Modeling method for SSC prediction in pomelo using Vis-NIRS with wavelength selection and latent variable updating

Hao Tian[1,2,3], Shuai Wang[1,2], Huirong Xu[1,2], Yibin Ying[1,2*]

(1. *College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China*;
2. *Key Laboratory of On-Site Processing Equipment for Agricultural Products, Ministry of Agriculture and Rural Affairs, Hangzhou 310058, China*;
3. *College of Food Science, Shihezi University, Shihezi 832003, Xinjiang, China*)

Abstract: The aim of this study was in-line, rapid, and non-destructive detection for soluble solid content (SSC) in pomelos using visible and near-infrared spectroscopy (Vis-NIRS). However, the large size and thick rind of pomelo affect the stability of spectral acquisition and the biological variabilities affect the robustness of models. Given these issues, in this study, an efficient prototype in-line detection system in transmittance mode was designed and evaluated in comparison with an off-line detection system. Data from the years 2019 and 2020 were used for modeling and the external validation data were obtained by the in-line detection system in 2021. The wavelength selection methods of changeable size moving window (CSMW), random frog (RF), and competitive adaptive reweighted sampling (CARS) were used to improve the prediction accuracy of partial least squares regression (PLSR) models. The best performance of internal prediction was obtained by CARS-PLSR and the determination coefficient of prediction (R_p^2), root mean square error of prediction (RMSEP), and residual predictive deviation (RPD) were 0.958, 0.204%, and 4.821, respectively. However, all models obtained large prediction biases in external validation. The latent variable updating (LVU) method was proposed to update models and improve the performance in external validation. Ten samples from the external validation set were randomly selected to update the models. Compared with the recalibration method, LVU could effectively modify the original models which matched the SSC range of the external validation set. The CSMW-PLSR models were more robust in external validations. The off-line model with LVU performed best with a root mean square error of validation (RMSEV) of 0.599% and the in-line model with recalibration obtained RMSEV of 0.864%. These results demonstrated the application potential of the transmittance Vis-NIRS for in-line rapid prediction of SSC in pomelos and the modeling and updating methods could be applied to samples with biological variabilities.

Keywords: Vis-NIRS, in-line detection, external validation, wavelength selection, model updating, pomelo, SSC
DOI: 10.25165/j.ijabe.20241701.7491

Citation: Tian H, Wang S, Xu H R, Ying Y B. Modeling method for SSC prediction in pomelo using Vis-NIRS with wavelength selection and latent variable updating. Int J Agric & Biol Eng, 2024; 17(1): 251–260.

1 Introduction

Pomelo (*Citrus maxima* Merr.) is widely grown in South China as a crucial cash crop. It is rich in many nutrients and beneficial to human health[1]. As a kind of citrus, pomelo is deliciously sweet and sour and favored by consumers. Over the years, consumers have had increasingly stringent requirements for fruit quality, especially for internal attributes[2]. Soluble solids content (SSC) is an important index to evaluate the flavor of pomelo, which directly affects consumers' purchasing decisions[3]. Consequently, sorting according to the value of SSC is urgently needed in the process of postpartum processing and commercialization to help increase the added value of pomelos.

For the internal quality inspection of fruit, the techniques used in the existing research are near-infrared spectroscopy (NIRS)[4], NIRS-based systems such as multispectral and hyperspectral imaging[5], nuclear magnetic resonance imaging[6], X-ray computed tomography[7], acoustic vibration[8]. Visible and near-infrared spectroscopy (Vis-NIRS), an efficient and pollution-free technique, has been utilized for SSC assessment for many years[9,10].

The 'diagnostic, window' at 700-900 nm provides the greatest signal-to-noise ratio for the transmittance mode, and it provides a means of obtaining internal information about fruit with thick peel[11]. For citrus, exocarp, and endocarp were found to contribute mostly to the diffuse reflectance spectra[12], thus the transmittance mode was more applied to the internal quality detection of citrus, such as oranges[13,14]. For large fruits with thick rinds, such as watermelon and melon, transmittance spectral detection has also been used to evaluate the internal qualities[4,15,16]. However, pomelo contains a multi-layer structure with endocarp, exocarp, and pulp. The endocarp has a particular structure of gradient foam and fiber bundles with a thickness of approximately 1.5 cm[17]. The complex multi-layer tissue structure and large-size thick peel lead to low spectral intensity and poor signal-to-noise ratios. These problems would be more prominent in in-line rapid detection, which would affect the stability of the spectra and thus the prediction models. According to the literature search, there are few reports on the in-line and non-destructive detection of SSC in pomelos.

Establishing a stable and robust statistical model requires representative data sets. Outliers that are incorporated into a

Received date: 2022-03-08 **Accepted date:** 2022-07-13

Biographies: Hao Tian, PhD candidate, research interest: non-destructive quality detection of agro-products, Email: 804668643@qq.con; **Shuai Wang**, PhD candidate, research interest: non-destructive quality evaluation of agro-products, Email: 21713009@zju.edu.cn; **Huirong Xu**, PhD, Professor, research interest: agricultural robots and agro-product quality perception, Email: hrxu@zju.edu.cn
***Corresponding author: Yibin Ying**, PhD, Professor, research interest: agricultural robots and agro-product quality perception. 866 Yuhangtang Rd., Hangzhou 310058, Zhejiang, China. Tel/Fax: +86-571-88982282, Email: ybying@zju.edu.cn.

multivariate calibration model can significantly reduce the performance of the model[18]. In the spectral modeling of chemometrics, there are two reasons for outliers. One of them is the spectral anomaly (*X*). Light leakage, specular reflection, and sample abnormality during the spectra acquisition, especially in in-line detection, would cause spectral profile distortion. Therefore, the probability of spectral anomalies is a way to evaluate the stability of the spectral acquisition system. The other one is the response outlier (*Y*). In partial least squares regression (PLSR) models, the response is usually required to be continuous and obey the normal distribution[19]. Outliers of *Y* would produce leverage lead to deviation in variable extraction and distort the outcome and accuracy of a regression[20,21]. The Monte Carlo sampling (MCS) method could help to reduce the risk that the masking effect brings about and provide a feasible way to detect different kinds of outliers using the distribution of prediction errors of a test sample resulting from a population of sub-models[22-24].

The robustness of the model and the reproducibility of the results are the key to the successful application of the model. However, the variabilities of biological properties such as different seasons, origins, cultivars, and storage conditions have great influences on the robustness of models. Because different batches of samples may have differences in SSC, weight, ripeness, and other attributes. Increasing the range of data sets to make them sufficiently representative can increase the robustness of the model[25]. However, a large number of experiments are time-consuming and laborious. Moreover, the wavelength selection and model updating methods have been reported to address model performance in external validation[26,27]. Wavelength selection could extract key wavelengths and eliminate the interference of redundant information[28]. The model upgrading method could establish the correlation between the existing data model and external data, modify the model parameters, and prevent local overfitting. For the model transfer or model upgrading method of biological diversity, Fan et al.[26] reported the slope and bias correction method to correct the deviation of model prediction from data in different years. Mishra et al.[27] modified the SSC prediction by variable selection and recalibration to reduce the bias from −0.62% to 0.07% and the RMSEP from 0.90% to 0.63%. Sun et al.[29] achieved robustness to temperature change of models for intact mango fruit dry matter content with six methods in which the external parameter orthogonalization obtained the best result with RMSEP of 1.05% w/w. Appropriate external validation is necessary to test the robustness of calibration models. Meanwhile, selecting a subset of new samples to establish its association with the main model would help to improve the performance of external validation.

Overall, the aims of this study were the in-line detection system for pomelo and the robust calibration model for the detection of SSC based on Vis-NIRS. The main work carried out was as follows: 1) to test the in-line detection prototype system and evaluate the performance compared with the off-line system; 2) to establish PLSR calibration models for SSC combined with three methods of wavelength selection; 3) to evaluate the model robustness by the external validation in a different year; 4) to correct the deviation of external validation using model updating methods.

2 Materials and methods

2.1 Spectra acquisition

2.1.1 In-line detection system

An original transmittance system for in-line detection was designed with fruit cup conveying and a double-layer parallel light source, as shown in Figure 1. It was composed of a light source assembly, transmission unit, and electronic control unit. The light source assembly was composed of two light boxes in parallel divided into upper and lower layers. The light boxes were arranged on the adjusting mechanism with the angles adjustable. The light box contained four 150 W halogen tungsten lamps inside and two fans outside to dissipate heat. The two groups of light source assembly were symmetrically arranged on both sides of the conveyor belt. In other words, 16 halogen lamps were used as the initial incident light. The system adopted partial-transmittance mode with photons passed through Pomelo, entered the collimator below, and was captured by the spectrometer (QE65pro, Ocean Optics, USA). The laser trigger was used to trigger the spectrometer collecting signals when the fruit cup passed through. According to the demand for productivity, the speed of the conveyor belt was set at 1.5 m/s.

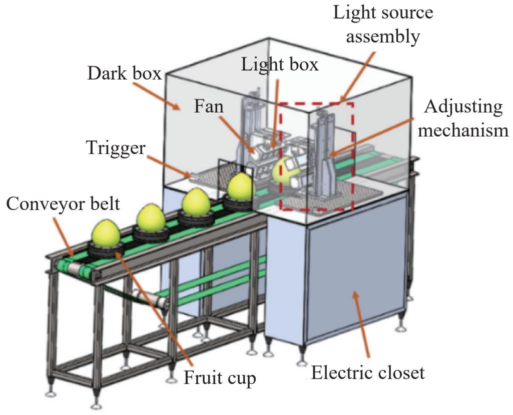


Figure 1 Schematic diagram of in-line detection system for pomelo

2.1.2 Off-line detection system

In previous studies, average spectra of multi-point detection could effectively eliminate the influence of light distribution on the model[30]. The off-line spectra were collected by a semi-transmission system with a rotatable fruit cup, as shown in Figure 2.

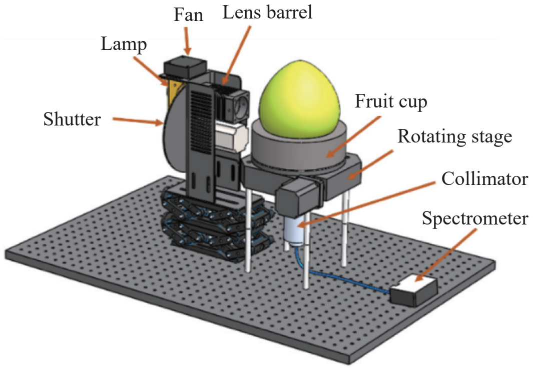


Figure 2 Schematic diagram of off-line detection system for pomelo

Briefly, the pomelos were placed on the fruit cup with the navel down. A 150 W tungsten halogen lamp was horizontally shining on the equatorial surface of pomelos. The photons passed through the pomelo and were received by the optical fiber collimator at the bottom of the fruit cup. Then they were transmitted by an optical fiber (P1000-2-Vis-NIR, Ocean Optics, USA) to a commercial fiber optic spectrometer (QE65pro, Ocean Optics., USA). The pomelo rotated around the vertical axis on the rotating stage, with intervals of 90° degrees. Four spectra of each pomelo were collected and averaged.

2.2 Samples and SSC measurement

Pomelos of the cultivar Guanximiyou (GX) were picked from an orchard (116.65°N, 24.37°E) in Meizhou, Guangdong, China. A total of 335 samples from three batches in three years were used for the SSC evaluation. The first batch (off-line detection) was picked in September 2019 and the second batch (in-line detection) was picked in September 2020. These two batches were used to establish the calibration models. The third batch was picked in September 2021 and acquired the spectra on the in-line system and it was used for external validation. Before the data acquisition, to eliminate the influence of the temperature, all samples were stored in a constant temperature and humidity chamber ((25±1)°C, 75%) for 48 h in the laboratory at Zhejiang University, China.

The SSC was measured by a digital refractometer (PAL-BX|ACID F5, ATAGO., Japan). Pomelos were first cut along the equatorial plane. The juice vesicles of the equatorial plane near the navel were taken out, and squeezed juice with a juicer, then dropped into the refractometer after filtering through the filter screen. As the unsymmetrical distribution of composition, the SSC was measured and averaged at three regions every 120° along the cross-section.

2.3 Spectral analysis and modeling

2.3.1 Spectra preprocessing

In this study, the spectral intensity from different systems and different batches was calibrated by a customized polytetrafluoroethylene (PTFE) cylinder. The relative transmittance (RT) was calculated for spectral modeling by the following equation:

$$\mathrm{RT} = \frac{I_s - D}{I_r - D} \times 100\% \qquad (1)$$

where, I_r is the transmittance intensity of PTFE reference; I_s is the transmittance intensity of samples; D is the dark current intensity of the spectrometer. Subsequently, the standard normal variate (SNV)[31] was used to eliminate amplitude differences.

2.3.2 Partial least squares regression

Partial least square regression (PLSR) is based on the computation of the optimal least-squares fit part of a correlation or covariance matrix to relate spectral data to quality attributes[32]. Cross-validation provides a way to find the best LVs and prevent overfitting. In this study, the optimal combination was chosen by 5-fold cross-validation, in which the maximum variable search was 20. Subset selection according to Kennard-Stone (KS) strategy[33] was used to split into two subsets: the calibration set (75%) and the prediction set (25%).

2.3.3 Outlier detection

Monte Carlo Sampling (MCS) method calculates the mean value and standard deviation of the prediction residuals of samples by multiple random sampling. In this study, combined with analyzing the spectral profiles and the SSC histogram distribution, outliers were manually selected through the scatter diagram distribution of MCS output. The number of outliers was strictly controlled within 5% of the sample size which belongs to small probability events in statistical analysis. The MSC ran under 5000 iterations with the ratio of the sample of 0.75. After removing outliers, the samples were re-divided into calibration set (75%) and prediction set (25%).

2.3.4 Evaluation of models

The determination coefficient of calibration (R_c^2), root mean square error of calibration (RMSEC), determination coefficient of prediction (R_p^2), root mean square error of prediction (RMSEP), residual predictive deviation (RPD) for the prediction set, and root mean square error of validation (RMSEV), were used to assess the predictive ability of models.

2.4 Wavelength selection

2.4.1 CSMW-PLSR

Changeable size moving window PLSR (CSMW-PLSR) aims at selecting the optimized interval from all possible spectral intervals within an informative region[34]. It moves a window across the whole wavelength range and builds a PLSR model for each window, as is shown in Figure 3. The windows with low root mean squared errors of fitting (RMSEF) values could be selected to build a final model.

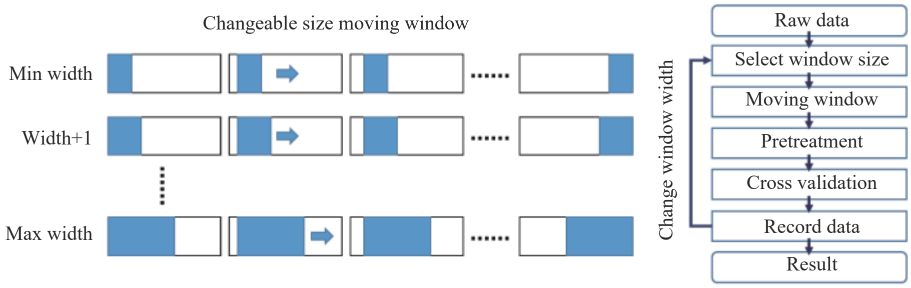


Figure 3 Scheme for explanation of CSMW-PLSR

Compared with the CSMW-PLSR of Du et al.[34], some improvements were made to obtain a stable wavelength interval, as shown in the flow chart in Figure 3. Firstly, the pretreatment results of SNV were changeable in different window ranges. Therefore, spectral pretreatment was added to the cycles of window scanning instead of before importing the spectra. Secondly, The minimum root mean square error of cross-validation (RMSECV) of 5-fold cross-validation was used as the basis for model selection to prevent overfitting and select the best LVs. The effective spectral wavelength range of pomelo was 500-1000 nm, with a total wavelength of 675. In the CSMW-PLSR selection, the minimum width was set to 300, and the maximum was 650.

2.4.2 RF-PLSR

Random frog is an efficient variable selection approach based on the reversible jump Markov Chain Monte Carlo-like for applications to gene selection and disease classification[35]. It is searching for a model space through the realization of both fixed-dimensional and trans-dimensional jumps between different models. A pseudo-MCMC chain is calculated to determine the selection probability of each variable to measure the relevance of variables, which can be used as a variable selection criterion. In this study, the number of iterations was set to 10 000 and the maximal number of LVs for cross-validation was 20.

2.4.3 CARS-PLSR

A competitive adaptive reweighted sampling technique coupled with PLSR (CARS-PLSR) is a promising procedure to eliminate the uninformative variables and/or conduct wavelength selection for building a high-performance calibration model[36]. This method starts by building a full model with all variables included, followed by iteratively eliminating variables of the least importance in a backward manner. The number of variables to eliminate at each iteration is controlled by an exponentially decreasing function and an adaptive reweighted sampling technique. At each iteration, performance evaluation is conducted for a subset of variables rather than for individual variables. The number of MCS runs was set to 100 and the number of variables to be selected was determined by 5-fold cross-validation.

The procedures of all modeling were written and performed in Matlab R2019b (Mathworks, USA) with the toolbox of libPLS1.98[31].

2.5 Model updating methods

When the samples to be predicted are measured on a different instrument or under differing environmental factors, the original models might be invalid. Various model transfer methods have been developed to enable a calibration model to be effectively transferred between two systems[37]. However, correcting the prediction deviation caused by different sample attributes on the same system, this step was called ‘model updating’.

Latent variables updating (LVU): As it was known, the selection of the LVs number is important when building a PLSR model. If more variables are selected, the model will easily result in overfitting, while the selection of fewer variables will cause underfitting[34]. The general process of determining the optimal number of LVs is cross-validation. However, the optimal LVs of cross-validation are based on the performance of the original data. For external validation, it might be inapplicable. According to this characteristic of PLSR, the LVU method was proposed to correct the optimal LVs for the external validation data. The specific process is as follows:

1) Select several representative new samples (such as 5, 10, 20 samples);

2) Calculate the predicted values and RMSEPs using the coefficients of the original model with different numbers of LVs (from 1 to the maximum LVs of cross-validation);

3) Select the number of LVs with minimum RMSEP of new samples as the model parameter to predict the remaining samples.

Recalibration method: Recalibration is to modify the loading matrix and score matrix in the original model by introducing a small number of new samples[27]. This method needs to carry out the modeling process again and select a new number of LVs through cross-validation. In this study, 10 samples were selected randomly and the number of LVs was determined by 5-fold cross-validation.

This is a process of reselecting LVs. Like the minimum RMSECV in cross-validation was used for modeling, the minimum RMSEP of external data was used for the correction of overfitting. The process of LVU is different from the recalibration method in that recalibration recalculates the loading and score matrix of a new model, while LVU only reselects LVs with the original coefficients. In other words, LVU attempts to use the original model to interpret external data, rather than directly modify its coefficient matrix. In this study, 10 samples were used to compare the two model updating methods.

3 Results and discussion

3.1 Statistics of pomelo attributes

The attribute statistics of pomelos in three batches are listed in Table 1. Among them, the SSC of pomelos harvested in September 2020 was significantly different. The SSC ranged from 6.90% to 17.00%, and the mean value was 13.55%. It was larger than the other 2 batches. Meanwhile, the validation set has a small standard deviation which was only 0.49%. In addition, samples in the external validation set had larger sizes and heavier mass compared with Batch 1 and Batch 2 which might make external validation more difficult with low spectra intensity.

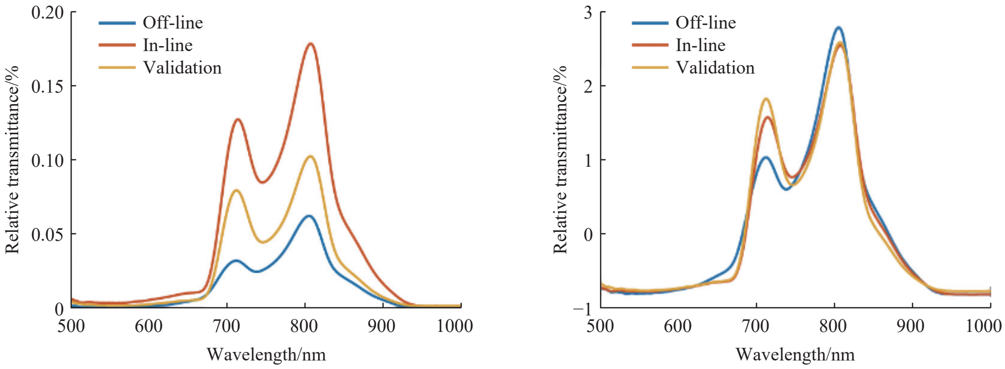
Table 1 Descriptive statistics of the pomelo datasets in three years

Batch	Data set	Harvest year	Data acquisition	No. of pomelos	Weight/g*	SSC/%
1	Off-line modeling	2019.09	Off-line	125	995±122^c	10.96±1.05^b
2	In-line modeling	2020.09	In-line	110	1187±120^b	13.55±1.17^a
3	External validation	2021.09	In-line	100	1595±166^a	10.79±0.49^b

Note: * ‘±’ indicates the mean value and standard deviation; lower case letters indicate the significant differences between the three batches (p<0.05); SSC: Soluble solid content.

3.2 Spectra profile

Due to the strong absorption of light by tissues, only in the range of 500-1000 nm has acceptable signal intensity. Therefore, the relative transmittance in the range of 500-1000 nm was employed to establish the models. The mean spectra in the range of 500-1000 nm of each batch are shown in Figure 4.



a. Raw data

b. Preprocessing with SNV

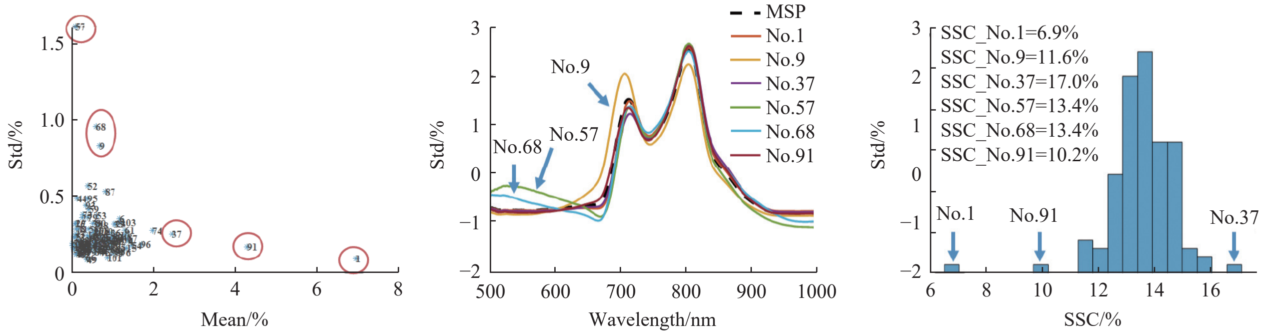
Note: SNV: Standard normal variate.

Figure 4 Spectra profile of raw data and preprocessing with SNV in different batches

As shown in Figure 4, there are two transmittance peaks at 712 nm and 808 nm for all spectra, and the peak at 808 nm is higher than that at 712 nm. The absorbance peak in the range of 720-750 nm might be relevant to sugar including the third overtones of OH stretching vibrations at 740 nm[38]. The visible light in the range of 650-675 nm might be absorbed by the pigments such as chlorophylls, and the short wave near-infrared light around 950 nm is related to water absorption, so the RT signals in these ranges are weak. The difference in spectral intensity mainly came from optical path length changes caused by sample size. The preprocessing method of SNV could standardize the spectra and eliminate the difference in light intensity. After pretreatment, the spectral peak positions of different batches were the same. The peak difference near 720 nm might be caused by the biological variabilities of pomelos.

3.3 Outliers detection

The boundary of outliers was determined manually by observing the error distribution of MCS. In this study, the in-line system was the focus of attention. Therefore, outlier detection of the in-line data set was taken as an example to describe the screening process. As shown in Figure 5a, the red circle shows the outliers in the error distribution map. Among them, No. 37, No. 91, and No. 1 show large mean error, while No. 9, No. 68, and No. 57 show excessive standard deviation of error. Figure 5b shows the spectral profiles of the six outliers and the average spectrum. It could be found that samples No. 9, No. 68, and No. 57 have spectral distortion while the other three spectra are normal. Similarly, in Figure 5c, No. 1, No. 91, and No. 37 with larger mean values are outliers of SSC. Therefore, these six points were eliminated. It is worth noting that there were three spectra showing anomalies, of which the baseline offset of No. 57 and No. 68 might be caused by light leakage of the system. However, the spectrum of No. 9 was only abnormal in the spectral shape, which could be derived from the attribute difference of the fruit itself. In other words, only two abnormal spectra were caused by the instability of the in-line system and the probability of abnormal spectra of the in-line system was 1.8% (2/110). It is acceptable as a prototype system. In addition, the outliers of SSC were eliminated to prevent leverage in the model of existing data. When the data set was wide enough to cover its SSC value, it could be added to the model. Therefore, when data sets were mixed, MCS procedures need to be repeated instead of simply merging outliers.



a. Plot of mean errors vs standard deviation (STD) errors of prediction

b. Spectral profile of outliers selected by MCS.

c. Histogram SSC of all samples with outliers markers

Note: The red circles mark the outliers; MSP means spectrum of all samples.

Figure 5 Monte Carlo Sampling (MCS) outlier detection of the in-line detection set

3.4 PLSR models combined with different wavelengths selection methods

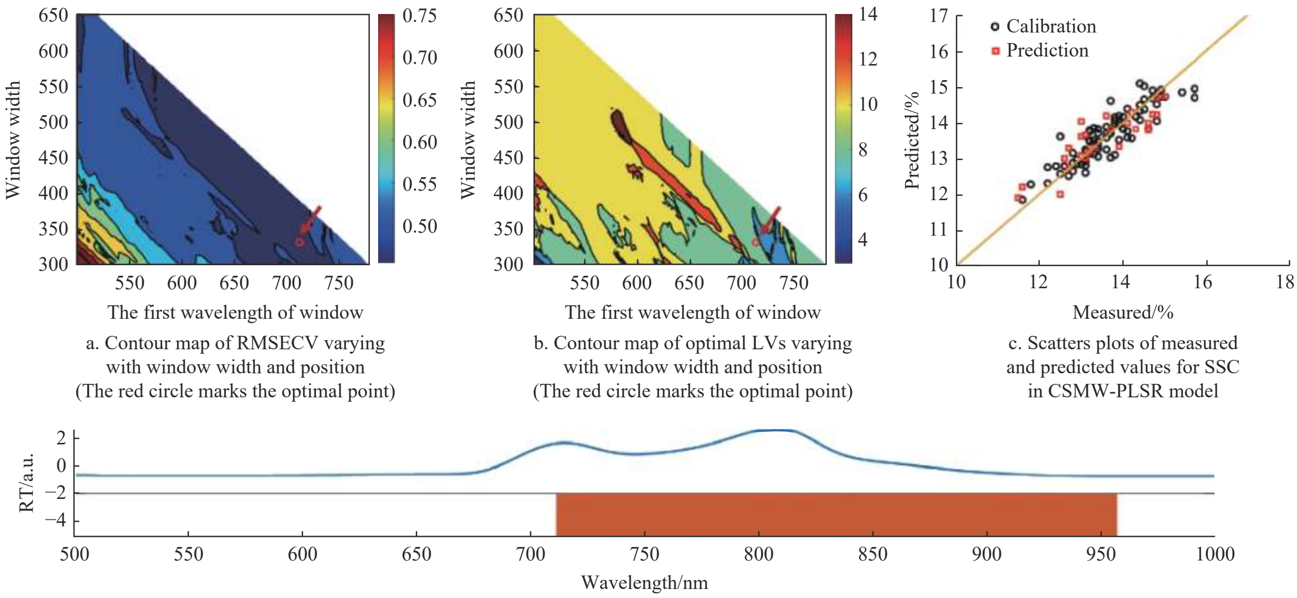
The wavelength optimization results of different data sets might be slightly different, but the method and process are the same. To avoid repetition, the wavelength optimization process of in-line models was described as an example.

3.4.1 CSMW-PLSR models

CSMW-PLSR adopted the window width ranging from 300 to 650, the moving step was 1, and the window size iterated with 1. The window moved from the first point until it covered the entire band. As shown in Figures 6a and 6b, the minimum RMSECV and optimal LVs with different window widths and positions were recorded and drawn on the contour map. In Figure 6a, there is a regular gradient change in the contour map, in which the value of RMSECV would decrease with the increase of window width and backward position. When the window width was small, such as 300-350, the window in the dark blue area started near the wavelength of 650 nm. Meanwhile, as the window size increases, the starting position of the dark blue area moves forward. It showed that the effective wavelength was distributed in the second half of the spectral range which was the near-infrared band. In Figure 6b, the distribution of optimal LVs also shows regionalization. When the window size is small, LVs change greatly in different window positions. When the window becomes larger, LVs tend to be stable. It should be noted here that when the RMSECV value is close, the smaller the LVs value is, the more stable the model is. Because the large number of LVs brings a risk of overfitting. Finally, the red circle marked the best point with the minimum RMSECV was 0.4%, optimal LVs was 8, and a window width was 332. The selected wavelength range was 711.54-956.48 nm, as shown in Figure 6d. This band covers wavelengths of 740 nm and 950 nm which are related to SSC and water. The PLSR model established using the selected band is shown in Figure 6c. The RMSEP was 0.490%, and RPD was 2.01.

3.4.2 RF-PLSR models

The results of RF wavelength selection are shown in Figure 7. Figure 7a shows the selection probability of each wavelength. Among them, the wavelengths around 650 nm, 740 nm, and 900 nm had higher selection probabilities. Moreover, the effective wavelength and the noise wavelength were mixed with no obvious boundaries. Therefore, the probability of wavelength selection was ranked and the more important wavelengths were selected successively for cross-validation iterative calculation. With the number of selected wavelengths increased, the changes in minimum RMSECV are shown in Figure 7b. When the number of selected wavelengths reached 120, the optimal value appeared. After that, the wavelength mixing brought noise information and reduced the performance of cross-validation. The selected wavelengths were dispersed over the entire wavelength range, as shown in Figure 7d. Among them, wavelength aggregation appeared in the main absorption peak bands such as 650, 730, 900, and 950 nm. The RF-PLSR model established using the selected wavelengths is shown in Figure 7c. The RMSEP was 0.339%, and RPD was 2.90.



a. Contour map of RMSECV varying with window width and position (The red circle marks the optimal point)

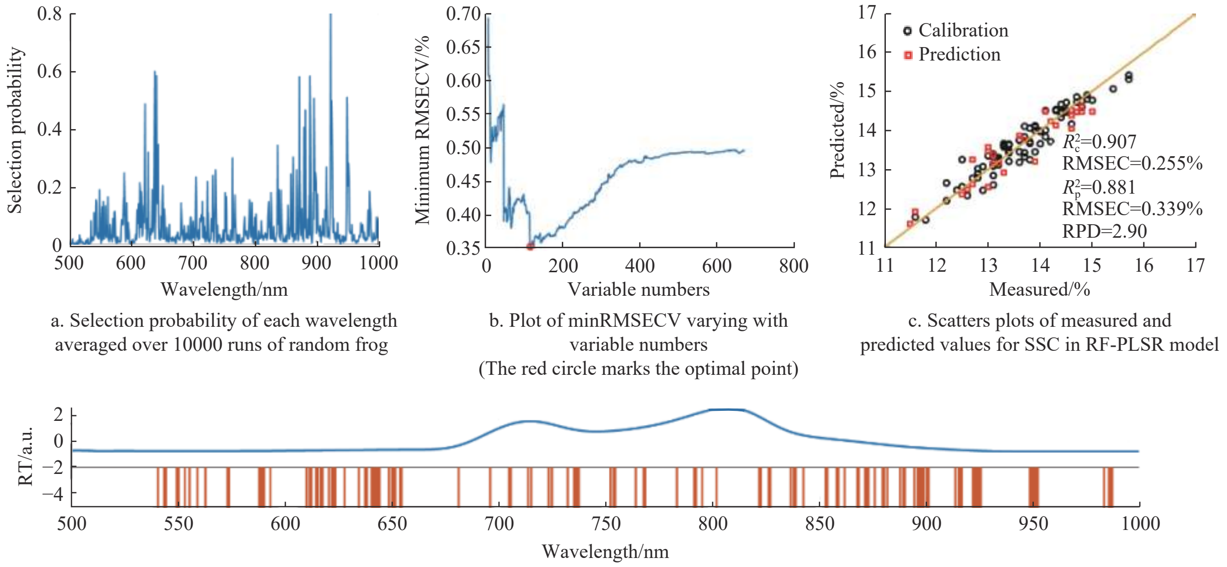
b. Contour map of optimal LVs varying with window width and position (The red circle marks the optimal point)

c. Scatters plots of measured and predicted values for SSC in CSMW-PLSR model

d. Optimal wavelength points selected by CSMW-PLSR

Note: RT, and normalized relative transmittance (The orange area is the selected band)

Figure 6 Results of wavelength selection by CSMW-PLSR of in-line data



a. Selection probability of each wavelength averaged over 10000 runs of random frog

b. Plot of minRMSECV varying with variable numbers (The red circle marks the optimal point)

c. Scatters plots of measured and predicted values for SSC in RF-PLSR model

d. Optimal wavelength points selected by RF-PLSR

Note: RT, and normalized relative transmittance (Orange vertical lines are the selected wavelengths)

Figure 7 Results of wavelength selection by RF-PLSR of in-line data

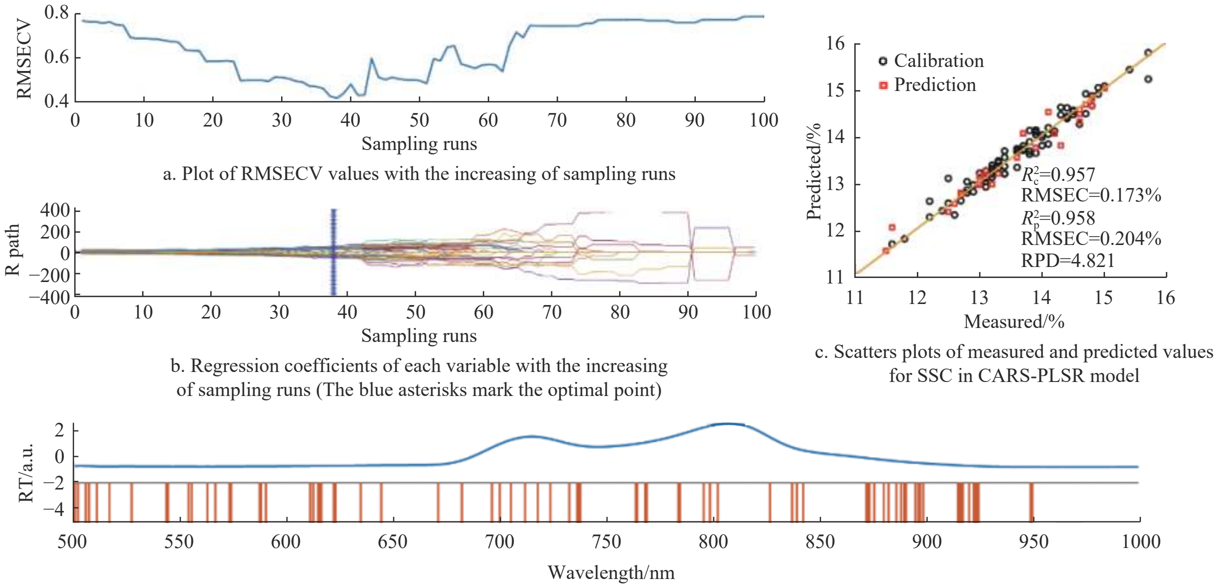
3.4.3 CARS-PLSR models

Figure 8 shows 5-fold RMSECV values (Plot *a*), and the regression coefficient path of each variable (Plot *b*) with the increasing of sampling runs from one CARS running. The RMSECV values descend before 38 runs which should be ascribed to the elimination of uninformative variables, and finally increased fast because of the loss of information caused by eliminating some key variables from the optimal subset (denoted by asterisks). The 77 selected wavelengths selected by CARS-PLSR are shown in Figure 8d. The wavelength distribution was more scattered than RF-PLSR except that the aggregations were around 880 and 920 nm. The number was less than that of RF-PLSR, however, the model performance of CARS-PLSR was better than the RF-PLSR with RMSEP of 0.204 % and RPD of 4.821, as shown in Figure 8c.

3.5 Comparison of PLSR models with wavelength selection

Table 2 lists the performances of PLSR models established with different wavelength selection methods. Both off-line detection and in-line detection have achieved good model performance. Among them, the RMSECVs of the in-line models were between 0.287%-0.496%. The model performance was slightly better than that of the off-line models with RMSECVs of 0.375%-0.572%. In other words, the in-line detection prototype system could obtain stable spectral data and be used to establish a prediction model equal to or slightly better than off-line detection. Meanwhile, the mixture model combining the two data sets could get the model performance of RMSECVs in 0.461%-0.586%. The model accuracy decreased with the data increased. It should be noted that the mixing makes the SSCs have a wider distribution and the robustness and representativeness of the model might improve. However, biological variabilities and system differences were mixed in, which reduced the accuracy of the models. The Models established by RF-PLSR and CARS-PLSR obtained better performances than the CSMW-PLSR. The best model was obtained by the CARS-PLSR in in-line detection, with RMSECV of

0.287%, R_p^2 of 0.958, RMSEP of 0.204%, and RPD of 4.821. Moreover, ranking the numbers of selected wavelengths, CSMW-PLSR>RF-PLSR>CARS-PLSR. All models obtained good performance for the non-destructive prediction for the SSC in pomelo. The PRDs of most models were obtained greater than 2. With methods of RF and CARS, some PRDs exceed 3 or even 4. The wavelength selection method significantly improved the performance of models in the internal calibrations and predictions. The prediction abilities were satisfactory with RMSEPs between 0.204% to 0.570%. The prediction accuracy was similar to the SSC detection results of Citrus reported in previous literature, such as oranges[13,39].



a. Plot of RMSECV values with the increasing of sampling runs

b. Regression coefficients of each variable with the increasing of sampling runs (The blue asterisks mark the optimal point)

c. Scatters plots of measured and predicted values for SSC in CARS-PLSR model

d. Optimal wavelength points selected by CARS-PLSR

Note: RT, normalized relative transmittance (Orange vertical lines are the selected wavelengths)

Figure 8 Results of wavelength selection by CARS-PLSR of in-line data

Table 2 Model performance of PLSR with different wavelength selection methods in the range of 500-1000 nm

Batch	Samples (Outliers)	Modeling methods	EWs (Proportion)	LVs	RMSECV/%	R_c^2	RMSEC/%	R_p^2	RMSEP/%	RPD
Off-line	125 (2)	PLSR	675 (100%)	13	0.572	0.807	0.419	0.719	0.447	1.900
		CSMW-PLSR	304 (45%)	12	0.444	0.885	0.323	0.871	0.302	2.814
		RF-PLSR	100 (15%)	16	0.447	0.902	0.298	0.915	0.245	3.466
		CARS-PLSR	61 (9%)	17	0.375	0.924	0.263	0.856	0.336	2.529
In-line	110 (5)	PLSR	675 (100%)	11	0.496	0.798	0.376	0.781	0.476	2.069
		CSMW-PLSR	332 (49%)	8	0.456	0.776	0.395	0.753	0.490	2.008
		RF-PLSR	120 (18%)	13	0.404	0.907	0.255	0.881	0.339	2.902
		CARS-PLSR	77 (11%)	17	0.287	0.957	0.173	0.958	0.204	4.821
Mixture	235 (7)	PLSR	675 (100%)	15	0.586	0.920	0.446	0.878	0.570	2.801
		CSMW-PLSR	331 (49%)	13	0.520	0.927	0.424	0.889	0.530	3.013
		RF-PLSR	59 (9%)	15	0.499	0.933	0.407	0.928	0.427	3.735
		CARS-PLSR	20 (3%)	12	0.461	0.921	0.443	0.931	0.434	3.673

Note: EWs: Effective wavelengths; LVs: Latent variables; CSMW-PLSR: Change size moving window PLSR; RF-PLSR, random frog PLSR; CARS-PLSR, competitive adaptive reweighted sampling PLSR. Same below.

However, due to the selection strategy of iterative optimization, the optimization results of all methods showed randomness, which is usually related to the number of iterations and the data structure of the sample itself. Meanwhile, the model has a risk of overfitting by selecting the minimal RMECV to screen the optimal LVs in the cross-validation. In fact, except for the CSMW-PLSR models in in-line detection and mixture data set, all other models have appeared with RMESPs lower than RMSECVs. These models needed external validation to evaluate their practicality.

3.6 External validation and model updating

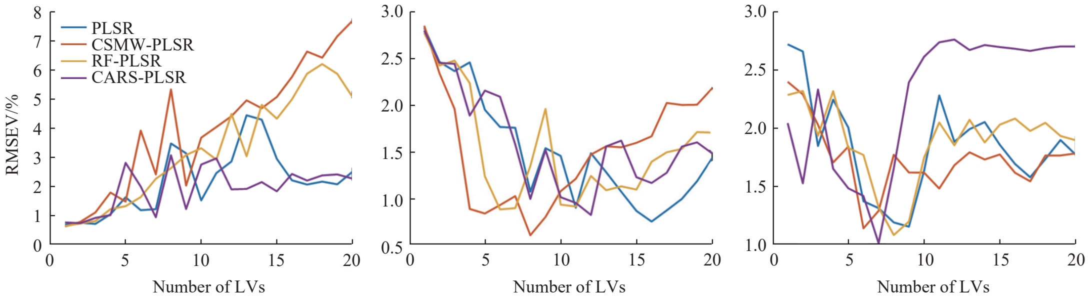
Table 3 describes the performance of external validation before and after model updating with different wavelength selection methods. For the external validation without model updating, the best result was obtained by the CSMW-PLSR of the in-line model with RMSEV of 0.920%. The results of in-line models were better than mixture models, and mixture models were better than off-line models. However, all models obtained high prediction bias with a mean RMSECV of 2.51%. There were many reasons for the deviation, such as biological variability and the difference in detection systems. This result could not meet the prediction requirement that the residual of SSC in in-line detection should be less than 1%. Therefore, it was necessary to update the models.

As is known, model updating depends on representative samples. However, it is difficult to manually and non-destructive screen samples with representative internal qualities. Therefore, ten samples were randomly selected and applied to both model updating methods. Figure 9 shows the LVs selection process of the LVU updating method. RMSEVs fluctuated violently with the number of LVs but showed a trend overall. Interestingly, RMSEVs first decreased and then increased in the in-line model and mixture model. However, one or two LVs obtained the best performance in the off-line models. It showed that the number of LVs has a great impact on the prediction performance of the model. This might be related to the interpretation degree of LVs to the model.

Table 3 External validation performance of PLSR models with wavelength selection and model updating methods

Batch	Modeling methods	EWs (Proportion)	Original models			Updated models			
			LVs	RMSECV/%	RMSEV/%	LVs of recalibration	RMSEV with recalibration/%	LVs of LVU	RMSEV with LVU/%
Off-line	PLSR	675 (100%)	13	0.572	4.421	13	1.384	1	0.664
	CSMW-PLSR	304 (45%)	12	0.444	4.390	11	1.817	1	0.599
	RF-PLSR	100 (15%)	16	0.447	4.970	16	1.696	1	0.613
	CARS-PLSR	61 (9%)	17	0.375	2.176	18	1.601	2	0.710
In-line	PLSR	675 (100%)	11	0.496	1.287	12	0.923	16	1.385
	CSMW-PLSR	332 (49%)	8	0.456	0.920	8	0.864	8	0.920
	RF-PLSR	120 (18%)	13	0.404	1.722	12	0.978	6	1.047
	CARS-PLSR	77 (11%)	17	0.287	1.748	12	0.965	12	1.350
Mixture	PLSR	675 (100%)	15	0.586	1.854	16	1.279	17	1.576
	CSMW-PLSR	331 (49%)	13	0.520	1.790	12	1.263	11	1.479
	RF-PLSR	59 (9%)	15	0.499	2.027	14	1.311	9	1.197
	CARS-PLSR	20 (3%)	12	0.461	2.757	10	1.279	7	1.010

Note: LVU, Latent variables updating.

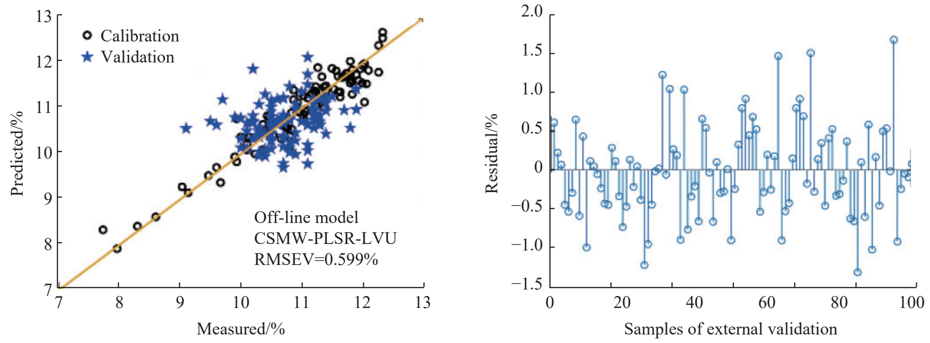


a. Off-line models b. In-line models c. Mixture models

Figure 9 Plots of RMSEV with increasing numbers of LVs (1-20) using 10 samples randomly selected in the external validation set

Both methods reduced the RMSEV of external validation data in each model, as listed in Table 3. However, the correction effects of the two methods showed differences due to their different principles. First, recalibration was to build a new coefficient matrix by mixing in new samples. Therefore, the selection of optimal LVs was similar to the original model. Its correction effect in the in-line model was better than that of LVU. The optimal performance was obtained by the CSMW-PLSR model of in-line data with RMSEV of 0.864%. In contrast, LVU optimized the number of LVs to match the original model with external validation data. This method corrected the overfitting or underfitting phenomenon and attempted to use the LVs of the original model to explain the external data. It showed a better correction effect in off-line models. The corrected RMSEVs of off-line models were less than 0.710%, and the minimum was 0.599% with the CSMW-PLSR model. For the mixture models, the correction effect was not as good as the off-line and in-line models, and the results of the two methods were similar. This might be caused by unstable factors in the two data sets that were added at the same time.

The wavelength selection methods performed randomness in the results of different update methods and different data sets. In contrast, CSMW-PLSR showed better robustness in in-line models and obtained the best performance in the off-line model with LUV correction. The best performance of all models was the CSMW-PLSR of the off-line model after LVU correction with RMSEV of 0.599%. Figure 10 shows the results of the performance of the best model. The residuals of 100 samples of external validation could meet the needs of practical applications of in-line detection with 91% residuals less than 1.0% and 60.0% less than 0.5%.



a. Scatters of measured values and predicted values of the calibration and external validation

b. Stem plots of residuals of samples in the external validation

Figure 10 Results of the best performance of external validation using model updating methods

4 Discussion

Compared with in-line detection, off-line static spectra acquisition was stable and controllable. However, due to the differences in operation state and integration time, the model or data in the off-line system was difficult to be directly applied to the in-line detection. In this study, the in-line system was designed based on the off-line system. The two systems had the same hardware composition, including fruit cup, collimator, tungsten halogen lamp, optical fiber, spectrometer, etc. Therefore, the off-line model was used as a standard to evaluate the performance of the in-line system. Moreover, to match the symmetrical illumination in in-line detection, the average spectrum with multi-point detection in the off-line system was used for modeling. Finally, the in-line detection prototype system could obtain stable spectral data with a spectral anomaly rate of less than 2% and could be used to establish a prediction model equal to or slightly better than the off-line detection. The transmission speed of 1.5 m/s fully met the needs of real-time detection. Moreover, the mixture model has good performance that proved the homology of data in the two systems. An exciting finding was that the off-line model could predict the in-line data through the LVU correction methods. This indicated that the laboratory data could participate in the improvement or even be directly applied to the in-line detection system. This was of great significance to reduce the modeling labor and improve data utilization of fruit.

The wavelength selection methods could find the internal law and improve the interpretability of models. For the CSMW-PLSR, the optimized band (711.54-956.48 nm) was covered the main absorption peaks of water and SSC, which was consistent with the prior knowledge of Vis-NIRS detection. However, the noise information might be mixed into the CSMW model since the strong correlation between adjacent wavelengths in the wide wavelength band. Therefore, CSMW was worse than CARS and RF in the improvement effect. The optimal wavelengths selected by CARS and RF usually appeared independently and distributed at different absorption peak positions. It is worth mentioning that both CARS and RF selected the wavelengths in the baseline part (500-550 nm and 960-1000 nm), such as 500 nm in CARS and 986 nm in RF. The baseline band was mainly dark noise in the transmittance spectra with a low signal-to-noise ratio. However, after pretreatment with SNV, the spectra were normalized to standard spectra with mean values of 0 and standard deviations of 1. At the same time, the dark baseline was given a value, which might be related to the spectral mean value and spectral change trend. These wavelengths might be beneficial for the models.

For the external validation, the model improvement of wavelength selection methods showed indeterminacy. This might be due to the overfitting caused by the large number of LVs. The model based on a small amount of data would prone to overfitting. When validated in a broader data set, the overfitting phenomenon would be amplified. Because, as the number of LVs decreased, the external validation performance improved. Especially when LVs were less than 10, the performance of external validation was relatively better. Therefore, it was recommended that the LVs should be controlled within 10.

The LVU method eliminates overfitting or underfitting in the local data modeling by correcting the number of LVs in the original models. The LVU method was verified in all models in which performed best in off-line models. Since the LVU was modified original models, these models needed to be well interpretable to external data. It is worth mentioning that the good performance of the LVU method for correcting off-line data might be related to the matched SSC ranges with the external validation set (the SSC ranges are listed in Table 1). For the in-line data set, LVU also worked, but it was not as effective as the off-line data. Ascribed to the new model mixing with information from external data, the recalibration method performed better than LVU for the model with SSC mismatch. In addition, the correction performance of recalibration was related to the selected new samples. The more the better, the more representative the better. It was a continuous process of change. However, for LVU, the correction result was discrete and the optimal solution was unique for a specific data set.

5 Conclusions

To achieve the real-time and nondestructive prediction of SSC in pomelo, this study explored the process from model optimization, external validation to model updating. The outlier detection method of MCS was used to eliminate abnormal samples and evaluate the stability of in-line spectra acquisition. The spectral anomaly rate of in-line detection was less than 2%. Subsequently, the wavelength selection methods significantly improved the performance of models in internal prediction with RMSEP of 0.240%-0.570%. However, large prediction deviations caused by overfitting were shown in external validation from a different year. Finally, a model updating method called LVU was proposed and verified for correcting the bias of external validations. Compared with recalibration, LVU used the original model to interpret external data, which was more suitable for data with SSC matched. Overall, the best external performance of inline detection was obtained by CSMW-PLSR-LVU with RMSEV of 0.599% in a transmission speed of 1.5 m/s. The residuals of 100 samples of external validation could meet the needs of practical applications of in-line detection with 91% residuals less than 1.0% and 60.0% less than 0.5%.The methods and models are of great significance for the internal quality detection of large fruit with thick rinds.

Acknowledgements

This work was financially supported by the key research and development projects of Zhejiang province (Grant No. 2022C02021).

[References]

[1] Goh R M V, Lau H, Liu S Q, Lassabliere B, Guervilly R, Sun J C, et al. Comparative analysis of pomelo volatiles using headspace-solid phase micro-extraction and solvent assisted flavour evaporation. LWT-Food Science And Technology, 2019; 99: 328–345.

[2] Cortés V, Blasco J, Aleixos N, Cubero S, Talens P. Monitoring strategies for quality control of agricultural products using visible and near-infrared spectroscopy: A review. Trends in Food Science & Technology, 2019; 85: 138–148.

[3] Obenland D, Collin S, Mackey B, Sievert J, Fjeld K, Arpaia M L. Determinants of flavor acceptability during the maturation of navel oranges. Postharvest Biology and Technology, 2009; 52(2): 156–163.

[4] Jie D F, Zhou W H, Wei X. Nondestructive detection of maturity of watermelon by spectral characteristic using NIR diffuse transmittance technique. Scientia Horticulturae, 2019; 257: 108718.

[5] Lu Y, Saeys W, Kim M, Peng Y, Lu R. Hyperspectral imaging technology for quality and safety evaluation of horticultural products: A review and celebration of the past 20-year progress. Postharvest Biology and Technology, 2020; 170: 111318.

[6] Srivastava R K, Talluri S, Beebi S K, Kumar B R. Magnetic resonance imaging for quality evaluation of fruits: A review. Food Analytical Methods, 2018; 11(10): 2943–2960.

[7] Pereira L F A, Janssens E, Cavalcanti G D C, Ren T I, Van Dael M,

Verboven P, et al. Inline discrete tomography system: Application to agricultural product inspection. Computers and Electronics in Agriculture, 2017; 138: 117–126.

[8] Zhang W, Lyu Z Z, Xiong S L. Nondestructive quality evaluation of agro-products using acoustic vibration methods - A review. Critical Reviews in Food Science and Nutrition, 2018; 58(14): 2386–2397.

[9] Xie L J, Wang A C, Xu H R, Fu X P, Ying Y B. Applications of near-infrared systems for quality evaluation of fruits: A review. Transactions of the ASABE, 2016; 59(2): 399–419.

[10] Arendse E, Fawole O A, Magwaza L S, Opara U L. Non-destructive prediction of internal and external quality attributes of fruit with thick rind: A review. Journal of Food Engineering, 2018; 217: 11–23.

[11] Magwaza L S, Opara U L, Nieuwoudt H, Cronje P J R, Saeys W, Nicolaï B. NIR spectroscopy applications for internal and external quality analysis of citrus fruit - A review. Food and Bioprocess Technology, 2012; 5(2): 425–444.

[12] Sun C J, Aernouts B, Van Beers R, Saeys W. Simulation of light propagation in citrus fruit using Monte Carlo multi-layered (MCML) method. Journal of Food Engineering, 2021; 291: 110225.

[13] Zhang H L, Zhan B S, Pan F, Luo W. Determination of soluble solids content in oranges using visible and near infrared full transmittance hyperspectral imaging with comparative analysis of models. Postharvest Biology and Technology, 2020; 163: 111148.

[14] Song J, Li G L, Yang X D, Liu X W, Xie L. Rapid analysis of soluble solid content in navel orange based on visible-near infrared spectroscopy combined with a swarm intelligence optimization method. Spectrochimica Acta Part A:Molecular and Biomolecular Spectroscopy, 2020; 228: 117815.

[15] Jie D F, Xie L J, Fu X P, Rao X Q, Ying Y B. Variable selection for partial least squares analysis of soluble solids content in watermelon using near-infrared diffuse transmission technique. Journal of Food Engineering, 2013; 118(4): 387–392.

[16] Tian H Q, Wang C G, Zhang H J, Yu Z H, Li J K. Measurement of soluble solids content in melon by transmittance spectroscopy. Sensor Letters, 2012; 10(1-2): 570–573.

[17] Li T-T, Wang H Y, Huang S-Y, Lou C-W, Lin J-H. Bioinspired foam composites resembling pomelo peel: Structural design and compressive, bursting and cushioning properties. Composites Part B:Engineering, 2019; 172: 290–298.

[18] Pell R J. Multiple outlier detection for multivariate calibration using robust statistical techniques. Chemometrics and Intelligent Laboratory Systems, 2000; 52(1): 87–104.

[19] Metz M, Abdelghafour F, Roger J-M, Lesnoff M. A novel robust PLS regression method inspired from boosting principles: RoBoost-PLSR. Analytica Chimica Acta, 2021; 1179: 338823.

[20] Zhou J L, Zhang S L, Wang J. A dual robustness projection to latent structure method and its application. IEEE Transactions on Industrial Electronics, 2021; 68(2): 1604–1614.

[21] Cappozzo A, Duponchel L, Greselin F, Murphy T B. Robust variable selection in the framework of classification with label noise and outliers: Applications to spectroscopic data in agri-food. Analytica Chimica Acta, 2021; 1153: 338245.

[22] Zhang L X, Wang D, Gao R R, Li P W, Zhang W, Mao J, et al. Improvement on enhanced Monte-Carlo outlier detection method. Chemometrics and Intelligent Laboratory Systems, 2016; 151: 89–94.

[23] Li H-D, Liang Y-Z, Cao D-S, Xu Q-S. Model-population analysis and its applications in chemical and biological modeling. TrAC Trends in Analytical Chemistry, 2012; 38: 154–162.

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

[25] Nordey T, Joas J, Davrieux F, Chillet M, Léchaudel M. Robust NIRS models for non-destructive prediction of mango internal quality. Scientia Horticulturae, 2017; 216: 51–57.

[26] Fan S X, Li J B, Xia Y, Tian X, Guo Z M, Huang W Q. Long-term evaluation of soluble solids content of apples with biological variability by using near-infrared spectroscopy and calibration transfer method. Postharvest Biology and Technology, 2019; 151: 79–87.

[27] Mishra P, Woltering E, Brouwer B, Hogeveen-van Echtelt E. Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with variable selection and model updating approach. Postharvest Biology and Technology, 2021; 171: 111348.

[28] Yun Y-H, Li H-D, Deng B-C, Cao D-S. An overview of variable selection methods in multivariate analysis of near-infrared spectra. TrAC Trends in Analytical Chemistry, 2019; 113: 102–115.

[29] Sun X D, Subedi P, Walsh K B. Achieving robustness to temperature change of a NIRS-PLSR model for intact mango fruit dry matter content. Postharvest Biology and Technology, 2020; 162: 111117.

[30] Tian H, Xu H R, Ying Y B. Can light penetrate through pomelos and carry information for the non-destructive prediction of soluble solid content using Vis-NIRS? Biosystems Engineering, 2022; 214: 152–164.

[31] Li H-D, Xu Q-S, Liang Y-Z. libPLS: An integrated library for partial least squares regression and linear discriminant analysis. Chemometrics and Intelligent Laboratory Systems, 2018; 176: 34–43.

[32] Boardman A E, Hui B S, Wold H. The partial least squares-fix point method of estimating interdependent systems with latent variables. Communications in Statistics - Theory and Methods, 1981; 10(7): 613–639.

[33] Kennard R W, Stone L A. Computer aided design of experiments. Technometrics, 1969; 11(1): 137–148.

[34] Du Y P, Liang Y Z, Jiang J H, Berry R J, Ozaki Y. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. Analytica Chimica Acta, 2004; 501(2): 183–191.

[35] Li H D, Xu Q S, Liang Y Z. Random frog: An efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification. Analytica Chimica Acta, 2012; 740: 20–26.

[36] Li H D, Liang Y Z, Xu Q S, Cao D S. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Analytica Chimica Acta, 2009; 648(1): 77–84.

[37] Feudale R N, Woody N A, Tan H W, Myles A J, Brown S D, Ferre J. Transfer of multivariate calibration models: A review. Chemometrics and Intelligent Laboratory Systems, 2002; 64(2): 181–92.

[38] Xu H R, Qi B, Sun T, Fu X P, Ying Y B. Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears. Journal of Food Engineering, 2012; 109(1): 142–147.

[39] Cayuela J A. Vis/NIR soluble solids prediction in intact oranges (*Citrus sinensis* L.) cv. Valencia Late by reflectance. Postharvest Biology and Technology, 2008; 47(1): 75–80.