Nitrogen content diagnosis of apple trees canopies using hyperspectral reflectance combined with PLS variable extraction and extreme learning machine

Shaomin Chen[1], Lihui Ma[2], Tiantian Hu[1*], Lihua Luo[1], Qiong He[1], Shaowu Zhang[1]

(1. *Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas of Ministry of Education, Northwest A&F University, Yangling 712100, Shaanxi, China*;
2. *Institute of Soil and Water Conservation, Northwest A&F University Yangling 712100, Shaanxi, China*)

Abstract: Nitrogen (N) is an important mineral element in apple production. Rapid estimation of apple tree N status is helpful for achieving precise N management. The objective of this work was to explore partial least squares (PLS) regression in dimensional reduction of spectral data and build the diagnostic model. The spectral reflectance data were collected from Fuji apple trees with 4 levels of N fertilizer treatment in the Loess Plateau in 2018 and 2019 using an ASD portable spectroradiometer, and leaf total N content was obtained at the same time. The raw spectra were pretreated using Savitzky-Golay (SG) smoothing and a combination of SG and first-order derivative (SG_FD) or second-order derivative (SG_SD). The samples were divided into a calibration dataset and a prediction dataset using SPXY. Based on 4 factors of PLS regression, including latent variables (LVs), *X*-loading, variable importance in projection (VIP) and regression coefficients (RC), the 6 methods (LVs, *X*-loading, VIP_01, VIP_02, RC_01 and RC_02) were derived and used for variable extraction, based on which PLS model and ELM model were established. The results indicated that the spectral data processed by SG_FD had the highest signal-to-noise ratio and was selected for subsequent analysis. The amounts of variables extracted by LVs, *X*-loading, VIP_01, VIP_02, RC_01 and RC_02 were 6, 11, 18, 305, 26 and 88, respectively. The method of extracting variables with an RC threshold based on the minimum RMSEP (RC_02) could effectively avoid the omission of effective information. The RC_02 method was recommended for related research which required accurate wavelength information as a variable. The variable extraction method based on LVs generated an ELM model with a simple structure. The prediction results showed that the ELM model outperformed the PLS model. The PLS(LVs)_ELM model was the best; R^2_P, RMSEP and RPD were 0.837, 2.393 and 2.220, respectively.

Keywords: partial least square, variable extraction method, extreme learning machine, hyperspectral reflectance, apple tree, canopy nitrogen content
DOI: 10.25165/j.ijabe.20211403.6157

Citation: Chen S M, Ma L H, Hu T T, Luo L H, He Q, Zhang S W. Nitrogen content diagnosis of apple trees canopies using hyperspectral reflectance combined with PLS variable extraction and extreme learning machine. Int J Agric & Biol Eng, 2021; 14(3): 181–188.

1 Introduction

Apple production is one of the important industries in terms of economic income in China. The Loess Plateau is the largest advantageous apple production area in China and the world, accounting for 51.52% and 47.14% of the domestic planting area and fresh fruit production, respectively[1]. Previous studies have shown that nitrogen (N) is closely related to apple quality and yield[2]. The traditional laboratory method had poor timeliness in diagnosing N nutrition of plant. At present, Fast, non-destructive and accurate N nutrition diagnosis of crops is one effective method for N management, based on hyperspectral technology[3,4].

In the process of establishing diagnostic models of crop N nutrition, previous studies found that the sensitive bands of N content were greatly different under the influence of different crops[5-7]. With the development of artificial intelligence technology, CARS, UVE, GA, Random Frog and other methods were used to eliminate invalid variables and select sensitive wavelengths[8-10]. These methods were usually called computer-aided variable selection. However, these methods had a common feature, namely, multiple operations based on random sampling strategy[11], and their portability was limited. If the extracted results were missing the key information, the final calibration model might be damaged. To avoid such a situation, some scholars coupled different variable selection methods[8,12], which might increase the computational burden.

Partial least squares (PLS) regression is a standard calibration method for analyzing spectral data. The factors in the process of PLS regression analysis were used to select variables or features of spectral data[13,14]. Gao et al.[15] adopted latent variables (LVs) as

Received date: 2020-09-11 **Accepted date:** 2021-04-06
Biographies: Shaomin Chen, PhD candidate, research interest: agricultural engineering, Email: shaomin_ly24@126.com; **Lihui Ma**, PhD, Associate Researcher, research interest: efficient utilization of water and soil resources, Email: gjzmlh@126.com; **Lihua Luo**, MS candidate, research interest: agricultural engineering, Email: llh@nwafu.edu.cn; **Qiong He**, MS, research interest: efficient utilization of water and fertilizer resources, Email: 1244566299@qq.com; **Shaowu Zhang**, PhD candidate, research interest: efficient utilization of water and fertilizer resources, Email: zhangshaowu@nwsuaf.edu.cn.
***Corresponding: Tiantian Hu**, PhD, Professor, research interest: plant nutrients, precision agriculture. Northwest A&F University, No.23, Weihui Road, Yangling 712100, Shaanxi, China. Tel: +86-29-87082901, Email: hutiant@nwsuaf.edu.cn.

the input for machine learning (BPANN and SVM) and established a model to qualitatively identify the different fruit waxes on an apple's surface. Zhang et al.[16] used LVs as the input for machine learning (SVM) to conduct a quantitative study on the adulteration rate of edible gelatin. Ye et al.[17] used wavelength information corresponding to the crest or trough of the *X*-loading or variable importance in projection (VIP) curve to predict the N content of apple leaves. Cheng et al.[18] used the threshold method to select wavelength information based on the *X*-loading or regression coefficient (RC) curve, and used the wavelength information as the input for machine learning (SVM) to qualitatively identify liquor quality. Zhang[19] predicted soil organic matter content based on LVs and the corresponding wavelength information of the RC wave peak or trough. Obviously, the factors identified in the process of PLS regression can be used to reduce the dimensions of hyperspectral data, and the method can be transferred to other applications. Few studies, however, have focused on the systematic analysis of the 4 factors (LVs, *X*-loading, VIP and RC) applied in variable extraction. Moreover, some studies have shown the randomness of the application of some factors. For example, Wang et al.[20] used 0.02 as the threshold for selecting wavelength variables of the *X*-loading curve; Cheng et al.[18] used 1.6 and 0.2 as thresholds for the selection wavelength information of the RC and the *X*-loading curve, respectively. None of these studies provided a method to determine the threshold.

The aims of this study were (1) to extract variables of the canopy scale hyperspectral data from the spring-shoot-growing stage to the fruit enlargement stage using the 6 methods derived from 4 PLS-related factors; (2) to propose a method to determine the threshold value of RC based on minimum RMSEP and extract the key wavelengths; (3) to compare the performance of the PLS regression and extreme learning machine (ELM) model based on the variables from (1), so as to provide theoretical support for PLS assisted ELM to diagnose the apple tree canopies' N content.

2 Materials and methods

2.1 Experimental site

During the apple tree growing seasons from March 10 through September 15 of 2018 and March 13 through September 22 of 2019, a field experiment was conducted at the Luochuan Apple Experimental Station of Northwest A&F University (109°21′40″E, 35°47′8″N), Shaanxi Province, China. The site is located in the central part of the Loess Plateau apple production region and is one of the world's recognized advantageous apple-producing areas. This region has a warm temperate monsoon and semi-humid climate, with an average altitude of 1072 m, a mean annual precipitation of 610 mm, a mean average temperature of 9.2 ℃, annual sunshine 2525 h, a sunshine rate of 58%, 180 frost-free days, annual total radiation of 554 KJ/cm^2 and an accumulated temperature of 3040 ℃ (above 10 ℃). The soil in apple orchard is dark loessial soil.

2.2 Data acquisition

The apple trees were Fuji cultivar planted in 2012. The planting mode was high density dwarfing with the row-by-stand spacing of 4 m by 2 m. The apple trees were spindle-shaped with a height of 4 m. The 4 levels of N (0 kg/hm^2, 120 kg/hm^2, 240 kg/hm^2, and 360 kg/hm^2) were used in this study and each treatment replicated two times, resulting in 8 plots in total. There were 15 apple trees in each plot, and there was a row of trees between each plot as a buffer. There were three apple trees marked for measurement in each plot, so a total of 24 trees were measured each time. N fertilizer (urea) was supplied by fertigation according to the schedule in Table 1.

Table 1 Nitrogen fertilizer schedule

Growth stage	Fertilizer time	Fertilizer amount/kg hm^{-2}			
		0	120	240	360
Base fertilizer	Late September	0	36	72	108
Germination stage	Mid-March	0	12	24	36
Blooming stage	Early April	0	24	48	72
Spring-shoot-growing stage	Early May	0	24	48	72
Young fruit stage	Early June	0	12	24	36
Early stage of fruit enlargement	Late June	0	6	12	18
Middle stage of fruit Enlargement	Mid-July	0	6	12	18

Note: The base fertilizer was applied after the previous year's apple harvest, usually at the end of September.

2.2.1 Remote sensing data

Field reflectance spectra were measured over apple tree canopies using an ASD FieldSpec 3 portable spectroradiometer (Analytical Spectral Devices, Inc., St. Boulder, Co., USA) with a wavelength range of 350-1830 nm. The portable spectroradiometer was equipped with optical fiber with a length of 1.5 m and 25 ° field-of-view (FOV). The sampling intervals were 1.4 nm @ 350-1000 nm and 2 nm @ 1000-1830 nm, finally the output data was given in 1-nm interval by instrument automatically interpolated. All measurements were made under cloud-free or near cloud-free conditions between 10:00 and 14:00 local time with the help of a special platform[21]. Measurements were made about 1 m above the canopy top with the sensor pointing vertically downward. The FOV covered the projection area of the canopy. Ten scans were made of each sample tree after dark current reading and white reference panel calibration. Canopy scale hyperspectral information of apple trees was collected 4 times in 2018 and 5 times in 2019 from the spring-shoot growth stage to fruit enlargement stage, respectively, resulting in 216 samples were collected in total.

2.2.2 Canopy nitrogen content

More than 25 mature leaves (without petiole) were collected from a sample tree. These leaves, from the middle of the elongated branch in the upper and lower parts of the canopy, were free of pests and mechanical damage. The leaves were de-enzymed under 105 ℃ for 30 min and dried to a constant weight under 75 ℃, then used to measure total N using the Kjeldahl method[22].

2.3 Sample data division

Monte Carlo second detection method (MC2) was used to identify outliers in the dataset before the sample division. To obtain a representative prediction dataset and improve the prediction ability and accuracy of the model, an extensively used method, Sample set Partitioning based on joint X-Y distance (SPXY), was used to divide the samples[23]. The one-third samples were labeled as the prediction dataset and the remainders were used as the calibration dataset.

2.4 Spectral pretreatment

In the spectral curve, the ranges of 1350-1450 nm and 1801-1830 nm were excluded, due to the strong water absorption band near 1400 nm and the strong edge noise at 1830 nm[24,25]. So, this study focused on the ranges of 350-1349 nm and 1451-1800 nm. Hyperspectral data are affected by canopy geometry, soil cover, leaf water content and atmospheric absorption. In this study, to reduce various noises and improve the signal-noise ratio (SNR), a total of 4 methods, including raw spectrum (RS),

Savitzky-Golay (SG, with a square polynomial of seven spans) smoothing, a combination of SG and first-order derivative (SG_FD, gap=5) and a combination of SG and second-order derivative (SG_SD, gap=5), were used to preprocess the spectral data.

2.5 Variable extraction

Variable extraction is an effective method for dimensionality reduction of hyperspectral data, which can simplify the model and reduce the computational burden. PLS regression is a common calibration method for spectral data, which can effectively solve the multicollinearity problem between spectral variables[21]. In the PLS regression process, data dimensionality reduction, information integration and screening techniques are adopted to extract the latent variables (LVs) with the best interpretation ability for the model. The brief PLS model is as follows:

$$X=TP'+E,\ Y=UQ'+F,\ U=\beta T \quad (1)$$

where, X and Y are the predictors matrix and responses matrix; T and U are the X-scores and Y-scores matrices; P and Q are the X-loading and Y-loading matrices; E and F are the residual matrices for X and Y; β is the regression coefficient matrix.

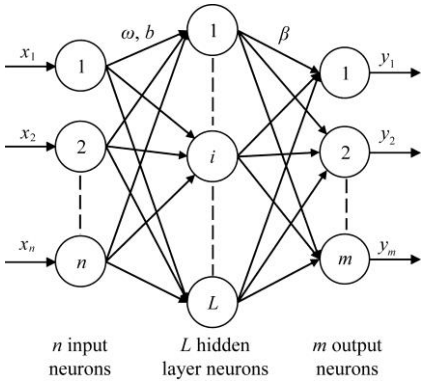
In this study, the 6 methods were derived from 4 PLS-related factors (X-loading, RC, VIP and LVs). In addition to the LVs method, the other methods were defined as follows. Extraction of the corresponding wavelengths based on the peak or trough of the X-loading curve (X-loading). Two methods were derived from VIP, namely extraction of the corresponding wavelengths based on peak or trough of VIP curve (VIP_01), and extraction of the wavelengths which VIP values were greater than 1 (VIP_02). Two methods were derived from RC, namely extraction of the corresponding wavelengths based on peak or trough of RC curve (RC_01), and extraction of the wavelength based on a threshold (RC_02).

One important issue in PLS regression is to determine the number of LVs, so that the best predictive ability of the developed model could be achieved. An insufficient number of LVs would result in lower prediction accuracy. Too many LVs will lead to an over-fitting of the PLS model. In this work, to avoid the impact of redundant information on the model, the optimal number of LVs was determined by both the internal and external validation[16]. The internal validation was used to obtain the first minimum of the root mean square error of 10-fold cross validation[14] (RMSECV), and the external validation was used to achieve the first minimum of the root mean square error of prediction (RMSEP).

2.6 Modeling method

In this work, linear model (PLS) and nonlinear model (ELM) were used to build the calibration model.

Extreme learning machine (ELM) was first proposed by Huang et al[26]. As a specific type of single feed-forward neural network (SFFNN), ELM contains three layers (an input layer, one hidden layer and an output layer), as shown in Figure 1. The weights and biases between the input layer and the hidden layer nodes of ELM were randomly chosen and fixed by the continuous probability density function. The output weights were analytically determined using the Moore–Penrose generalized inverse. In this way, the ELM model has been proven to be effective for classification or regression tasks and has a higher generalization performance[27-29]. In this study, the default 'sigmoid' was adopted as the activation function of the hidden layer neurons. The number of hidden layer nodes (HLNs) within the range of [5, 100] gradually increased at an interval of 1. Each model was operated 5000 times, and the number of HLNs was determined by the optimal result trained.



Note: ω and b are the weight and bias vectors respectively; β is the weight vector.

Figure 1 Network structure of ELM

2.7 Evaluation metrics and software

Determination coefficients (R^2), root mean squared error of calibration (RMSEC) and prediction (RMSEP), and ratio of performance to deviation (RPD) were used to evaluate the performance of the models. Generally, the larger the R^2 and RPD, the smaller the RMSEC and RMSEP, indicating the superior performance of that model. In particular, when the RPD value is greater than 2.0, it means that the model could be used for quantitative analysis[30].

In this study, spectral data preprocessing, elimination of the outliers, sample division, variables extraction, model building and scientific drawing were conducted in MATLAB 2017a Professional version (The MathWorks, Natick, MA, USA.).

3 Results and discussion

3.1 Elimination of the outliers

The number of Monte Carlo cross validation (MCCV) sampling times and sampling ratio were set to 2500 and 0.8, respectively. According to the Monte Carlo random sampling algorithm and PLS regression, the mean (MEAN) and standard deviation (STD) of the predicted residuals for each sample were calculated. The scatter diagram of MEAN-STD was obtained (Figure 2). Outliers were defined as points that fell outside the threshold values of 2.5 times the MEAN and 2.5 times the STD denoted by the blue dotted line in Figure 2 below[14,31]. The samples numbered 178, 156, 142, 137, 116, 49 and 43, which clearly fell outside of the blue dotted line, were regarded as significant outliers. The samples marked in red in Figure 2 near the blue dotted line were regarded as suspicious outliers.

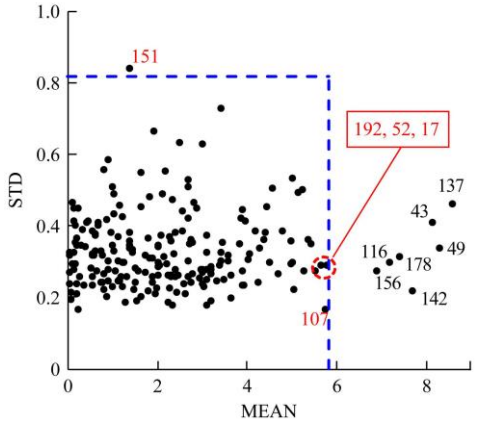


Figure 2 Results of outlier detection by MCCV

The 10-fold cross validation PLS regression model was established after eliminating the significant outliers. The

determination coefficients of cross-validation (R^2_{CV}) and RMSECV were obtained. The suspicious outliers were eliminated one by one, and the results are shown in Table 2. Obviously, as the suspicious outliers were removed one by one, R^2_{CV} gradually increased and RMSECV gradually decreased. A total of 12 outliers were removed, and 204 samples remained in the dataset.

Table 2 Process and results of outlier identification by Monte Carlo second detection method

Step name	Outliers	10-fold cross-validation	
		R^2_{CV}	RMSECV
Remove significant outliers	**178, 156, 142, 137, 116, 49, 43**	0.777	2.285
Second detection of suspicious samples	178, 156, 142, 137, 116, 49, 43, **192**	0.778	2.274
	178, 156, 142, 137, 116, 49, 43, 192, **151**	0.779	2.272
	178, 156, 142, 137, 116, 49, 43, 192, 151, **107**	0.784	2.250
	178, 156, 142, 137, 116, 49, 43, 192, 151, 107, **52**	0.786	2.223
	178, 156, 142, 137, 116, 49, 43, 192, 151, 107, 52, **17**	0.793	2.191

Note: The bold number is the newly removed abnormal sample in the current step.

3.2 Sample data division

In the process of dividing the samples, SPXY can consider both the predictors matrix (X) and responses matrix (Y) when calculating the distance between samples[32]. This method can effectively cover the multi-dimensional vector space and improve the prediction ability of the model. The one-third samples were separated out as the prediction dataset and the remainders were used as the calibration dataset. The descriptive statistics results are shown in Table 3.

Table 3 Descriptive statistics of canopy N content

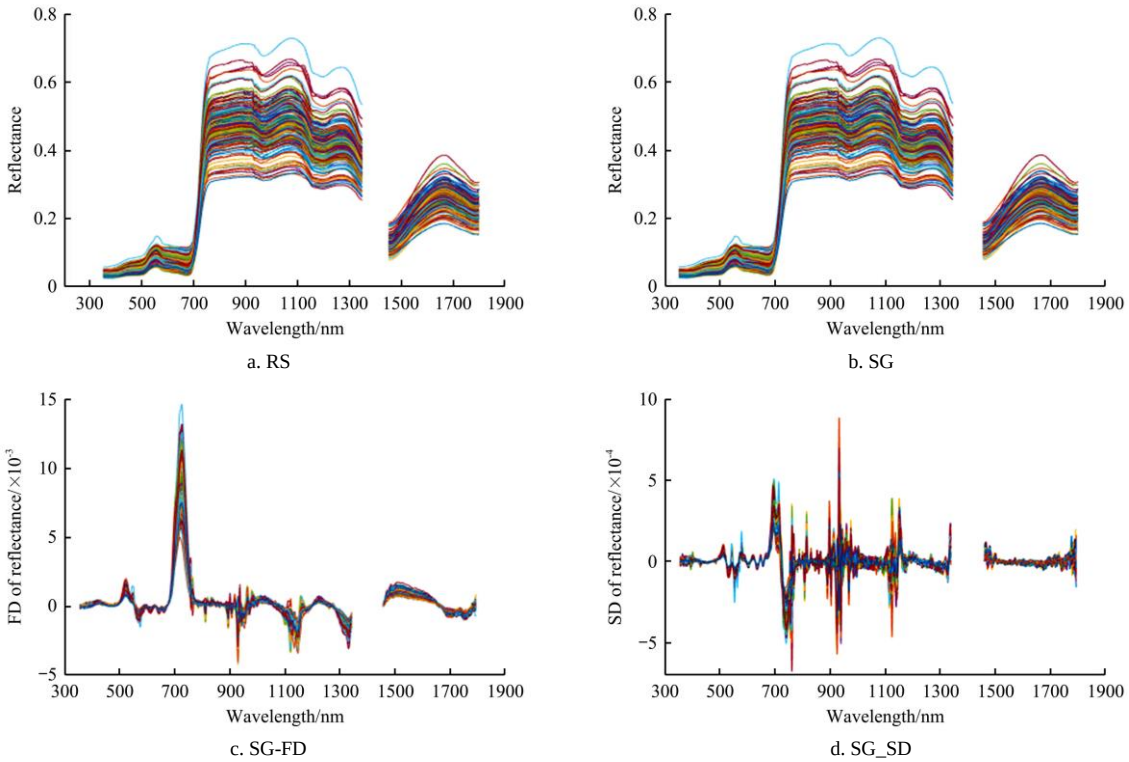
Dataset	Sample size	Min /g kg^{-1}	Max /g kg^{-1}	Average /g kg^{-1}	STD /g kg^{-1}
Calibration dataset	136	12.63	34.32	22.05	4.58
Prediction dataset	68	11.54	35.12	22.10	5.31

3.3 Screening of the optimal spectral preprocessing method

The preprocessed spectra are shown in Figure 3. Compared with the RS, the spectra pretreated with SG showed no obvious change; the slight changes in the spectra were magnified by SG_FD and SG_SD. PLS regression models of N content in apple tree canopies were established based on the spectral data after pretreatment (Table 4). Compared with RS, SG and SG_FD preprocessed spectra improved the model prediction ability, while SG_SD preprocessed spectra decreased the prediction ability. The prediction accuracy of the model was improved after simple smoothing (SG), which indicating that there was a little noise in the RS. The PLS regression model preprocessed with SG_FD spectra had the maximum R^2_P (0.791) and the minimum RMSEC (2.421), as well as the fewest LVs required for modeling (LVs=6). The results indicated that spectra preprocessed by SG_ FD enhanced the spectral SNR, so the spectral data processed by SG_FD, which contained 1326 band information, should be selected for the following study.

Table 4 PLS regression model of canopy N content based on different pretreatment methods

Pretreatment method	LVs	Calibration dataset		Prediction dataset	
		R^2_C	RMSEC	R^2_P	RMSEP
RS	10	0.801	1.902	0.768	2.888
SG	13	0.825	1.781	0.774	2.804
SG_FD	6	0.792	2.082	0.791	2.421
SG_SD	10	0.949	1.036	0.643	3.176

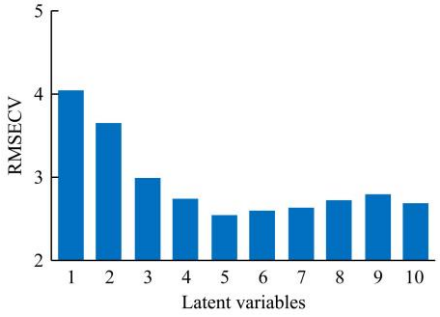


a. RS b. SG c. SG-FD d. SG_SD

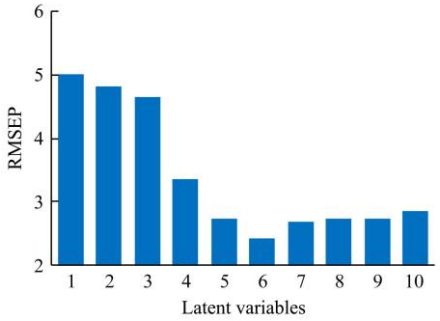
Figure 3 Spectra with different pretreatment methods

3.4 Variable extraction

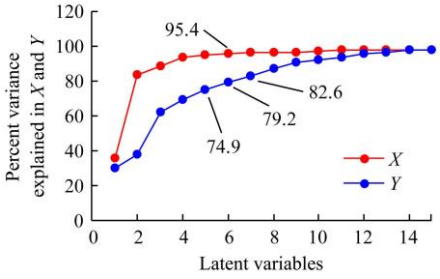
The number of LVs was determined using internal and external validation. The changes in RMSECV and RMSEP with the first ten LVs are illustrated in Figure 4a and Figure 4b, respectively. The RMSECV continuously decreased from a single latent variable until 5 LVs were included in the PLS regression model, when it reached the minimum value (Figure 4a). Similar to the changes in RMSECV, when 6 LVs were included in the PLS regression model, the RMSEP reached the minimum value (Figure 4b). The change in the cumulated variance explained by the first 15 LVs is shown in Figure 4c. The first 6 LVs explained 95.4% and 79.2% of variance in the *X* (SG-FD spectral data) and *Y* (N content) variables, respectively. Compared with the first 5 LVs, the relative increment of cumulative variance of the *Y* variable explained by the first 6 LVs is 5.74%. Similar, compared with the first 6 LVs, the relative increment of cumulative variance of the *Y* variable explained by the first 7 LVs is 4.29%. To avoid over-fitting caused by including redundant LVs, only when the relative increment of the cumulative variance explained is greater than 5%, the LVs are added to the PLS regression model. Therefore, it was reasonable to determine the optimal number of LVs as 6 in this study[16,33].



a. Changes in the RMSECV with the first 10 LVs



b. Changes in the RMSEP with the first 10 LVs



c. The accumulated variance explained in X and Y by the first 15 LVs

Figure 4 Determining the optimal number of LVs

In the preceding PLS regression, the *X*-loadings that indicated the importance of the wavelengths to the latent variables were calculated. The wavelengths with a higher value in the *X*-loadings of a latent variable could be considered more important than other wavelengths in contributing to the corresponding latent variable. There is a feature of PLS algorithm. The first few factors usually shows a high correlation between the *X*-scores and *Y*-scores, and the degree of correlation usually decreases from one factor to the next[34]. Therefore, only the *X*-loadings in the first few LVs need to be considered rather than those in all LVs. In this work, the corresponding *X*-loadings of the first 3 LVs were selected for analysis (Figure 5). As the number of LVs increased, the *X*-loading curve slightly shifted to the right-hand longer wavelengths. The key wavelengths were extracted based on the peak or trough (local maxima or minima) of LV1 curve in the *X*-loadings spectrum for the latent variable. This method extracted 11 key wavelengths, accounting for 0.83% of the SG_FD data band (Figure 8).

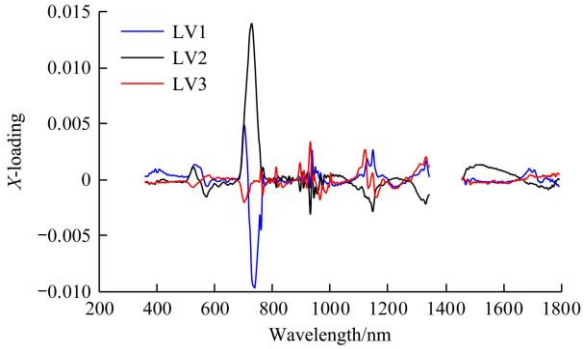


Figure 5 *X*-loadings of the first 3 LVs obtained by PLS regression

In the process of PLS regression, the VIP index was used to construct a new low-dimensional data space. The VIP value was related to the contribution of wavelength information and the calculation theory of VIP value was detailed introduce in References [35] and [36]. In general, a VIP value larger than 1 means that the contribution of the wavelength is significant, and the larger the VIP value, the greater the contribution of the wavelength to the model is. Therefore, the VIP value can be used to select the key wavelength. The key wavelength was extracted based on the peak or trough of the VIP curve in Figure 6 and marked as VIP_ 01. This method extracted 18 key wavelengths, accounting for 1.36% of the SG_FD data band (Figure 8).

On the other hand, the average squared VIP value equals 1. Therefore, 'the greater than one rule' was generally used as a criterion for variable selection[35]. Here, this method was marked as VIP_02 and a total of 305 key wavelengths were extracted, accounting for 22.98% of the SG_FD data band (Figure 8).

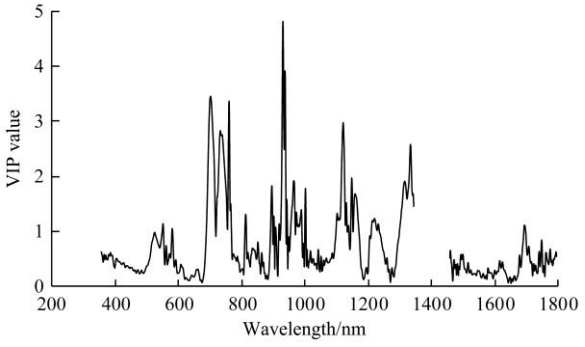
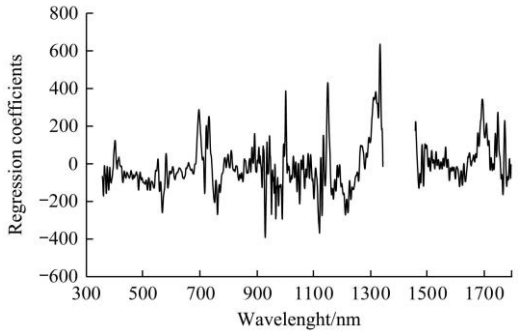


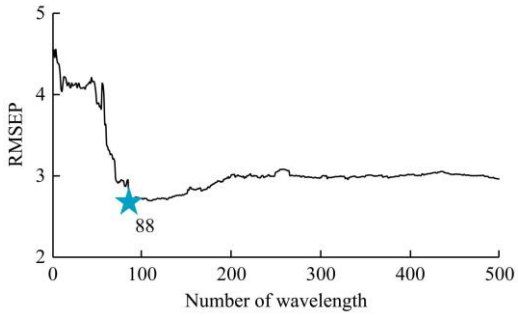
Figure 6 VIP value of PLS regression model

The regression coefficient (RC) curve was obtained by the preceding PLS regression (Figure 7a). The key wavelengths were extracted based on the rule that the information contained in the wavelength at the extreme value of the peak or trough is significantly related to the retrieval objects[19]. Here, this method

was marked as RC_01 and a total of 26 key wavelengths were extracted, accounting for 1.96% of the SG_FD data band (Figure 8).

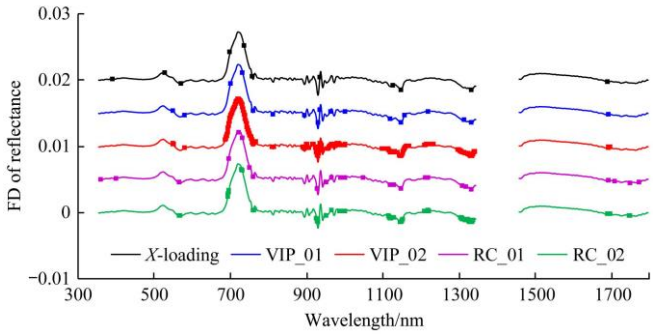


a. Regression coefficients of PLS



b. Determination of threshold

Figure 7 Key wavelengths selected by regression coefficients of PLS



Note: The first derivative spectra were gradually increased by 0.005 for clearer discrimination.

Figure 8 Wavelengths extracted by different methods

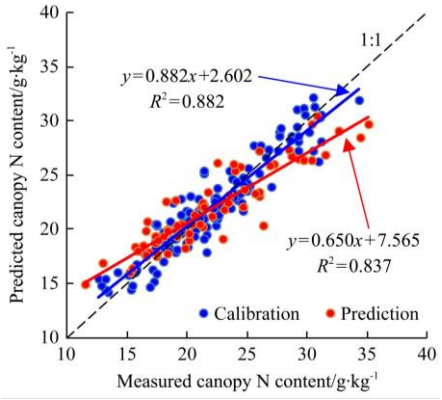
To propose a method to determine the threshold value of RC based on minimum RMSEP and avoid missing the effective wavelength, all wavelengths were arranged in descending order of RC value, and then added to PLS regression model in turn. Figure 7b shows the change in RMSEP with the number of selected wavelengths. When the top 88 wavelengths were chosen for the PLS regression model, the RMSEP reached the minimum value of 2.71 (marked as a blue asterisk in Figure 7b). The corresponding threshold of RC was 244.57. Here, this method was marked as RC_02 and a total of 88 key wavelengths were extracted, accounting for 6.63% of the SG_FD data band (Figure 8).

3.5 Model establishment and evaluation

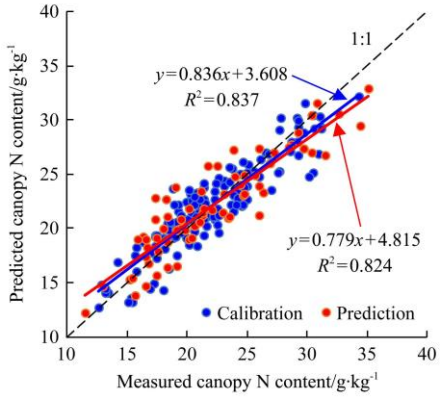
Variables in Section 3.4 extracted by the 6 methods were used to establish the PLS regression model and the ELM model, and the prediction results of the models are shown in Table 5. According to Table 5, the R^2 and RPD values indicate that the prediction accuracy and prediction ability of the ELM model are better than those of the PLS regression model. Based on the same variables, the ELM model can obtain a better prediction ability (RPD>2)[37]. The R^2_P of the ELM model established by the key wavelengths extracted by *X*-loading and VIP_01 were 0.782 and 0.784, respectively. Compared with the full-spectrum PLS regression model, the prediction accuracy was not improved. The R^2 and RPD of the ELM model established by the key wavelengths extracted by VIP_02 were 0.799 and 2.245, respectively. The prediction accuracy and prediction ability were similar to the full-spectrum with PLS regression model. The prediction accuracy and prediction ability were enhanced by the ELM model established by the variables extracted by LVs, RC_01 and RC_02. Compared with the full-spectrum PLS regression model, the R^2_P were 0.837, 0.823 and 0.824, an increased of 5.82%, 4.05% and 4.17%, respectively; the RPD were 2.22, 2.366 and 2.382, an increased of 1.19%, 7.84% and 8.57%, respectively; the RMSEP were 2.393, 2.245 and 2.230, a decrease of 1.16%, 7.27% and 7.89%, respectively. The ELM model established by the variables extracted by LVs, RC_01 and RC_02 had potential applicability. The number of variables extracted by LVs method was the least (6), R^2_P was the largest (0.837), and the ELM model structure was the simplest (number of HLNs=18). Therefore, PLS(LVs) could be used as the preferred method for variable extraction, and the result of PLS(LVs)_ELM is shown in Figure 9a.

Table 5 Modeling results based on variables extracted

Calibration Model	Variable selection method	No. of variables	Calibration dataset		Prediction dataset			Parameter
			R^2_C	RMSEC	R^2_P	RMSEP	RPD	
PLS	full-spectrum	1326	0.792	2.082	0.791	2.421	2.194	LVs=6
	LVs	6	0.867	1.664	0.792	2.585	2.055	LVs=6
	X-loading	11	0.787	2.108	0.713	2.925	1.816	LVs=10
	VIP_01	18	0.764	2.216	0.712	2.957	1.796	LVs=10
	VIP_02	305	0.780	2.142	0.715	2.888	1.839	LVs=6
	RC_01	26	0.756	2.253	0.755	2.636	2.015	LVs=5
	RC_02	88	0.791	2.084	0.742	2.729	1.946	LVs=6
ELM	LVs	6	0.882	1.567	0.837	2.393	2.220	HLNs=18
	X-loading	11	0.809	1.995	0.782	2.533	2.097	HLNs =35
	VIP_01	18	0.790	2.076	0.784	2.469	2.151	HLNs=32
	VIP_02	305	0.852	1.753	0.799	2.366	2.245	HLNs=60
	RC_01	26	0.861	1.698	0.823	2.245	2.366	HLNs=38
	RC_02	88	0.837	1.844	0.824	2.230	2.382	HLNs=38



a. Variables extracted by LVs



b. Variables extracted by RC_02

Figure 9 Scatter diagrams of the measured and predicted values by ELM

In this work, the method to determine the threshold value of RC based on minimum RMSEP (RC_02) has been proposed. Figure 9b illustrates that this method combined with the ELM model has a good prediction accuracy (R^2_P=0.824). The RMSEP and RPD of PLS(RC_02)_ELM were better than PLS(LVs)_ELM to a certain degree. This method avoided the random sampling strategy effectively and had good portability. Compared with the random selection of thresholds (or lack theoretic support) in previous studies[18,20], the determination of threshold in this study has a sufficient theoretical basis. Further, PLS(RC_02) was recommended to be used in related research which required accurate wavelength information as a variable (the variables extracted by PLS(LVs) were comprehensive variables, not wavelength information).

4 Conclusions

In this study, visible and near-infrared hyperspectral technology combined with PLS-ELM algorithm was used to predict the N content of apple tree canopies. The following conclusions were drawn:

(1) A total of 12 outliers were removed by Monte Carlo second detection method. The signal-to-noise ratio (SNR) of spectral data preprocessed by SG_FD was enhanced, which improved the prediction accuracy.

(2) The numbers of variables extracted by the 6 methods (LVs, *X*-loading, VIP_01, VIP_02, RC_01 and RC_02) derived from 4 PLS-related factors were 6, 11, 18, 305, 26 and 88, which was a great reduction in comparison to the full-spectrum.

(3) The PLS(LVs)_ELM model yielded the optimal prediction results for apple tree canopies N content from the spring-shoot-growing stage to the fruit enlargement stage, and R^2_P, RMSEP and RPD were 0.837, 2.393 and 2.220, respectively.

(4) The method of extracting variables with regression coefficient threshold based on minimum RMSEP was proposed (RC_02). This method could effectively avoid the omission of relevant information.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2017YFD0201508).

[References]

[1] Statistics Bureau of China. China statistical yearbook 2018. Beijing, China Statistics Press, 2018. (in Chinese)

[2] Wang G Y, Zhang X Z, Wang Y, Xu X F, Han Z H. Key minerals influencing apple quality in Chinese orchard identified by nutritional diagnosis of leaf and soil analysis. Journal of Integrative Agriculture, 2015; 14(5): 864–874.

[3] Zhu X C, Gao L L, Fang X Y, Zhao G X, Wang L. Estimating canopy nitrogen contents of an apple tree using hyperspectral remote sensing. Remote Sensing Science, 2016; 4(2): 42–50.

[4] Gao L L, Zhu X C, Li C, Cheng L Z. Evaluation of the nitrogen content during the new-shoot-growing stage in apple leaves using two-dimensional correlation spectroscopy. PLoS ONE, 2017; 12(10): e0186751. doi: 10.1371/journal.pone.0186751.

[5] Fernandez S, Vidal D, Simon E, Sugranes L S. Radiometric characteristics of Triticum aestivum cv, Astral under water and nitrogen stress. International Journal of Remote Sensing, 1994, 15(9): 1867–1884.

[6] Zhu Y, Li Y X, Zhou D Q, Tian Y C, Yao X, Cao W X. Quantitative relationship between leaf nitrogen concentration and canopy reflectance spectra in rice and wheat. Acta ecologica sinica, 2006; 26(10): 3463–3469.

[7] Xue L H, Cao W X, Luo W H, Zhang X. Correlation between leaf nitrogen status and canopy spectral characteristics in wheat. Chinese Journal of Plant Ecology, 2004; 28(2): 172–177. (in Chinese)

[8] Yu L, Hong Y S, Zhou Y, Zhu Q, Xu L, Li J Y, et al. Wavelength variable selection methods for estimation of soil organic matter content using hyperspectral technique. Transactions of the CSAE, 2016; 32(13): 95–102. (in Chinese)

[9] Zou X H, Hao Z Q, Yi R X, Guo L B, Shen M, Li X Y, et al. Quantitative analysis of soil by laser-induced breakdown spectroscopy using genetic algorithm-partial least squares. Chinese Journal of Analytical Chemistry, 2015; 43(2): 181–186. (in Chinese)

[10] Chen L D, Zhao Y R. Measurement of water content in biodiesel using visible and near infrared spectroscopy combined with Random-Frog algorithm. Transactions of the CSAE, 2014; 30(8): 168–173. (in Chinese)

[11] Zou X B, Zhao J W, Povey M J W, Holmes M, Mao H P. Variables selection methods in near-infrared spectroscopy. Analytica Chimica Acta, 2010; 667(1-2): 14–32.

[12] Zhu Y X, Yu L, Hong Y S, Zhang T, Zhu Q, Li S D, et al. Hyperspectral features and wavelength variables selection methods of soil organic matter. Scientia Agricultura Sinica, 2017; 50(22): 4325–4337. (in Chinese)

[13] Liu F, He Y, Wang L. Comparison of calibrations for the determination of soluble solids content and pH of rice vinegars using visible and short-wave near infrared spectroscopy. Analytica Chimica Acta, 2008; 610(2): 196–204.

[14] Zhang R R, Wen Y, Li L L, Chen L P, Xu G, Huang Y B, et al. Method for UAV spraying pattern measurement with PLS model based spectrum analysis. Int J Agric & Biol Eng, 2020; 13(3): 22–28.

[15] Gao J F, Zhang H L, Kong W W, He Y. Nondestructive discrimination of waxed apples based on hyperspectral imaging technology. Spectroscopy and spectral analysis, 2013; 33(7): 1922–1926. (in Chinese)

[16] Zhang H, Wang S, Li D X, Zhang Y Y, Hu J D, Wang L. Edible gelatin diagnosis using laser-induced breakdown spectroscopy and partial least square assisted support vector machine. Sensors, 2019; 19: 4225. doi: 10.3390/s19194225.

[17] Ye X J, Abe S, Zhang S H. Estimation and mapping of nitrogen content in apple trees at leaf and canopy levels using hyperspectral imaging. Precision Agric, 2020; 21: 198–225.

[18] Cheng P Y, Fan W L, Xu Y. Quality grade discrimination of Chinese

strong aroma type liquors using mass spectrometry and multivariate analysis. Food Research International, 2013; 54(2): 1753–1760.

[19] Zhang H L. Soil nutrition content and type measurement based on NIR spectrum and hyper spectra image technology and design portable instrument. Doctoral dissertation. Hangzhou: Zhejiang University, 2015; 145p. (in Chinese)

[20] Wang Y, Gao Y, Yu X, Wang Y Y, Deng S, Gao J M. Rapid determination of *Lycium barbarum* polysaccharide with effective wavelength selection using near-infrared diffuse reflectance spectroscopy. Food Analytical Methods, 2016; 9: 131–138.

[21] Zhang C, Liu J G, Shang J L, Cai H J. Capability of crop water content for revealing variability of winter wheat grain yield and soil moisture under limited irrigation. Science of the Total Environment, 2018; 631–632: 677–687.

[22] Zhu W J, Li J Y, Li L, Wang A C, Wei X H, Mao H P. Nondestructive diagnostics of soluble sugar, total nitrogen and their ratio of tomato leaves in greenhouse by polarized spectra–hyperspectral data fusion. Int J Agric & Biol Eng, 2020; 13(2): 189–197.

[23] Li X X, Zhou J, Tang H, Sun L Q, Cao X M, Zhang X S. Rapid determination of total nitrogen in aquaculture water based on ultraviolet spectroscopy. Spectroscopy and spectral analysis, 2020; 40(1): 195–201. (in Chinese)

[24] Ollinger S V. Sources of variability in canopy reflectance and the convergent properties of plants. New Phytologist, 2011; 189: 375–394.

[25] Shi Z, Liang Z Z, Yang Y Y, Guo Y. Status and prospect of agricultural remote sensing. Transactions of CSAM, 2015; 46(2): 247–260. (in Chinese)

[26] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: Theory and applications. Neurocomputing, 2006; 70: 489–501.

[27] Ouyang Q, Chen Q S, Zhao J W, Lin H. Determination of amino acid nitrogen in soy sauce using near infrared spectroscopy combined with characteristic variables selection and extreme learning machine. Food Bioprocess Technol, 2013; 6: 2486–2493.

[28] Huang G B, Zhou H M, Ding X J, Zhang R. Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems Man & Cybernetics Part B, 2012; 42(2): 513–529.

[29] Czarnecki W M. Weighted tanimoto extreme learning machine with case study in drug discovery. IEEE Computational Intelligence Magazine, 2015; 10(3): 19–29.

[30] Kamruzzaman M, Elmasry G, Sun D W, Allen P. Prediction of some quality attributes of lamb meat using near-infrared hyperspectral imaging and multivariate analysis. Analytica Chimica Acta, 2012; 714: 57–67.

[31] Cao D S, Liang Y Z, Xu Q S, Li H D, Chen X. A new strategy of outlier detection for QSAR/QSPR. Journal of Computational Chemistry, 2010; 31(3): 592–602.

[32] Galvão R K H, Araujo M C U, José G E, Pontes M J C, Silva E C, Saldanha T C B. A method for calibration and validation subset partitioning. Talanta, 2005; 67(4): 736–740.

[33] Shan P, Zhao Y H, Wang Q Y, Sha X P, Lyu X Y, Peng S L, et al. Stacked ensemble extreme learning machine coupled with Partial Least Squares-based weighting strategy for nonlinear multivariate calibration. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019; 215: 97–111.

[34] Ramadan Z, Hopke P K, Johnson M J, Scow K M. Application of PLS and back-propagation neural networks for the estimation of soil properties. Chemometrics & Intelligent Laboratory Systems, 2005; 75(1): 23–30.

[35] Chong I G, Jun C H. Performance of some variable selection methods when multicollinearity is present. Chemometrics & Intelligent Laboratory Systems, 2005; 78(1-2): 103–112.

[36] Wen P F. Monitoring the vertical distribution of nitrogen status at leaf and canopy scales with remote sensing data in maize. Doctoral dissertation. Yangling: Northwest A& F University, 2019; 124p. (in Chinese)

[37] Guo P T, Su Y, Cha Z Z, Lin Q H, Luo W, Lin Z M. Prediction of leaf phosphorus contents for rubber seedlings based on hyperspectral sensitive bands and back propagation artificial neural network. Transactions of the CSAE, 2016; 32(Supp. 1): 177–183. (in Chinese)