Estimating soil total nitrogen contents from visible/near-infrared reflectance based on successive projections algorithm

X M Shu1, T Z Shi2, Y Liu1, H W Peng3, W X Gao4, and C Zhang1
1 Changjiang waterway survey center, 430010, Wuhan, China
2 School of Resource and Environmental Science & Key Laboratory of Geographic Information System of the Ministry of Education, Wuhan University, 430079, Wuhan, China
3 The Second Artillery Command College, 430012, Wuhan, China
4 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

tiezhushi@whu.edu.cn.

Abstract. The Visible/near-infrared reflectance (Vis/NIR) spectroscopy combined with multiple regression methods have been employed to assess various soil properties quantitatively. However, useless or irrelevant information in some bands worsen the performances of the calibrated models. Successive projections algorithm (SPA) is an effective technique for selecting informative variables, and its application potential in estimating soil contents using visible/near-infrared reflectance (Vis/NIR) spectroscopy has not been explored. This paper aimed to explore the effect of SPA combined with support vector machine regression (SVMR) on soil total nitrogen (TN) contents estimation using Vis/NIR spectroscopy. For this purpose, SPA-SVMR was compared with full spectrum based SVMR (FS-SVMR) for soil total nitrogen (TN) estimation, respectively. The results showed that the SPA-SVMR model ($R^2_v = 0.58$, RMSEV = 0.364) performed better than the FS-SVMR model ($R^2_v = 0.49$, RMSEV = 0.394). We concluded that the SPA could eliminate the uninformative variables and improve the performance of SVMR on estimating TN contents. In further study, using the SPA-SVMR method to estimate other soil properties is necessary to test this finding.

1. Introduction
The Visible/near-infrared reflectance (Vis/NIR) spectroscopy combined with multiple regression methods have been employed to estimate various soil properties. However, the Vis/NIR spectra are typically consisted of vast, weak, non-specific and over-lapped bands [1], and some bands may contain useless or irrelevant information for model calibration like noise and background, which might worsen the performances of the calibrated models. Some studies indicated that the better models might be obtained by selecting the characteristic bands or wavelengths including component-specific information instead of the full bands [1].

Successive projections algorithm (SPA) is proposed as a novel, easy and efficient variable selection strategy for multivariate calibration [2]. Multiple linear regression (MLR) combined with SPA (SPA-MLR) were comparable to or superior than the PLSR in terms of estimation accuracies, and they better performed than the SMLR in most cases [3, 4]. As a new machine learning method based on the support vector machine (SVM) method developed by Vapnik [5], support vector machine regression (SVMR) can handle nonlinear data effectively. Many authors have estimated various soil properties...
with SVMR and obtained good results [6]. However, there was no report that combining SVMR and SPA to estimation soil properties.

This study, with the Yixing region of Jiangsu Province as the study area, aimed to explore the influence of SPA on the accuracy of soil content estimation and its effect on SVMR modelling from Vis/NIR reflectance. For this purpose, the SVMR employing SPA (SPA-SVMR) was compared with the SVMR with full spectral bands (FS-SVMR) for soil total nitrogen (TN) estimation, respectively.

2. Materials and methods

2.1. Study area and Field sampling

Yixing (119°31′–120°03′E, 31°07′–31°37′N) is situated in Jiangsu Province, China. The soil types of the Yixing region are mainly yellow brown, krasnozem, paddy, cardalzeye, calcareous alluvial, limestone, purple and boggy soil [7]. A total of 100 soil samples were collected from the southern part of the Yixing region on 11–14 August 2010. The soil samples were heterogeneous and mainly comprised the typical soil types within the Yixing region.

2.2. Laboratory analyses and measurements

After air-drying at an indoor temperature and removing stones and plant residues, the 100 soil samples were ground with an agate mortar and passed through a 20-mesh grid sieve (< 2 mm). An ASD FieldSpec®3 portable spectroradiometer with a wavelength range of 350–2500 nm was used to measure the spectral reflectance. After spectral measurements, the soil TN contents were determined using the semi-micro Kjeldahl method.

2.3. Pre-processing transformations

The spectral data were combined with the soil TN content values. The whole dataset was divided into calibration and validation set randomly. The reflectance curves were first reduced to 400–2450 nm by removing the wavelengths with high noise effects at the edges of spectra. The reflectance curves were then smoothed with a moving window of 9 nm using the Savitzky-Golay smoothing method. Each reflectance spectrum was resampled to 10 nm intervals, and thus it covered 205 spectral variables. The first derivative was used to reduce baseline variation and to enhance spectral features.

2.4. Statistical analysis.

The FS-SVMR and the SPA-SVMR were calibrated using Matlab version 2008b. The calibrated SVMR and SPA-SVMR models were used to estimate the soil TN contents of the validation dataset. The model performances were described by the coefficient of determination ($R^2_v$) of the regression line between the measured and estimated values in validation dataset and the root mean square error of validation (RMSEV).

3. Results and discussions.

3.1. Wavelengths selected by SPA.

The best variable subset generated by SPA was matrix $X_{SPA} = \{x_{1930}, x_{1880}, x_{2410}, x_{570}, x_{2380}, x_{2160}, x_{2190}\}$. Most selected wavelengths were associated to the peak positions of the first derivations (Fig. 1 (a), indicated by black squares). The correlation coefficients ($r$) of the soil TN content against these SPA-selected wavelengths are displayed in Fig. 1 (b) (indicated by black squares), and the selected wavelengths are mostly located around the peak points of the correlation coefficient ($r$) curve.
3.2. FS-SVMR and SPA-SVMR comparison.

![Figure 1](image1.png)

**Figure 1.** The SPA-selected (indicated by black squares) (a) and correlation coefficients ($r$) of the soil TN content against these wavelengths (b).

![Figure 2](image2.png)

**Figure 2.** Scatter plots of the measured vs. estimated total nitrogen (TN) contents (g kg$^{-1}$) for the FS-SVMR (a), SPA-SVMR (b) methods (the solid line is the regression line between the estimated and measured values, and the dashed line is the 1:1 line).

The $R^2$ of FS-SVMR and SPA-SVMR were 0.85 and 0.68, respectively. When the calibrated FS-SVMR and SPA-SVMR models were employed to estimate the soil TN contents of the validation dataset, the results showed that the SPA-MLR model ($R^2=0.58$, RMSEV=0.364) (Fig. 2 (a)) performed better than the FS-SMVR model ($R^2=0.49$, RMSEV=0.394) (Fig. 2 (b)). The SPA-SVMR model avoided over-fitting effectively ($R^2=0.68$, $R^2=0.58$) by using SPA-selected variables. In addition, we found that, because of using less spectral variables, the SPA-SVMR model was more parsimonious than the FS-SVMR.

Although the SPA is an effective method for selecting informative spectral variables in soil TN estimation, there still exists shortages. The number of selected wavelengths, which cannot be larger than the number of calibration samples, is a limitation of SPA[2]. If more spectral variables were needed to estimate soil components, a large number of soil samples would be required to perform the
calibration, which is major handicap in the application. It is necessary to pay attention to these shortcomings when the SPA is used for wavelengths selection/elimination.

In this study, the SPA method was used to eliminate the uninformative variables and improved the accuracy for soil TN content estimation. The combination of SPA method with soil diffuse reflectance will be further explored for estimating other soil properties (soil organic matter, organic carbon, slit, clay, heavy metal content et al.), while the SPA will be compared with other uninformative elimination methods (like uninformative variable eliminate (UVE), genetic algorithm (GA)) to explore their relative performances in order to find an optimal variable selection/elimination technique.

4. Conclusions.
In this paper, the SPA-SVMR model was more parsimony and obtained better estimation accuracies than the FS-SVMR. We concluded that the SPA might be an effective method to select informative variables and minimize the multi-collinearity among variables for soil TN content estimation.

Reference
[1] Thomas E V 1994 Anal Chem 66 795A-804A
[2] Araújo M C U, Saldanha T C B, Galvão R K H, Yoneyama T, Chame H C and Visani V 2001 Chemometr Intell Lab Syst 52 65-73
[3] Araújo H F, Galvão R K H, Pimentel M F, Neto B B, Araújo M C U and Carvalho F R 2005 Chemometr Intell Lab Syst 76 65-72
[4] Galvão R K H, Araújo M C U, Fragoso W D, Silva E C, Jose G E, Soares S F C and Paiva H M 2008 Chemometr Intell Lab Syst 92 83-91
[5] Vapnik V 1995 The nature of statistical learning theory (New York: Springer)
[6] Vohland M, Besold J, Hill J and Frund H C 2011 Geoderma 166 198-205
[7] Li R H, Zhou S L, Song J B, Ye F and Zhu Q 2004 Act Pedologica Sinica 41 517-521