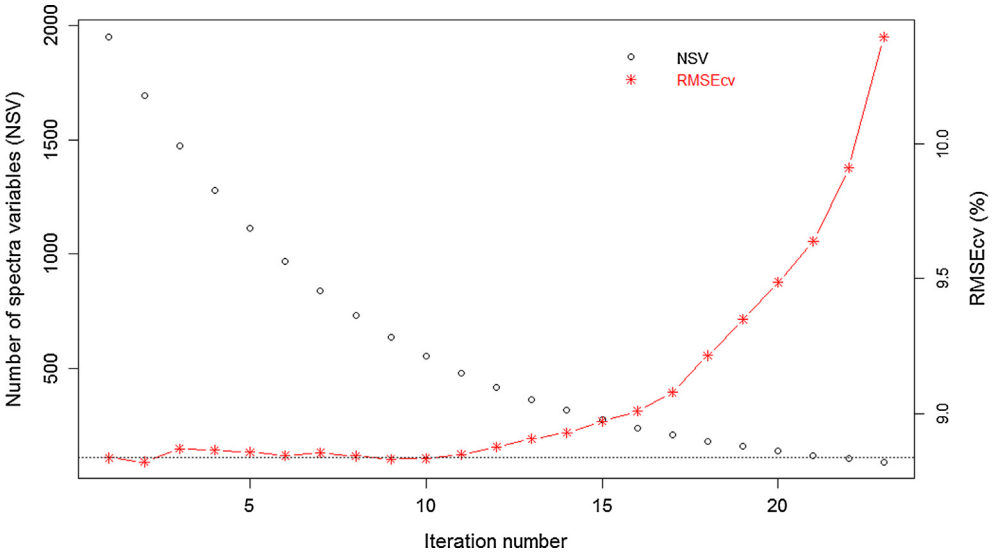

Fig. 4. Illustration for the model performance with ordered predictor selection and exponential decreasing function approach to predict sand content using PLS model with b-BIC vector. As the iteration number increases, the ratio of the number of wavelengths variable to be kept decreases (smaller NSV). The red asterisks represent the model performance for each subset models based on the RMSEcv. The dotted line represents the model performance when the full spectra is used to create the regression model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Important wavelengths utilized by the Cubist and PLSR model in determining various soil properties.

Properties	Important wavelengths for Cubist model (nm)	Important wavelengths for PLSR model (nm)	Common wavelengths between the two models (nm)
CEC	512, 538, 551, 577, 590, 668, 694, 707, 1654, 1927, 1966, 2078, 2117, 2156, 2195, 2234, 2390, 2450	512, 538, 551, 577, 590, 629, 668, 694, 1654, 1927, 1966, 2000, 2039, 2117, 2156, 2195, 2234, 2390, 2424	512, 538, 551, 577, 590, 668, 694, 1654, 1927, 1966, 2117, 2156, 2195, 2234, 2390
CLAY	590, 616, 629, 655, 746, 772, 785, 811, 824, 850, 863, 1732, 1810, 1849, 1888, 1966, 2039, 2078, 2234, 2390	590, 616, 629, 655, 746, 772, 785, 811, 824, 850, 1732, 1849, 1966, 2039, 2078, 2234, 2390, 2411	590, 616, 629, 655, 746, 772, 785, 811, 824, 850, 1732, 1849, 1966, 2039, 2078, 2234, 2390
pH	1927	1927	1927
Sand	551, 668, 707, 733, 746, 1693, 1732, 1810, 2044, 2078, 2156	551, 668, 707, 733, 746, 1654, 1693, 1732, 1810, 2078, 2156	551, 668, 707, 733, 746, 1693, 1732, 1810, 2078, 2156
Total C*	1006, 1097, 1123, 1654, 2005, 2117, 2122, 2156, 2161, 2229, 2234, 2239, 2307, 2346, 2351, 2424, 2429	1097, 1123, 1654, 1927, 2117, 2156, 2161, 2234, 2239, 2312, 2424, 2429	1097, 1123, 1654, 2117, 2156, 2161, 2234, 2239, 2424, 2429

* For total Carbon, the wavelengths selections were done based on those that showed up at least in the 40 repetitions.

N-th run, $r_N = 2/p$. This function allows the 'fast selection' followed by 'refined selection' in its successive iterations (Li et al., 2009).

2.5. Simulation datasets

A soil spectra library with a total number of 379 samples from 68 different profiles were used for evaluating the performance of the variable selection algorithm. The samples were collected from each of the horizons taken up to 1 m depth from the study by Geeves et al. (1995), representing soils in the wheat-belt of southern New South Wales (NSW) and northern Victoria (VIC). The soil groups found in this area are Chromosols, Dermosols, Sodosols and Kandosols (Australian Soil Classification) or Alfisols and Inceptisols (USDA Soil Taxonomy) or Luvisols, Cambisols, Solonetz, and Lixisols (WRB). Samples were collected from different soil horizons up to 1 m depth. Samples were selected non-randomly to represent full management system within the areas while keeping away from area such as stock tracks, dams, drainage, heavy traffic zones, isolated trees and other features that were considered as unrepresentative of the areas (Geeves et al., 1995).

All the samples were air-dried, ground, and sieved to < 2-mm particle size diameter. The samples were subjected to laboratory analyses of the physical and chemical properties including clay and sand content, pH, cation exchange capacity (CEC) and total carbon (TC). The sand content was determined through wet sieving of chemically dispersed sample while the clay content was determined using hydrometer method. CEC was determined using silver thiourea method. Soil pH was measured in 0.1 M $CaCl_2$ in 1:5 soil to solution ratio after 1 h of rotational shaking and 0.5 h of settling. While the TC was determined in sediments dried using LECO CR-12 combustion furnace with an infrared detector.

The total carbon values were subjected to natural log-transformed before modelling to correct for its skewness. The summary statistics of the parameters used for this study are shown in Fig. 2.

2.5.1. Spectra collection

The infrared spectra were recorded using an AgriSpec (Analytical Spectral Devices, Boulder, CO, USA) with a spectra range of 350–2500 nm at 1 nm interval. The samples were illuminated with a 4.5 W halogen lamp, and the reflected light was transmitted to the spectrometer through a fibre optic bundle. A Spectralon (Labsphere Inc., North Sutton, NH, USA) was used as the white base reflectance standard and scanned after every five samples. Each spectrum was obtained as the mean of three replicates.

2.5.2. Spectra Pre-processing

The reflectance spectra collected were pre-processed to reduce irrelevant information which may decrease the performance of the fitted prediction models. Spectra between 350 and 499 and 2451–2500 nm were removed due to their low signal to noise ratio resulting in 1951 spectra values. The remaining spectra were converted to absorbance (log 1/reflectance) followed by Savitzky-Golay transformation (Savitzky and Golay, 1964) with a window size of 11 and polynomial order 2 and followed with the Standard Normal Variate (SNV) transformation. The spectra data after the pre-processing are presented in

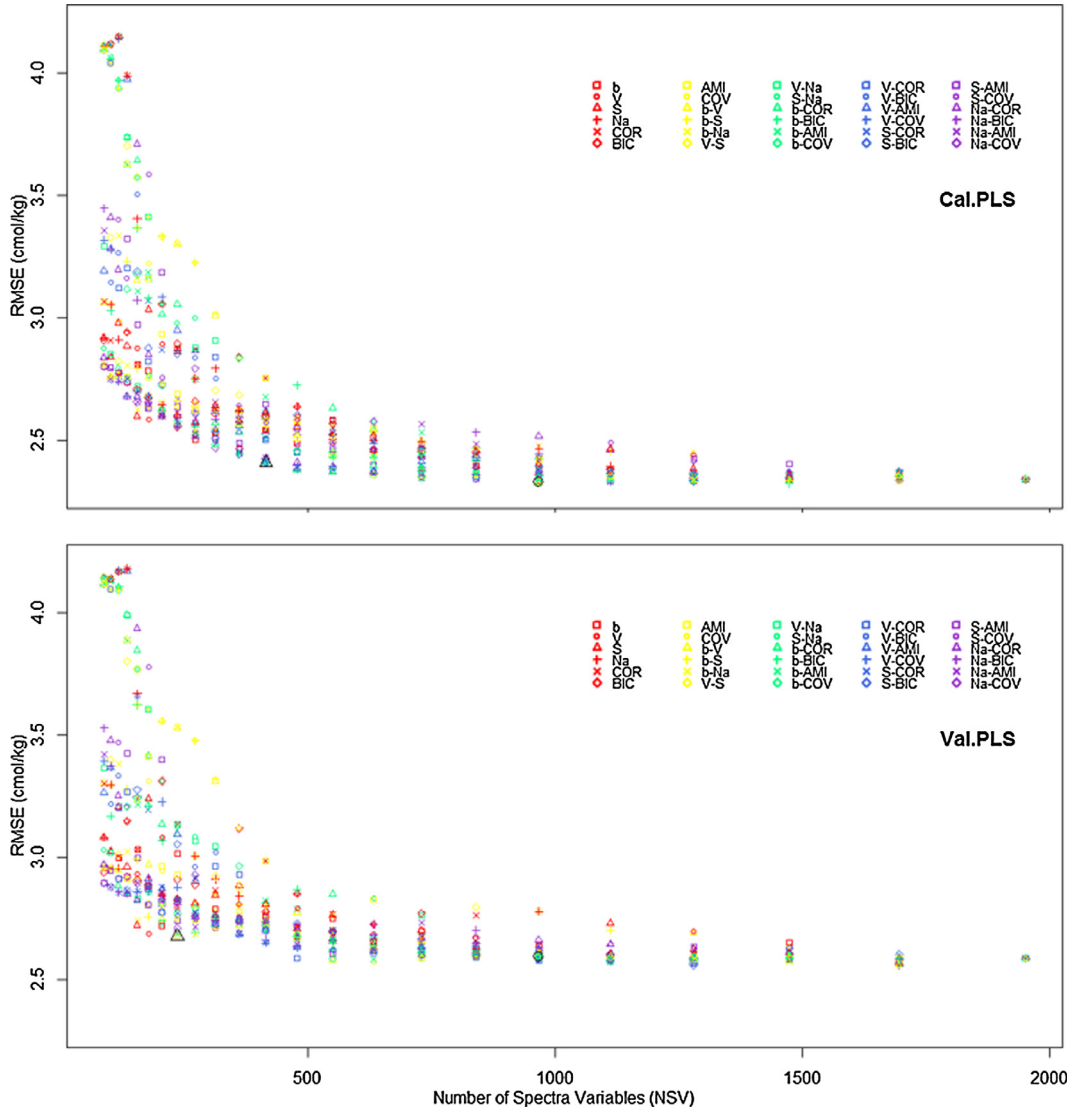

Fig. 5. Plot of the effect of various informative vectors on model performance in predicting cation exchange capacity (CEC) as a function of number of spectra variables (NSV). Each shape in the plot represents a particular informative vector. The black triangular shape (△) indicates the best subset model with the lowest root mean squared error (RMSE), while the black round shape (○) indicates the best subset model that is also better than the full-spectra model.

Fig. 3.

2.6. Chemometrics

The spectra were modelled for prediction of clay and sand content, pH, CEC, and TC. The variable selection procedure and regression analyses were repeated fifty times, each time the data were randomly partitioned into calibration (70%) and validation (30%) sets. The partial least squares regression (PLSR; Mevik et al., 2016) and Cubist regression tree model (Kuhn et al., 2016) were used for modelling. The optimal number of latent variables utilized in the PLSR was determined by doing cross-validation and selecting those that give the lowest value of $RMSE_{CV}$ calibration. The $RMSE_{CV}$ is calculated as follows:

$$RMSE_{CV} = \sqrt{\frac{\sum_{i=1}^{n} (\tilde{y}_i - y_i)^2}{n}} \quad (2.10)$$

where n is the number of samples, y_i is the reference measurement of sample i, and $\tilde{y}_i$ is the estimated result for the sample i when the sample i is removed.

Cubist is a rule-based regression tree model which was found to be useful for building NIR calibration models. Cubist creates rules by splitting the data based on its independent variables minimising within-class variation. After that, it builds a linear model of the absorbance spectra for each rule. The detail of the algorithm is presented in Quinlan (1993).

The performance of the model was then compared using the coefficient of determination (R^2) and root-mean-square error ($RMSE_v$) values of the validation dataset.

3. Results and discussion

The model performance in predicting five soil properties from the full spectra (Number of Spectra Variables, NSV = 1951) in comparison to the variable selection subset models are summarized in Table 1.

By utilizing the full spectra, both the Cubist and PLS regression models gave relatively accurate predictions for all the properties. An illustrative figure of the model performance for the prediction of sand content using OPS and the *b-BIC* vector is shown in Fig. 4. The plot shows the NSV retained by the algorithm for each iteration, which follows an exponential function. Typical variations in the RMSE value of the validation dataset for the subset models using different NSV are shown in the plot. The performance of the first few iterations is similar

to those from the full spectra models, even after ten iterations (NSV = 551). After that, the RMSE values seem to be increasing for each subsequent iteration (reduced number of spectra variables). This is most likely due to the loss of information. In this example, an optimum number of spectra variables kept based on the lowest RMSE would be 1695 (at the second iteration). However, since the RMSE for the subsequent iterations with a lower number of spectra variables (NSV) are within 5% proximity of the lowest RMSE, it would also be considered. In this case, the number of spectra variables kept for this model is 551 (achieved after ten iterations).

Using the full spectra, the regression statistics from both PLS regression and Cubist model were compared (Table 2). In general, the PLS model performed slightly better than the Cubist model based on the R^2 and RMSE values for all soil properties even though the Cubist model seemed to perform better based on the model calibration dataset.

3.1. Performance of the subset models

To select the optimal subset models with reduced NSV, three different options will be compared.

3.1.1. Determining NSV based on the best calibration model

The best subset models were determined by those that had $RMSE_{cv}$ values within 5% proximity to the lowest RMSE calibration and the least number of spectra variables (NSV) of the calibration data-set as suggested by Sarathjith et al. (2016).

Generally, the subset of the pre-processed spectra models generated acceptable regression models. The R^2 and RMSE values for the subset models using Cubist regression were similar to those from the full spectra models; however with much smaller NSV (362-730), or on average 27% of the full spectra. The same trends can also be observed by using PLS regression with NSV ranging from 156 to 416, or on average 15% of the full spectra.

Comparing the Cubist and PLS regression model side by side, the performance was pretty much similar; PLS gave a slightly better model performance for the prediction of all other properties while using a smaller number of spectra variables.

3.1.2. Determining NSV based on the best validation model

The best subset models were determined by those that had $RMSE_v$ values within 5% proximity to the lowest RMSE validation and the least NSV of the validation data-set. For Cubist models, the prediction of CEC and total C were improved by using lower NSV (416 and 135, respectively) in comparison to the best calibration model. Although the model performance for the pH, sand and clay content did not improve, the NSV used in the model were much lower (on average 11% of the full spectra), which can be a positive outcome in some respects through the reduction in computational time. Similar to the Cubist regression, the best validation model for the prediction of CEC using PLS regression can be further improved by using lower NSV (Table 1). For the total C prediction, by utilizing different informative vector, the model performance was improved. When the Cubist and PLS regression are compared, the overall performance of all properties was similar. PLS gave a slightly better prediction for clay content and pH with similar NSVs. CEC and sand content was predicted slightly better with the Cubist model. However, the NSVs used is almost double than those in the PLS model for CEC prediction.

3.1.3. Determining NSV based on the best calibration model II

The selection for these subset models was based on those that have $RMSE_{cv}$ values within 5% proximity as well as lower $RMSE_{cv}$ than those of the full models. The overall subset model performance was similar to those using the full spectra models with less NSV, on average 35% of the full spectra. For Cubist subset models, prediction performance can be further increased by retaining NSV ranging from 416 to 967. The NSV retained for the PLS subset models to perform similarly if not better than the full PLS model ranges from 416 to 1280 (see Table 1). Overall, the performances of both regression approaches were similar.

3.1.4. Determining best informative vectors

The best informative vectors will be derived from the best subset calibration model II. For the prediction of pH and CEC, *b* (and its combinations) vectors were found to be the most appropriate by using either PLS or Cubist regression. No common informative vectors were found to predict the total carbon, sand and clay content. Within the PLS regression models, it may be generalized that the *b* (and its combinations). The best informative vectors utilized in the Cubist models varied slightly, with some improvements being found when *S* (and its combinations) for the predictions of clay content, sand content, and total carbon and b (and its combinations) for the prediction of pH and CEC content.

As stated by Sarathjith et al. (2016), informative vectors that provided inferior predictions in one case may provide superior predictions in another case. An example of the evaluation of model performance by using the PLS model for the prediction of CEC using various informative vectors is included in Fig. 5. As discussed above, the figure indicated that the wavelengths can be reduced to about 500 or one-quarter of the full spectra, and still achieve the same results using the full wavelength.

3.1.5. General discussion

We have demonstrated that the models that utilized the optimum wavelengths (about 25% of the full spectra) can perform as well as the model that utilized the full spectra model. Different wavelengths were selected as important for different soil properties. The top 400 wavelengths that are common among the 50 repetitions based on the most informative vectors for the best calibration II method are tabulated in Table 2. If there are no common wavelengths that exist, the rule is flexed so that the most important wavelengths can be identified. Some of the wavelengths that were used in the Cubist model were not utilized in the PLSR model. As an effort to determine the effectiveness of the variable selection process, only the common wavelengths found between the two models will be discussed.

The most important wavelengths for the CEC content estimations are 512, 538, 551, 577, 590, 668, 694, 1654, 1927, 1966, 2117, 2156, 2195, 2234 and 2390 nm. Because CEC is correlated to the clay content and clay content affects the colour of the soil (Gomez et al., 2008), the use of the absorption bands between 400 and 700 nm could potentially contribute to the CEC prediction as this region corresponds to colour information. Other wavelengths found were also reported by Xu et al. (2018), particularly absorption at 680, 890, 1410, 1900, 2210 and 2400 nm. The absorption band at 1654 is related to the first overtone of $-CH_2$ and $-CH_3$ bonds (Hourant et al., 2000).

The absorption between the regions of 590, 616, 629, 655, 746, 772, 785, 811, 824, 850, 1732, 1849, 1966, 2039, 2078, 2234 and 2390 nm are deemed to be important for the prediction of clay content. Aside from the visible range (400–700 nm), Gomez et al. (2008) also found that absorptions near 981, 1400, 1800, 1900 and 2350 nm were important in predicting clay content. Some of these wavelengths were also confirmed in our study. Nonetheless, our models also utilized wavelengths between 750 and 850 nm and ~2060 nm which is attributed to iron oxides organics in the soils (overtones of amines and alkyls) (Viscarra Rossel and Behrens, 2010).

For pH content, both models have an agreement that the absorptions at 1927 nm is the most important. The absorption in this region is associated with H−O−H bend and O−H stretching vibration (Viscarra Rossel and Behrens, 2010). If we reduce the frequency limit to 48 instead of 50, our findings is in agreement to the finding with the most important wavelengths between the regions of 668, 785, 1693, 1927, 2117, 2195 and 2424 nm. This finding is in similar range to those found by Xu et al. (2018) who found that the absorption bands of 480, 780, 1120, 1910, 2200 and 2390 nm were important.

Sand content can be determined using the absorption bands at 551,

668, 707, 733, 746, 1693, 1732, 1810, 2078 and 2156 nm. Nonetheless, these are slightly different to those wavelengths reported by Xu et al. (2018) at 480, 920, 1910 and 2200 nm. Our predictions depended on the absorbance at visible range (400–700 nm) and the organics absorbance (~750 and 1650–1850 nm).

In this study, the most important wavelengths to determine the total carbon concentration are 1097, 1123, 1654, 2117, 2156, 2161, 2234, 2239, 2424 and 2429 nm. Chang et al. (2001) reported that multiple absorption bands for organic carbon in the similar region between 2100 and 2400 nm. Other studies reported bands around 1100, 1600, 1700–1800, 2000, and 2200 to 2400 nm to be important which in agreement to the finding in our study (Dalal and Henry, 1986; Stenberg et al., 2010).

4. Conclusions

Variable selection is an essential process in providing reliable variables for model calibration. This process is important because spectra contain collinear information. Here, an algorithm for variable selection using the combination of informative vectors with OPS approach and EDF for various soil properties prediction was evaluated. This algorithm was assessed with both PLS and Cubist regression models. Before the variable selection process, all the spectra underwent pre-treatment processes to remove unwanted noise. Overall, the validation results from both subset models were similar, with PLS performing slightly better than Cubist models. However, the number of spectra variables used in the subset model (ranging from 416 to 1280 variables) was much lower compared to the full models (1951 variables) which would reduce the computational time. The results show that the variable selection is a beneficial procedure that can be used to improve model performance in terms of computational time. From the dataset used in this study, the best informative vector found to provide optimum subset models are b (and its combinations) and S (and its combinations). Although the findings might be valid only for this dataset, the availability of the algorithms allows experts from various disciplines to test on different datasets.

For future study, the combination of three informative vectors could also be investigated. Wavelength selection is an existing and still-growing field in chemometrics. Various algorithms approaches have been developed for various data, and thus difficult to obtain the best algorithm that would fit all data. Users are suggested to use what is best for their dataset.

Acknowledgments

The authors acknowledge the Sydney Informatics Hub and the University of Sydney's high performance computing cluster Artemis for providing the high performance computing resources that have contributed to the research results reported within this paper. This work is funded by the ARC Linkage Project LP150100566, Optimised field delineation of contaminated soils.

References

Araujo, M.C.U., Saldanha, T.C.B., Galvao, R.K.H., Yoneyama, T., Chame, H.C., Visani, V., 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometr. Intell. Lab. Syst. 57 (2), 65–73.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132 (3–4), 273–290.

Centner, V., Massart, D.L., deNoord, O.E., deJong, S., Vandeginste, B.M., Sterna, C., 1996. Elimination of uninformative variables for multivariate calibration. Anal. Chem. 68 (21), 3851–3858.

Chang, C.W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65 (2), 480–490.

Dalal, R.C., Henry, R.J., 1986. Simultaneous determination of moisture, organic-carbon, and total nitrogen by near-infrared reflectance spectrophotometry. Soil Sci. Soc. Am. J. 50 (1), 120–123.

Daniel, K.W., Tripathi, X.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). Aust. J. Soil Res. 41 (1), 47–59.

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L., Huvenne, J.-P., 2009. Support vector machines (SVM) in near infrared (NIR) spectroscopy: focus on parameters optimization and model interpretation. Chemometr. Intell. Lab. Syst. 96 (1), 27–33.

Faber, N.M., 1998. Efficient computation of net analyte signal vector in inverse multivariate calibration models. Anal. Chem. 70 (23), 5108–5110.

Geeves, G.W., Cresswell, H.P., Murphy, B.W, Gessler, P.E., Chartres, C.J., Little, I.P., 1995. The physical, chemical and morphological properties of soils in the wheat-belt of southern NSW and northern Victoria. NSW Department of Conservation and Land Management/CSIRO Div. Soils occasional rep., CSIRO, Australia.

Gomez, C., Lagacherie, P., Coulouma, G., 2008. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. Geoderma 148 (2), 141–148.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Horta, A., Malone, B., Stockmann, U., Minasny, B., Bishop, T.F.A., McBratney, A.B., Pallasser, R., Pozza, L., 2015. Potential of integrated field spectroscopy and spatial analysis for enhanced assessment of soil contamination: a prospective review. Geoderma 241, 180–209.

Hourant, P., Baeten, V., Morales, M.T., Meurens, M., Aparicio, R., 2000. Oil and fat classification by selected bands of near-infrared spectroscopy. Appl. Spectrosc. 54 (8), 1168–1174.

Inagaki, T., Shinoda, Y., Miyazawa, M., Takamura, H., Tsuchikawa, S., 2010. Near-infrared spectroscopic assessment of contamination level of sewage. Water Sci. Technol. 61 (8), 1957–1963.

Islam, K., Singh, B., McBratney, A., 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. Aust. J. Soil Res. 41 (6), 1101–1114.

Kemper, T., Sommer, S., 2002. Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. Environ. Sci. Technol. 36 (12), 2742–2747.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., 2016. Cubist: rule- and instance-based regression modeling. R package version 0.0.19. Available at: < https://CRAN.R-project.org/package=Cubist > .

Kurz, C., Leitenberger, M., Carle, R., Schieber, A., 2010. Evaluation of fruit authenticity and determination of the fruit content of fruit products using FT-NIR spectroscopy of cell wall components. Food Chem. 119 (2), 806–812.

Li, H.D., Liang, Y.Z., Xu, Q.S., Cao, D.S., 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal. Chim. Acta 648 (1), 77–84.

Li, T.-H., Lucasius, C.B., Kateman, G., 1992. Optimization of calibration data with the dynamic genetic algorithm. Anal. Chim. Acta 268 (1), 123–134.

Li, X.L., Sun, C.J., Luo, L.B., He, Y., 2015. Determination of tea polyphenols content by infrared spectroscopy coupled with iPLS and random frog techniques. Comput. Electron. Agric. 112, 28–35.

Marquetti, I., Link, J.V., Lemes, A.L.G., Scholz, M.B.D., Valderrama, P., Bona, E., 2016. Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee. Comput. Electron. Agric. 121, 313–319.

Mehmood, T., Liland, K.H., Snipen, L., Saebo, S., 2012. A review of variable selection methods in Partial Least Squares Regression. Chemometr. Intell. Lab. Syst. 118, 62–69.

Mevik, B.-H., Wehrens, R., Liland, K.H., 2016. pls: Partial Least Squares and Principal Component Regression. R package version 2.6-0. Available at: < https://CRAN.R-project.org/package=pls > .

Miller, G.A., 1955. Note on the bias on information estimates. Information Theory in Psychology; Problems and Methods II-B, 95-100.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl. Spectrosc. 54 (3), 413–419.

Pascoa, R.N.M.J., Lopo, M., dos Santos, C.A.T., Graca, A.R., Lopes, J.A., 2016. Exploratory study on vineyards soil mapping by visible/near-infrared spectroscopy of grapevine leaves. Comput. Electron. Agric. 127, 15–25.

Quinlan, J.R., 1993. C4. 5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Mateo, California.

R Core Team. 2016. R: A language and environment for statistical computing. Available at: < https://www.R-project.org/ > .

Reinikainen, S.P., Hoskuldsson, A., 2003. COVPROC method: strategy in modeling dynamic systems. J. Chemometr. 17 (2), 130–139.

Rodgers, J.L., Nicewander, W.A., 1988. Thirteen ways to look at the correlation coefficient. Am. Statistician 42 (1), 59–66.

Sarathjith, M.C., Das, B.S., Wani, S.P., Sahrawat, K.L., 2016. Variable indicators for optimum wavelength selection in diffuse reflectance spectroscopy of soils. Geoderma 267, 1–9.

Savitzky, A., Golay, M.J.E., 1964. Smoothing + differentiation of data by simplified least squares procedures. Anal. Chem. 36 (8), 1627–2000.

Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66 (3), 988–998.

Shrestha, S., Deleuran, L.C., Gislum, R., 2017. Separation of viable and non-viable tomato (*Solanum lycopersicum* L.) seeds using single seed near-infrared spectroscopy. Comput. Electron. Agric. 142, 348–355.

Song, L., Langfelder, P., Horvath, S., 2012. Comparison of co-expression measures: mutual

information, correlation, and model based indices. *Bmc*. Bioinformatics 13.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science.
Teofilo, R.F., Martins, J.P.A., Ferreira, M.M.C., 2009. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. J. Chemometr. 23 (1–2), 32–48.
Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1–2), 46–54.
Viscarra Rossel, R.A., Lark, R.M., 2009. Improved analysis and modelling of soil diffuse reflectance spectra using wavelets. Eur. J. Soil Sci. 60 (3), 453–464.
Viscarra Rossel, R.A., Raphael, A., McBratney, A.B., Minasny, B., 2010. Proximal soil sensing. Progress in soil science. Springer, Dordrecht; New York.
Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2005. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1–2), 59–75.
Vohland, M., Ludwig, M., Thiele-Bruhn, S., Ludwig, B., 2014. Determination of soil properties with visible to near- and mid-infrared spectroscopy: effects of spectral variable selection. Geoderma 223, 88–96.
Wilcox, R.R., 2012. Introduction to Robust Estimation and Hypothesis Testing. Elsevier, New York, NY, pp. 3rd..
Wold, S., Johansson, E., Cocchi, M., 1993. PLS – partial least-squares projections to latent structures, 3D QSAR in drug design. In: Kubinyi, H. (Ed.), Theory Methods and Applications. ESCOM Science Publishers, Leiden.
Wu, D., Feng, S., He, Y., 2008. Short-wave near-infrared spectroscopy of milk powder for brand identification and component analysis. J. Dairy Sci. 91 (3), 939–949.
Xu, D., Ma, W., Chen, S., Jiang, Q., He, K., Shi, Z., 2018. Assessment of important soil properties related to Chinese Soil Taxonomy based on vis–NIR reflectance spectroscopy. Comput. Electron. Agric. 144, 1–8.
Xuemei, L., Hailiang, Z., and Xudong, S., 2010, 26-28 June 2010. NIR sensitive wavelength selection based on different methods. In: 2010 International Conference on Mechanic Automation and Control Engineering.
Zou, X., Zhao, J., Li, Y., 2007. Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of 'Fuji' apple based on BiPLS and FiPLS models. Vib. Spectrosc. 44 (2), 220–227.
Zou, X.B., Zhao, J.W., Povey, M.J.W., Holmes, M., Mao, H.P., 2010. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta 667 (1–2), 14–32.