Geoderma 267 (2016) 1–9

Contents lists available at ScienceDirect

Geoderma

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

Variable indicators for optimum wavelength selection in diffuse reflectance spectroscopy of soils

M.C. Sarathjith [a,*], Bhabani Sankar Das [b], Suhas P. Wani [c], Kanwar L. Sahrawat [c]

[a] International Crops Research Institute for the Semi-Arid Tropics, Bamako BP-320, Mali
[b] Agricultural and Food Engineering Department, Indian Institute of Technology Kharagpur, West Bengal 721302, India
[c] International Crops Research Institute for the Semi-Arid Tropics, Patancheru, Hyderabad, Telangana 502324, India

ARTICLE INFO

Article history:
Received 27 June 2015
Received in revised form 23 December 2015
Accepted 29 December 2015
Available online 7 January 2016

Keywords:
Diffuse reflectance spectroscopy
Variable indicator
Ordered predictor selection
Partial least squares regression
Spectral variables

ABSTRACT

Diffuse reflectance spectroscopy (DRS) operating in 350–2500 nm wavelength range is fast emerging as a rapid and non-invasive technique for analyzing multiple soil attributes. Because the spectral reflectance values in this range of wavelengths are highly co-linear, it is important to select relevant spectral information from the reflectance spectra to build a robust spectral algorithm. The objective of this study is to examine the utility of different variable indicators such as partial least squares regression (PLSR) coefficients (β), variable influence on projection, squared residual (*SqRes*), correlation coefficient (r), biweightmidcorrelation (*bicor*), mutual information based adjacency value (*AMI*), signal-to-noise ratio (*StN*), covariance procedures (*CovProc*) and their combinations in conjunction with an ordered predictor selection (OPS) approach for selecting optimum number of spectral variables (NSV) which could improve DRS model performance. The approach was tested with the PLSR models of pH, organic carbon, extractable iron (Fe), sand and clay contents and geometric mean diameter in Vertisols and Alfisols. The prediction accuracy of best models selected via OPS approach was found to be superior to full-spectrum (NSV = 2048) model for all the soil attributes. The percent decrease in RMSE value was found to be highest for Fe (14%, NSV = 79) in Alfisols followed by pH (9%, NSV = 660) in Vertisols while it varied between 3 and 8% for other soil attributes. Although the results were not conclusive in favor of one specific variable indicator, the *CovProc* and *bicor* were found to be more appropriate for accurate and moderate DRS models in this study, respectively. The overall results of this study advocate the use of OPS approach with variable indicators and their combinations as a promising strategy to develop simple and effective DRS models for soils.

1. Introduction

Over the last few decades, diffuse reflectance spectroscopy (DRS) has been recognized as a rapid and non-invasive technique for the measurement of multiple soil attributes. The DRS approach is also widely adapted as a digital soil mapping tool across the globe (Ben-Dor and Banin, 1995; Soriano-Disla et al., 2014). Typically, an efficient multivariate regression model is developed between targeted soil attributes and spectral reflectance values in visible to near- and shortwave-infrared (VisNIR) range of wavelengths (350–2500 nm) in the DRS approach. Both linear and non-linear chemometric and data mining algorithms such as principal components regression, partial least squares regression (PLSR), support vector regression (Thissen et al., 2004), regression trees (Brown et al., 2006), multivariate adaptive regression splines (Shepherd and Walsh, 2002), committee trees (Vasques et al., 2009b), artificial neural networks (Daniel et al., 2003; Goldshleger et al., 2012) have been examined in soil DRS studies. Among these, the PLSR approach seems most frequently used because of its ability to address multicollinearity of spectral variables, interpretability and computational performance (Viscarra Rossel et al., 2006; Stenberg et al., 2010). Performance of these models relies on their capability to extract important spectral characteristics or features (e.g., electronic transitions, overtones and combinations of fundamental vibrations in the mid-infrared frequencies) relevant to the soil attribute of interest (Viscarra Rossel et al., 2006; Viscarra Rossel and Lark, 2009).

A general practice in the DRS approach is the use of either entire (full-spectrum) or selected reflectance values as spectral variables for building a DRS model. The VisNIR response is generally weak and consists of complex absorption features (Stenberg et al., 2010). Hence, the selection of either a full or a part of the spectrum without a proper guideline often leads to have redundant or irrelevant information in the DRS model affecting its performance. The selection of appropriate and optimum number of spectral variables is expected to reduce model complexity and improve robustness (Xiaobo et al., 2010) and prediction accuracy of calibration models (Jouan-Rimbaud et al., 1995; Nadler and Coifman, 2005). Fernández Pierna et al. (2009) suggested that a robust variable selection method should yield a small set of variables capable of providing better, or at least, equivalent model performance to those obtained by the original set of variables. Hence, variable selection

* Corresponding author.
E-mail addresses: sarathjith.mc@gmail.com, mc.sarathjith@cgiar.org (M.C. Sarathjith).

http://dx.doi.org/10.1016/j.geoderma.2015.12.031
0016-7061/

should be included as a critical step in DRS data analysis routine to accomplish the aforesaid advantages. A few sophisticated variable selection approaches (Xiaobo et al., 2010) have already been examined in the spectroscopic studies including successive projections algorithm (Araújo et al., 2001), uninformative variable elimination (Centner et al., 1996), simulated annealing (Kirkpatrick et al., 1983), genetic algorithms (Leardi et al., 1992), moving window partial least square (Chen et al., 2011), interval partial least squares (Norgaard et al., 2000), backward variable selection for PLSR (Fernández Pierna et al., 2009), wavelet transformation (Ge and Thomasson, 2006) among others. Recently, Li et al. (2009) developed a competitive adaptive reweighted sampling (CARS) as a strategy for spectral variable selection using regression coefficient (β) of PLSR model. Vohland et al. (2014) successfully implemented the CARS approach in the soil dataset, and concluded that the approach is simple, accurate, and involves reasonable and parsimonious variable selection. However, no unique solution exists for this approach, mainly because of the Monte Carlo strategy and random numbers used in CARS. The issue may be resolved with the use of 'variable indicators' or 'informative vectors' in conjunction with an ordered predictor selection (OPS) approach, as suggested by Teófilo et al. (2009). In addition, the OPS approach has the following advantages: simple, flexible, effective in parsimonious selection and interpretability of spectral variables. The OPS approach has not been tested with soil datasets for a multitude of variable indicators.

In general, the variable indicators are descriptors of the relationships between predictor (spectral variables) and response (soil attribute) variables. The information on the predictors–response relationship conveyed by each variable indicator differs by the underlying mathematical principle or operation that guides their calculation. Thus, variable indicators may be considered appropriate for optimum spectral variable selection. In spectroscopy, several variable indicators exist (Teófilo et al., 2009), which may be broadly classified into PLSR-dependent and PLSR-independent categories. The β, variable influence on projection (*VIP*), squared residual vector (*SqRes*) and net analyte signal (*NAS*) are PLSR-dependent variable indicators, while correlation vector (r), signal-to-noise vector (*StN*) and covariance procedures vector (*CovProc*) are independent of PLSR model in their calculation. The coefficient vector β is a linear measure that represents the expected change in the response per unit change in the predictor variable (Mosteller and Tukey, 1977), whereas *VIP* (Wold et al., 1993) represents the importance of a predictor variable on the model based on the weighted PLSR coefficients. Variable indicator *SqRes* (Teófilo et al., 2009) represents the difference between the original and reconstructed spectra, which has relevant information on the important spectral variables. Variable indicator *NAS* is defined as the part of the spectrum unique to the attribute of interest (Ferré and Faber, 2003), and is similar to β for inverse calibration algorithms (Teófilo et al., 2009). The indicator r represents the Pearson correlation coefficients. The *StN* (Brown, 1992) denotes signal-to-noise statistics for each variable generated by least squares fit between predictor variables to the response variable. The indicator *CovProc* (Reinikainen and Hoskuldsson, 2003) represents the diagonal values of covariance matrix as a measure of strength between predictors and response variable. New vectors could be generated by combining different variable indicators following normalization (Teófilo et al., 2009).

To the best of our knowledge, the utility of variable indicators in spectral variables selection has been limited to β (Vohland et al., 2014), while *VIP* and r have been mainly used for feature visualization in soil DRS studies. Vohland et al. (2014) have cross-validated the use of β and emphasized the need for an independent validation for its use in optimum variable selection. In addition, the elemental values of β are highly dependent on the number of latent variables used in the model (Teófilo et al., 2009), and hence assumed to be less stable compared to the PLSR-independent counter parts. The utility of other PLSR-dependent, independent and their combinations in spectral variables selection is rarely examined, and thus warrants further investigation. Thus, the objectives of this study are a) to evaluate the performance of OPS approach in the optimum spectral variables selection using different variable indicators, and b) to identify the best variable indicator for optimum spectral variable selection for each soil attribute.

2. Materials and methods

2.1. Soil samples and their analyses

Soil samples examined in this study were those used by Sarathjith et al. (2014a, 2014b). Briefly, the surface (0–10 cm) soil samples were collected from 25 contiguous villages from the northern Karnataka (sampled area: 9839 km^2) and 25 villages from southern Karnataka (sampled area: 2602 km^2). In general, soils in northern Karnataka are classified as Vertisols and those in the southern Karnataka as Alfisols. Vertisols in Karnataka generally occur as Vertisols with intergrades and a mixture of Vertic Inceptisols. These soil groups are distinctly different with regard to pH, iron oxides, clay mineral, cation exchange capacity, silica-sesquioxide ratio and parent material (Lotse et al., 1972). The chemical, physical and spectral attributes of soils were estimated using that fraction which sifted through 2 mm sieve after air drying and manual grinding. Soil samples were subjected to the chemical analyses routine for the determination of pH by potentiometric means using a 1:2.5 soil/water ratio; organic carbon (OC) by the dichromate oxidation method (Walkley and Black, 1934); and extractable iron (Fe) content using inductively coupled plasma optical emission spectrometry (ICP-OES, HD Prodigy, Leeman Labs, New Hampshire, USA). The physical attributes examined in this study include soil particle size (clay and sand content) measured by pipette method (Gee and Bauder, 1986) and geometric mean diameter (GMD) by dry sieving of soil samples in a stack of eight sieves (Sarathjith et al., 2014a). These soil attributes cover a range of chemical and physical chromophores frequently estimated in the DRS approach.

A portable spectroradiometer (Field Spec 3 FR, Analytical Spectral Devices Inc.) equipped with a contact probe of 10 mm spot size was used to record the proximal spectral reflectance (350–2500 nm) from a leveled surface (Mouazen et al., 2010) of about 50 g soil sample placed in an aluminum moisture box (10 cm diameter). Soil reflectance was measured from each quadrant of the moisture box. White reference spectrum from a Spectralon (Labsphere) panel (9.2-cm diameter) was acquired before scanning each soil sample (Sarathjith et al., 2014b).

Table 1
Descriptive statistics of soil attributes.

Soil attribute	Calibration			Validation		
	n	Mean	Range	*n*	Mean	Range
Vertisols						
pH	175	8.57 (5)[a]	6.63–9.60	58	8.56 (5)	6.65–9.23
OC, %	175	0.39 (37)	0.14–0.93	58	0.39 (36)	0.15–0.76
Fe, mgL^{-1}	175	7.22 (78)	1.70–29.60	59	7.03 (76)	1.70–28.30
Sand, %	178	66.39 (13)	44.51–84.82	60	66.18 (13)	44.71–84.21
Clay, %	176	14.45 (30)	4.47–35.27	59	14.52 (32)	6.43–33.30
GMD	176	0.31 (23)	0.17–0.49	59	0.31 (24)	0.17–0.49
Alfisols						
pH	175	6.68 (21)	4.30–9.50	58	6.65 (20)	4.40–8.80
OC, %	174	0.37 (33)	0.11–0.75	58	0.37 (33)	0.12–0.70
Fe, mgL^{-1}	175	14.87 (86)	2.00–104.80	58	14.19 (75)	2.60–40.00
Sand, %	174	78.85 (9)	53.30–91.60	58	78.73 (9)	55.00–90.40
Clay, %	178	12.72 (51)	3.70–34.30	59	12.49 (49)	3.90–28.50
GMD	175	0.21 (25)	0.13–0.45	58	0.21 (24)	0.13–0.37

n: Number of soil samples.
[a] Values in parentheses are the coefficients of variation (%).