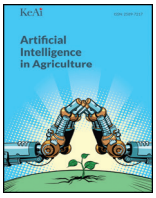
Artificial Intelligence in Agriculture 4 (2020) 48–57

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Nondestructive determining the soluble solids content of citrus using near infrared transmittance technology combined with the variable selection algorithm

Xi Tian [a,b], Jiangbo Li [b], Shilai Yi [c], Guoqiang Jin [d], Xiaoying Qiu [e], Yongjie Li [d,*]

[a] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
[b] Beijing Research Center of Intelligent Equipment for Agriculture, Beijing 100097, China
[c] Citrus Research Institute, Southwest University, Chongqing 400712, China
[d] Co-innovation Center of Citrus Industrial Technology in Linhai, Linhai 31700, China
[e] Station of Popularizing Speciality Technology in Linhai, Linhai 31700, China

ARTICLE INFO

Article history:
Received 22 April 2020
Received in revised form 6 May 2020
Accepted 6 May 2020
Available online 11 May 2020

Keywords:
Full transmittance spectrum
Spectral preprocessing
Thick-skin fruits
Soluble solids content
Variable selection algorithm

ABSTRACT

Nondestructive determination the internal quality of thick-skin fruits has always been a challenge. In order to investigate the prediction ability of full transmittance mode on the soluble solid content (SSC) in thick-skin fruits, the full transmittance spectra of citrus were collected using a visible/near infrared (Vis/NIR) portable spectrograph (550–1100 nm). Three obvious absorption peaks were found at 710, 810 and 915 nm in the original spectra curve. Four spectral preprocessing methods including Smoothing, multiplicative scatter correction (MSC), standard normal variate (SNV) and first derivative were employed to improve the quality of the original spectra. Subsequently, the effective wavelengths of SSC were selected from the original and pretreated spectra with the algorithms of successive projections algorithm (SPA), competitive adaptive reweighted sampling (CARS) and genetic algorithm (GA). Finally, the prediction models of SSC were established based on the full wavelengths and effective wavelengths. Results showed that SPA performed the best performance on eliminating the useless information variable and optimizing the number of effective variables. The optimal prediction model was established based on 10 characteristic variables selected from the spectra pretreated by SNV with the algorithm of SPA, with the correlation coefficient, root mean square error, and residual predictive deviation for prediction set being 0.9165, 0.5684°Brix and 2.5120, respectively. Overall, the full transmittance mode was feasible to predict the internal quality of thick-skin fruits, like citrus. Additionally, the combination of spectral preprocessing with a variable selection algorithm was effective for developing the reliable prediction model. The conclusions of this study also provide an alternative method for fast and real-time detection of the internal quality of thick-skin fruits using Vis/NIR spectroscopy.

1. Introduction

Fruits quality characteristics, including external quality (such as color, size, and shape) and internal quality (such as soluble solids, titratable acid, and firmness), are the basic factors affecting the consumers' desire of purchasing (Nghia et al., 2014). Nowadays, consumers' pursuit of fruits quality is not only limited to external quality, but also extends to internal quantity. Citrus is planted widely in the world for its high nutrition and unique flavor. Soluble solids content (SSC) is one the most crucial internal attribute determining consumers' purchasing desire also plays an important role in guiding orchard management. However, the traditional testing technology of SSC is destructive, time-consuming and high energy-consumption, which is not suitable for the demand of fast and large amounts of fruits grading industry (Ireri et al., 2019).

Nondestructive testing technology has the advantages of real-time, low energy consumption and minimal sample preparation. In recent years, it has been extensively studied and accepted as an effective method for the detection of fruits maturity, shelf life, quality sorting, and even internal disease. In general, nondestructive testing technologies include electronic nose (Brezmes et al., 2001; Saevels et al., 2003; Sanaeifar et al., 2014), visible/near infrared (Vis/NIR) spectroscopy (Sinelli et al., 2008; Tian et al., 2018; Li et al., 2020), ultrasonic sensing technique (Kim et al., 2009; Aboudaoud et al., 2012), machine vision sensing technique (Blasco et al., 2007; Elmasry et al., 2009; Lopezgarcia et al., 2010) and the fusion of multiple non-destructive sensing methods (Hong and Wang, 2014; Mendoza et al., 2012). The

* Corresponding author at: Co-innovation Center of Citrus Industrial Technology in Linhai, 219# dongfang avenue, Linhai 31700, China.
E-mail address: lyonjie@sina.com (Y. Li).

https://doi.org/10.1016/j.aiia.2020.05.001

Vis/NIR spectroscopy technology provides a promising approach for the quality detection of agricultural products due to its low implementation cost, high detection accuracy, and fast detection speed.

The Vis/NIR spectrum belongs to the high-energy vibrational spectrum performed in the range of 400–2500 nm. When Vis/NIR electromagnetic radiation reaches the fruits, the difference of physical features (such as the skin thickness and firmness) and internal chemical compositions (such as SSC and titratable acid) of fruits will perform various absorption or reflection spectrum at a different wavelength (Pasquini, 2018). Generally, three different spectra acquisition modes including diffuse reflectance, interactive (or semi-transmittance), and full transmittance modes are commonly used according to the difference of optical property of tested fruits. The diffuse reflectance spectrum is mainly associated with the superficial layer within about 4 mm (Lammertyn et al., 2000), rather than the internal tissue of fruits. Additionally, the specular reflection caused by the spherical shape of fruits could easily interference with the reflectance spectra, resulting in a poor prediction model (Fu et al., 2007). Citrus belongs to thick-skin fruits, the spectra information collected from a superficial layer using diffuse reflectance mode mainly relates with citrus peel, which is indirectly related to the flesh tissue. The spectrograph and the lamp locate opposite both sides of the tested fruit for directly acquiring the internal component information in full transmittance mode (Cayuela, 2008). Therefore, the full transmittance mode may assist in getting a better result for the prediction of SSC in thick-peel fruits (Wang et al., 2014).

Furthermore, citrus is a complicated natural product composed of a great variation of external and internal structures, the SSC and skin thickness are significant from proximal to distal of fruits, as well as the peel color in sunlit and shaded also. Diffuse reflectance spectrum is mainly collected from local position or superficial layer region, the non-uniform distribution of chemical components would weak the prediction accuracy of intact fruit. Hence the full transmittance mode can overcome the inhomogeneous condition to some extent through acquiring more internal tissue information. Dull and Birth (1989) argued that the peel in the light path would not change the fundamental dependencies between the spectrum data and chemical components. Therefore, full-transmittance mode maybe is a promising technology for the determination of SSC with high accuracy and robust in citrus.

The Vis/NIR spectrum usually has hundreds or thousands of wavelengths, the developed prediction model would be too complicated if all spectra wavelengths are used directly. Moreover, some irrelevant information with the target component exists in original variables, which could weaken the prediction ability of the developed model (Wu et al., 2010; Xu et al., 2012). Thus, variable selection algorithms, such as Monte Carlo based uninformative variable elimination (MC-UVE), genetic algorithm (GA), competitive adaptive reweighted sampling (CARS), successive projection algorithm (SPA) and random frog (RF), are usually applied to extract effective variables from full wavelengths. Those 'uninformative' variables that cannot enhance the overall performance of the prediction model are eliminated, while those 'informative' variables that correlate to interested parameters are remained through variable selection algorithm performing (Pasquini, 2018). Fan et al. (2015) developed a better prediction model of SSC and firmness of pear based on the 46 and 41 variables selected from full variables using CARS method, respectively. Zhang et al. (2018) found the prediction model of sugar in pear built by effective wavelengths was faster than the full wavelength. In a word, the selected variables contribute to building a simpler and faster regression model by revealing the relationship between the original spectra and the target property.

The variable selection methods have been widely applied to predict the internal quality of thin-skin fruits such as pears (Fan et al., 2015; Zhang et al., 2018) and apples (Tian et al., 2019; Zhang et al., 2019), but there have been few reports on thick-skin fruits such as citrus. The thicker skin and complex structure of citrus interfere with the light radiation to acquire the information of internal tissue, and increase the difficulty of effective variables selection and prediction model development, resulting in the internal quality prediction of intact citrus developed slowly (Krivoshiev et al., 2000; Fraser et al., 2003). Therefore, the specific objectives of this study were to: (1) investigate the feasibility of full transmittance mode on the prediction of SSC in thick-skin fruits; (2) compare the performance of different spectral preprocessing methods; (3) analyze the distribution of effective variables selected by different variable selection algorithms; (4) compare the accuracy of prediction models built by the combination of different spectral preprocessing methods and variable selection algorithms.

2. Material and methods

2.1. Sample preparation

A total of 224 'Gannan' navel orange fruit without any apparent defects were purchased from a local fruit supermarket in Beijing, China. Then all the samples were stored in the laboratory within 24 h at 21 °C and relative humidity of 65% to reduce the effect of temperature on the prediction results (Li et al., 2020). Kennard-Stone is the most widely used algorithm for sample assignation due to its excellent performance of ensuring the representativeness of calibration and prediction sets (Barbin et al., 2013; Fan et al., 2016). Therefore, a subset of 168 samples was selected as calibration set for developing the calibration model, and the remaining 56 were used as prediction set to verify the performance of developed model according to a proportion of 3:1 with the algorithm of Kennard-Stone, respectively.

2.2. Spectra collection

The transmittance spectra of citrus were collected using a commercial Vis/NIR portable spectrophotometer covering 550-1100 nm with an interval of 0.45 nm (AvaSpec-ULS2048XL-EVO, Inc., Apeldoorn, Netherlands). The optimal integration time was set at 10 ms for obtaining appropriate spectral intensity. The sample was placed manually on the platform between the spectrometer and the light source with the orientation of the stem-calyx axis vertical (Fig. 1), three separate spectra were measured around the equator (120°) for each fruit. Therefore, a total of 672 transmittance spectra were collected from equator positions for 224 fruit.

2.3. Quality attributes

After spectra collecting, SSC of each fruit was measured immediately using the traditional destructive method. The juice was squeezed from the overall fruit and then dropped 1 ml onto the fruit sugar refractometer (PAL-1, ATAGO, Japan) after stirring.

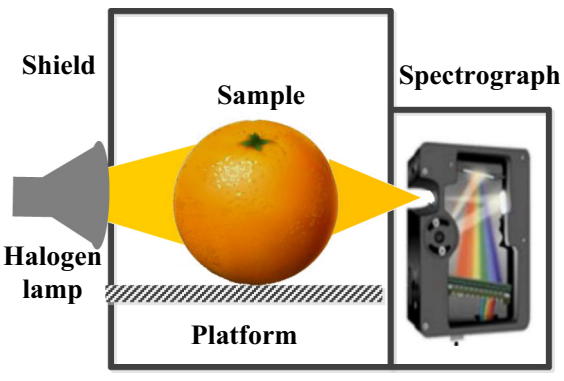


Fig. 1. The schematic of full transmittance spectra measurement system.

2.4. Spectral data preprocessing

Some irrelevant information including noise, baseline translation and stray light caused by instrument or external environment affects the identification of valuable spectra information in the process of modeling (Liu et al., 2010; Zhang et al., 2019). Spectra preprocessing methods are helpful to develop a reliable and robust model by eliminating the disturbances of irrelevant information (Magwaza et al., 2012). Therefore, four kinds of spectra preprocessing methods, including Smoothing (9 points), multiplicative scatter correction (MSC), standard normal variate (SNV), and the first derivative was used to process the original spectra data. The Smoothing could eliminate random noise or high-frequency random error to improve the signal-to-noise ratio. MSC was designed to separate scattering spectral signals from chemical information in the original spectra using mathematical methods. SNV was usually used to eliminate the effect of solid particle size and optical path change in the original spectra (Barnes et al., 1989), first derivative was effective for reducing baseline shifts and superposed peaks (Tian et al., 2018). The spectra preprocessing procedure was performed in the Unscramblerv 9.7 (CAMO PROCESS AS, Oslo, Norway).

2.5. Effective variable selection

Effective variable extraction can decrease spectral dimension and simplify the process of modeling, additionally, the relationship between the selected variable and the interested parameters also can be clearly explained. In this study, three variable selection algorithms including SPA, CARS, and GA were employed to select the effective wavelength of SSC of citrus from full wavelengths.

SPA starts from one wavelength and calculates its projection onto the unselected wavelength in each cycle. The wavelength corresponding to the maximum projection value is considered to be the initial value of the next cycle until the number of wavelengths set. The new variable selected by SPA is the maximum projected value of the previously selected variable. The SPA is employed to obtain these variables with small collinearity and the minimum redundant information, the number of variables used in modeling is greatly reduced, and the accuracy and speed of the prediction model are improved (Liu et al., 2014). In the process of variable optimization, different MLR calibration models are established based on different variable subsets. The variable subset with the lowest value of root mean-square error of cross validation (RMSECV) will be considered as the optimal variables. More details on the steps involved in SPA application can be found in previous studies (Araújo et al., 2001; Galvão et al., 2008).

CARS is a novel algorithm proposed by Li et al. (2009), in order to retain the wavelength variables with a large absolute value of regression coefficient in the PLS model, CARS employs the adaptive weighted sampling technique to evaluate the importance of each variable, N variable subsets are obtained through repeated screening, finally, the variables closely related to the measured components are extracted after cross validation for each variable subset. According to the importance level of each wavelength, the Monte Carlo (MC) sampling is used to randomly select *N* sample subsets with the principle of iterative and competitive for developing *N* PLS calibration modes, and the variable subsets with the lowest RMSECV are defined as the key variables. As a new variable selection method, the CARS performs better when dealing simultaneously with uninformative and collinear variables (Zhang et al., 2018). Details of the CARS methodology could be referred to the previous literature (Li et al., 2009; Wu and Sun, 2013).

GA is an adaptive global optimization search algorithm developed by natural selection and evolution, which realizes the recombination and iterative optimization of individual structures in a population through crossover, selection, and variation. When the iteration is completed, the frequency selected by the wavelength variable is modeled one by one from high to low. The detailed process of the GA procedure is available in the study of Leardi and Gonzalez (1998). GA has been successfully used for multivariate correction analysis, such as GA-PLS (Andersen and Bro, 2010), GA-least squares support vector machine (LS-SVM) and GA-artificial neural network (ANN) (Zhang et al., 2018) indicating that GA is useful for effective variable screening and adapt to different modeling techniques.

2.6. Multivariate model developing

The linear PLS model is the most widely applied and robust algorithm for its insensitive to collinear variables and tolerant to large numbers of variables. PLS could explain the correlation between the original spectrum and the target properties using the no more than 20 latent variables (LVs) extracted from the original wavelength, so it is commonly employed to analyze the internal content of fruit. In this research, the PLS models were developed based on the full wavelengths spectra and the effective variables selected by SPA, CARS, and GA, then the performance of all developed models was compared for evaluating the feasibility of full transmittance model on the prediction of SSC in thick-skin fruits. The root mean square error of calibration set (RMSEC) and prediction set (RMSEP), the correlation coefficient of calibration set (R_{cal}) and prediction set (R_{pre}), were selected as the evaluation parameters for these PLS models. A good prediction model should have a higher R_{cal} and R_{pre}, and a lower RMSEC and RMSEP. The higher the R_{pre} and the smaller the RMSEP, the stronger the prediction ability of the model is. The residual predictive deviation (RPD) also was introduced to further evaluate the performance of the prediction results. RPD is the standard deviation of the predicted sample reference data divided by the predicted standard error (SEP), which provides the standardization for SEP. Generally, the RPD value is divided into three levels to define the performance of prediction models, 1.5 to 2 indicates that the model can distinguish high or low values of the target variable, 2 to 2.5means that rough quantitative prediction can be made, and an excellent prediction accuracy would occur on 2.5 to 3 or above (Li et al., 2014a).

3. Results and discussion

3.1. SSC of sample

Table 1 shows the overview of SSC for 224 samples, it could be seen that the SSC ranges were 7.00 to 14.6°Brix and 7.80 to 13.00°Brix with standard deviations (S.D.) of 1.50 and 1.45°Brix for calibration and the prediction sets, respectively. It is worth noting that the range of calibration set was wider than prediction set, which is helpful to build a robust model for SSC prediction of citrus fruit.

3.2. Spectral features and preprocessing

Only the spectral range of 600–1000 nm (928 bands) was used to build prediction model in this study due to the lower signal to noise ratio below 600 nm and up 1000 nm. The mean value of three spectra measured from the same one fruit is used as the transmittance spectrum of each fruit. The transmittance spectra of all 224 fruits are shown in Fig. 2. Three significant absorption peaks around 710, 810, and 915 nm are found from original transmittance spectra, respectively. 710 nm is well known as the red spectral region associated with

Table 1
Statistic results of SSC (°Brix) in different data sets.

Data sets	No. of samples	Min.	Max.	Mean	S.D.
Calibration set	168	7.00	14.60	10.57	1.50
Prediction set	56	7.80	13.00	10.42	1.45
Total	224	7.00	14.60	10.53	1.47

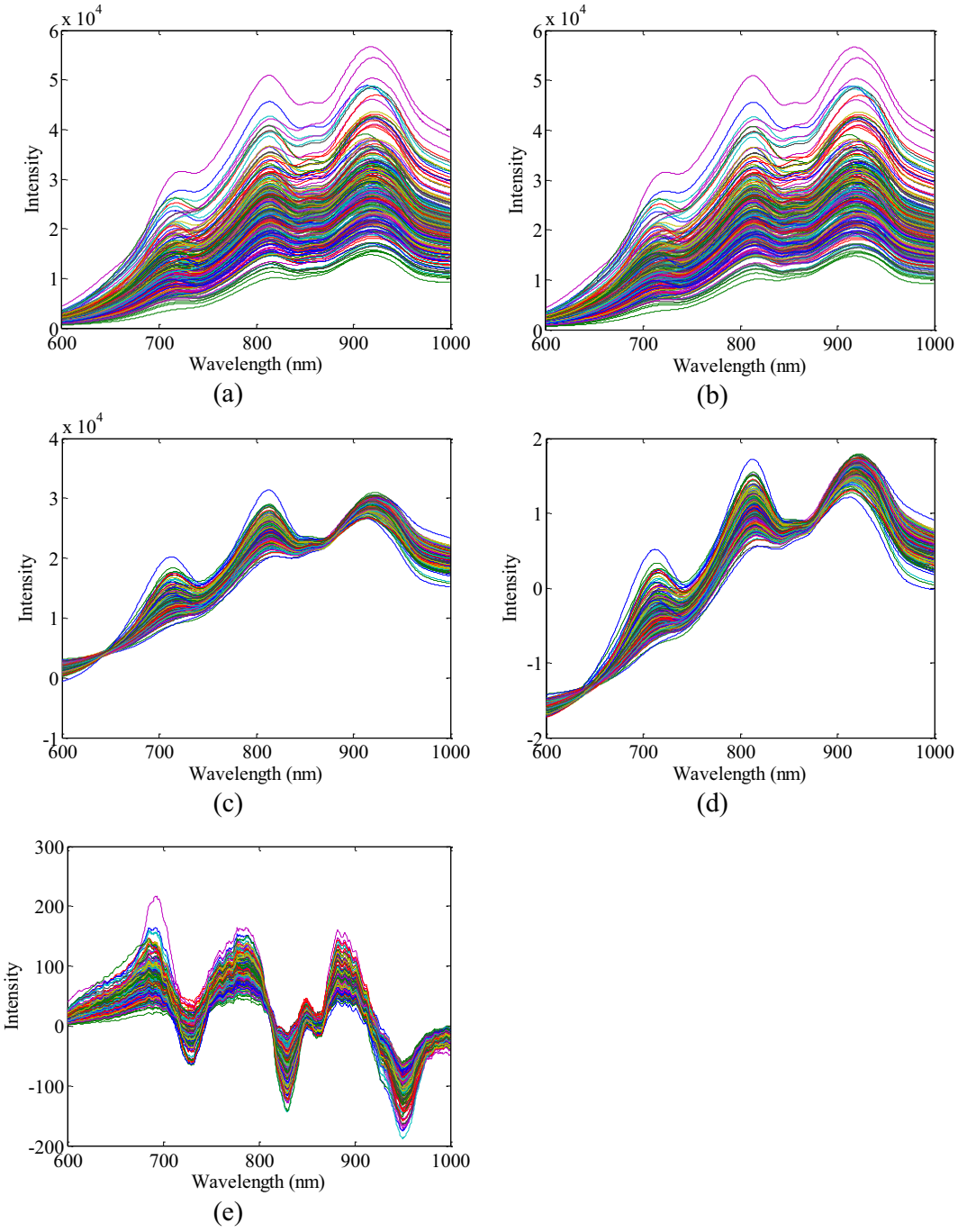


Fig. 2. The original spectra (a) and preprocessed spectra of Smoothing (b), MSC (c), SNV (d) and first derivative (e).

maturity, which can be used to determine the chlorophyll content of fruit (Zude-Sasse et al., 2002). Water is the major component of citrus, 810 nm is mostly triggered by the third overtone stretch of O—H in water (Martinsen and Schaare, 1998). 915 nm is mostly related to the third overtones of C—H in sugar (Guthrie et al., 2005). The absorbance spectrum stayed relatively flat in the range of 680–910 nm for diffuse reflectance mode of citrus (Gómez et al., 2006; Wang et al., 2014), however, three absorption peaks were found around 710, 810 and 915 nm in transmittance spectrum of citrus, which were less reported. Therefore, these characteristic peaks may contribute to predict the internal quality of citrus fruit.

Four kinds of preprocessing methods including Smoothing, MSC, SNV, and first derivative were applied in this research. It could be found from Fig. 2, the spectra curves treated by MSC and SNV were more flat than the original and other preprocessed spectra. MSC and SNV were effective to highlight the peak and valley of the spectra profile and balance the distribution of variables by correcting the difference of optical path. However, the Smoothing method had no significant effect on noise reduction, and the first derivative method seems to greatly amplify the noise. In conclusion, MSC and SNV performed better on noise reduction and simultaneously emphasized the spectral feature, both methods might be optimal for improving the transmittance spectral quality of citrus.

3.3. Prediction models build by full wavelength

Table 2 shows the performance of prediction models based on full wavelengths data of original and processed spectra. It could be seen

Table 2
Results of the prediction models based on the full wavelengths spectra pretreated by different methods.

Pretreatment methods	LVs	Calibration set		Prediction set		
		R_{cal}	RMSEC (°Brix)	R_{pre}	RMSEP (°Brix)	RPD
None	12	0.9105	0.7609	0.9034	0.6163	2.3169
Smoothing	12	0.9088	0.7613	0.9033	0.6168	2.3148
MSC	13	0.9439	0.5924	0.9089	0.5966	2.3935
SNV	14	0.9523	0.5554	0.9215	0.5546	2.5746
First derivative	9	0.9062	0.7871	0.8518	0.7419	1.9248

that the Smoothing had no significant difference with original spectra because their latent variables (LVs), R_{pre}, RMSEP and RPD values were close, indicating that Smoothing unable to remove undesirable signal interference in original spectra. The first derivative is not applicable to preprocess the original transmittance spectra of citrus, because the performance turned worse than the original spectra, the result was similar to Liu et al. (2010) and Liu et al. (2015). The best performance of the prediction model was derived from the pretreated spectra of SNV with R_{pre} of 0.9215, RMSEP of 0.5546°Brix, and RPD of 2.5746, respectively. The closer RMSEC and RMSEP value, the more stable the model is. The absolute difference between RMSEC and RMSEP were 0.0008 and 0.0042°Brix for the prediction models built by pretreated spectra of SNV and MSC, respectively, which were smaller than that of original and other pretreated spectra, suggesting that the method of SNV or MSC could significantly improve the robustness of the prediction model. Overall, SNV performed the best performance on eliminating undesirable interference of original spectra and obtained the best model for predicting SSC of citrus fruit.

3.4. Prediction models build by effective variables

To simplify the prediction model and improve the prediction speed, three variable selection algorithms including SPA, CARS, and GA were applied to select the effective wavelength of SSC from full wavelengths. Taking spectra pretreated by smooth to describe detail the procedure of variable selection of SPA, CARS, and GA.

3.4.1. Effective variable selection of SPA

Fig. 3a shows the curves of the RMSECV changing with the number of selected variables in the SPA. The RMSECV value reduced rapidly firstly and then gradually decreased with the growth of selected variables. When the RMSECV obtained the optimal value, 16 variables were selected to predict the SSC of citrus (marked in an open red square). Fig. 3b shows the distribution map of the 16 selected variables (marked in an open red square), included 622.62, 648.53, 664.53, 676.94, 706.92, 732.34, 746.74, 758.92, 787.07, 813.76, 831.76, 847.13, 878.16, 906.47, 964.72 and 981.37 nm. It could be found that the 16 selected variables almost evenly distributed in the range of 600–1000 nm. It was worth noting that half of the variables belonged to the Vis region (400–780 nm), indicating that the color variances was helpful to evaluate the SSC in citrus accurately, the result agreed with previous studies for apple (Tian et al., 2019) and pear (Li et al., 2014a). Comparing the distribution of effective variables selected by different preprocessing methods in Fig. 6 (vertical solid red lines), 17, 22, 10 and 25 effective variables were selected as effective variables from full wavelengths (928 bands) for original, MSC, SNV and first derivative spectra, respectively. The number of effective variables varying with the difference of preprocessing methods, proving that the spectral preprocessing has a significant influence on the results of variable selection.

3.4.2. Effective variable selection of CARS

Before the CARS operating, the number of Monte Carlo sampling runs was set to 50 and the 10-fold cross validation was used to choose the optimal variable for SSC prediction (Fan et al., 2015). It could be seen from Fig. 4a, the number of sampled variables decreases sharply at an exponential rate with the number increasing of sampling runs firstly. Subsequently, the selection speed of sampled variables gradually slows down when the number of sampling runs reaches 15 (marked in an open red square), which reflects the process from rough selection to fine selection in the CARS algorithm. Fig. 4b shows the transformation trends of RMSECV value with the sampling runs, further describing the elimination procedure of useless variables. The RMSECV value firstly reduces significantly due to the elimination of useless variables and then increases with the loss of some characteristic wavelengths of SSC. The RMSECV reaches the minimum or optimal value when the sampling runs are 15. Fig. 4c shows the regression coefficient path of PLS modeling for each variable, each line records the change of regression coefficient at different sampling runs. These variables with a larger absolute coefficient would be selected from each sampling runs to form a subset that is also considered as the subset of effective wavelengths (marked by the vertical blue star line). Finally, 161 characteristic variables were extracted from the full wavelengths and would be set as the inputs to

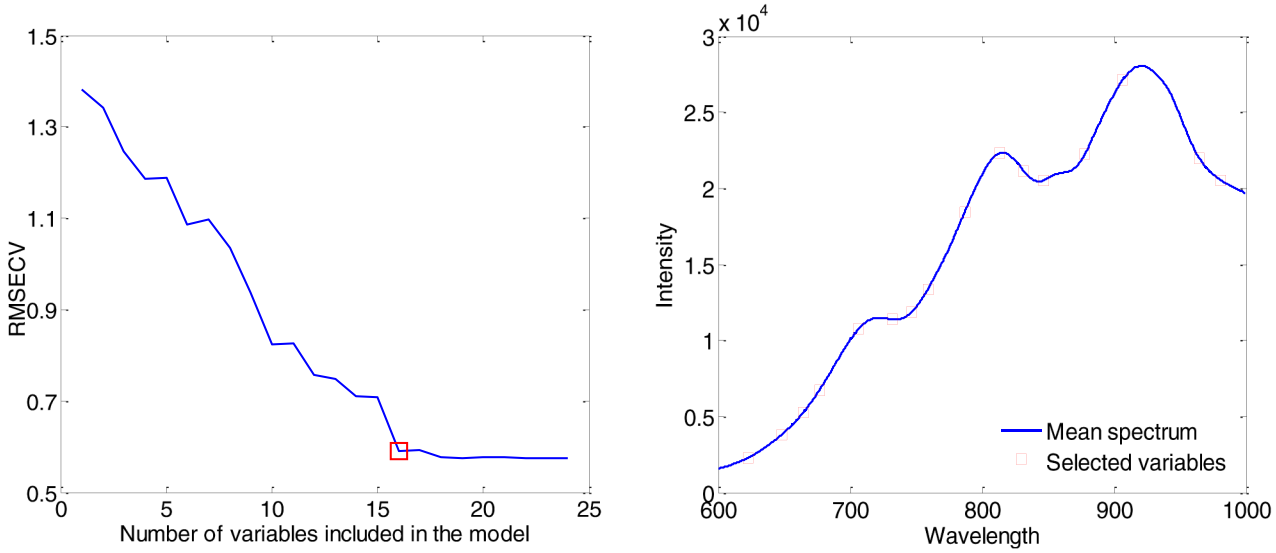


Fig. 3. RMSEP plot (a) and the distribution map of selected variables (b) in SPA algorithm.

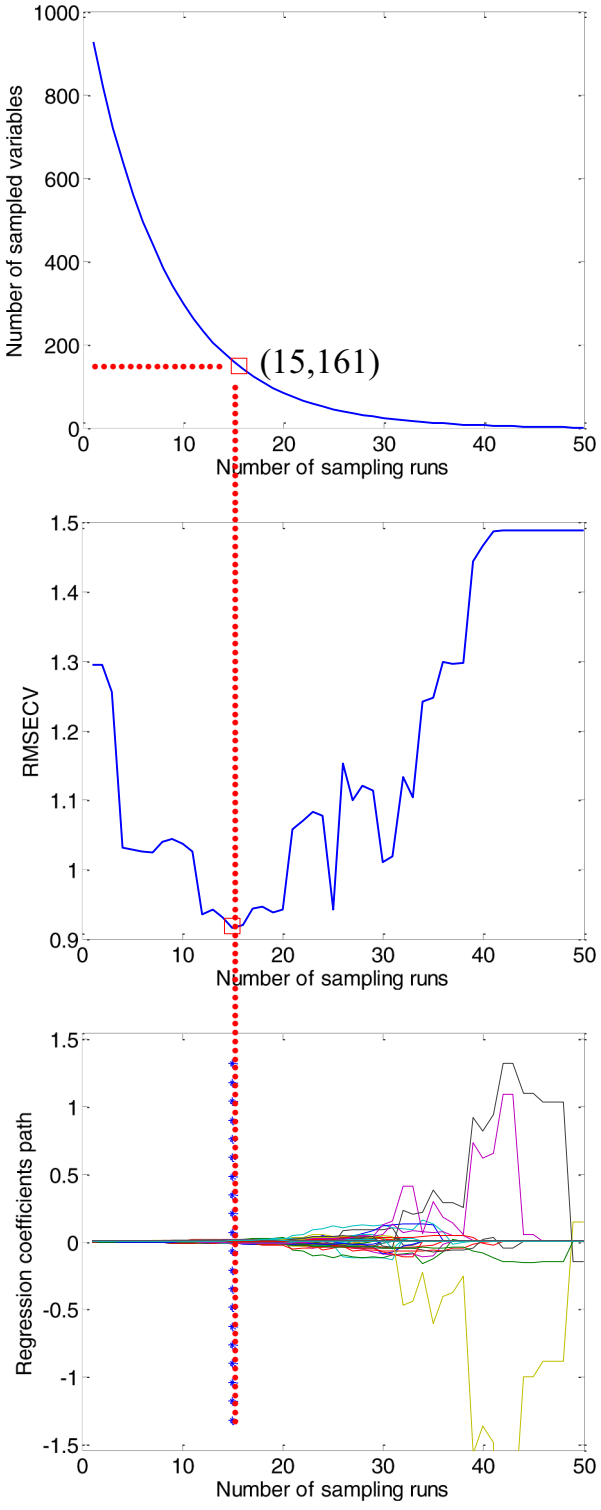


Fig. 4. Changing trend of the number of sampled wavelengths (a), 10-fold RMSECV values (b) and regression coefficients of each wavelength (c) with the increasing of sampling runs. The line (marked by red dot) denotes the optimal point where 10-fold RMSECV values are the lowest (161 bands were selected when the number of sampling runs reaches 15).

develop the model of citrus SSC, which accounted for 17.3% of the 928 bands.

The same procedure also was applied to the original and other pretreated spectra. 97, 125, 97 and 110 bands were selected as effective wavelengths for the original, Smoothing, MSC, SNV and first derivative

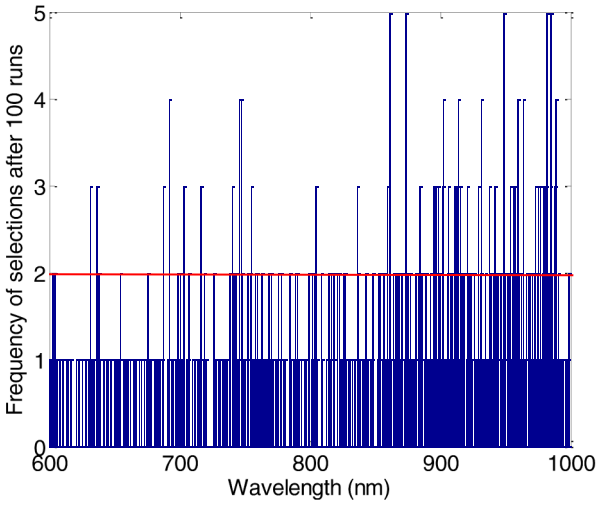


Fig. 5. Frequency of variable selection after 100 runs by GA algorithm.

spectra, and the distribution of these selected bands are marked by the vertical solid blue lines below the curves in Fig. 6. The number of the selected variable by CARS was significantly larger than that of SPA, which mainly caused by the difference of principle within different variable selection algorithms.

3.4.3. Effective variable selection of GA

The parameter of GA was set as follows before calculation: the initial population size was 50, 5 variables as the chromosome, the crossover probability was 0.5, and the variation probability was 0.01, the backward progressive iteration times was 100 (Zhang et al., 2018). During the implementation process of the GA, due to the randomness of the selection of the initial population, the selection of genetic operators, the crossover and mutation process, those variables selected at high frequency would have a certain consistency. Fig. 5 shows the selected frequency for each variable after 100 runs, the variables selected more than two times would be considered as the effective wavelength to build the prediction models, therefore 170 variables were selected by GA for the spectra pretreated by Smoothing, which accounted for 18.3% of the full wavelengths. 181, 194, 198, and 141 bands were selected as effective wavelengths for the original, MSC, SNV and first derivative spectra, and their distribution maps (marked by the black dots on the spectrum curve) can be seen in Fig. 6. It was worth mentioning that the selected bands by GA almost uniformly distributed throughout the range of 600–1000 nm.

Comparing the results of different variable selection methods, the number of variables selected by SPA was significantly lower than CARS and GA. The distribution of selected variables varied with the difference of variable selection algorithms. The difference in size and distribution of selected variables is mainly caused by the difference in the selecting mechanism of algorithms and the collinearity of pretreated spectra (Li et al., 2014b). SPA is designed to obtain variables with small collinearity preferentially, caused the number and correlation of selected variables to be less and lower, respectively. CARS was firstly used to eliminate the uninformative variables existed in full spectra (Fan et al., 2015), however, GA is an algorithm performing the searching principle of random and global, it still carries out the global search for avoiding local minimization in the later iteration period, leading to some redundancy variables were often included in the optimal subset (Durand et al., 2007).

3.4.4. Prediction models built by effective wavelength

In order to evaluate the performance of the effective variables selected by SPA, CARS, and GA, the selected variables were used to develop

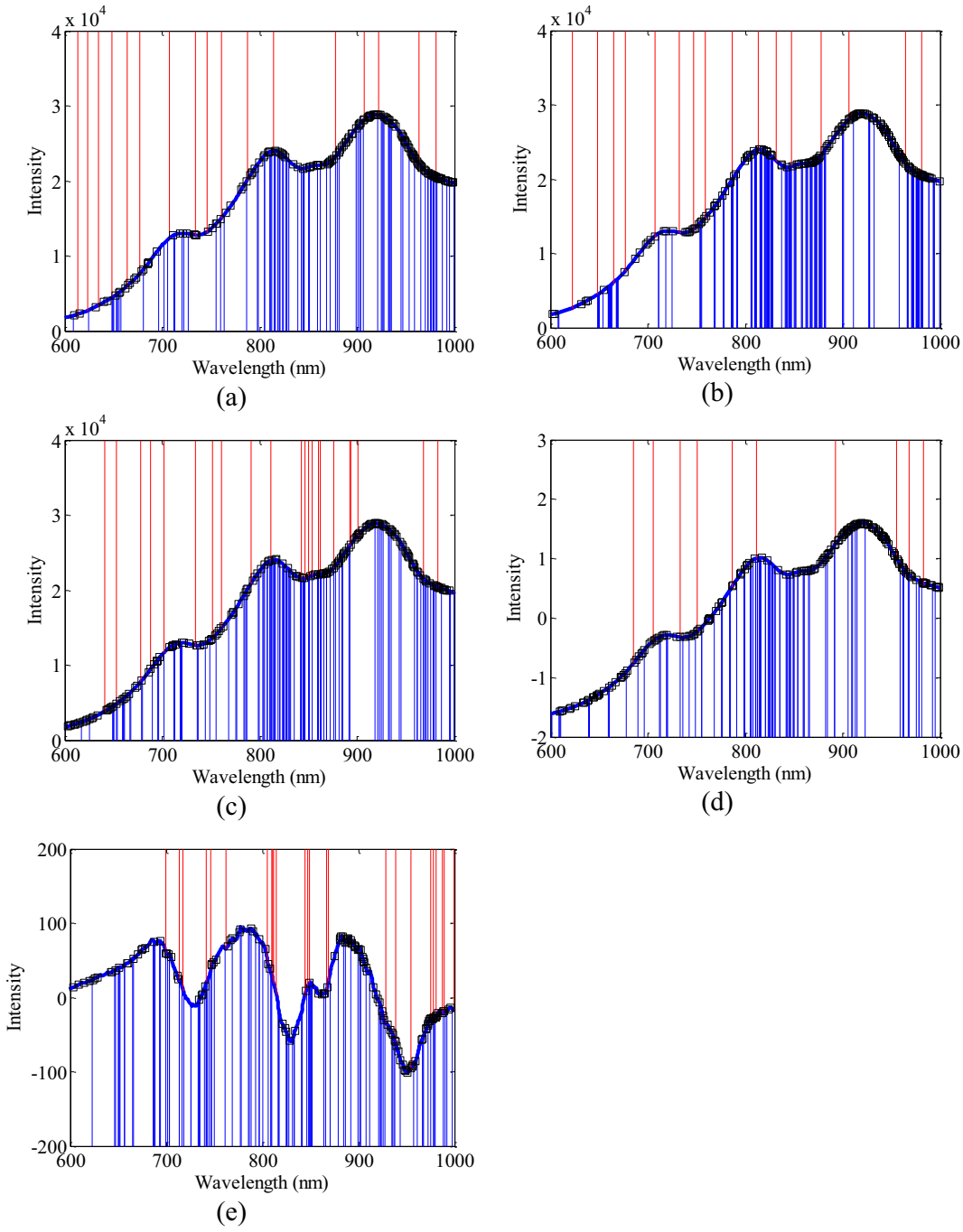


Fig. 6. The distribution of effective variables selected by SPA, CARS and GA algorithms based on original (a), smoothing (b), MSC (c), SNV (d) and first derivative (e) spectra. —— SPA; —— CARS; –□– GA.

the prediction models and Table 3 shows the performance of all developed models. Comparing with the CARS or GA, the variables selected by SPA performed the best performance for almost all models except the spectra pretreated by first derivative. The SNV-SPA performs best based on only 10 characteristic variables with R_{pre} of 0.9165, RMSEP of 0.5684°Brix, and RPD of 2.5120, respectively. It is also worth noting that models developed by SNV-GA also gave an acceptable accuracy considering the R_{pre}, RMSEP and RPD were 0.9164, 0.5776°Brix and 2.4271, but 198 characteristic variables were selected as inputs. It indicated that GA could well extract the useful informative from the spectra pretreated by SNV, to build a better prediction model. The variables selected by CARS yielded worst prediction results, and the difference between RMSEC and RMSEP also was greater than that of SPA and GA, indicating that some redundant and uninformative variables were still retained in the effective spectra selected by CARS.

In the models built by full wavelengths, the best performance was derived from the pretreated spectra of SNV with R_{pre} of 0.9215, RMSEP of 0.5546°Brix and RPD of 2.5746, respectively. Although the model built by the SNV-SPA (R_{pre} = 0.9165, RMSEP = 0.5684°Brix, RPD = 2.5120) was slightly inferior to the full wavelengths model, the number of characteristic variables of the SNV-SPA model decrease by 98.9%. Therefore, the elimination of redundant information can reduce the complexity and increase the performing speed while maintaining the determination accuracy. The scatter plots of measured versus

Table 3
Results of the prediction models based on effective bands selected by different algorithms.

Variable selection algorithms	Pretreatment methods	LVs	No. of selected bands	Calibration set		Prediction set		
				R_{cal}	RMSEC (°Brix)	R_{pre}	RMSEP (°Brix)	RPD
SPA	None	14	17	0.8963	0.7845	0.9162	0.5770	2.4747
	Smoothing	13	16	0.8971	0.7683	0.9148	0.5802	2.4612
	MSC	14	22	0.9355	0.6048	0.8988	0.6211	2.2990
	SNV	**10**	**10**	**0.9133**	**0.6468**	**0.9165**	**0.5684**	**2.5120**
	First derivative	8	25	0.8774	0.7835	0.8494	0.7529	1.8964
CARS	None	15	97	0.9749	0.4989	0.8279	0.8220	1.7371
	Smoothing	15	161	0.9524	0.6254	0.8108	0.8559	1.6682
	MSC	14	125	0.9797	0.4540	0.8635	0.7599	1.8790
	SNV	14	97	0.9817	0.3888	0.8526	0.7894	1.8088
	First derivative	15	110	0.9924	0.3409	0.7895	0.9608	1.4861
GA	None	14	181	0.9186	0.7657	0.8947	0.6433	2.2196
	Smoothing	12	170	0.8981	0.7869	0.8536	0.7401	1.9292
	MSC	13	194	0.9438	0.5955	0.9053	0.6112	2.3364
	SNV	14	198	0.9508	0.5733	0.9164	0.5776	2.4721
	First derivative	11	141	0.9239	0.8023	0.8643	0.7168	1.9920

The boldfaced line represents the optimal model for SSC prediction of citrus.

predicted SSC are shown in Fig. 7, the point are evenly and closely aligned along the line and the difference between RMSEC and RMSEP also is very smaller.

Wang and Xie (2014) developed SSC prediction model of 'Tangelo' citrus and achieved the performance of R_{pre} = 0.893, RMSEP = 0.436°Brix based on the GA–SPA–MLR model, which was slightly inferior to this study. On the other hand, the previous results achieved by diffuse reflectance mode were R_{pre} = 0.946, RMSEP = 0.308°Brix and RPD = 2.94 for grapefruit (Ncama et al., 2017), were R = 0.742, SECV = 0.74°Brix, and RPD = 1.49 for 'Clemevilla' mandarin (Cavaco et al., 2018), were R_{pre} = 0.87, RMSEP = 0.47°Brix, and RPD = 2.34 for 'Gannan' navel orange (Liu et al., 2010), were R_{pre} = 0.858, RMSEP = 0.7113°Brix, RPD = 1.95 for 'Nanfeng' mandarin (Zhang et al., 2011). The performance of full transmittance mode achieved by this study was slightly superior or similar to those modes established by diffuse reflectance mode, proving the transmittance mode was feasible to determinate SSC of citrus.

The best prediction model of SNV-SPA-PLS were built based on 10 effective variables including 685.34, 705.61, 733.21, 750.66, 787.07, 811.61, 892.55, 955.13, 968.47 and 983.03 nm. The 685.34, 705.61 nm are associated with the red spectrum region, the color difference is considered to be indirect and important indicators of fruit ripeness (Gómez et al., 2006). The 733.21 and 750.66 nm belong to the third overtone of O—H and the fourth overtone of C—H stretching (Wang et al., 2014). The 787.07 and 811.61 nm are covered by the range of 760–850 nm, which are mostly associated with O—H functional groups of water (Martinsen and Schaare, 1998), both 787 nm (Cavaco et al., 2018; Wang and Xie, 2014) and 811 nm (Li et al., 2014a) also were reported as important variables to predict fruit quality and yielded good performance. It is worth noting that 810 nm is one of the absorption peak of the original transmittance spectrum, its adjacent band (811.61 nm) was selected as an effective band by SPA, simultaneously. Therefore, the effective variables selected by SPA are closely related to water and carbohydrate in navel orange fruit, which helped to explain the relationship between the developed model and SSC parameter.

SPA was considered as the best variable selection algorithm in this study, which may be contradictory to previous reports. The difference in spectral acquisition mode and the fruits variety may contribute to this result. Although variable selection algorithm can eliminate most of the noise, interference and useless information of the spectra, spectral preprocessing is still the powerful tool to eliminate spectral differences caused by optical path changes between samples, to correct scattering differences caused by various sizes of particle, and to improve SNR (Signal Noise Ratio). Therefore the combination of spectral preprocessing with a variable selection algorithm will be the indispensable link in the analysis of Vis/NIR spectroscopy.

4. Conclusion

The full transmittance spectra of citrus were collected using a Vis/NIR portable spectrograph (550–1100 nm) to investigate the feasibility of full transmittance mode on the prediction of SSC in thick-skin fruits. Four kinds of spectral preprocessing methods including Smoothing, MSC, SNV, and first derivative were employed to improve the quality

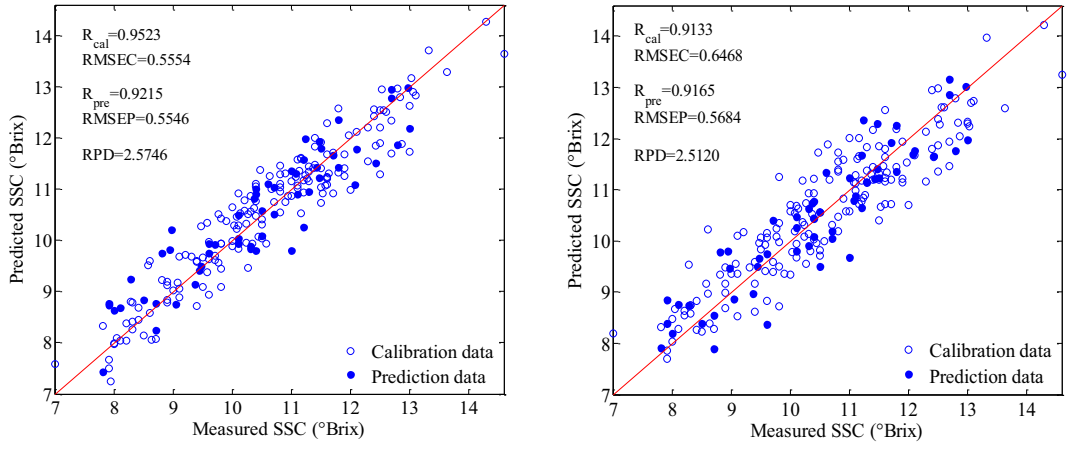


Fig. 7. The scatter plots of measured and predicted SSC based on the full wavelength preprocessed by SNV (a) and the effective variables selected by SPA (b).

of original spectra. Subsequently, the effective wavelengths of SSC were selected from the original and pretreated spectra with the algorithms of SPA, CARS, and GA. Finally, the prediction model of SSC was constructed based on full wavelengths and effective wavelengths. The results showed that the SPA performed best on eliminating the useless information variable and optimizing the number of effective variables. The optimal prediction model was established based on the 10 characteristic variables selected from the spectra pretreated by SNV with the algorithm of SPA, with R_{pre} of 0.9165, RMSEP of 0.5684°Brix, and RPD of 2.5120 respectively. Overall, the full transmittance mode was feasible to predict the internal quality of thick-skin fruits, like citrus. Additionally, the combination of spectral preprocessing with a variable selection algorithm was effective for developing a reliable prediction model. The conclusions of this study also provide an alternative method for fast and real-time detection of the internal quality of thick-skin fruits using Vis/NIR spectroscopy.

Funding

This study was supported by National Key Research and Development Program (2016YFD0200104), Beijing Talents Foundation (2018000021223ZK06) and National Natural Science Foundation of China (Grant No. 31671927).

CRediT authorship contribution statement

Xi Tian: Writing - original draft, Formal analysis, Investigation. **Jiangbo Li:**Methodology.**Shilai Yi:**Funding acquisition, Project administration.**Guoqiang Jin:**Data curation, Formal analysis.**Xiaoying Qiu:**Investigation.**Yongjie Li:**Writing - review & editing.

Declaration of competing interest

The authors declare no conflict of interest.

References

Aboudaoud, I., Faiz, B., Aassif, E., Izbaim, D., Abassi, D.E., Malainine, M., 2012. The maturity characterization of orange fruit by using high frequency ultrasonic echo pulse method. Iop Conference. 42, 012038.

Andersen, C.M., Bro, R., 2010. Variable selection in regression-a tutorial. J. Chemom. 24, 728–737.

Araújo, M.C.U., Saldanha, T.C.B., Galvã, R.K.H., Yoneyama, T., Chame, H.C., Visani, V., 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemom. Intell. Lab. Syst. 57 (2), 65–73.

Barbin, D.F., ElMasry, G., Sun, D.W., Allen, P., Morsy, N., 2013. Non-destructive assessment of microbial contamination in porcine meat using NIR hyperspectral imaging. Innovative Food Sci. Emerging Technol. 17 (17), 180–191.

Barnes, R., Dhanoa, M., Lister, S.J., 1989. Standard normal variate transformation and detrending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43 (5), 772–777.

Blasco, J., Aleixos, N., Gomez, J., Molto, E., 2007. Citrus sorting by identification of the most common defects using multispectral computer vision. J. Food Eng. 83 (3), 384–393.

Brezmes, J., Llobet, E., Vilanova, X., Orts, J., Saiz, G., Correig, X., 2001. Correlation between electronic nose signals and fruit quality indicators on shelf-life measurements with pinklady apples. Sens. Actuators, B. 80(1), 41–50.

Cavaco, A.M., Pires, R., Antunes, D.M., Panagopoulos, T., Brazio, A., Afonso, A., Silva, L., Lucas, M.R., Cadeiras, B., Cruz, S.P., Guerra, R., 2018. Validation of short wave near infrared calibration models for the quality and ripening of 'Newhall' orange on tree across years and orchards. Postharvest Biol. Technol. 141, 86–97.

Cayuela, J.A., 2008. Vis/NIR soluble solids prediction in intact oranges (Citrus sinensis L.) cv. Valencia Late by reflectance. Postharvest Biol. Technol. 47 (1), 75–80.

Dull, G.G., Birth, G.S., 1989. Nondestructive evaluation of fruit quality: use of near infrared spectrophotometry to measure soluble solids in intact honeydew melons. HortScience 24, 754.

Durand, A., Devos, O., Ruckebusch, C., Huvenne, J.P., 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.

Elmasry, G., Wang, N., Vigneault, Clément, 2009. Detecting chilling injury in red delicious apple using hyperspectral imaging and neural networks. Postharvest Biol. Technol. 52 (1), 1–8.

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 Method. 8 (8), 1936–1946.

Fan, S., Guo, Z., Zhang, B., Huang, W., Zhao, C., 2016. Using Vis/NIR diffuse transmittance spectroscopy and multivariate analysis to predicate soluble solids content of apple. Food Anal Method. 9 (5), 1333–1343.

Fraser, D.G., Jordan, R.B., Künnemeyer, R., Mcglone, V.A., 2003. Light distribution inside mandarin fruit during internal quality assessment by NIR spectroscopy. Postharvest Biol. Technol. 27 (2), 185–196.

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.

Galvão, R.K.H., Araújo, M.C.U., Fragoso, W.D., Silva, E.C., José, G.E., Soares, S.F.C., Paiva, H.M., 2008. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. Chemom. Intell. Lab. Syst. 92 (1), 83–91.

Gómez, A.H., He, Y., Pereira, A.G., 2006. Non-destructive measurement of acidity, soluble solids and firmness of Satsuma mandarin using vis–NIR spectroscopy techniques. J. Food Eng. 77, 313–319.

Guthrie, J.A., Walsh, K.B., Reid, D.J., Liebenberg, C.J., 2005. Assessment of internal quality attributes of mandarin fruit. 1. NIR calibration model development. Aust. J. Agric. Res. 56 (4), 405–416.

Hong, X., Wang, J., 2014. Detection of adulteration in cherry tomato juices based on electronic nose and tongue: comparison of different data fusion approaches. J. Food Eng. 126, 89–97.

Ireri, D., Belal, E., Okinda, C., Makange, N., Ji, C., 2019. A computer vision system for defect discrimination and grading in tomatoes using machine learning and image processing. Artificial Intelligence in Agriculture 2 (C).

Kim, K. B., Lee, S., Kim, M. S., Cho, B.K., 2009. Determination of apple firmness by nondestructive ultrasonic measurement. Postharvest Biol. Technol. 52(1), 44–48.

Krivoshiev, G.P., Chalucova, R.P., Moukarev, M.I., 2000. A possibility for elimination of the interference from the peel in nondestructive determination of the internal quality of fruit and vegetables by VIS/NIR spectroscopy. LWT-Food Sci Technol. 33 (5), 344–353.

Lammertyn, J., Peirs, A., Baerdemaeker, J.D., Bart, N., 2000. Light penetration properties of NIR radiation in fruit with respect to non-destructive quality assessment. Postharvest Biol. Technol. 18 (2), 121–132.

Leardi, R., Gonzalez, A.L., 1998. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemom. Intell. Lab. Syst. 41 (2), 195–207.

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 (1), 77–84.

Li, J., Zhao, C., Huang, W., Zhang, C., Peng, Y., 2014a. A combination algorithm for variable selection to determine soluble solid content and firmness of pears. Anal. Methods 6 (7), 2170–2180.

Li, J., Huang, W., Chen, L., Fan, S., Zhang, B., Guo, Z., 2014b. Variable selection in visible and near-infrared spectral analysis for noninvasive determination of soluble solids content of 'ya' pear. Food Anal Method. 7 (9), 1891–1902.

Li, Y., Jin, G., Jiang, X., Yi, S., Tian, X., 2020. Non-destructive determination of soluble solids content using a multi-region combination model in hybrid citrus. Infrared Phys. Technol. 104, 103138.

Liu, Y., Sun, X., Zhou, J., Zhang, H., Yang, C., 2010. Linear and nonlinear multivariate regressions for determination sugar content of intact gannan navel orange by vis–nir diffuse reflectance spectroscopy. Math. Comput. Model. 51 (11—12), 1438–1443.

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 (2), 307–323.

Liu, C., Yang, S.X., Deng, Lie, 2015. Determination of internal qualities of Newhall navel oranges based on NIR spectroscopy using machine learning. J. Food Eng. 161, 16–23.

Lopezgarcia, F., Andreugarcia, G., Blasco, J., Aleixos, N., Valiente, J., 2010. Automatic detection of skin defects in citrus fruits using a multivariate image analysis approach. Comput Electron Agr. 71(2), 189–197.

Magwaza, L.S., Opara, U.L., Nieuwoudt, H.H., Cronje, P.J., Saeys, W., Nicolai, B., 2012. NIR spectroscopy applications for internal and external quality analysis of citrus fruit—a review. Food Bioprocess Technol. 5 (2), 425–444.

Martinsen, P., Schaare, P., 1998. Measuring soluble solids distribution in kiwifruit using near-infrared imaging spectroscopy. Postharvest biol. Technol. 14 (3), 271–281.

Mendoza, F., Lu, R., Cen, H., 2012. Comparison and fusion of four nondestructive sensors for predicting apple fruit firmness and soluble solids content. Postharvest Biol. Technol. 73, 89–98.

Ncama, K., Opara, U.L., Tesfay, S.Z., Fawole, O.A., Magwaza, L.S., 2017. Application of Vis/NIR spectroscopy for predicting sweetness and flavour parameters of 'Valencia' orange (Citrus sinensis) and 'star ruby' grapefruit (Citrus x paradisi Macfad). J. Food Eng. 193, 86–94.

Nghia, N.D.T., 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.

Pasquini, C., 2018. Near infrared spectroscopy: a mature analytical technique with new perspectives—a review. Analytica Chimica Acta 1026, 8–36.

Saevels, S., Lammertyn, J., Berna, A.Z., Veraverbeke, E., Natale, C. D., Nicolai, B., 2003. Electronic nose as a non-destructive tool to evaluate the optimal harvest date of apples. Postharvest Biol. Technol. 30(1), 3–14.

Sanaeifar, A., Mohtasebi, S.S., Ghasemi-Varnamkhasti, M., Ahmadi, H., Lozano, J., 2014. Development and application of a new low cost electronic nose for the ripeness monitoring of banana using computational techniques (PCA, LDA, SIMCA and SVM). Czech J. Food Sci. 32 (6), 538–548.

Sinelli, N., Spinardi, A., Egidio, V.D., Mignani, I., Casiraghi, E., 2008. Evaluation of quality and nutraceutical content of blueberries (Vaccinium corymbosum L.) by near and mid-infrared spectroscopy. Postharvest Biol. Technol. 50 (1), 31–36.

Tian, X., Wang, Q., Li, J., Peng, F., Huang, W., 2018. Non-destructive prediction of soluble solids content of pear based on fruit surface feature classification and multivariate regression analysis. Infrared Phys. Technol. 92, 336–344.
Tian, X., Fan, S., Xia, Y., Huang, W., Zhao, C., 2019. Comparison and optimization of models for SSC on-line determination of intact apple using efficient spectrum optimization and variable selection algorithm. Infrared Phys. Technol. 102, 102979.
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.
Wang, A., Hu, D., Xie, L., 2014. Comparison of detection modes in terms of the necessity of visible region (VIS) and influence of the peel on soluble solids content (SSC) determination of navel orange using VIS–SWNIR spectroscopy. J. Food Eng. 126, 126–132.
Wu, D., Sun, D.W., 2013. Potential of time series-hyperspectral imaging (TS-HSI) for non-invasive determination of microbial spoilage of salmon flesh. Talanta 111, 39–46.
Wu, D., He, Y., Nie, P., Cao, F., Bao, Y., 2010. Hybrid variable selection in visible and near-infrared spectral analysis for non-invasive quality determination of grape juice. Anal. Chim. Acta 659 (1), 229–237.
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.
Zhang, L., Xue, L., Liu, M. H., Li, J., 2011. Nondestructive detection of soluble solids content of nanfeng mandarin orange using VIS-NIR spectroscopy. Adv. Mater. Res. 361–363, 1634-1637.
Zhang, D., Xu, L., Liang, D., Xu, C., Jin, X., Weng, S., 2018. Fast prediction of sugar content in Dangshan pear (Pyrus spp.) using hyperspectral imagery data. Food Anal Method. 11 (8), 2336–2345.
Zhang, D., Xu, Y., Huang, W., Tian, X., Xia, Y., Xu, L., Fan, S., 2019. Nondestructive measurement of soluble solids content in apple using near infrared hyperspectral imaging coupled with wavelength selection algorithm. Infrared Phys. Technol. 98, 297–304.
Zude-Sasse, M., Truppel, I., Herold, B., 2002. An approach to non-destructive apple fruit chlorophyll determination. Postharvest Biol. Technol. 25 (2), 123–133.