ISSN : 0974-7419 Volume 12 Issues 6

Trade Science Inc.

Analytical CHEMISTRY

An Indian Journal

Full Paper

ACAIJ, 12(6) 2013 [214-221]

Application of successive projections algorithm on spectral monitoring of rice leaves nitrogen contents

Ming-Bo Liu, Yan-Lin Tang*, Xiao-Li Li, Jia Lou

College of Sciences, Guizhou University, Guiyang 550025, (CHINA)

E-mail : tylgzu@163.com; 0315-7243468@163.com

ABSTRACT

Visible-NIR reflective spectrum was used to predict the nitrogen contents of rice leaves. Different preprocessing methods were used in pretreatment of the original spectra. The effective wavelengths were selected by successive projections algorithm (SPA) for original spectra and pretreated spectra. Multiple linear regression (MLR) models and Partial least squares regression (PLS) models were built respectively. SPA could reduce the dimensions of spectral matrix efficiently. In the models established on SPA effective wavelength, MLR model and PLS model based on multiplicative scatter correction (MSC) pretreated spectrum had the best predicting effect with r=0.7943 and *RMSE*=0.4558. In PLS models established on all wavelengths, the best predicting effect model was that based on MSC pretreated spectrum with r=0.8470 and *RMSE*=0.3953.

KEYWORDS

Rice leaf;
Nitrogen content;
Visible-NIR spectrum;
Successive projections algorithm;
Effective wavelengths.

INTRODUCTION

Rice is one of the staple foods around the world, especially in the Asian regions. The monitoring of rice growth is important agricultural information. Nutrients play an important role on rice growth, deficiency of nutrients will lead to short and weak bodies, influence the color, shape and structure of leaves, and then influence rice growth and yield[1]. For those reasons, the monitoring of rice leaves nitrogen content becomes an important aspect in the monitoring of rice growth. The traditional method for nitrogen determination is Kjeldahl method. Its advantage is the high accuracy, and disadvantages are the complexity of processing, time consuming, vulnerability of instrument, the high level of professional requirements to the operator, and the high cost for large number of samples. Spectral analysis is a high-speed, low cost, and nondestructive analytical method. It could finish the determination of large number of samples in a short time. In addition, it could simultaneously generate the content information of a number of ingredients in samples during a determination, and has obvious advantages compared with traditional methods[2-7]. Successive projections algorithm (SPA) is a forward variable selection algorithm that minimize the colinearity of vector space. Its advantage lies in extracting a few characteristic wavelengths and eliminating the redundant information in original spectral data matrices. It can be used for the selection of characteristic wavelengths[8-13].

In recent years, some researchers have used SPA for the selection of effective wavelengths when using

spectral analysis method for the determination of certain important components in crops and food. There are researches about the application of SPA to nondestructive determination of pork pH value[8], and application of SPA in nondestructive determination of total amino acids in oilseed rape leaves[9]. Some researchers exploring near and mid-infrared spectroscopy to predict trace iron and zinc contents in powdered milk combined with SPA for effective wavelengths selection[10], and application of SPA as a variable selection in a QSPR study to predict the octanol/water partition coefficients (K_{OW}) of some halogenated organic compounds[11]. However, there are few researches about the application of SPA on spectral monitoring of rice leaves nitrogen contents.

The objective of this study was to investigate the application of SPA on spectral monitoring of rice leaves nitrogen contents. 5 segments moving average, baseline correction, area normalization, 1st derivative, multiplicative scatter correction (MSC) was used to preprocess the original reflectance spectrums. Multiple linear regressions (MLR) and partial least squares regressions (PLS) were applied to establish regression models of rice leave nitrogen contents.

MATERIALS AND METHODS

Design of experiments

Experimental field is located in Guizhou University at 26° 342 N, 104° 342 E, and covers an area of 420.48 m^2, it was divided into 15 equal regions, which were separated by balk covered with plastic film. Two varieties of rice were planted in every region, 0.3 m wide space was leaved to separate those two varieties, inter plant distance is 0.13 m × 0.17 m, transplanting rice shoots individually, 30 sample points, 3 repetitions, and completely random arrangement. The varieties are local glutinous rice (glutinous rice, full growth period 140 days, mark as S1) and Xiang You 109 (hybrid indica rice, full growth period 145 days, mark as S2). It was seeding on April 28 and transplanting on June 2. Fertilizer was designed to 5 levels, which are 0, 120, 240, 360, 480 kg N ha^{-1} respectively, covered situations of famine, lack, suitable, over dose, and severely over dose. Calcium superphosphate was used 533.3 kg ha^{-1} in each region as basic fertilizer. Potash muriate was used 300 kg ha^{-1} as earring fertilizer. Experimental soil is sand loam. Total soil nitrogen 1.15 g kg^{-1}, available nitrogen 188.5 mg kg^{-1}, total phosphorus 1.21 g kg^{-1}, total potassium 72.7 mg kg^{-1}, organic matter 9.96 g kg^{-1}, pH 6.78.

Spectral measurements

Collect samples of each variety in every region at tillering, jointing, boot, heading and milk stage of rice growth period. Reflectance spectra of the first and third leaf of rice plant were measured in doors with an AvaSpec-2048 Visible-NIR spectrophotometer (Avantes, Netherlands, scanning region 332-1100 nm, scanning intervals 2.4 nm). Before starting measurements, zero reflectance was adjusted by blocking the light path, and a standard whiteboard was used for 100% reflectance (so the spectrum obtained is dimensionless relative reflectance). 10 spectra were obtained each sample, then they were averaged into one sample spectrum, 300 sample spectrums were obtained in the end[14, 15].

Measurement of nitrogen contents of rice leaves

After measurement of spectra, the leaves were left into paper bags and killed out in an oven at 105°C for half hours, then dried at 70°C for 36 hours. Semi-micro Kjeldahl method was carried out to the dried leaves for nitrogen content determination.

Dividing of train set and prediction set

Combine reflectance spectra data with nitrogen content data respectively. Afterward, 15 abnormal samples were eliminated. The remaining 285 samples were used for spectral analysis. Those 285 samples were randomly divided into train set and prediction set. The training set had 190 samples, and prediction set 95 samples.

Spectral preprocessing

To minimize the errors come from environment and operating process, spectral preprocessing was taken. We use 5 segments moving average, baseline correction, area normalization, 1st derivative and multiplicative scatter correction (MSC) to preprocess the original spectral data[16]. Their model performances and effective wavelengths selected by SPA were compared.

Full Paper

Variable selection by SPA

SPA is a forward variable selection algorithm that minimizes colinearity of vector space. It was first posed by Bregman in 1965 to solve the convex feasibility problem, and is now widely used in the fields of biomedical imaging, computer tomography, signal processing and spectrum metrology[8]. SPA includes three phases[11-13]:

At first, the algorithm builds candidate subsets of variables on the basis of a colinearity minimization criterion. Those steps are described below, assuming that the first variable, $k(0)$, and the number N are given.

Step 1: Before the first iteration (n=1), let $X_j = j$ th column of X_{cal}; $j = 1, \ldots, J$

Step 2: Let S be the set of variables which have not been selected yet. That is: $S = \{j$ such that $1 \leq j \leq J$ and $j \notin \{k(0), \ldots, k(n\text{-}1)\}\}$

Step 3: Calculate the projection of X_j on the subspace orthogonal to $X_{k(n\text{-}1)}$ as: $Px_j = x_j - (x^T_j x_{k(n\text{-}1)}) x_{k(n\text{-}1)} (x^T_{k(n\text{-}1)} x_{k(n\text{-}1)})^{-1}$ for all $j \in S$, where P is the projection operator.

Step 4: Let $k(n) = \arg(\max\|Px_j\|, j \in S)$

Step 5: Let $x_j = Px_j$, $j \in S$

Step 6: Let $n = n + 1$. If $n < N$ go back to step 2.

End: The resulting variables are $\{k(n); n = 0,..., N\text{-}1\}$

The number of projection operations performed in the selection process can be shown to be (N-1) (J-N/2).

Then the best candidate subset is chosen based on minimum root mean square error (*RMSE*) obtained through a validation procedure.

Finally, the selected subset is subjected to an elimination procedure to determine if any variables can be removed without significant loss of prediction ability.

SPA was realized by MATLAB 7.0 (The Math Works, Natick, United States).

Modeling by MLR and PLS

Multiple linear regressions (MLR) are a widely used regression method, which is simple and easy to understand. But this method is usually interrupted by the colinearity between variables, and the number of input variables should be larger than the response chemical variable number and less than the sample number[17]. Normally, the spectral matrix has thousands of wavelength varieties. There is severe colinearity among wavelength varieties, and the number of wavelength varieties is far more than the number of samples. Therefore, for MLR modeling, selection of effective wavelengths is necessity.

Partial least squares regression (PLS) is a kind of multivariable correction method based on factor analysis which is mostly used multivariable correction method in spectral analysis now. PLS reduce the dimensions of spectral matrix and takes the influences of properties matrix into account at the same time. The decomposition of spectrum and contents is implemented simultaneously. The content information is imported into the spectral data decomposition process. Scores of spectra and contents are exchanged before the computation of every new principal component. In this way, relevance was built between spectral principal components and the content of interested components in samples. PLS was applied in chemical research from 1980s, and now become one of the most popular multivariable correction methods in chemometrics[17, 18].

Cross validation is a statistical analysis approach used to validate the regression model. Its basic idea is to divide the train set into k parts, and develop regression model using k-1 parts, the remaining one is used as validation set to validate the performance of the model built previously. This process is implemented k times until every set has been used as validation set. Finally, statistical results of all k validations are used as the assessment of original model. Leave-one-out cross validation is one of them. In the validation process, every sample will be used as validation set for one time. Its advantages are that most samples are used in regression each loop, which is close to the sample distribution of whole train set, and the validation result is more reliable, especially in the case of small amount of samples. In our research, leave-one-out cross validation is used to validate the regression model developed by train set.

The performances of regression models are evaluated through the computation of correlation coefficients r and root mean square error (*RMSE*).

RESULTS AND DISCUSSION

Statistical description of sample nitrogen content

The statistical description of nitrogen content in train, prediction, and all samples sets are shown in TABLE 1. The ranges of nitrogen content of all samples are 0.9848-4.8940%, statistical parameters of train set are the same

or very close to the statistical parameters of all samples, that indicates a good representativeness of train set. As shown in Figure 1, frequency distributions of train set and prediction set are almost the same, which indicates that all samples are randomly divided into train set and predicting set, which is very important for subsequent regression model establishing.

TABLE 1 : Statistical description of samples nitrogen content

Data Set	Number Of Samples	Range (%)	Mean (%)	Variance
Train	190	0.9848 ~ 4.8940	3.2849	0.5237
Prediction	95	1.1428 ~ 4.8742	3.3015	0.5567
All	285	0.9848 ~ 4.8940	3.2904	0.5328

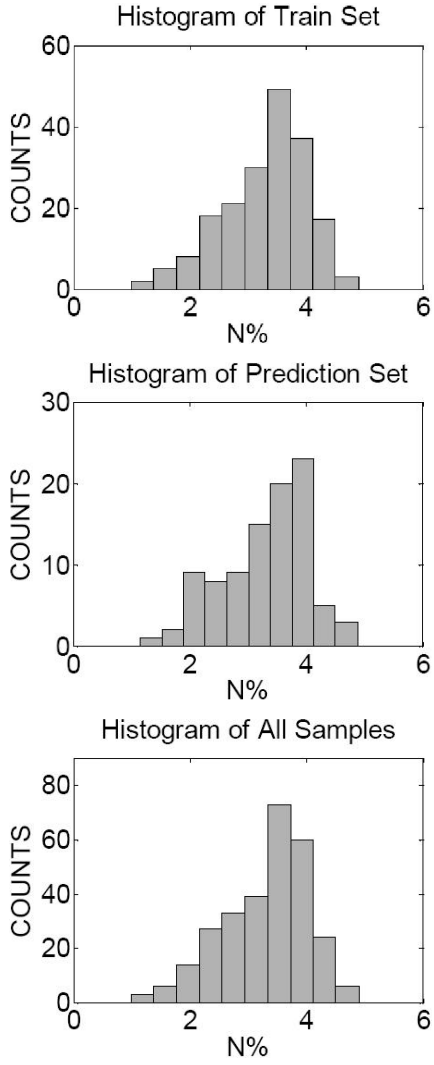

Figure 1 : Histogram of nitrogen content in train set, prediction set, and all samples

Results of spectral preprocessing and effective wavelength selected by SPA

Original spectra and preprocessed spectra are shown in Figure 2. As indicated in original spectra, the reflectance spectra of rice leaves have the general spectral characteristics of most green plants in this region of wavelength. In between 400 nm and 700 nm waveband, the reflectance is lower than 30%. There is a strong peak near 550 nm. The reflectance has a sharp increase in wave range of 680-760 nm, and becomes gradual in the infrared region. At the ends of spectra (<400 nm and >1000 nm), the intensity of light source becomes weak, so the noises become obvious. As indicated in preprocessed spectra, there are not many differences between 5 segments moving average spectra, baseline correction spectra and original spectra. In area normalization spectra, the reflectance around 550 nm is amplified, and the reflectance in infrared region is compressed. The 1st derivative spectra emphasized the differences in the wave range of 680-760 nm between samples, but amplified the noises at the same time. Spectra preprocessed by MSC are a lot of different from the original spectra, the differences between samples in infrared wave range become small, but they are large in visible wave range.

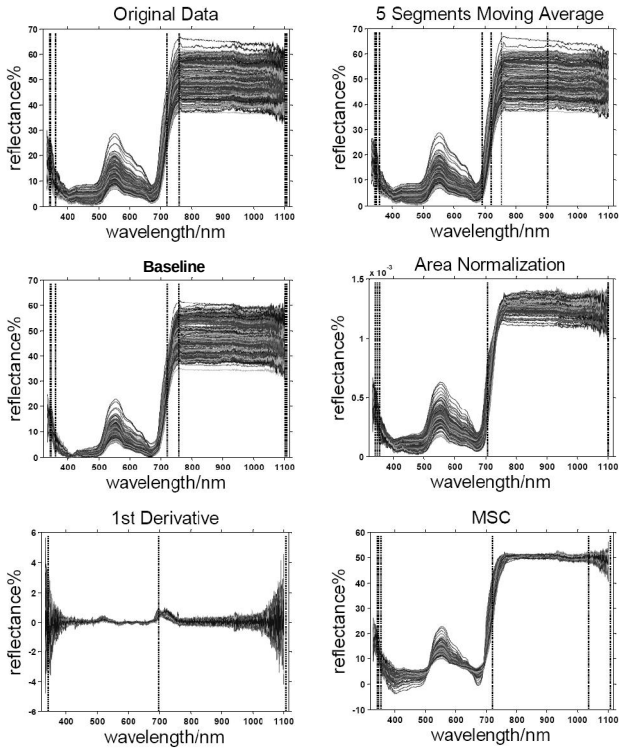

Figure 2 : Preprocessed spectra and effective wavelengths selected by SPA

Full Paper

TABLE 2 : Effective wavelengths selected by SPA

Preprocess Method	Wavelengths Selected by SPA								
Original Data	336	337	354	714	752	1095	1096	1100	
5 Segments Moving Average	338	341	352	684	713	745	897		
Baseline	336	337	354	714	752	1095	1096	1098	1100
Area Normalization	333	342	349	699	1091	1094			
1st Derivative	336	690	1100						
MSC	337	342	349	713	1029	1100			

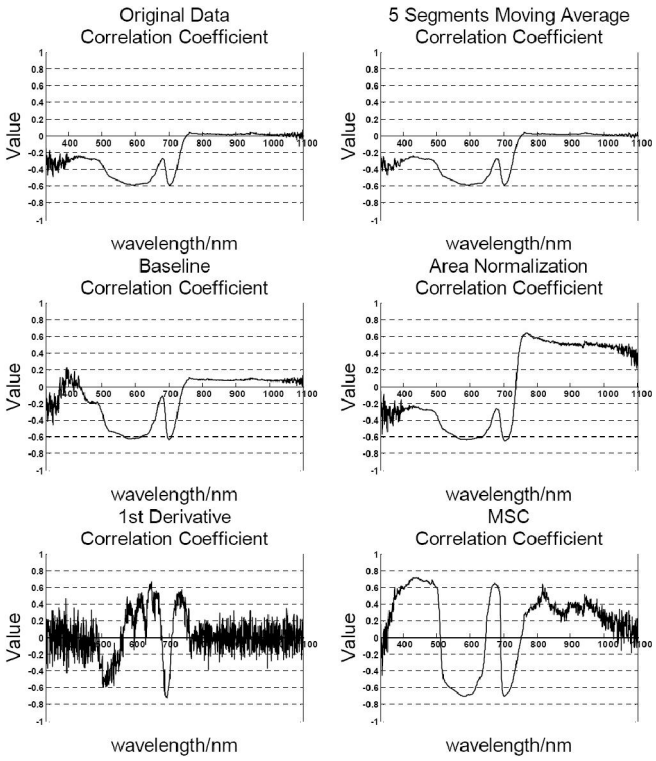

Figure 3 : Correlations between nitrogen content and different preprocessed spectra in train set

Dotted lines in spectra of Figure 2 show positions of effective wavelengths selected by SPA. As indicated in the figure and TABLE 2, although the positions of effective wavelengths are different in different preprocessed spectra, their similarities are distinct. In original and preprocessed spectra, five of them select wavelengths around 350 nm, Spectra preprocessed by 5 segment moving average, area normalization and 1st derivative select wavelengths around 690 nm, original spectra and spectra preprocessed by 5 segment moving average, baseline correction and MSC select wavelengths at 713 nm or 714 nm, original spectra and spectra preprocessed by 5 segment moving average and baseline correction select wavelengths around 750 nm. The selected wavelengths ahead of 342 nm and after 1090 nm can be attributed to the algorithm itself, for the edges of spectra may be liable to be selected, which can be deleted in future applications. In general, the positions of effective wavelengths selected by SPA are basically concentrated on four regions: 350 nm, 690 nm, 713 nm and 750 nm. Those regions are frequently used in the construction of plant spectral indexes[19, 20].

Correlations between nitrogen content and different preprocessed spectra in train set are shown in Figure 3. The positions of peaks and valleys in the figure indicate good relations between nitrogen content and corresponding wavelengths. Compare Figure 3 against Figure 2, it can be seen that effective wavelengths selected by SPA have good correlations with nitrogen content, which prove the effectiveness of SPA to some extent.

Results of MLR and PLS models

Results of MLR and PLS models are shown in TABLE 3. SPA-MLR and SPA-PLS are MLR and PLS models based on effective wavelengths selected by SPA. For original and all preprocessed spectra, SPA-MLR calibration performances are better than SPA-PLS. The best calibration performance of SPA-MLR models comes from the model based on baseline correction with r=0.8356 and *RMSE*=0.3965. The best calibration performance of SPA-PLS models comes from the model based on baseline correction with r=0.8348 and *RMSE*=0.3974. The best cross validation performance of SPA-MLR models comes from model based on baseline correction with r=0.8152 and *RMSE*=0.4184. The best cross validation performance of SPA-PLS models comes from model based on baseline correction with r=0.8155 and *RMSE*=0.4180. The cross validation performances of SPA-PLS models based on original and baseline correction spectra are better than those of SPA-MLR models. But comparing the prediction performances of SPA-MLR and SPA-PLS models, it can be seen that PLS models perform better than MLR models. The best prediction performances in both SPA-MLR and SPA-PLS models come from spectra preprocessed by MSC with r=0.7943 and *RMSE*=0.4558. Full Spectrum-PLS means the PLS models based on all wavelengths, its performances of calibration, cross validation, and prediction are better than the two kinds of models based on SPA effective wavelengths. 1st derivative spectra has

the best calibration performance with *r*=0.9476 and *RMSE*=0.2305. Spectra preprocessed by area normalization has the best cross validation performance with *r*=0.8655 and *RMSE*=0.3619. MSC spectra has the best prediction performance with *r*=0.8470 and *RMSE*=0.3953. Although the performances of SPA models are worse than the performances of models based on all wavelengths, SPA models are still valuable for the number of SPA effective wavelengths are merely about 1/200 of all 1335 data collection points, which reduce the dimensions of spectra matrixes effectively.

TABLE 3 : Results of MLR and PLS models

	SPA-MLR					
	Calibration		Validation		Prediction	
	r	RMSE	r	RMSE	r	RMSE
Original data	0.8082	0.4250	0.7874	0.4452	0.7020	0.5322
5 segments moving average	0.8108	0.4224	0.7912	0.4418	0.7761	0.4695
Baseline	0.8356	0.3965	0.8152	0.4184	0.7128	0.5267
Area normalization	0.7905	0.4420	0.7758	0.4556	0.7385	0.5031
1st derivative	0.7862	0.4460	0.7760	0.4554	0.6766	0.5491
MSC	0.8227	0.4103	0.8081	0.4254	0.7943	0.4558
	SPA-PLS					
	Calibration		Validation		Prediction	
	r	RMSE	r	RMSE	r	RMSE
Original data	0.8077	0.4255	0.7893	0.4434	0.7007	0.5334
5 segments moving average	0.8052	0.4281	0.7862	0.4463	0.7761	0.4695
Baseline	0.8348	0.3974	0.8155	0.4180	0.7154	0.5238
Area normalization	0.7892	0.4433	0.7740	0.4571	0.7445	0.4981
1st derivative	0.4329	0.6506	0.4043	0.6605	0.6766	0.5491
MSC	0.8129	0.4203	0.7974	0.4358	0.7943	0.4558
	Full Spectrum-PLS					
	Calibration		Validation		Prediction	
	r	RMSE	r	RMSE	r	RMSE
Original data	0.9039	0.3087	0.8562	0.3734	0.8249	0.4219
5 segments moving average	0.9126	0.2951	0.8464	0.3866	0.8080	0.4417
Baseline	0.9177	0.2867	0.8614	0.3676	0.8276	0.4180
Area normalization	0.9087	0.3013	0.8655	0.3619	0.8344	0.4095
1st derivative	0.9476	0.2305	0.8224	0.4137	0.8225	0.4223
MSC	0.9079	0.3026	0.8627	0.3656	0.8470	0.3953

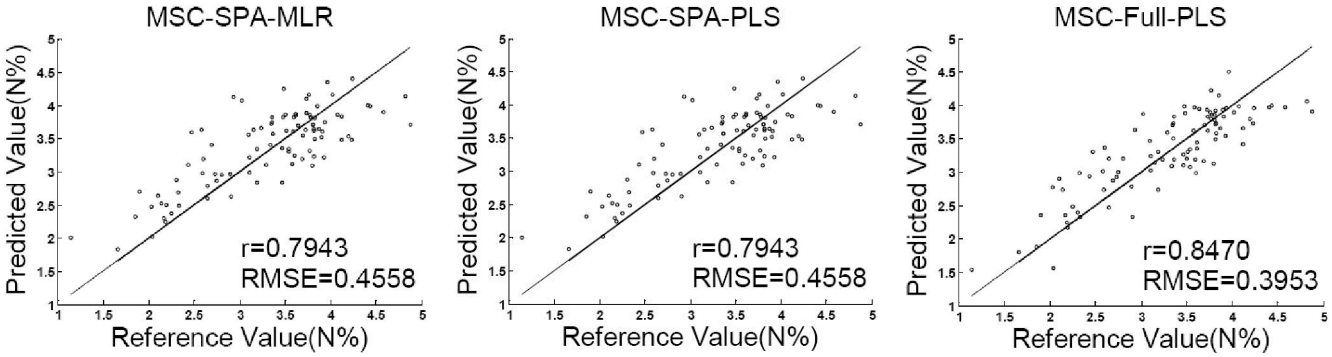

Figure 4 : Prediction performances of 3 models based on MSC spectra preprocessing

Comparing the regression performances of different preprocessing methods, MSC spectra have a good

performance. Figure 4 shows the scatter diagrams of reference value against predicted value of nitrogen content in prediction set.

CONCLUSIONS

The nitrogen content of rice leaves was successfully determined using Visible-NIR spectroscopy combined with SPA for effective wavelengths selection. Wavelengths selected by SPA in different preprocessed spectra mostly distributed near 350 nm, 690 nm, 713 nm, 750 nm, and there was good correlation between SPA effective wavelengths and nitrogen content. SPA could reduce the dimensions of spectral matrix efficiently, the results of MLR and PLS models based on SPA effective wavelengths were worse than that of PLS models based on all wavelengths. In models established on SPA effective wavelengths, MLR models and PLS models based on MSC pretreated spectrum had the best predicting effect with r=0.7943 and *RMSE*=0.4558. In PLS models established on all wavelengths, the best predicting effect model was that based on MSC preprocessed spectra with r=0.8470 and *RMSE*=0.3953.

ABBREVIATIONS USED

Visible-NIR, visible and near-infrared; SPA, successive projections algorithm; MSC, multiplicative scatter correction; MLR, multiple linear regression; PLS, partial least-squares; RMSE, root mean squares error.

REFERENCES

[1] K.Wang, Z.Q.Shen, R.C.Wang; Vegetation nutrient condition and spectral feature, Remote sensing for Land & resources, **39**, 9-14 **(1999)**.

[2] B.G.Osborne; Near-infrared spectroscopy in food analysis, in encyclopedia of analytical chemistry, R.A.Meyers (Ed); John Wiley & Sons Ltd, Chichester, 1-14 **(2000)**.

[3] A.Clément, M.Dorais, M.Vernon; Nondestructive measurement of fresh tomato lycopene content and other physicochemical characteristics using visible-NIR spectroscopy, J.Agric.Food Chem., **56**, 9813-9818 **(2008)**.

[4] A.Clément, M.Dorais, M.Vernon; Multivariate approach to the measurement of tomato maturity and gustatory attributes and their rapid assessment by Vis-NIR spectroscopy, J.Agric.Food Chem., **56**, 1538-1544 **(2008)**.

[5] S.E.Kays, F.E.Barton; Rapid prediction of gross energy and utilizable energy in cereal food products using near-infrared reflectance spectroscopy, J.Agric.Food Chem.,**50**, 1284-1289 **(2002)**.

[6] A.Lucas, D.Andueza, E.Rock, B.Martin; Prediction of dry matter, Fat, pH, Vitamins, Minerals, Carotenoids, Total antioxidant capacity, and color in fresh and freeze-dried cheeses by visible-near-infrared reflectance spectroscopy, J.Agric.Food Chem., **56**, 6801-6808 **(2008)**.

[7] M.Forouzangohar, D.Cozzolino, R.S.Kookana, R.J.Smernik, S.T.Forrester, D.J.Chittleborough; Direct comparison between visible near- and mid-infrared spectroscopy for describing diuron sorption in soils, Environ.Sci.Technol, **43**, 4049-4055 **(2009)**.

[8] Y.T.Liao, Y.X.Fan, F.Cheng, X.Q.Wu; Application of successive projections algorithm to nondestructive determination of pork pH value, Transactions of the CSAE, **26(1)**, 379-383 **(2010)**.

[9] F.Liu, F.Zhang, H.Fang, Z.L.Jin, W.J.Zhou, Y.He; Application of successive projections algorithm to nondestructive determination of total amino acids in oilseed rape leaves, Spectroscopy and spectral analysis, 29(11), 3079-3083 **(2009)**.

[10] D.Wu, Y.He, J.H.Shi, S.J.Feng. Exploring near and midinfrared spectroscopy to predict trace iron and zinc contents in powdered milk, J.Agric.Food Chem., 57, 1697-1704 **(2009)**.

[11] N.Goudarzi, M.Goodarzi; Application of successive projections algorithm (SPA) as a variable selection in a QSPR study to predict the octanol. Water partition coefficients (KOW) of some halogenated organic compounds, Anal.Methods, **2**, 758-764 **(2010)**.

[12] M.C.U.Araújo, T.C.B.Saldanha, R.K.H.Galvão, T.Yoneyama, H.C.Chame, V.Visani; The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom.Intell.Lab.Syst., **57**, 65-73 **(2001)**.

[13] R.K.H.Galvão, M.C.U.Araújo, W.D.Fragoso, E.C.Silva, G.E.José, S.F.C.Soares, H.M.Paiva; A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm, Chemom.Intell.Lab.Syst., **92**, 83-91 **(2008)**.

[14] J.Lou, Y.L.Tang, S.H.Cai, J.F.Huang; Correlation

between transmittance spectra and nitrogen content for rice, Chinese Agricultural Science Bulletin, **25(24)**, 544-548 **(2009)**.

[15] J.Lou; Nitrogen nutrition monitoring and diagnosis of rice using spectral analysis method. Master dissertation of guizhou university, Guiyang, China, **(2010)**.

[16] F.Liu, Y.He, G.M.Sun; Determination of protein content of auricularia auricula using near infrared spectroscopy combined with linear and nonlinear calibrations, J.Agric.Food Chem., **57**, 4520-4527 **(2009)**.

[17] J.X.Liu; Applied near infrared spectrometry. Science press, Beijing, China, 100, 131 **(2008)**.

[18] J.L.Rodriguez-Otero, M.Hermida, J.Centeno; Analysis of dairy products by near-infrared spectroscopy: A Review, J.Agric.Food Chem., **45(8)**, 2815-2819 **(1997)**.

[19] C.Y.Yan, Z.Niu, J.H.Wang, L.Y.Liu, W.J.Huang; The assessment of spectral indices applied in chlorophyll content retrieval and a modified crop canopy chlorophyll content retrieval model, Journal of Remote Sensing, **9(6)**, 742-750 **(2005)**.

[20] Y.X.Li, Y.Zhu, Y.C.Tian, X.T.You, D.Q.Zhou, W.X.Cao; Relationship of grain protein content and relevant quality traits to canopy reflectance spectra in wheat, Scientia agricultura sinica, **38(7)**, 1332-1338 **(2005)**.