Geoderma 268 (2016) 92–99

Contents lists available at ScienceDirect

Geoderma

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

Recursive variable selection to update near-infrared spectroscopy model for the determination of soil nitrogen and organic carbon

Shengyao Jia [a], Hongyang Li [a,b], Yanjie Wang [a], Renyuan Tong [a], Qing Li [a,*]

[a] College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou, PR China
[b] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, PR China

ARTICLE INFO

Article history:
Received 26 July 2015
Received in revised form 10 January 2016
Accepted 16 January 2016
Available online 5 February 2016

Keywords:
Near infrared spectroscopy
Variable selection
Variable importance in the projection
Recursive partial least squares regression

ABSTRACT

Variable selection is widely accepted as an important step in quantitative analysis of near-infrared (NIR) spectroscopy. However, the variables preselected based on the calibration set might not be representative of the effective variables in future prediction process due to the large variability among soil sample sets. In this work, variable-updating methods (i.e., update both the model coefficients and effective variables in the prediction process) have been applied to support the robustness of the calibration model when it used to predict heterogeneous samples. Partial least squares regression (PLSR), recursive PLSR (RPLSR), and three variable-updating methods, namely variable importance in the projection combined with PLSR (VIP-PLSR), VIP-RPLSR, and uninformative variable elimination combined with PLSR (UVE-PLSR) were used to construct calibration models for the prediction of soil nitrogen (N) and organic carbon (OC) based on NIR spectroscopy. The entire data set was split into calibration set and prediction set according to soil type. The model VIP-RPLSR achieved the optimal performance for soil N and OC. The values of residual prediction deviation (RPD) were 2.9 and 2.8 for N and OC respectively. The results indicated that VIP-RPLSR was able to learn the information from the latest samples by adapting both model coefficients and effective variables at every sample interval. The proposed method VIP-RPLSR has the advantages of wider applicability and better performance for NIR prediction of soil N and OC in comparison with PLSR, RPLSR, VIP-PLSR and UVE-PLSR modeling techniques.

1. Introduction

During the last two decades, near-infrared (NIR) spectroscopy has been widely employed as an effective tool for the analysis of soil properties. Compared with traditional wet chemistry analysis, NIR analysis is rapid, cost effective, non-destructive, requires minimal sample preparation and can be used in situ. More importantly, it permits a quantitative assessment of several properties from a single measurement (Viscarra Rossel et al., 2006a, 2009). This technique mainly measures overtones and combinations of fundamental vibrational bands for O—H, N—H and C—H bonds from the mid-infrared region (Wetzel, 1983). Numerous studies for the measurement of soil nitrogen (N) and organic carbon (OC) have been reported using this technique (Rodionov et al., 2015; Kuang et al., 2015; Xie et al., 2011).

In order to cover the full spectra with a high resolution, spectrometers often provide large sets of wavelength variables for a single measurement. However, the full spectra may include wavelength variables which contribute more collinearity, redundancies and noise than relevant information to calibration models (Liu and He, 2009). Moreover, a calibration process based on full spectra is time-consuming and not convenient to fulfill the high speed features of spectroscopic techniques. Hence, variable selection methods have been proposed for the purpose of improving prediction performance and reducing the complexity of calibration models, such as variable importance in the projection (VIP; Chong and Jun, 2005), genetic algorithm (GA; Jouan-Rimbaud et al., 1995), successive projections algorithm (SPA; Galvão et al., 2008), uninformative variable elimination (UVE; Centner et al., 1996), Monte Carlo uninformative variable elimination (MC-UVE; Cai et al., 2008) and competitive adaptive reweighted sampling (CARS; Tong et al., 2015). Generally, the variable selection algorithm is performed on NIR calibration set, and then the selected variables are used for modeling and prediction. Nevertheless, the feature variables selected from the calibration set might not be representative of the effective variables in future prediction process, because soil spectral predictive mechanisms may vary from one sample set to another depending on the soil type, moisture content, surface roughness and the nature of the compounds present in soil (Mouazen et al., 2007; Waiser et al., 2007). If the samples to be measured contain interfering effects which are not included in the calibration set, the existing models may be invalid to predict them. On the other hand, the development of a multivariate calibration model is often time-consuming and costly, involving selection and preparation of a large number of standard samples, measurement of spectra. It is preferable if the calibration models can be

* Corresponding author at: College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou 310018, PR China.
E-mail address: lqdarlbar@163.com (Q. Li).

http://dx.doi.org/10.1016/j.geoderma.2016.01.018
0016-7061/© 2016 Elsevier B.V. All rights reserved.

used for an extended period. Therefore, updating the effective variables needs to be considered so as to capture the information from the latest samples. This has a great significance to maintain the robustness of the established models.

To prevent degradation in the accuracy and reliability of multivariate calibration models, various methods have been reported in NIR applications. A straightforward way is to recalculate the model coefficients using partial least squares regression (PLSR) with the addition of a few new samples to the old calibration set (Greensill et al., 2001; Wu et al., 2012). This approach requires a substantial number of samples to obtain satisfactory results. Another widely used method is spectral response standardization, also known as calibration transfer (Tan and Brown, 2001; Du et al., 2011). In this method, a transformation matrix is calculated to transform the spectra data obtained from current condition into the initial calibration condition. So the original calibration model can be still used without a significant degradation in prediction accuracy. However, this method requires the spectra of a few samples to be measured under both initial calibration condition and current test condition, which is impossible in some practical applications. Recursive PLSR (RPLSR) is another way used to update NIR model coefficients (Dayal and MacGregor, 1997a). Other than accumulating a certain number of new samples, RPLSR expands the calibration set by adding every sample available and continuously recalculated the model coefficients. Because of the frequent updating of the model coefficients, RPLSR is able to capture the information from the latest sample rapidly (Dayal and MacGregor, 1997a; Haavisto et al., 2008). In summary, among the model-updating methods (i.e., update the model coefficients) mentioned above, variable-updating (i.e., not only update the model coefficients, but also update the feature variables) has been seldom considered. Up to our knowledge, no literature adopted variable-updating method for the analysis of soil properties.

In this work, a heterogeneous set of soil samples covered a relatively wide range of soil types, soil textures and soil use has been adopted. The combinations of VIP-RPLSR, VIP-PLSR and UVE-PLSR were applied as variable-updating methods. The objectives of this work were to: (i) compare the performances of PLSR, RPLSR and variable-updating methods for the prediction accuracy of soil N and OC using NIR spectroscopy, and (ii) investigate the feasibility of using variable-updating methods to maintain the robustness of the calibration model when they used to predict heterogeneous samples.

2. Material and methods

2.1. Soil samples

A total of 205 soil samples (Table 1) collected from nine towns in Wencheng county, Zhejiang province, P. R. of China, were used in this work. They were chosen from the 5–25 cm layer between April 2010 and March 2012. According to the classification and codes for Chinese soil, the soil samples belong to three soil orders, namely, Ferralsols, Anthrosols and Primarosols. All the samples were air dried and sieved to pass through a 2 mm mesh. After removing identifiable crop residues and stones, the samples were air-dried again at 40 °C for 48 h. About 50 g of each sample was sent to the agricultural testing center of Zhejiang Provincial Academy of Agricultural Sciences (ZPAAS) for soil chemical analyses. The remaining samples were used for spectrometer measurements and data analysis.

2.2. Laboratory reference measurement

Laboratory analyses of soil N and OC were performed by ZPAAS using standard procedures. Soil N was determined using the Kjeldahl method and OC using the Walkley–Black method, both described in Hesse (1971). Soil N and OC expressed in percentage of their weight to the total weight of dry soil.

2.3. NIR spectrum acquisition and data pre-processing

The diffuse reflectance spectra of soil samples were measured by a Fourier-type NIR spectrometer (Matrix-I, Bruker Optics Inc., Germany). The light source of Matrix-I irradiated samples from down to up through a quartz window, which was embedded in the top of the spectrometer. Soil sample was packed in a sample cup. Then the cup was fixed on the quartz window by swivel bracket. When measuring, soil cup was spinning around for the purpose of getting the averaged spectrum of each sample. Background spectrum measurement was taken once every ten sample measurements. The spectral resolution was 8 cm^{-1} and sampling interval was 3.86 cm^{-1}. Spectral absorbance was recorded in the wavelength range of 1000–2500 nm for a total of 1555 wavelength variables per spectrum. Each reading was an average of 64 successive measurements in 40 s, and this was used for spectra pre-processing and model establishment.

The same pre-processing was carried out for both soil N and OC. The spectra were first smoothed by averaging five successive wavelengths. Then, standard normalized variate (SNV) was used to reduce baseline offset and noise of the spectra. Finally, z-score normalization (Rahman et al., 2009) was used to get all the data to approximately the same scale. In order to examine the structure of the spectral data and distinguish soil types, a principal components analysis (PCA) was carried out on the wavelength matrix, and the PCA scores were submitted to Fisher's linear discriminant analysis (LDA). Software used for pretreatment was the Unscrambler 9.7 (CAMO Inc., Oslo, Norway), whereas LDA was implemented using PASW Statistics 18 (SPSS Inc., Chicago, United States).

In order to examine the robustness and reliability of calibration models, the pretreated spectra were divided into two subsets according to soil type: the calibration set was composed of 120 Anthrosols samples, while the prediction set included the remaining 24 Anthrosols samples and 61 samples of Ferralsols and Primarosols. Sample statistics were summarized in Table 2. For the method of PLSR, calibration model was established based on the calibration set and did not change during the prediction process. The prediction set was used for independent prediction of the established model. For the method of RPLSR, VIP-RPLSR, VIP-PLSR and UVE-PLSR, the calibration set was used to establish a database and initialize the original calibration model. Once a sample from the prediction set was predicted by the current calibration model, it would be added to the database. Then a new calibration model was reconstructed using the updated database.

Table 1
Basic information about experimental soil samples.

No.	Soil classification[a]	Soil subgroup name	Dominant crop	NS[b]	Soil texture[c]				
					Silty loam	Loam	Sandy loam	Clay loam	Clay
1	Ferralsols	Red and yellow soil	Soybean, rice, potato	38		11	22	5	
2	Anthrosols	Paddy soil	Rice, potato, vegetable, tea	144	2	69	32	30	11
3	Primarosols	Purple soil	Potato, vegetable, strawberry, pear,	23		4	15	4	

[a] According to the classification and codes for Chinese soil (GB/T 17296-2009).
[b] Number of samples.
[c] According to the standard of United State Department of Agriculture (USDA) for soil texture.