

Using Vis/NIR Diffuse Transmittance Spectroscopy and Multivariate Analysis to Predicate Soluble Solids Content of Apple

Shuxiang Fan^{1,2,3,4} · Zhiming Guo^{1,2,3,4} · Baohua Zhang^{1,2,3,4} · Wenqian Huang^{1,2,3,4} · Chunjiang Zhao^{1,2,3,4}

Received: 12 August 2015 / Accepted: 9 September 2015 / Published online: 23 September 2015
© Springer Science+Business Media New York 2015

Abstract The objectives of this research were to compare the effect of different fruit orientations on the quality of acquired spectra and to provide a suitable calibration model for further online determination of soluble solids content (SSC) of “Fuji” apples using visible and near-infrared (Vis/NIR) diffuse transmittance. The diffuse transmittance spectra between 650 and 910 nm were collected with the designed spectrum measurement system in two fruit orientations: stem-calyx axis horizontal (T1) and stem-calyx axis vertical (T2). Area change rate (ACR) was used to evaluate the stability of spectra collected in two fruit orientations. Results showed that the fruit orientation T1 was better for our designed spectrum measurement system. Then, the performance of partial least squares (PLS) models based on spectral data after the pretreatment of several preprocessing methods was analyzed and compared. Finally, the modified competitive adaptive reweighted sampling (MCARS), successive projection algorithm (SPA), and their combination were investigated to select the effective variables for the determination of SSC. It concluded that the MCARS-SPA-PLS model based on the spectra after preprocessing of Savitzky-Golay (SG) smoothing achieved better results for SSC prediction. The correlation coefficients

between measured and predicted SSC were 0.962 and 0.946, and the root mean square errors were 0.510 and 0.527°Brix for calibration and prediction set, respectively. Moreover, the physicochemical properties of 27 variables selected by MCARS-SPA were discussed to obtain a better interpretation of the calibration model. The overall results indicated that the designed diffuse transmittance spectrum measurement system together with the PLS calibration model with 27 effective variables selected by MCARS-SPA method had a potential application for online SSC detection of apple.

Keywords Apple · Soluble solids content · Fruit orientation · Diffuse transmittance · Variable selection · Area change rate

Introduction

High-quality requirements for commercial fresh fruit on the global produce market have ever been increasing: Fruit should not only be nutritious but also have appropriate texture and taste to meet consumer demands (Do Trong et al. 2014). Apple is an important and a widespread agricultural commodity (Mendoza et al. 2014). In particular, apple is a good source of antioxidant components, such as ascorbic acid and polyphenolic, which exert protective effects against various degenerative diseases (Giovannelli et al. 2014). Soluble solids content (SSC) is one of the most important properties that influence the consumer purchasing decision on fresh apple fruit (Lu 2004) and determine the fruit maturity and harvest time (Peng and Lu 2007). Consequently, nondestructive and rapid detection of SSC of apple is of great value in ensuring high quality, consistent apples for the consumer. In the last decades, visible and near-infrared (Vis/NIR) spectroscopy has been proposed as a fast, easy to use, and nondestructive analytical technique (Nicolai et al. 2007). The technique, coupled with an appropriate

✉ Wenqian Huang
huangwq@nrcita.org.cn

¹ Beijing Research Center of Intelligent Equipment for Agriculture, Beijing 100097, China

² National Research Center of Intelligent Equipment for Agriculture, Beijing 100097, China

³ Key Laboratory of Agri-informatics, Ministry of Agriculture, Beijing 100097, China

⁴ Beijing Key Laboratory of Intelligent Equipment Technology for Agriculture, Beijing 100097, China

calibration method, has been successfully used to measure the SSC of apple (Peirs et al. 2001; Liu and Ying 2005; Mendoza et al. 2014). Recently, much attention has been paid on online detection of SSC using Vis/NIR spectroscopy. Several studies about online SSC determination using diffuse transmittance mode were reported for fruits such as pear (Sun et al. 2009; Xu et al. 2012) and watermelon with thick skin (Jie et al. 2014), which indicated that diffused transmittance mode was suitable for internal quality determination and was a viable option for high-speed fruit measurement. Therefore, a prototype of diffused transmittance system was realized in our laboratory to provide some reference for the online detection of apple SSC.

The fruit orientation is an important factor that affects the quality of acquired spectra. Fu et al. (2007) compared two fruit orientations (stem-calyx axis vertical, stem-calyx axis horizontal) in diffuse transmission mode for detecting brown heart in pears, and better results were obtained based on the stem-calyx axis horizontal orientation. Fan et al. (2009) investigated the effect of fruit orientation on the prediction results for detecting the SSC and firmness of apples. However, they found that the best fruit orientation was the stem-calyx axis vertical. Recently, a new surface scanning technology invented by Schmutzler and Huck (2014) was shown to result in improved calibration models by the measurement of hundreds of spectra over the apple surface. But, it was time-consuming and complicated to use this technology for online detection. Therefore, the present work was using area change rate (ACR) to select the best fruit orientations from two commonly used fruit orientations (stem-calyx axis vertical, stem-calyx axis horizontal) in the SSC determination by comparing the stability of the collected spectrum instead of just depending on the prediction or classifying results.

For online detection of SSC, a calibration model is needed to be mainly considered. Partial least squares (PLS) regression has been widely used to develop calibration models for determining SSC of apple and other fruits. When used for online purpose, the complex calibration models developed with the whole spectrum will not be applicable because of useless or irrelevant information (Andersen and Bro 2010). Furthermore, the modern spectroscopy instrumentations usually possess high resolution, with hundreds or thousands of spectral variables including collinearity, redundancies, and noise (Wang and Xie 2014). Therefore, many variable selection methods such as genetic algorithms (GAs) (Durand et al. 2007), Monte Carlo uninformative variable elimination (MC-UVE) (Cai et al. 2008), competitive adaptive reweighted sampling (CARS) (Li et al. 2009) and successive projection algorithm (SPA) (Araújo et al. 2001) have been proposed to solve this problem. Some published papers had reported the application of Vis/NIR spectroscopy combined with variable selection methods for online prediction of internal quality of fruits. Xu et al. (2012) compared four variable selection methods (stepwise multi-linear regression (SMLR), GA, interval PLS (iPLS), and GA-SPA) for the analysis of SSC of pear in the spectra range 533–929 nm. It was found that the MLR calibration

model built using GA-SPA on 18 selected wavelengths exhibited coefficient of determination $r_{\text{pre}}^2=0.880$ and root mean square error of prediction (RMSEP)= 0.459°Brix for the prediction set. Jie et al. (2014) investigated Vis/NIR diffuse transmission spectrum of 687–920 nm region for online determination of SSC of watermelon. They found that the MC-UVE-SMLR calibration model with baseline offset correction pretreatment was the best with r_{pre} of 0.70 and RMSEP of 0.33°Brix for the prediction set. In these studies, the elimination of uninformative variables enhanced the model prediction, reduced measurement costs, and facilitated model interpretation. However, in most of the scientific works about apple SSC detection, calibration with the full-range variables is time-consuming and the irrelevant information within spectra would affect the accuracy and robustness of the model. In order to meet the needs of online detection, variable selection is conducted to simplify the model and improve detection efficiency. In our study, the modified CARS (MCARS), SPA, and their combination were conducted to select effective variables for SSC determination. In addition, the physicochemical properties of the selected variables were also discussed to obtain a better interpretation of the calibration model.

The overall goal of this study was to provide references for online determination of SSC of “Fuji” apple in terms of fruit orientation and model foundation by using Vis/NIR diffuse transmittance. Specific objectives of the research were to (1) investigate the performance of the designed device that collects the diffuse transmittance spectrum of apple; (2) evaluate the stability of spectra acquired in two fruit orientations using ACR; and (3) pick out the most effective variables and discuss the physicochemical properties of the selected variables for further online determination of SSC of apple using Vis/NIR diffused transmittance spectroscopy.

Materials and Methods

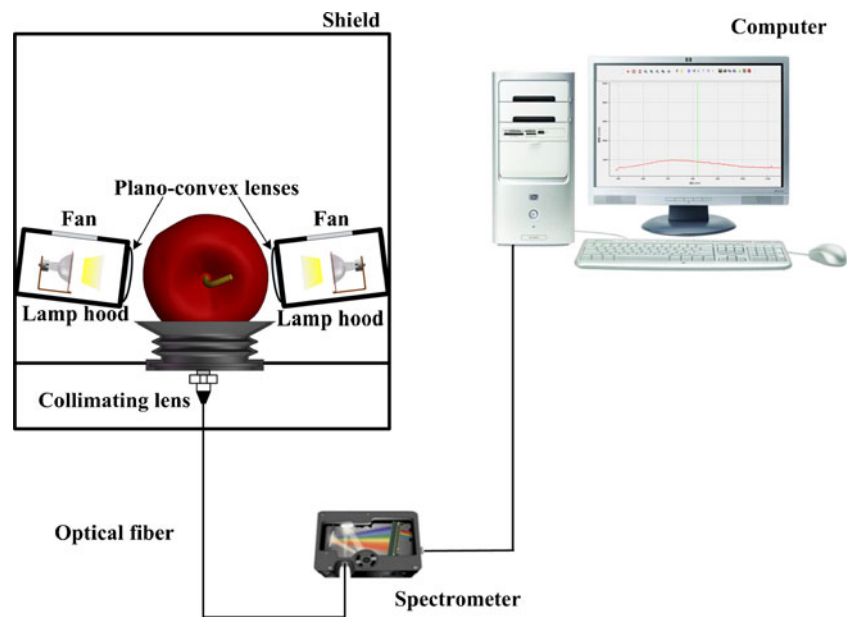
Samples

A total of 130 Fuji apples free of visual defects (such as scars, cuts, shrivel, etc.) were purchased from a local fruit market in Beijing, China. The equatorial diameter range of apples was 70–80 mm, and all samples were individually washed and numbered and then stored in laboratory (temperature 20°C , relative humidity 60 %) for 24 h before the experiment to allow the samples to reach room temperature to reduce the effect on the prediction accuracy by the temperature of samples (Fan et al. 2015).

Spectra Collection and ACR Analysis

The diffuse transmittance spectra of samples were obtained by a spectrum measurement system designed by ourselves (Fig. 1). The measurement consisted of a fruit tray for holding

Fig. 1 Schematic of the diffuse transmittance spectrum measurement system



the fruits and blocking the light leaking through the surface between sample and tray, light source mounted on both sides of the holder, collimating lenses embedded optical fiber designed to accumulate transmitted light penetrating from different parts of a sample through the center hole of fruit tray, and an optical fiber used to connect the collimating lens and a commercial portable fiber spectrometer (Model USB2000+, Ocean Optics Inc., USA) with wavelength range from 487 to 1148 nm. The light source was composed of two halogen lamps (100 W, 12 V), two sets of plano-convex lens (Edmund Optics Inc., Barrington, USA) for focusing the lamp light on the fruit embedded with the lamp hoods, and two fans mounted in the lamp hoods to radiate the heat from the lamps. All components were fixed inside a dark chamber to avoid any stray light that might affect the spectrum of sample.

The reference spectrum (R_{white}) was collected from a standard Teflon plate, and the dark spectrum (R_{dark}) was collected when all lamps were turned off. They were measured and stored before collecting spectra of samples. Two fruit orientations (T1 and T2) were investigated in this paper, and the sketches of the two modes are shown in Fig. 2.

- T1: Fruit stem-calyx axis horizontal; irradiated from cheek by the light source and detected from the equator position by the optic fiber
 T2: Fruit stem-calyx axis vertical; irradiated from cheek by the light source and detected from the calyx by the optic fiber

The spectral data were measured in two fruit orientations using software SpectraSuite (Ocean Optics Inc., USA) with the integration time of 100 ms. The spectral data were first acquired in fruit orientation T1. Each sample was placed centrally and steadily on the fruit tray by hand, with the stem-

calyx axis horizontal and the equator position facing the detector. Only spectral data between 650 and 910 nm which composed of 785 wavelength variables were retained for further analysis, and other wavelength ranges were eliminated because of some sharp noises and irrelative information. Then, the raw spectra (R_{raw}) were converted to absorbance values ($\log(1/T)$) according to the following equation:

$$\log(1/T) = \log\left(\frac{R_{\text{white}} - R_{\text{dark}}}{R_{\text{raw}} - R_{\text{dark}}}\right) \quad (1)$$

In order to decrease the error of operator and instrument and improve signal-to-noise ratio (SNR), the process was repeated three times for each sample with the same position facing the detector to acquire a mean spectrum by averaging the three absorbance spectra. All the spectral data were stored in a computer for further analysis. The measurement progress for fruit orientation T2 was the same as that for T1. Finally, three absorbance spectra and a mean spectrum of each apple were obtained in fruit orientation T2.

The spectral area means the area under a certain curve between two wavelength bands, which is measured by adding the counts of many different data points together (Smith 2003). The spectra collected under the same condition should be exactly the same if there is no interference and noise. Therefore, the smaller the change of the area, the more stable the spectra is (Zhang et al. 2014). So, the ACR is selected as one of the important indicators for spectra stability evaluation. In this paper, the ACR was approximated as the root mean square deviation (RMSD) of the area of the three absorbance spectrum between 650 and 910 nm in any fruit orientation (T1 and T2). Before calculating ACR, the Y coordinate of the spectrum was normalized

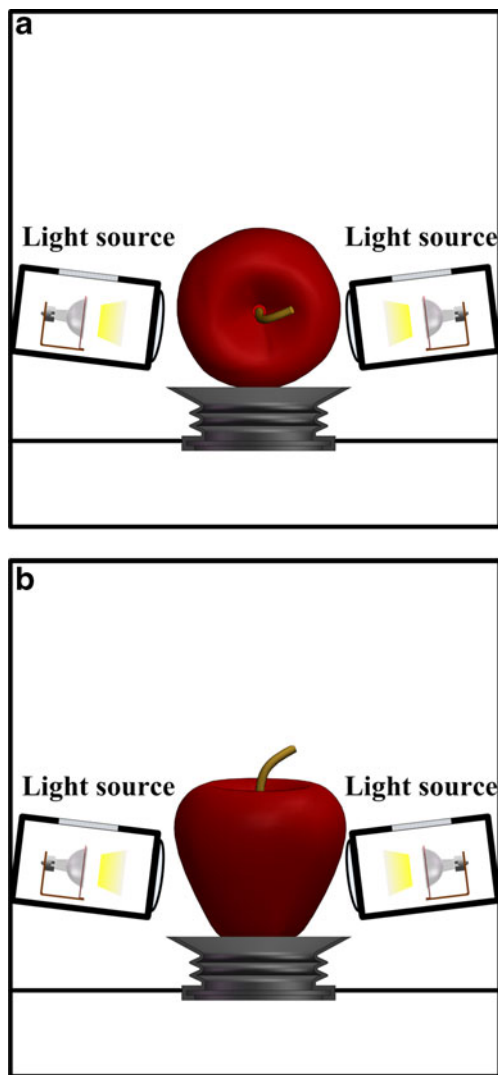


Fig. 2 General sketches of two fruit orientation modes: stem-calyx axis horizontal mode T1 (a) and stem-calyx axis vertical mode T2 (b)

using vector normalization. RMSD is depicted as follows:

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - Y_{\text{mean}})^2} \quad (2)$$

where Y_i is each spectral area between two selected wavelength bands, Y_{mean} is the average of total spectral area, and N is the number of the spectra. Then, the ACR values of all samples were calculated for fruit orientation T1 and T2, respectively. The spectral normalization and area calculation were processed by Matlab2012a (The Math Works, Natick, USA).

SSC Measurement

Immediately after the spectra collection and analysis of ACR, the SSC was determined using traditional destructive test.

According to Schmutzler and Huck (2014), for different parts of the same apple, SSC values were found with significant variations higher than 2°Brix. In order to make the spectrum and SSC correspond more appropriately, tissue sample was cut from each sample at the location of the spectrum measurement in fruit orientation T1 or T2 according to the results of ACR analysis. Then, juice was squeezed and dropped onto a refractometer (ARIAS 500, Reichert Technologies, New York, USA) to record the SSC value.

Spectra Preprocessing

The spectrum acquired from spectrometer might contain background information or noise besides sample information, so it is necessary to preprocess the spectral data to establish a reliable, accurate, and stable calibration model (Cen and He 2007). Different spectral preprocessing methods including SG smoothing (Savitzky and Golay 1964), multiplicative scattering correction (MSC) (Helland et al. 1995), standard normal variate transformation (SNV) (Barnes et al. 1989), baseline offset correction, and second derivative (Demetriades-Shah et al. 1990) were tried to preprocess the raw spectral data. Because SNV and MSC had the similar function of reducing the (physical) variability between samples due to scattering and adjusting for baseline shifts between samples (Jie et al. 2013), so only MSC was used in our study. SG smoothing was an average algorithm that fits a polynomial to the data points and was necessary to optimize the signal-to-noise ratio (Perkin et al. 1988). Baseline offset correction was often used to adjust the spectral offset by adjusting the data to the minimum point in the data. The second derivative spectra were calculated using the method of Savitsky-Golay smoothing algorithm to correct for additive and multiplicative effects in the spectra. These spectral preprocessing treatments were performed in the Unscrambler v9.7 (CAMO PRECESS AS, Oslo, Norway).

Regression Analysis

PLS regression is a powerful multivariate calibration method that is insensitive to collinear variables and tolerant to large numbers of variables and widely employed in chemometric analysis (Huang et al. 2008). PLS has the potential to consider not only variable matrix X (spectral data) but also variable matrix Y (the properties of interest). Generally, PLS is applied to extract no more than top 20 latent variables (LVs) from a large set of highly correlated and collinear original spectral data (Leiva-Valenzuela et al. 2013). The LVs can explain the variance and reduce the dimensionality of the original spectra. In this study, calibration models between spectral data of tested fruits and their quality attributes (SSC) were developed using PLS. In the development of a PLS model, the optimal

number of LVs was determined by a full cross-validation of the calibration samples.

Variable Selection Methods

MCARS

CARS is an innovative and useful variable selection algorithm first proposed by Li et al. (2009), which has the potential to select an optimal combination of the effective variables existing in the full spectrum coupled with PLS regression. Absolute values of regression coefficients of PLS model are used as an index for evaluating the importance of each variable. Then, CARS sequentially selects N subsets of variables from N Monte Carlo sampling run in an iterative and competitive manner according to the importance level of each variable. In each sampling run, some samples are first randomly chosen in a fixed ratio to build a calibration model. Next, the exponentially decreasing function (EDF) and adaptive reweighted sampling (ARS) process are adopted to select the key variables based on the regression coefficients. Finally, the subset with the lowest root mean square error of cross-validation (RMSECV) is chosen. The procedure of CARS was performed in the Matlab2012a with libPLS toolbox available at <http://www.libpls.net/>. However, Yun et al. (2014) thought that CARS is a very fast method but not always stable due to the Monte Carlo sampling. Therefore, in order to improve the stability of CARS behavior and guarantee the reliability of the model, the MCARS method was proposed in our study. Firstly, the CARS was conducted 500 times to obtain the selected frequency of each variables. Then, the variables were added to develop PLS models according to the frequency of selections (i.e., in the model with n variables, the n most frequently selected variables were selected), and the corresponding RMSECV were calculated. The optimal subset was obtained with the lowest value of RMSECV.

SPA

SPA is a forward selection method applying vector projection operations in a vector space for the selection of relevant variables for multivariate calibration, which begins with one wavelength variable and then incorporates a new one at each iteration, until it obtains variables with a minimum of collinearity. The principle of variable selection by SPA is that the candidate variables selected by SPA has the maximum projection value on the orthogonal subspace of the previous selected variables. In the algorithm, candidate subsets of variables with minimum collinearity are first generated and evaluated by the value of root mean square error (RMSE) obtained from validation set of MLR calibration, and then, the uninformative variables are removed by a variable elimination procedure without significant loss of prediction capability. Details of

the SPA methodology could be referred to the previous literature (Araújo et al. 2001; Liu et al. 2009). The variable selection procedure was carried out in the Matlab2012a with a graphical user interface for SPA (GUI_SPA) which was downloaded from www.ele.ita.br/~kawakami/spa/.

Evaluation of the Performance of Models

In developing a calibration model, 100 samples were selected randomly from 130 apples as calibration set for developing the calibration model. To ensure the SSC of calibration set covered that of all prediction set, two samples of the highest and the lowest concentrations were put into the calibration set manually. The remaining 30 were selected as prediction set to verify the prediction power of the calibration model. No single sample was used in calibration set and validation set at the same time. The performance of the calibration model was evaluated in terms of correlation coefficient of calibration (r_{cal}) and prediction (r_{pre}), root mean square error of calibration (RMSEC) and RMSEP. The calculations of r_{cal} , r_{pre} , RMSEC, and RMSEP are defined in the following equations (Liu et al. 2010):

$$r_{cal} = \sqrt{\frac{\sum_{i=1}^{n_c} (y_{pi} - y_{mi})^2}{\sum_{i=1}^{n_c} (y_{pi} - y_{mean})^2}} \quad (3)$$

$$r_{pre} = \sqrt{\frac{\sum_{i=1}^{n_p} (y_{pi} - y_{mi})^2}{\sum_{i=1}^{n_p} (y_{pi} - y_{mean})^2}} \quad (4)$$

$$RMSEC = \sqrt{\frac{1}{n_c} \sum_{i=1}^{n_c} (y_{pi} - y_{mi})^2} \quad (5)$$

$$RMSEP = \sqrt{\frac{1}{n_p} \sum_{i=1}^{n_p} (y_{pi} - y_{mi})^2} \quad (6)$$

where y_{pi} is the predicted value of SSC in fruit number i , y_{mi} is the measured value of SSC in fruit number i , y_{mean} is the mean value of SSC in the calibration or prediction set, and n_c and n_p are the number of fruits in the calibration set and prediction set, respectively. Generally, a good model should be selected based on not only the higher r_{cal} and r_{pre} values and lower RMSEC and RMSEP values but also a small difference between RMSEC and RMSEP (Li et al. 2014).

The calibration and prediction results may vary depending on how the calibration and prediction samples were actually selected (Huang et al. 2014). Therefore, the calibration and prediction procedure described above was repeated 10 times by electing a random set of samples. Then, the results (i.e., r_{cal} , r_{pre} , RMSEC, and RMSEP) were averaged to estimate the

final performance of the models. Finally, *t* test was performed on the average RMSEPs for the 10 runs to compare the statistical differences of different variable selection methods for predicting SSC of apple.

Results and Discussion

Spectral Features and ACR Analysis

The raw diffuse transmittance spectra of apple samples ranging from 650 to 910 nm acquired in fruit orientation T1 are shown in Fig. 3. As can be seen, the trends of these spectra were similar. There were two strong absorption peaks around 750 and 850 nm. The absorption peak around 750 nm was associated with the third overtone of the H₂O, and the absorption peak around 850 nm was due to the third overtone of the C–H functional group (Jamshidi et al. 2012). Similar spectral features were observed for the spectra acquired in fruit orientation T2.

Figure 4 shows the ACR values of all 130 fruit samples calculated from the Eq. (2) for fruit orientation T1 and T2, with the mean ACR values of 0.0125 and 0.0333, respectively. As shown in Fig. 4, for most of the samples, the ACR values for fruit orientation T2 were much bigger than those for T1. Moreover, the ACR values fluctuated drastically for fruit orientation T2. These results showed that the variation of the spectral curve was small for the fruit orientation T1 and the stability of spectra acquired in fruit orientation T1 was better for our designed diffused transmittance spectrum measurement system. Consequently, the mean spectra acquired in fruit orientation T1 were used for further preprocess and calibration analysis.

Statistics of SSC

After the analysis of ACR, tissue sample cut from the equator position of each sample at the location of the spectrum

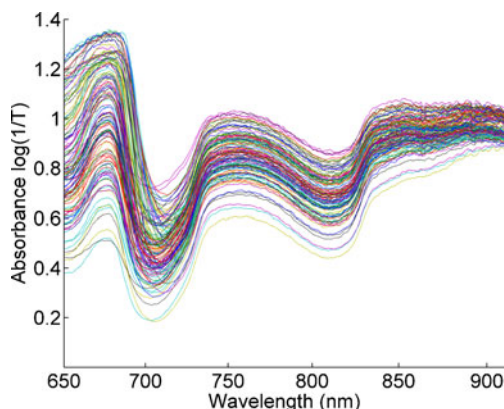


Fig. 3 Raw diffuse transmittance spectra of apple samples acquired in fruit orientation T1

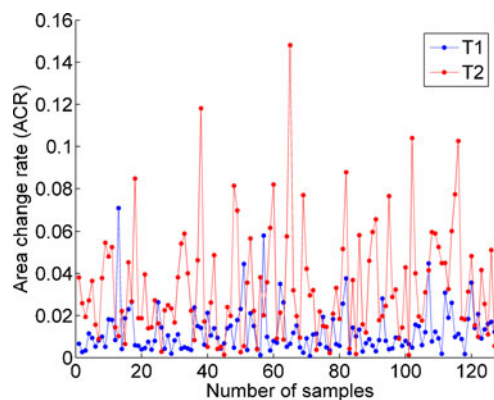


Fig. 4 Area change rate (ACR) values of all 130 fruit samples

measurement in fruit orientation T1 was used to measure real SSC value. The distribution of SSC of all apple samples is presented in Fig. 5. The SSC measurements of 130 samples were fairly normally distributed around the mean value (14.07°Brix) with standard deviations of 1.09. The SSC values varied between 8.78 and 19.49°Brix, covering a large enough range. More importantly, the range of calibration sets were bigger than prediction sets by putting the two samples of the highest and the lowest SSC values into the calibration sets manually. These features were helpful to develop a good calibration model (Li et al. 2013).

Calibration Models of SSC with Full Spectra

The full-spectrum PLS models were developed using raw spectra and preprocessed spectral data pretreated by SG smoothing, MSC, baseline correction, and second derivative, respectively. The results showing a comparison of these preprocessing methods in the SSC prediction by PLS are presented in Table 1. It can be found that the prediction results were significantly improved by using the spectra pretreated by SG smoothing (21-point window size and third-order polynomial) than those using raw spectra for PLS calibration models ($p < 0.05$). In comparison to the results obtained by raw spectra, the model performance based on other processing methods

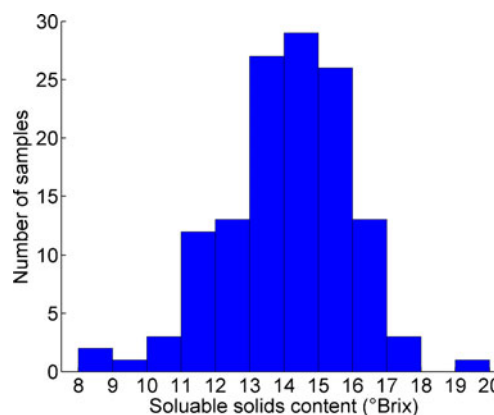


Fig. 5 Distribution of soluble solids content of all 130 fruit samples

Table 1 Average of 10 calibration and prediction results for SSC by PLS with different preprocessing methods

Pretreatment	LVs	Calibration		Prediction	
		r_{cal}	RMSEC	r_{pre}	RMSEP
Raw	12	0.963	0.411	0.920	0.768 a
SG	10	0.957	0.541	0.928	0.606 b
MSC	14	0.979	0.271	0.874	0.794 a
Second derivative	9	0.943	0.616	0.870	0.810 a
Baseline	11	0.953	0.477	0.874	0.796 a

RMSEPs with different letters represent statistical difference ($p < 0.05$) of prediction models between the raw spectra and tested preprocessing method

including MSC, second derivative (31-point window size and third-order polynomial), and baseline correction was not improved and no significant statistical difference was found ($p > 0.05$). So, the best result of PLS model for SSC prediction was obtained by using spectra after SG smoothing preprocessing with r_{pre} of 0.928 and RMSEP of 0.606. Therefore, further analysis was conducted based on the spectra data after SG smoothing preprocessing.

PLS Models with Effect Variables

The MCARS and SPA variable selection methods were used for PLS regression to pick out the most effective variables for SSC prediction based on one date set which was randomly selected from 10 calibration and prediction sets.

Variables Selected by MCARS

For MCARS method, the CARS procedure was run 500 times to improve the stability behavior. In this study, for each running of CARS, the number of Monte Carlo sampling runs was set to 100 and the number of variables to be selected was determined by 10-fold cross-validation. Figure 6 shows the changing trend of the number of sampled variables (Fig. 6a) and 10-fold RMSECV values (Fig. 6b) with the increasing of sampling runs from one CARS running. As can be seen in Fig. 6a, the number of sampled variables decreased fast at the first stage of EDF and then slowly at the second stage of EDF, which displayed the fast selection and refined selection of CARS. On the other hand, Fig. 6b gives the corresponding 10-fold RMSECV values for SSC. According to the minimal 10-fold RMSECV value in the 17th sampling run which was marked by the open square in Fig. 6b, the optimal variable subsets were determined for SSC prediction, while the corresponding number of sampled variables was 130 (red solid dot in Fig. 6a) and they were selected as effective variables in this CARS running.

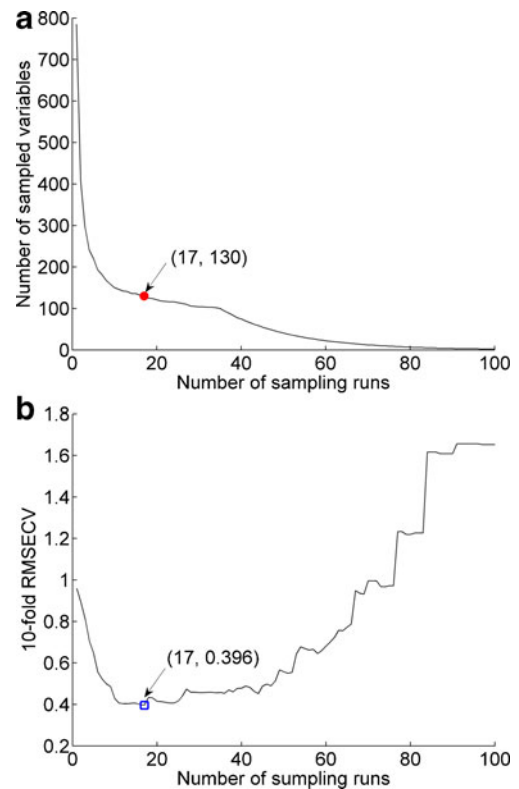


Fig. 6 The changing trend of the number of sampled variables (a) and 10-fold RMSECV values (b) with the increasing of sampling runs from one CARS running

After the process of MCARS, the selected frequency of each variable by running was shown in Fig. 7. Then, the RMSECVs were calculated through the PLS models with the variables added according to the frequency of selections (i.e., in the PLS model developed with n variables, these n variables were most frequently selected). Clearly, when 164 variables were added to the PLS model, the RMSECV reached the minimum, corresponding to the cutoff threshold in Fig. 7. Then, the 164 variables in the subset were selected as the key variables for determining SSC of apple by using MCARS. The selected variables were set as the inputs to develop PLS models for all the 10 calibration and prediction sets to determine the SSC of apple. The average of 10 calibration and

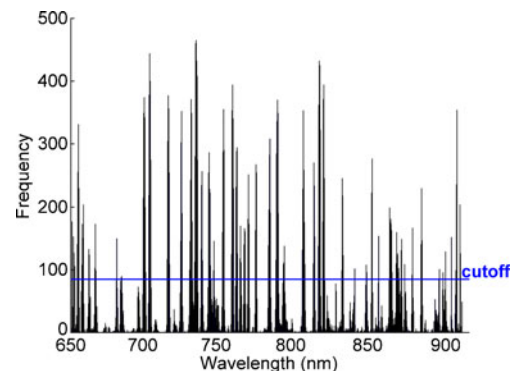


Fig. 7 The selected frequency of each variable by running MCARS

Table 2 Average of 10 calibration and prediction results for SSC by different PLS models

Model	No. of variables	LVs	Calibration		Prediction	
			r_{cal}	RMSEC	r_{pre}	RMSEP
Full-spectra PLS	785	10	0.957	0.541	0.928	0.606 a
MCARS-PLS	164	11	0.975	0.395	0.951	0.513 b
CARS-PLS	130	15	0.989	0.271	0.938	0.584 a
SPA-PLS	60	10	0.959	0.529	0.930	0.595 a
MCARS-SPA-PLS	27	10	0.962	0.510	0.946	0.527 b

RMSEPs with different letters represent statistical difference ($p < 0.05$) of prediction models between the full-range spectra and tested variable selection method

prediction results of MCARS-PLS models using 164 variables for SSC prediction were shown in Table 2. Furthermore, for comparison with the performance of MCARS-PLS models, the PLS models were developed using 130 variables selected by one CARS running which was stated above. The average of 10 calibration and prediction results by CARS-PLS models were also shown in Table 2. As can be seen, the performance of MCARS-PLS model was significantly improved ($p < 0.05$) with r_{pre} of 0.951 and RMSEP of 0.513°Brix compared with the results of full-spectrum PLS model with r_{pre} of 0.928 and RMSEP of 0.606°Brix, whereas relatively poor results were obtained for CARS-PLS models ($r_{pre} = 0.938$, RMSEP = 0.584°Brix). In addition, small difference between calibration and validation was found in MCARS-PLS models. These results indicated that MCARS was a very effective variable selection method to improve the prediction accuracy of the calibration models for the SSC determination of apple, and such improvement was even achieved by using only about 21 % of variables of full-range spectra (164 vs. 785).

Variables Selected by SPA

After the MCARS processing, the number of variables decreased to 164, but with respect to the online detection of the SSC, there were still too many variables for application. In addition, some collinear variables which contain a number of redundant information might still exist in the spectra data. Therefore, in order to further simplify the model and improve the robustness, SPA was carried out on 164 selected variables for further variable selection. Meanwhile, the full spectrum was also employed as the input of the SPA to investigate whether the MCARS method would have effect on the SPA-PLS model.

SPA was firstly performed to select effective variables from the full spectra for the prediction of SSC. During performing the SPA, a cross-validation procedure was used for the calculation of a sequence of RMSE values using the selected variable subsets. The optimal number of selected variables was determined by this process. After SPA, 60 variables were selected from the full spectrum for the prediction of SSC. The results of PLS models developed with the selected variables

were shown in Table 2. Compared with the full-spectrum PLS model, although the number of variables decreased sharply after using the SPA, the predicting ability of SPA-PLS models was not improved significantly ($p > 0.05$) with r_{pre} of 0.930 and RMSEP of 0.595°Brix. It might be caused by that SPA was operated on the full spectrum which contained some uninformative variables. In addition, Liu et al. (2014) thought that one disadvantage of variable selection by SPA was its low S/N or insufficiency in multivariate calibration, which could negatively affect the accuracy of the model prediction. To reduce this limitation, it might be possible to improve the SPA performance by conducting SPA operation after using MCARS. Therefore, SPA was carried out on 164 variables selected by MCARS for further variable selection.

As a result of MCARS-SPA, 27 variables that were obtained from the full spectra for prediction of SSC included 733.21, 755.48, 818.10, 791.47, 737.61, 760.51, 809.26, 650.08, 795.11, 655.34, 745.72, 777.21, 821.69, 772.21, 852.78, 834.37, 864.32, 868.48, 878.66, 906.43, 871.98, 748.75, 849.24, 899.83, 841.82, 766.87, and 700.43 nm. In order to estimate the performance of variables obtained by MCARS-SPA, PLS calibration models were developed by using the 27 variables for the prediction of SSC. The results were also shown in Table 2. As can be seen, MCARS-SPA-PLS models showed better prediction ability ($r_{pre} = 0.946$, RMSEP = 0.527°Brix) than full-spectrum PLS models ($p < 0.05$). In contrast with the MCARS-PLS models, the MCARS-SPA-PLS models used far less variables, making a great help for the simplification of the prediction model and the satisfaction of the requirement of online detection. It is also worth mentioning that the absolute difference values between RMSEC and RMSEP of the MCARS-PLS and MCARS-SPA-PLS models were 0.118 and 0.017 for SSC, showing that the established models using the variables selected by MCARS-SPA were more robust than those with the variables selected by only MCARS. The results implied that it was effective to adopt MCARS method to eliminate the variables with irrelative information for modeling before applying the SPA procedure. The proposed MCARS-SPA method which combines the advantage of the MCARS and SPA would be an effective

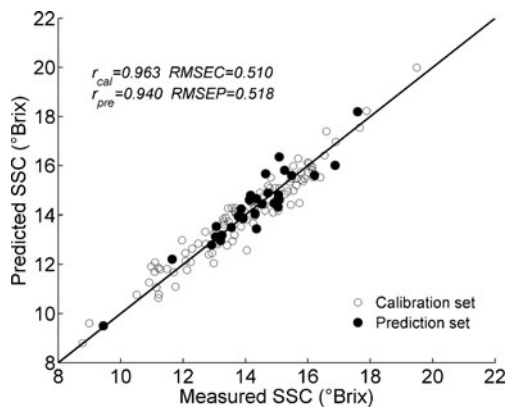


Fig. 8 Scatter plot of measured versus predicted SSC by the PLS calibration models combining MCARS and SPA methods

variable selection method. Figure 8 shows the scatter plots of measured versus predicted SSC obtained by PLS combining MCARS and SPA methods for one of the 10 calibration and prediction sets. The solid line represents the ideal regression line, as the closer the points are to this line, the better the model is (Xie et al. 2011).

Discussion

The selected variables in our study at 650.08 nm was due to the chlorophyll b (Jamshidi et al. 2012), 733.21 and 737 nm were due to the O–H stretching third overtone (Liu et al. 2014), and 745.72 and 748.75 nm were associated with H₂O third overtone (Fu and Ying 2014). In addition, 791.47 and 795.11 nm (N–H stretching third overtone); 841.82 (O–H combinations); and 849, 852, and 899 nm (C–H stretching third overtone) (Jamshidi et al. 2012; Fu and Ying 2014) were also included in the MCARS-SPA-PLS model which was no longer a black box model but had a physicochemical background. The selected variables at 700.43, 745.72, 748.75, 809.26, 821.69, 841.82, and 906.43 nm in our study were the same as variables selected by Qing et al. (2007) for the prediction of SSC of Fuji apple, while the variables at 650.08, 655.34, 700.43, 755.48, 760.51, 766.87, and 849.24 nm were similar to the variables selected by semi-supervised affinity propagation method using hyperspectral scattering imaging technique for determining the SSC of “Golden Delicious” apples (Zhu et al. 2013). The variables selected in our study for predicting the SSC of Fuji apples were not totally identical with those selected in the literature stated above and might be due to the differences of the measurement spectrum system, the processing methods (preprocessing methods and variable selection method), or the variety of apple samples. Moreover, the variables at 737.61, 748.75, 791.47, 809.26, 864.32, and 878.66 nm were also selected for the prediction of SSC of pears (Xu et al. 2012), while the variables at 650.08, 748.75, and 849.24 nm were included in the model to determine the SSC of citrus fruit (Wang and Xie 2014). These results indicated that some variables were

associated with the prediction of SSC of fruits. Therefore, an additional analysis for discovering chemical compounds matching to the selected variables may be an interesting future subject.

As for the SSC prediction of Fuji apple, the results obtained based on 27 effective variables in this study are comparable with those obtained by Liu et al. (2007) with RMSEP = 0.77°Brix using FT-NIR diffuse reflectance technique in 12, 500–4000 cm⁻¹ spectrum region Qing et al. (2007), with RMSECV = 0.90°Brix for SSC based on 16 variables which were selected from the diffuse reflectance spectrum in the region of 600–1100 nm. On the other hand, the results in our study are a little worse than those obtained by Xiaobo et al. (2007) with $r_{pre} = 0.936$ and RMSEP = 0.414°Brix using 44 variables selected from the full FT-NIR interaction spectrum (11, 000–3800 cm⁻¹) and those obtained by Liu and Ying (2005) with $r_{pre} = 0.968$ and RMSEP = 0.455°Brix with the full FT-NIR spectrum. Better results also have been found in Fuji apple with $r_{pre}^2 = 0.982$ and RMSEP = 0.277°Brix for SSC prediction using the wavelength range of 650–920 nm in diffused transmittance mode (Fan et al. 2009). As can be seen, the SSC values of Fuji apple could be well predicted using Vis/NIR technique in literatures stated above. In addition, the effective variables could simplify the prediction model and improve the model efficiency. By comparing the results obtained in the literatures about the SSC prediction of Fuji apples, it can be concluded that the diffused transmittance technology could be more suitable for the online determination. In order to meet the needs of online detection, it is meaningful to combine the technique with variable selection methods to improve efficiency. Although the results in our study are a little inferior to several studies stated above, the slight difference of prediction results between different studies has beyond what a consumer may perceive for SSC. Moreover, in postharvest quality sorting and grading, we normally do not need to determine the SSC for each apple exactly; instead, we only need to sort apples into different classes according to their SSC values. So, the RPD values, which are defined as the ratio between the sample standard deviation and RMSEP, were also calculated to measure the ability of a model for classification (Nicolăi et al. 2007). The average RPD value of the MCARS-SPA-PLS models developed using the spectra from 10 calibration and prediction sets was 3.680, which means that the model was good to excellent prediction accuracy and could be used for sorting and grading apple fruits based on their SSC values. These results implied that the PLS calibration method based on the 27 variables selected by MCARS-SPA in this work was reasonable and applicable.

Conclusion

The determination of SSC of apple was studied using visible and near-infrared (Vis/NIR) diffuse transmittance. The ACR

was proposed to investigate the stability of spectra collected in two fruit orientations, with stem-calyx axis being horizontal (T1) and with stem-calyx axis being vertical (T2), which showed that the fruit orientation T1 was better for our designed spectrum measurement system. Then, the performance of PLS calibration methods based on several preprocessing methods was analyzed and compared. Finally, in order to establish an accurate, robust, and simplified model, MCARS, SPA, and their combination were used to select the optimal variables for future online application. It concluded that the MCARS-SPA-PLS model based on the spectra after preprocessing of SG smoothing achieved better results for SSC prediction ($r_{\text{pre}}=0.946$, $\text{RMSEP}=0.527^\circ\text{Brix}$) and MCARS combined with SPA was an effective approach for selecting variables. Moreover, the physicochemical properties of 27 selected variables were discussed to obtain a better interpretation of the calibration model. The overall results indicated that the designed diffuse transmittance spectrum system together with PLS calibration model with 27 effective variables selected by MCARS-SPA method had a potential application for online detection of apple SSC. Future work will be focused on the online detection of apple SSC detection using the diffuse transmittance spectrum system. However, the limitation of our research is that only a small portion of each individual apple was assessed. In order to evaluate the quality of SSC comprehensively, more research is needed to gain more spectral information by increasing spectral measurement portion along the peel to get a more accurate and robust calibration model.

Acknowledgments The authors gratefully acknowledge the financial support provided by Beijing Municipal Natural Science Foundation (No. 6144024).

Conflict of Interest Shuxiang Fan declares that he has no conflict of interest. Zhiming Guo declares that he has no conflict of interest. Baohua Zhang declares that he has no conflict of interest. Wenqian Huang declares that he has no conflict of interest. Chunjiang Zhao declares that he has no conflict of interest.

Compliance with Ethics Requirements This article does not contain any studies with human participants performed by any of the authors.

Informed Consent Informed consent is not applicable in this study.

References

- Andersen CM, Bro R (2010) Variable selection in regression—a tutorial. *J Chemom* 24:728–737
- Araújo MCU, Saldanha TCB, Galvão RKH, Yoneyama T, Chame HC, Visani V (2001) The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom Intell Lab Syst* 57:65–73
- Barnes R, Dhanoa M, Lister SJ (1989) Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl Spectrosc* 43:772–777
- Cai W, Li Y, Shao X (2008) A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom Intell Lab Syst* 90:188–194
- Cen H, He Y (2007) Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends Food Sci Technol* 18:72–83
- Demetriades-Shah TH, Steven MD, Clark JA (1990) High resolution derivative spectra in remote sensing. *Remote Sens Environ* 33:55–64
- Do Trong NN, Erkinbaev C, Tsuta M, De Baerdemaeker J, Nicolai B, Saeys W (2014) Spatially resolved diffuse reflectance in the visible and near-infrared wavelength range for non-destructive quality assessment of ‘Braeburn’ apples. *Postharvest Biol Technol* 91:39–48
- Durand A, Devos O, Ruckebusch C, Huvenne J (2007) Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton–viscose textiles. *Anal Chim Acta* 595:72–79
- Fan G, Zha J, Du R, Gao L (2009) Determination of soluble solids and firmness of apples by Vis/NIR transmittance. *J Food Eng* 93:416–420
- Fan S, Huang W, Guo Z, Zhang B, Zhao C (2015) Prediction of soluble solids content and firmness of pears using hyperspectral reflectance imaging. *Food Anal Methods* 8:1936–1946
- Fu X, Ying Y (2014) Food safety evaluation based on near infrared spectroscopy and imaging: a review. *Crit Rev Food Sci Nutr* (just accepted)
- Fu X, Ying Y, Lu H, Xu H (2007) Comparison of diffuse reflectance and transmission mode of visible-near infrared spectroscopy for detecting brown heart of pear. *J Food Eng* 83:317–323
- Giovanelli G, Sinelli N, Beghi R, Guidetti R, Casiraghi E (2014) NIR spectroscopy for the optimization of postharvest apple management. *Postharvest Biol Technol* 87:13–20
- Helland IS, Næs T, Isaksson T (1995) Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemom Intell Lab Syst* 29:233–241
- Huang H, Yu H, Xu H, Ying Y (2008) Near infrared spectroscopy for on-line monitoring of quality in foods and beverages: a review. *J Food Eng* 87:303–313
- Huang M, Wang Q, Zhang M, Zhu Q (2014) Prediction of color and moisture content for vegetable soybean during drying using hyperspectral imaging technology. *J Food Eng* 128:24–30
- Jamshidi B, Minaei S, Mohajerani E, Ghassemian H (2012) Reflectance Vis/NIR spectroscopy for nondestructive taste characterization of Valencia oranges. *Comput Electron Agric* 85:64–69
- Jie D, Xie L, Fu X, Rao X, Ying Y (2013) Variable selection for partial least squares analysis of soluble solids content in watermelon using near-infrared diffuse transmission technique. *J Food Eng* 118:387–392
- Jie D, Xie L, Rao X, Ying Y (2014) Using visible and near infrared diffuse transmittance technique to predict soluble solids content of watermelon in an on-line detection system. *Postharvest Biol Technol* 90:1–6
- Leiva-Valenzuela GA, Lu R, Aguilera JM (2013) Prediction of firmness and soluble solids content of blueberries using hyperspectral reflectance imaging. *J Food Eng* 115:91–98
- Li H, Liang Y, Xu Q, Cao D (2009) Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal Chim Acta* 648:77–84
- Li J, Huang W, Zhao C, Zhang B (2013) A comparative study for the quantitative determination of soluble solids content, pH and firmness of pears by Vis/NIR spectroscopy. *J Food Eng* 116:324–332

- Li J, Huang W, Chen L, Fan S, Zhang B, Guo Z, Zhao C (2014) Variable selection in visible and near-infrared spectral analysis for noninvasive determination of soluble solids content of ‘Ya’ Pear. *Food Anal Methods* 7:1891–1902
- Liu Y, Ying Y (2005) Use of FT-NIR spectrometry in non-invasive measurements of internal quality of ‘Fuji’ apples. *Postharvest Biol Technol* 37:65–71
- Liu Y, Ying Y, Fu X, Lu H (2007) Experiments on predicting sugar content in apples by FT-NIR technique. *J Food Eng* 80:986–989
- Liu F, Jiang Y, He Y (2009) Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: a case study to determine soluble solids content of beer. *Anal Chim Acta* 635:45–52
- Liu Y, Sun X, Ouyang A (2010) Nondestructive measurement of soluble solid content of navel orange fruit by visible–NIR spectrometric technique with PLSR and PCA-BPNN. *LWT-Food Sci Technol* 43:602–607
- Liu D, Sun D-W, Zeng X-A (2014) Recent advances in wavelength selection techniques for hyperspectral image processing in the food industry. *Food Bioprocess Technol* 7:307–323
- Lu R (2004) Multispectral imaging for predicting firmness and soluble solids content of apple fruit. *Postharvest Biol Technol* 31:147–157
- Mendoza F, Lu R, Cen H (2014) Grading of apples based on firmness and soluble solids content using Vis/SWNIR spectroscopy and spectral scattering techniques. *J Food Eng* 125:59–68
- Nicolai BM, Beullens K, Bobelyn E, Peirs A, Saeys W, Theron KI, Lammertyn J (2007) Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review. *Postharvest Biol Technol* 46:99–118
- Peirs A, Lammertyn J, Ooms K, Nicolai BM (2001) Prediction of the optimal picking date of different apple cultivars by means of VIS/NIR-spectroscopy. *Postharvest Biol Technol* 21:189–199
- Peng Y, Lu R (2007) Prediction of apple fruit firmness and soluble solids content using characteristics of multispectral scattering images. *J Food Eng* 82:142–152
- Perkins J, Tenge B, Honigs D (1988) Resolution enhancement using an approximate-inverse Savitzky-Golay smooth. *Spectrochim Acta B* 43:575–603
- Qing Z, Ji B, Zude M (2007) Wavelength selection for predicting physicochemical properties of apple fruit based on near-infrared spectroscopy. *J Food Quality* 30:511–526
- Savitzky A, Golay MJ (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem* 36:1627–1639
- Schmutzler M, Huck CW (2014) Automatic sample rotation for simultaneous determination of geographical origin and quality characteristics of apples based on near infrared spectroscopy (NIRS). *Vib Spectrosc* 72:97–104
- Smith BC (2003) *Quantitative spectroscopy: theory and practice*. Elsevier Science, Academic Press, pp. 66
- Sun T, Lin H, Xu H, Ying Y (2009) Effect of fruit moving speed on predicting soluble solids content of ‘Cuiguan’ pears (*Pomaceae pyrifolia* Nakai cv. Cuiguan) using PLS and LS-SVM regression. *Postharvest Biol Technol* 51:86–90
- Wang A, Xie L (2014) Technology using near infrared spectroscopic and multivariate analysis to determine the soluble solids content of citrus fruit. *J Food Eng* 143:17–24
- Xiaobo Z, Jiewen Z, Xingyi H, Yanxiao L (2007) Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of ‘Fuji’ apple based on different PLS models. *Chemom Intell Lab Syst* 87:43–51
- Xie L, Ye X, Liu D, Ying Y (2011) Prediction of titratable acidity, malic acid, and citric acid in bayberry fruit by near-infrared spectroscopy. *Food Res Int* 44:2198–2204
- Xu H, Qi B, Sun T, Fu X, Ying Y (2012) Variable selection in visible and near-infrared spectra: application to on-line determination of sugar content in pears. *J Food Eng* 109:142–147
- Yun YH et al (2014) A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration. *Anal Chim Acta* 807:36–43
- Zhang L, Xu H, Gu M (2014) Use of signal to noise ratio and area change rate of spectra to evaluate the Visible/NIR spectral system for fruit internal quality detection. *J Food Eng* 139:19–23
- Zhu Q, Huang M, Zhao X, Wang S (2013) Wavelength selection of hyperspectral scattering image using new semi-supervised affinity propagation for prediction of firmness and soluble solid content in apples. *Food Anal Methods* 6:334–342