Research Article
Measurement of Available Phosphorus and Potassium Contents in Soil using Visible-near-infrared Spectroscopy in Conjunction with SPA-LS-SVM Methods

Zhang Lei and Zhang Rong-Biao
School of Electric Information Engineering, Jiangsu University, Zhenjiang 212013, China

Abstract: Applying Near Infrared Reflectance Spectroscopy (NIRS) on farmlands can effectively estimate the available phosphorus and potassium contents of soil online. Spectral preprocessing, including Savitzky Golay (SG), Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC) and SG 1st derivative, aimed to eliminate system noise and external interference. A correction model was created using respectively Radial Basis Function (RBF) and Least Squares Support Vector Machine (LS-SVM) methods with input from the characteristic wavelengths obtained using Successive Projections Algorithm (SPA). The results of predicting available phosphorus and potassium contents in soil using these two modeling methods were evaluated and the better model was selected. The results showed that the LS-SVM method with input from the characteristic wavelengths obtained using SPA had an advantage over the RBF modeling method. In SPA-LS-SVM models, the correlation coefficient and mean square error of prediction for available phosphorus were 0.8625 and 8.67 and those for available potassium were 0.7843 and 13.42, respectively. This indicates that SPA-based visible-near-infrared spectroscopy using LS-SVM for modeling can be used as a method to accurately measure available phosphorus and potassium contents in soil.

Keywords: LS-SVM, RBF, Spectroscopy, SPA

INTRODUCTION

Soil nutrient is one of the major limiting factors that affect crop growth. Crops need water and various nutrients from agricultural soils to grow and a large amount of soil phosphorus and potassium is needed by plants. These elements are one of the major factors that decide plant growth and soil productivity, as well as an important basis to guide scientific and balanced fertilization. It is therefore important to gain information about distribution of phosphorus and potassium in agricultural soil so as to provide guidance for agricultural production (Rossel *et al*., 2006; Gomez *et al*., 2008; Reeves III and Smith, 2009; Pan *et al*., 2009; Reeves III, 2010; Yu, 2011; Wang *et al*., 2013). Traditionally, available phosphorus and potassium contents in soil are measured using a chemical analytical method, which is time-consuming, cumbersome and can hardly meet the need for rapid monitoring of these contents. In recent years, near-infrared spectroscopy technology has been gradually becoming the new powerful tool for measuring available phosphorus and potassium contents in soil due to its fast speed, simplicity and reliability (Bogrekci and Lee, 2005a, 2005b; 2006; Bao *et al*., 2007; Barthes *et al*., 2008; Brichlemyer and Brown, 2010; Reeves III, 2010; Minasny *et al*., 2011).

Recent study using VIS/NIRS technology to build regression model for estimating organic matters, total nitrogen, total phosphorus and total potassium in soil showed that it is possible to rapidly estimate these figures. Confalonieri *et al*. (2001) utilized VIS/NIRS technology to estimate available phosphorus and potassium contents in different soils, but the estimation result was unsatisfactory. Li *et al*. (2007) used Partial Least Squares (PLS) and Artificial Neural Network (ANN) methods to create models for estimating alkali-hydrolyzable nitrogen, available phosphorus and potassium contents in soil. The result showed that it is feasible to use NIRS technology to estimate alkali-hydrolyzable nitrogen content in soil, but the feasibility of estimating available phosphorus and potassium contents still requires further study.

Near-infrared spectroscopy mainly reflects multiple and frequency absorption of organic matters. Different matters have complicated overlapping bands, causing the entire band spectrum to contain a lot of redundant information and noise, thus affecting the predicting performance of the model. Wavelength extraction method includes wavelet algorithm, genetic algorithm, Uninformative Variable Elimination (UVE), SPA, etc. The correlation coefficient method and effective wavelength method often select thresholds based on subjective experience without any systematic and effective threshold selection criteria, while UVE methods such as Monte Carlo method and genetic algorithm require time-consuming and intensive computation during the search process but yet generate unstable result. SPA can significantly reduce the number of variables

Table 1: Statistics of sample properties

Property	Max.	Min.	Mean	S.D.
P/(mg/kg)	72.6	5.6	38	4.54
K/(mg/kg)	89	25	66	15.86

Max.: Maximum; Min.: Minimum; S.D.: Standard deviation

used in modeling to an extent less than that using algorithm such as Monte Carlo UVE, genetic algorithm and wavelet algorithm. It also improves the speed and efficiency of modeling. SPA has therefore been widely applied in selecting the characteristic band of a near-infrared spectrum. This study examined the feasibility of applying SPA in conjunction with the LS-SVM modeling method to select and optimize modeling variables for soil near-infrared spectrum. With 160 experimental soil samples collected, SPA was employed to select and optimize the modeling variables. Then the optimized samples were used to create a model for predicting available phosphorus and potassium contents in soil.

EXPERIMENTS

Sample collection and treatment: The samples were collected from the agricultural soil located in the city of Zhenjiang along the Yangtze River, at north latitude 32º 6́6́ and east longitude 119º 2́4́, within the monsoon climate zone in the southern part of the northern subtropical. Due to its location in the Middle and Lower Yangtze Valley Plain, most of the soil is yellow-brown soil formed by the impact of the Yangtze River. In order for the samples to contain a wide range of nutrient contents, sampling points were selected using cross method at a depth of 0-20 cm. When sampling soil, the gravels and plants on the soil surface were removed and a 20 cm deep rectangular pit was dug. Then a 5 cm thick layer of soil was shoveled off from the side wall of the pit and collected in bags, resulting in 160 soil samples. In order for the samples to reflect the actual conditions of the agricultural land as far as possible, the soil was not ground but dried in natural conditions and then passed through a 2 mm sieve for future use. Then the soil was placed in and fill 9 cm diameter and 1 cm deep petri dishes, with the surface leveled using a ruler. The samples were divided into two parts, one for chemical analysis and the other for near infrared analysis. Available phosphorus content in soil was measured using a spectro-photometer with the method of colorimetry. Phosphorus in acid soil is present mainly in the forms of Fe-P and A1-P. The ability of fluorine ion to complex Fe^{3+} and Al^{3+} in acidic solution can be utilized to successively activate and release the iron aluminum phosphate that is relatively active in such soil. Meanwhile, due to the effect of H^+, some Ca-P that is relatively active can also be dissolved out. Then the content can be measured using the Mo-Sb colorimetric method. Available potassium content in soil was measured using the flame photometry method. When
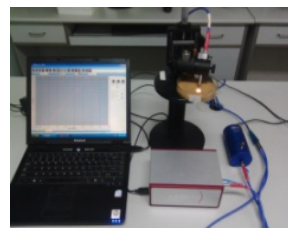

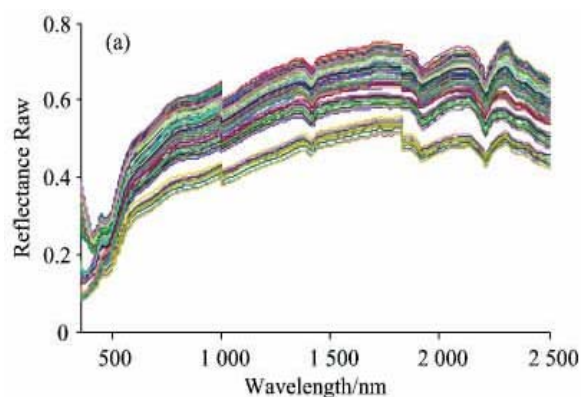Fig. 1: Spectral acquisition instrument

Fig. 2: Spectral curve

neutral 1 molL-1NH4Ac solution is used to leach the soil, NH_4^+ is exchanged with K^+ on the surface of soil colloid and goes into the solution along with the water-soluble K^+. K in the leaching solution can be measured directly using the flame photometry method. The parameters of the test sample soil are shown in Table 1.

Spectral acquisition instrument and spectral preprocessing: Fourier transform near-infrared spectrometer of Swiss Arcspectro FT Rocket type was used for spectroscopic measurement with spectral wavelength ranging between 350-2500 nm. The instrument is equipped with a 35 W halogen light source and a fiber optical probe, both mounted on the built-in bracket of the spectrometer at an included angle of 8°. The distance between the fiber optic probe and soil sample is adjustable and the probe is connected through an optical fiber to the spectrometer which parses the reflectance spectrum and sends it to a computer via USB, as shown in Fig. 1.

Each sample was taken at an interval of 1nm for 5 times. The spectral curve after averaging is shown in Fig. 2.

Averaging was done for every five consecutive wavelengths to reduce the dimensions of the spectrum. Then SNV, MSC and 1st derivative preprocessing were respectively performed on the spectrum in order to compare and analyze the effect of different spectral preprocessing methods on model prediction result. The results are shown in Table 2. By comparing the characteristic wavelengths, extraction was performed

Table 2: Different spectral preprocessing methods of model prediction results

Parameter	Preprocessing	Calibration		Prediction	
		r	RMSEC	r	RMSEP
P/(mg/kg)	None	0.6850	11.61	0.6830	11.69
	SG	0.6868	11.45	0.6825	11.52
	SNV	0.6874	11.23	0.6856	11.46
	MSC	0.6986	11.15	0.6925	11.35
	SG+1st derivative	0.7542	10.75	0.7328	11.02
K/(mg/kg)	None	0.6552	17.65	0.6450	18.46
	SG	0.6574	16.88	0.6532	17.35
	SNV	0.6596	15.79	0.6586	16.27
	MSC	0.6768	15.57	0.6735	15.87
	SG+1st derivative	0.7543	15.35	0.7328	15.80

Table 3: Comparison of results in calibration model with different height to soil surface

Height of sampling/cm	P/(mg/kg)		K/(mg/kg)	
	r	RMSEC	r	RMSEC
5	0.6386	10.68	0.5396	15.68
7	0.7292	9.88	0.6192	14.88
10	0.7985	9.29	0.6965	14.30
12	0.7652	9.52	0.6552	14.52
15	0.6996	10.33	0.5897	15.33

on the basis of the original spectrum using the SG 1st derivative preprocessing method to produce the optimal prediction model result.

PREDICTION MODELING

Effect of sampling height on spectrum: Reflectance spectrum in soil measurement is affected by factors such as measured component structure, stability and probe mounting height, so it is necessary to examine the effect of probe mounting height on measuring accuracy in order to increase the accuracy of online measurement. In order to facilitate field measurement, the spectral characteristics of soil sample when the height of the spectrometer's fiber optic probe were respectively 5, 7, 10, 12 and 15 cm, respectively to the measured object were selected and the spectral data collected were used in conjunction with available phosphorus and potassium contents in soil to create a PLS prediction model, the results is as shown in Table 3. The table shows that the probe height of 10 cm to the soil sample surface gave the best modeling result. The determination coefficient of modeling calibration set was 0.7985, 0.6965 and the root mean square error of modeling calibration set was 9.29, 14.30, so a sampling height of 10 cm was determined.

Selection of characteristic wavelength using SPA: Abnormal samples will seriously affect the prediction accuracy of the model created. This study applies the 3 σ edit rule and Principal Component Analysis (PCA) score plot to the values of matter content and spectral data to determine if any values of available phosphorus and potassium contents in soil or sample spectral data are abnormal. Since a large number of samples are required for quantitative analysis and overlapping spectral peak is present in the spectral data of each sample, resulting in redundant spectral information and

unobvious characteristic absorption peak, it is necessary to find effective wavelengths which play a key role to the model (Wang *et al.*, 2011). Successive Projections Algorithm (SPA) selects a small number of wavelengths from the original data by projecting and mapping spectral data so as to summarize as much sample spectral information as possible and avoid information overlapping to the greatest extent (Chen *et al.*, 2013). Competitive Adaptive Reweighted Sampling (CARS) method imitates the "survival of the fittest" principle in Darwinian evolution theory to perceive each wavelength variable as an individual and eliminate in adaptive individuals step by step. CARS allows "survival of the fittest and selection of the best" and eliminates useless variables at an exponentially decreasing speed to select the optimal wavelength (Sun *et al.*, 2012). Stability Competitive Adaptive Reweighted Sampling (sCARS) method takes stability into account on the basis of CARS and eliminates redundant information by using the stability of a variable as an indicator of its modeling capability. Random frog algorithm calculates the probability of each variable being selected by simulating a Markov chain that has a stationary distribution in the model space, thereby performing variable selection. In conjunction with the PLS regression algorithm, Random frog algorithm uses the absolute value of each variable regression coefficient in the model as the basis for deciding whether to eliminate the variable in each iteration process. It is a very effective method for selecting high-dimensional data variables (Mouazen *et al.*, 2005, 2006).

In order to find out the optimal wavelength selection method, a test for characteristic wavelength selection was conducted for the entire band spectral data using SPA, CARS, sCARS and Random frog respectively and a model was created for the selected

Table 4: Comparison of parameters in calibration model with different variable selection methods

Wavelength selection method	Characteristics of the wavelength number	P/(mg/kg)		K/(mg/kg)	
		r	RMSEC	r	RMSEC
RAW	500	0.8465	8.08	0.8025	12.08
SPA	30	0.7827	9.06	0.7357	13.06
CARS	30	0.6584	12.60	0.6886	14.80
sCARS	30	0.7236	11.62	0.6956	14.60
Random Frog	30	0.6258	11.33	0.6178	15.33

Table 5: Results of soil samples using SPA-RBF model and SPA-LS-SVM model

Model	Parameter	Calibration		Prediction	
		r	RMSEC	r	RMSEP
SPA-RBF	P/(mg/kg)	0.8626	8.65	0.8330	9.65
	K/(mg/kg)	0.7833	13.55	0.7505	14.12
SPA-LS-SVM	P/(mg/kg)	0.9126	7.68	0.8625	8.67
	K/(mg/kg)	0.7965	12.32	0.7843	13.42

wavelength points using PLS. The results are shown in Table 4. Table 4 shows that SPA can compress raw spectral data, thus greatly reducing the computational complexity of the model and improving its stability as well as prediction accuracy, provided that the number of characteristic wavelengths is equal. The algorithm attempts to find the wavelength data containing the minimum redundant information in the raw spectral data by sorting and filtering the selected wavelengths according to the contribution value of the test samples, thus preventing the selected wavelength data to contain overlapping information and removing redundant information. For these reasons, the SPA method was used for all subsequent prediction experiments. After the soil spectral data was compressed and filtered using SPA, nine contributing wavelength points were left, which were respectively 447, 655, 722, 1055, 1255, 1467, 1678, 1890 and 2246 nm, respectively.

Modeling methods: Common methods for quantitative analysis in near-infrared spectroscopy include linear analysis methods and nonlinear methods such as neural network. It was found through preliminary experiments that linear analysis methods produce poor prediction result with respect to rapidly available phosphorus content in soil, so this study compares and analyzes the modeling results of nonlinear methods such as Radial Basis Function (RBF) and Least Squares Support Vector Machine (LS-SVM) in an attempt to find out the optimal modeling method.

RBF network model: An RBF network consists of an input layer, a hidden layer and an output layer. The input layer has 120 neurons; the hidden layer contains a Gaussian radial basis function with 5 neurons as determined through comparative experiments; the output layer contains a linear transfer function with one neuron corresponding to the predicted value of available phosphorus or potassium content in soil. During training, 120 samples were selected as input using the SPXY (Sample set partitioning based on joint X-Y distance) method from the sample set in which all

abnormal samples had been removed (for each sample, the characteristic wavelengths were selected using the SPA method). Considering modeling and prediction accuracy on a comprehensive basis, an error target of 55 was determined through repeated tests and other parameters were set to default values. After the network had been trained, the remaining 30 soil samples were inputted into the trained RBF network as a prediction set for the purpose of prediction.

LS-SVM: LS-SVM is an improved support vector machine which maps the training set data from nonlinear space to high-dimensional characteristic space and substitutes equality constraints for inequality constraints to solve minimized loss function in a high-dimensional space so as to obtain a linear fitting function. It can significantly increase computational speed. During training, input X for LS-SVM was the 120 samples selected using the SPXY method; the number of dimensions of the input matrix was the number of characteristic wavelengths selected using SPA; output Y was the value of available phosphorus or potassium content in soil obtained from the 132 samples. During prediction, input X_t was the spectral information contained in the other 20 test samples; output Y_t was the predicted value of available phosphorus or potassium content in soil. It was determined through preliminary experiments that a regularization parameter (γ) of 0.72 and RBF kernel function parameter (σ^2) of 0.58 produce the best prediction result for phosphorus or potassium content in soil.

EXPERIMENTAL RESULTS AND ANALYSIS

Experimental results: Table 5 shows the results of correlation coefficient and mean square error of prediction for available phosphorus and potassium contents in soil samples using SPA-RBF model and SPA-LS-SVM model.

The relation between the predicted value and measured value of the prediction set samples when SPA
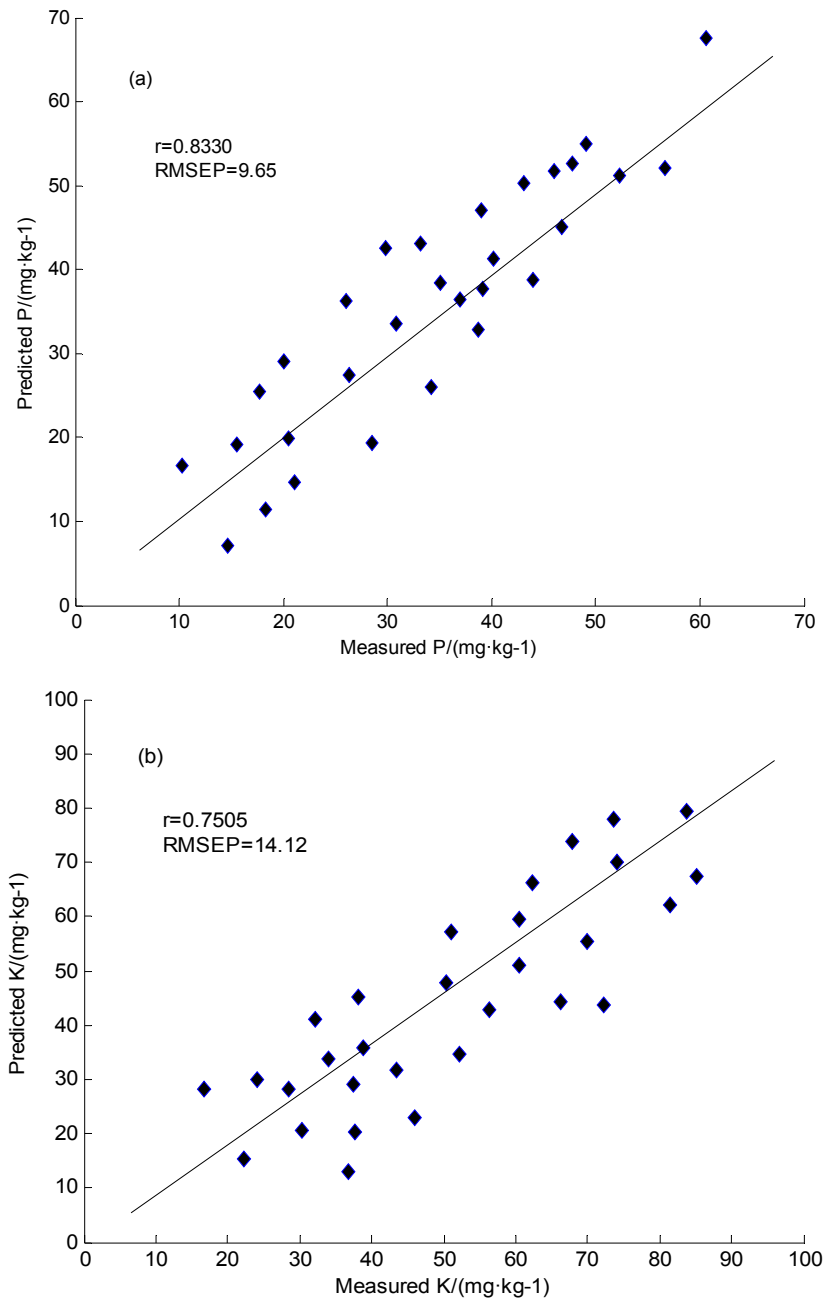
Fig. 3: Scatter plot by using SPA-RBF models; (a): available P; (b): available K

was used to select the optimal characteristic wavelength variables and available phosphorus and potassium contents in soil to create a RBF network prediction model is shown in Fig. 3. The determination coefficients of prediction for the prediction set of this model were respectively 0.8330 and 0.7505 and the root mean square errors of prediction were respectively 9.65 and 14.12. The relation between the predicted value and measured value of the 30 prediction set samples when SPA was used to select the optimal characteristic wavelength variables and LS-SVM was used for modeling is shown in Fig. 4. The determination coefficients of prediction were respectively 0.8625 and

0.7843 and the root mean square errors of prediction were respectively 8.67 and 13.42.

Experimental analysis: Due to the massive size of near-infrared spectral data for soil, severe collinearity is present in the data points and network computation can easily fall into local minimum and difficulty may be experienced in parsing. It can be seen from the model prediction result for available phosphorus and potassium contents in soil that the prediction result was satisfactory as no under-fitting or over-fitting phenomenon occurs in the two nonlinear models, indicating that spectroscopy is advisable for the
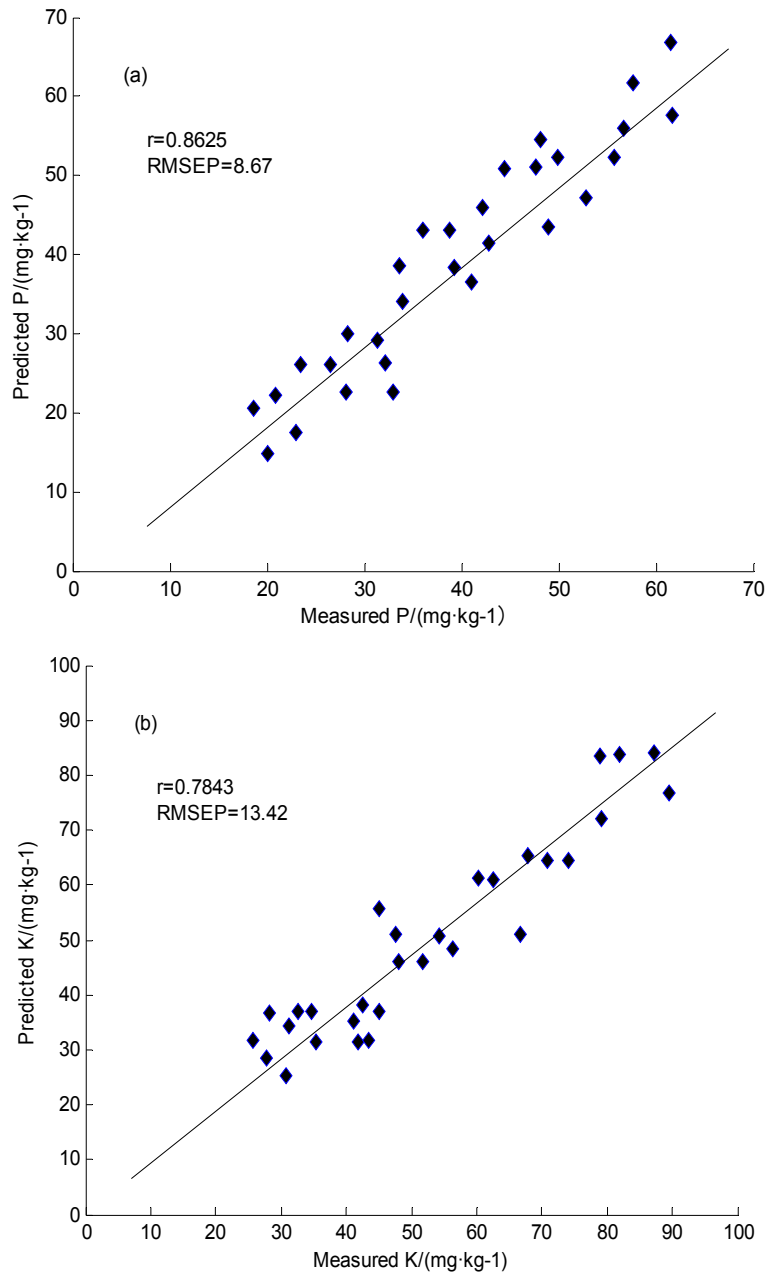
Fig. 4: Scatter plot by using SPA-LS-SVM models; (a): available P, (b): available K

prediction of available phosphorus and potassium contents.

To analyze the reliability and adaptability of different models, RBF and LS-SVM prediction models were compared and analyzed on a comprehensive basis. The result showed that the difference between their determination coefficients and root mean square errors was not obvious and both can meet the needs of practical application. However, LS-SVM using least squares linear system as loss function greatly reduces model complexity, increases computational speed and saves time. It is also more responsive to application of real-time field measurement. In summary, LS-SVM

features higher learning and prediction capabilities, so it can better solve the generalization problem of model prediction for available phosphorus and potassium contents in soil.

The prediction result for the available phosphorus and potassium contents in the soil samples was satisfactory in this test, probably because they were indirectly captured by near-infrared spectrum through some active ingredients and minerals in the soil. The successful prediction of available phosphorus and potassium contents was achieved using spectral modeling algorithms based on some kind of generative mechanism which was captured by near-infrared

spectrum signal. According to existing references, the characteristics of micronutrients such as available phosphorus, potassium, magnesium and calcium in soil, as well as the mechanisms in relation to prediction using near-infrared spectrum are still unknown and need further discussion and study. At present, the overall model accuracy is not high, because the study was based on a variety of soil types and the physical and chemical properties of different soils vary significantly with soil-forming factors such as climate, parent materials, topography and biology, as well as human activities. The diversity of soil constituents and unique spectral characteristics of each constituent give each soil spectrum its own features. In future studies, the accuracy of model can be improved by, for example, designing a more stable light source, the stability of which is important for ensuring the analysis accuracy and repeatability of measurement system, taking additional factors such as ambient temperature and soil moisture into account and increasing the number of samples for modeling to improve the stability and reliability of model.

CONCLUSION

After preprocessing of the original spectrum using SG 1st derivative, SPA was used to determine characteristic wavelengths as spectral data for the modeling set and prediction set, i.e., effective wavelengths. The following conclusions were made according to the study on measurement of available phosphorus and potassium contents in soil using SPA in conjunction with LS-SVM:

- Difference in sampling height significantly affected spectral data; the sampling height of 10 cm gave the highest accuracy. The division between abnormal samples and modeling set/prediction set samples also had a substantial impact on model accuracy.
- Of the several methods for selecting characteristic wavelengths, the SPA method extracted a few columns of data from a massive amount of raw spectral data to give a high-level overview of most sample spectral information, thus avoiding information overlapping, greatly reducing the computational complexity in calibration modeling and effectively compressing the time required for modeling.
- Comparison and analysis of the prediction results using two nonlinear modeling methods showed that SPA-LS-SVM gave the best prediction result, in which case the determination coefficients of model prediction were respectively 0.8625 and 0.7843 and the root mean square errors of prediction were respectively 8.67 and 13.42. The model was accurate enough to basically meet the requirements for actual measurement.

- It is feasible to employ spectral analysis to rapidly measure available phosphorus and potassium contents in soil. However, soil spectrum is complex and diffuse reflectance spectrum of soil is affected by soil quality, topography and the environment. Further study is still required to reduce and eliminate these interferences and create prediction models with better universality and robustness.

REFERENCES

Bao, Y., Y. He, H. Fang and A.G. Pereira, 2007. Spectral characterization and N content prediction of soil with different particle size and moisture content. Spectrosc. Spect. Anal., 27(1): 62-65.

Barthes, B.G., D. Brunet, E. Hien, F. Enjalric, S. Conche, G.T. Freschet, R. d'Annunzio and J. Toucet-Louri, 2008. Determining the distributions of soil carbon and nitrogen in particle size fractions using near-infrared reflectance spectrum of bulk soil samples. Soil Biol. Biochem., 40: 1533-1537.

Bogrekci, I. and W.S. Lee, 2005a. Spectral soil signatures and sensing phosphorus. Biosyst. Eng., 92(4): 527-533.

Bogrekci, I. and W.S. Lee, 2005b. Improving phosphorus sensing by eliminating soil particle size effect in spectral measurement. T. ASABE, 48(5): 1971-1978.

Bogrekci, I. and W.S. Lee, 2006. Effects of soil moisture content on absorbance spectra of sandy soils in sensing phosphorus concentrations using UV-VIS-NIR spectroscopy. T. ASABE, 49(4): 1175-1180.

Brichlemyer, R. and D. Brown, 2010. On-the-go VisNIR: Potential and limitations for mapping soil clay and organic carbon. Comput. Electron. Agric. 70(2): 209-216.

Chen, B., G. Liu and X.M. Zhang, 2013. Analysis on near infrared spectroscopy of water content in lubricating oil using successive projections algorithm. Infrared Laser Eng., 42(12): 3168-3174.

Confalonieri, M., F. Fornasier, A. Ursino, F. Boccardi, B. Pintus and M. Odoardi, 2001. The potential of near infrared reflectance spectroscopy as a tool for the chemical characterisation of agricultural soils. J. Near Infrared Spec., 9(2): 123-131.

Gomez, C., R.A. Viscarra Rossel and A.B. McBratney, 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma, 146(3-4): 403-411.

Li, W., S.H. Zhang, Q. Zhang *et al.*, 2007. Rapid prediction of available N,P and K content in soil using near-infrared reflectance spectroscopy. T. Chinese Soc. Agric. Eng., 23(1): 55-59.

Minasny, B., A.B. McBratney, V. Bellon-Maurel, J. Roger, A. Gobrecht, L. Ferrand and S. Joalland, 2011. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. Geoderma, 167: 118-124.

Mouazen, A.M., J. De Baerdemaeker and H. Ramon, 2005. Towards development of online soil moisture content sensor using a fibre-type NIR spectrophotometer. Soil Till. Res., 80: 171-183.

Mouazen, A.M., R. Karoui, J. De Baerdemaeker and H. Ramon, 2006. Characterization of soil water content using measured visible and near infrared spectra. Soil Sci. Soc. Am. J., 70(4): 1295-1302.

Pan, L., J. Wang, P. Li, Q. Sun, Y. Zhang and D. Han, 2009. Region optimization of SSC model for pyrus pyrifolia by genetic algorithm. Spectrosc. Spect. Anal., 29(05): 1246-1250.

Reeves III, J.B. and D.B. Smith, 2009. The potential of mid-and near-infrared diffuse reflectance spectroscopy for determining major-and trace-element concentrations in soils from a geochemical survey of North America. Appl. Geochem., 24(8): 1472-1481.

Reeves III, J.B., 2010. Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: where are we and what needs to be done? Geoderma, 158(1-2): 3-14.

Rossel, R.A.V., D.J.J. Walvoort, A.B. McBratney, L.J. Janik and J.O. Skjemstad, 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma, 131(1-2): 59-75.

Sun, T., W.L. Xu, J.L. Lin, M.H. Liu and X.W. He, 2012. Determination of soluble solids content in navel orange by Vis/NIR diffuse transmission spectra combined with CARS method. Spectrosc. Spect. Anal., 32(12): 3229-3233.

Wang, H., C. Li and M. Wang, 2013. Quantitative determination of onion internal quality using reflectance, interactance and transmittance modes of hyperspectral image. T. ASABE, 56(4): 1623-1635.

Wang, M., X. Xie, R. Zhou *et al.*, 2011. Determination of soil organic matter in red soils using VIS-NIR diffuse reflectance spectroscopy selection of optimal spectral bands. Acta Pedol. Sinica, 48(5): 1083-1089.

Yu, Y., 2011. Quantitative determination of parameters of substrate using near-infrared spectroscopy technique. Spectrosc. Spect. Anal., 31(11): 2928-2931.