NONDESTRUCTIVE TESTING OF SOLUBLE SOLIDS CONTENT IN CERASUS HUMILIS USING VISIBLE / NEAR-INFRARED SPECTROSCOPY COUPLED WITH WAVELENGTH SELECTION ALGORITHM

Bin Wang¹, Junlin He*¹, Shujuan Zhang*¹, Lili Li²
¹ College of Engineering, Shanxi Agricultural University, Taigu/China
² College of Information Science and Engineering, Shanxi Agricultural University, Taigu/China
Tel: +86-0354-6288400; E-mail: hejunlin26@126.com and zsujuan1@163.com
DOI: https://doi.org/10.35633/inmateh-61-28

Keywords: VIS-NIR spectroscopy, cerasus humilis, soluble solid, partial least square regression, nondestructive testing

ABSTRACT
Soluble solids content (SSC) is one of the most important quality attributes affecting the taste and maturity of fresh fruit. In this study, with the cerasus humilis fruit as the research object, a prediction model of soluble solid content (SSC) in cerasus humilis (CH) is established based on visible / near-infrared spectroscopy to explore a nondestructive testing method of the interior quality of CH. The visible / near-infrared spectral info (350-2500nm) of 160 CHs was collected to extract the reflection spectrum, establishing the linear model (PLSR) and non-linear model (LS-SVM) of CH’s spectral info and SSC. The prediction performance and stability of the model were justified using several statistical indicators namely correlation coefficient of the prediction set (Rp), the root mean square error of the prediction set (RMSEP), and the residual predictive deviation (RPD) index. Results showed that multiplicative scatter correction (MSC) was proved to be the best preprocessing method, UVE-CARS was the optimal method of dimension reduction, the quantities of characteristic wavelengths was 10 and the optimal model was UVE-CARS-PLSR, in which Rc is 0.8995, Rp is 0.8579, RMSEC is 0.8897, RMSEP is 0.9059, and RPD is 1.8766, indicating that the redundant data of the original spectrum can be reduced, the wavelength dimensions can be reduced, valid info can be retained and data processing can be simplified as UVE-CARS extracts characteristic wavelengths. Reference and theoretical basis are provided in this research for future research and development of portable detector and online sorting detection of CH internal quality.

INTRODUCTION
Cerasus humilis (Bge.) Sok. (CH) is a kind of rosaceae cherry. It is usually grown in sun-slope sandy land and mountain shrubs, or cultivated in gardens. The fruit ripens around August 20 every year. CH is usually found in the northern regions of Yellow River. CH has a strong reticular root system and is drought-resistant, making it not only fixate soil but also regulate ecosystem, and it is also known as the “fruit rich in calcium” for its high calcium content, and is the third generation exclusive in China. CH’s seed kernels are the main source of Yu Li Ren (Semen Pruni), an herb known for the idea of “homology of medicine and

1 Bin Wang, As. Ph.D. Stud. Eng.; Junlin He, Prof. Ph.D. Eng.; Shujuan Zhang, Prof. Ph.D. Eng.; 2 Lili Li, As. Ph.D. Stud. Eng.
food”. CH pulp can be consumed or made into juice, fruit wine, vinegar and preserved dried fruits, and it is good for health. The CH products have a unique taste and rich aroma and of high nutritional value, so it is known as a “super fruit”, and honoured as one of the three high-end fruits with American blueberry and Russian sea-buckthorn.

The internal qualities of CH are abundant. Soluble solid content (SSC) and titratable acidity (TA) will affect its taste and nutrition when it comes to its quality assessment, and they serve as the measurement standards and important indicators of CH’s maturity. Regular testing of CH’s internal qualities is destructive, complicated and time-consuming, and its tissues are often damaged, affecting its sales and edibility and lowering its values (Guo et al., 2010).

Near-internal spectroscopy has been successfully applied by domestic and overseas researchers to the fruit internal quality determination in recent years. There are in-depth studies on apples, mangos, tomatoes, kiwis, peaches, pears, strawberries and fresh dates. Ar et al., (2019) used near-infrared spectroscopy to predict the possibilities of internal qualities of persimmons such as SSC, Vc, total acid and hardness. The results showed that MSC is the optimal preprocessing method. With the establishment of a PLS calibration model, the optimal factor quantity of the SSC, Vc, total acid and hardness of persimmons is 17, 16, 12 and 12, respectively. Parpinello et al., (2013) studied a detection method, which combines near-infrared ray (NIR) measurement and glucose analysis, and the partial least squares (PLS) model based on cross-validation serves as the main statistical parameter where the prediction set determination coefficient is 0.82 and the standard error of prediction is 0.83%°Brix. Purwanto et al., (2015) used near-infrared spectroscopy to predict the SSC and acidity of the mango species known as “Gedong Gincu”. The results showed that different preprocessing methods play critical roles in terms of establishing accurate models of mango internal quality prediction. Maniwar et al., (2014) used visible light and short-wave near-infrared spectroscopy to establish a PLSR prediction model for the indicators of soluble solids content, titratable acid content, ascorbic acid content, ethanol concentration, and peel hardness of passion fruit. Studies showed that PLSR prediction model proves to have the best prediction performance on the SSC in passion fruit, and the prediction correlation coefficient is 0.923. Escribano et al., (2017) collected the NI spectra of sweet cherries within the wavelength of 729 to 975 nm, and established a PLS prediction model of SSC for sweet cherries under two temperature conditions. The results showed that the determination coefficient (R2) of SSC calibration set is 0.922 and 0.946 and the standard error is 0.612% and 0.792% when the temperature is 0°C and 23°C, respectively. Yu et al., (2017) tested the SSC in grapes based on NIR and RC, RMSEP and RMSEC of PLS are 0.83, 0.76 and 0.84 by using the orthogonal test. Sun et al., (2018) used visible / near-infrared semi-transmission spectroscopy to explore the influence of being unpeeled (complete) and peeled on the SSC detection accuracy of navel oranges. Studies showed that peel imposes significant impact on the SSC detection accuracy under the 5% confidence level. The correlation coefficient and root mean square error of the prediction set of the optimal PLS of the SSC in unpeeled and peeled navel oranges are 0.888 and 0.456% / 0.944 and 0.324%, respectively. Zhang et al., (2011) established the relations among visible light, near-infrared diffuse reflection spectrum (Vis/NIR) and the soluble tannin content in persimmons. The results showed that first derivative and detrending algorithms are the optimal preprocessing method, and the modified partial least squares (MPLS) demonstrated better prediction performance of the soluble tannin content in astringent persimmon, of which the RCV, RP2, RMSECV and RMSEP are 0.7227, 0.6785, 0.148 and 0.1763, respectively. The studies above show that it is feasible to use NIR to detect the internal qualities of fruits, but there is no research on the internal quality detection of CHs based on Vis/NIR spectroscopy.

With cerasus humilis “Nongda 5” as the research object, a prediction model of the SSC in Cerasus Humilis (CH) is established based on visible / near-infrared spectroscopy to explore a nondestructive testing method of the internal quality of CH, hoping to achieve a CH SSC prediction model of good stability and high prediction accuracy. ASD Field Spee3 spectrometer has been employed to collect DRS data, and a spectral preprocessing method has been selected optimally. Four dimension-reducing methods, which are PLSR regression coefficients (RC), competitive adaptive reweighted sampling (CARS), and successive projections algorithm (SPA) and uninformative variable elimination (UVE), are used to extract the characteristic wavelengths, and different prediction models are built combined with PLSR and LS-SVM.

They provide technical support for rapid, nondestructive, low-cost, and large-scale grading detection research of CH qualities.
MATERIALS AND METHODS

Cerasus humilis samples

Sampling was conducted on August 15, 2019. The sampling site was the Jinzhong Agricultural High-Tech Industrial Demonstration Zone Base in Taigu County, Shanxi Province, China (112°29'E, 37°23'N), and the variety was “Nongda 5”, sample growth state as shown in Fig. 1. Samples are of consistent maturity, shape and have no damage in order to minimize the influence of individual differences on experiment results. They were placed in a low-temperature fresh-keeping box which was transported to the laboratory the same day. The surface was wiped. Before the data acquisition, all samples were individually numbered and they were placed in an environment where temperature is 25°C and relative humidity is 20% for 6 hours to prevent the temperature from affecting the spectra and qualities.

![Fig. 1 - Growth state of cerasus humilis](image)

Table 1 shows the statistics of 160 samples.

| Sample parameters | Min. | Max. | Mean | Standard deviation | Variable coefficient (%) |
|-------------------|------|------|------|--------------------|--------------------------|
| Diameter [mm]     | 15.48| 28.66| 24.72| 3.62               | 14.64                    |
| Weight [g]        | 7.47 | 16.18| 11.39| 1.76               | 15.45                    |

Vis/NIR collection

A FieldSpec3 analytical spectral device (ASD, USA) was used to collect the VIS/NIR data of samples. The interval of spectral data is 1 nm, the number of scans is 30 times, the resolution is 3.5 nm, and the wavelength range is 350-2500 nm. Diffuse reflection was employed to sample the spectra. Each sample was scanned 3 times at an interval of 120 degrees above the equator, and its average value was taken as the final spectral data. The spectral experiment platform is shown in Fig. 2. The diffuse reflectance spectral data of a total of 160 CH samples were collected.

![Fig. 2 - Spectral experiment platform](image)

Determination of soluble solids

Refractive digital sugar content PR-101a (Atago, Japan) was employed to determine the SSC in CHs in accordance with GB/12295-1990. The scanned part of CHs was immediately sliced off after the spectra were collected for manual juicing, and the juice was filtered. Filtered juice was dripped onto the measurement window of the saccharimeter for SSC reading. Three measurements were taken and the mean value was used as the experimental value. The measurement window needs to be cleaned with distilled water and wiped clean after each measurement to avoid impact on experimental values.
Spectrum modeling
Spectral data preprocessing
Employing a feasible spectral preprocessing method can not only eliminate noise, but also minimize or reduce the impact of environmental factors such as lighting and instrumental factors such as random errors, ensuring the precision and validity of extracted spectral data (Zhang et al., 2012). Multiplicative scatter correction (MSC), Savitzky-Golay polynomial convolution smoothing (SG smoothing), de-trending, moving average (MA), median filtering (MF) and MSC + SG smoothing (5-point) were employed in this research, and PLSR model was used to assess the performance of different spectral preprocessed data.

Extraction of characteristic wavelengths
Four dimension-reducing methods, namely successive projections algorithm (SPA) (Jiang et al., 2016), competitive adaptive reweighed sampling (CARS) (Maniwar et al., 2014), PLSR regression coefficients (RC) (Liu et al., 2015) and uninformative variable elimination (UVE), were employed to extract characteristic wavelengths. The pros and cons of the prediction stability of the models built based on the aforementioned four methods were analysed, and they were put through secondary dimension reduction in combination with SPA and CARS for comparison of their pros and cons.

Modeling and prediction
Partial least squares regression (PLSR) is a multivariate linear modeling method that conducts linear fitting for curves with least square error sum, and it combines the advantages of correlation analysis, multivariate linear regression and principal components, and is widely applied in spectral modeling (Gao et al., 2019). It may comprehensively measure the sample spectral information and physical-chemical indicators at the same time to obtain the optimal model of calibration.

Least squares support vector machine (LS-SVM) is a modified and improved algorithm based on the principle of support vector machine (SVM). It is able to deal with the linear and non-linear problems in multivariate calibration modeling and resolving these relationships in a relatively fast way (Bao et al., 2015). Details of LS-SVM algorithm could be found in the literatures (Coen et al., 2006). LSSVM regression model was given as follows:

\[ y_{(x)} = \sum_{i=1}^{n} a_i K(x,x_i) + b \]  

(1)

Where:

- \( K(x,x_i) \) is the kernel function, \( x_i \) is the input vector, \( a_i \) is the Lagrange multiplier called support value, and \( b \) is the bias.

PLSR was established based on different preprocessing methods, and the best preprocessing method was selected in combination with a variety of characteristic wavelength extraction algorithms, establishing a full-spectrum (FS) and characteristic wavelength PLSR and LS-SVM models, respectively. To verify the prediction performance of different models, the prediction set samples were seen as input variables, and the prediction results of different models were compared and analysed to obtain the optimal prediction model.

Assessment of models
The following five indicators are usually selected to assess the precision and stability of models: the correlation coefficient of the calibration set (RC), the correlation coefficient of the prediction set (RP), the root mean square error of the calibration set (RMSEC), the root mean square error of the prediction set (RMSEP), and the residual predictive deviation (RPD); RPD is the ratio of standard deviation (SD) to RMSEP (Tamaki et al., 2015). The closer RC and RP are to 1, the smaller and closer the RMSEC and RMSEP are, the better the prediction performance and stability of the model, and the higher the precision. These assessment parameters were calculated as follows:

\[ R_c = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sqrt{\sum (\hat{y}_i - y)^2} \sqrt{\sum (\bar{y} - y)^2}} \]  

(2)

\[ R_p = \frac{\sum (\hat{y}_i - y)^2}{\sqrt{\sum (\hat{y}_i - y)^2} \sqrt{\sum (\hat{y}_i - y)^2}} \]  

(3)
### Results

#### Classification of Sample Sets

T2 ellipsometry (Galvão et al., 2015) was first used for abnormal sample detection before the classification, and no abnormal sample was detected. A total of 160 samples were classified into calibration set (120 samples) and prediction set (40 samples) randomly according to K-S. Table 2 shows the measurement statistics of the internal qualities of CH samples of both calibration and prediction sets.

| Sample quantity  | Indicators   | Min.  | Max.  | Mean ± Standard deviation |
|------------------|--------------|-------|-------|---------------------------|
| Calibration set  | SSC[^°Brix]  | 7.26  | 17.28 | 12.64±1.71                |
| Prediction set   | SSC[^°Brix]  | 9.42  | 15.44 | 12.56±1.70                |
| Total (160)      | SSC[^°Brix]  | 7.26  | 17.28 | 12.62±1.69                |

Table 2 shows that the ranges of SSC values of calibration sets and prediction sets are 7.26~17.28[^°Brix] and 9.42~15.44[^°Brix], and the SSC distribution of all samples is 7.26~17.28[^°Brix], and the SSC values of calibration sets and prediction sets are mean values of 12.64 and 12.56[^°Brix] with standard deviation (S.D.) of 1.71 and 1.70[^°Brix], respectively. Moreover, the SSC range of calibration set is bigger than that of prediction set, which is beneficial for the development of accurate and robust calibration models.

#### Analysis of Spectral Characteristics

Figure 3 shows the original near-infrared spectral curves of 160 CH samples, and it can be known that the trends of the curves of all samples have few differences and there is no significant abnormal sample.

![Fig. 3 - The original spectral diagrams of 160 CH samples](image)

Fig. 3 shows that the spectral curve is smooth in the range of 350~500nm where the reflectance value hardly changes. The reflectance of the sample rises rapidly after 690nm, and there are peaks at 610, 875, 1070, 1265, and 1570nm, and troughs at 680, 980, 1174, 1420, and 1660nm.
A significant absorption peak appeared at 680nm which was mainly the result of the fact that the chlorophyll on CH surface was absorbing spectra. Absorption peaks at 980, 1174, 1420, and 1660nm are the result of the strong absorption of water molecules, reflecting the MC of CHs within the waveband. 400-2450nm was chosen for experiment data processing since there are certain signal-to-noise ratio and low noise in the range of 350-399nm and 2451-2500nm.

**Comparison of PLSR modeling results of different preprocessing methods**

To eliminate the impact of exterior environmental factors as well as instrument noises on the DRS so that the collected spectra could have higher $SNR$ and the stability of prediction models could be improved, we have to look for the most effective preprocessing method, offering the best data for following modeling analysis. The spectral data obtained by different preprocessing methods are used as input variables of $PLSR$ to establish corresponding prediction models. The comparison of PLSR modeling based on different preprocessing methods is shown in Table 3.

| Pretreatment method       | Calibration set | Validation set | Prediction set | Factor quantity |
|---------------------------|-----------------|----------------|----------------|-----------------|
|                           | $R_c$           | RMSEC          | $R_cv$         | RMSECV          | $R_p$           | RMSEP          |                |
| Original spectrum         | 0.7302          | 0.9276         | 0.7223         | 1.3799          | 0.7406          | 1.3398         | 7              |
| MSC                       | 0.8511          | 0.8607         | 0.7587         | 1.3015          | 0.7939          | 1.2125         | 9              |
| S-G(5-point)              | 0.8302          | 0.9021         | 0.7231         | 1.3779          | 0.7406          | 1.3399         | 9              |
| De-trending               | 0.7979          | 1.1161         | 0.7217         | 1.3806          | 0.7304          | 1.3622         | 11             |
| MA                        | 0.8312          | 0.9176         | 0.7243         | 1.3754          | 0.7406          | 1.3401         | 9              |
| MF                        | 0.8351          | 0.9046         | 0.7305         | 1.3624          | 0.7464          | 1.3271         | 7              |
| MSC+S-G(5-point)          | 0.8302          | 0.9177         | 0.7158         | 1.3928          | 0.7282          | 1.3668         | 9              |

Based on how to assess a model, Table 3 shows that the $R_c$ is 0.8511, $R_p$ is 0.7939, and RMSEC of the model is 0.8607 when original spectra were processed with MSC, and the difference with RMSEP=1.2125 is minimal, which is 0.3158. Therefore, the model prediction performance is good, and MSC is proved to be the optimal preprocessing method. Fig. 4 shows the spectral curves processed with the MSC preprocessing method.

**Extraction of characteristic wavelengths**

**Successive projections algorithm (SPA)**

SPA was employed to select the characteristic wavelengths and a conclusion was drawn that the smaller the $RMSE$ values, the better the model's stability. Fig. 5 (a) shows the $RMSE$ distribution when different number of variables was selected by SPA. When the number of selected wavelength variables is 4, $RMSE$ is minimized, which is 1.0323; and Fig. 5 (b) shows the distribution of the number of characteristic wavelengths preferably selected by SPA. The four characteristic wavelengths selected are 1764, 615, 1259, and 2035nm, and their wavelength importance decreases in order.
Regression coefficient (RC)

Local extreme values of PLSR regression coefficients (RC) were used to select the number of characteristic wavebands. As shown in Fig. 6, 12 characteristic wavelengths were selected, which are 655, 695, 724, 777, 833, 931, 964, 993, 1102, 1187, 1334 and 1907nm.

Competitive adaptive reweighted sampling (CARS)

The process of CARS screening characteristic wavelengths is shown in Fig. 7. Monte Carlo Sampling was set 50 times and the number of cross-validation groups is 10. According to Fig. 7 (a), the wavelength number gradually decreased and stabilized at last as the number of sampling runs increased, which verified the rough and fine selection during wavelength screening. As Fig. 7 (b) shows, cross-validation RMSECV decreased gradually before it showed an increasing trend when the sampling runs increased to 34; when RMSECV decreased, it means that the null info among spectral info were eliminated; and when RMSECV increased, it means that valid info among spectral info were eliminated. Fig. 7 (c) shows that when the position of the line of “***” indicated the runs were 34, RMSECV was minimized, which is 0.9553. 19 characteristic wavelengths selected by CARS at this moment were 408, 531, 533, 652, 657, 728, 747, 940, 942, 943, 998, 1003, 1014, 1338, 2328, 2403, 2404, 2423, and 2435nm; these effective variables could be observed in Fig. 8.
Fig. 8 - Distribution of the optimal characteristic wavelength selected by CARS

Uninformative variable elimination (UVE)

UVE was employed to extract the characteristic wavelengths of CHs, and it was set as five interactive operation where different principal component quantities (6-17) were selected. RMSECV was minimized when the principal component quantity was 10, which was 1.0376, as shown in Fig. 9 (a). Fig. 9 (b) shows the stability distribution curve of UVE-PLSR when the principal component quantity was 10. There are the curves of 2,051 wavelength variable on the left of the vertical continuous line, and the curves of 2,051 randomly introduced variables on the right. The two horizontal dotted lines show the selection threshold of random variables (±26.03) where the threshold equals 99% of the maximized stability of random variables. Information bigger than the absolute threshold value is considered informative, namely the information in between the dotted lines are informative while the rest were uninformative. Therefore, 94 variables were determined to be effective variables that were shown in Fig. 9 (c) for SSC detection of cerasus humilis.

Fig. 9 - RMSECV distribution of different principal components (a), stability distribution curve of UVE-PLSR (b) and distribution of the effective variables in the raw spectral curve (c)

Distribution of the optimal characteristic wavelength selected by CARS

Extraction of characteristic wavelengths with secondary dimension-reduction

The dimensions of wavelengths were reduced and null info was eliminated after the use of the aforementioned 4 dimension-reducing methods to extract characteristic wavelengths, thus improving the stability and precision of models. The characteristic wavelengths extracted by UVE tend to have more variables, but there may be possible null info. SPA and CARS were combined, 5 and 10 characteristic wavelengths were selected respectively, which were 937, 1504, 992, 959 and 2163nm as well as 647, 652, 654, 660, 845, 923, 933, 940, 953 and 1082nm, these effective variables could be observed in Fig. 10 and Fig. 11, respectively.
Establishment of PLSR and LS-SVM models based on different variables

In order to compare the linear model (PLSR) and non-linear model (LS-SVM) for SSC prediction of CH. The characteristic wavelengths extracted by using full spectra by and 6 dimension-reducing methods (UVE, CARS, RC, SPA, UVE-SPA, UVE-CARS) were used as input to establish different PLSR and LS-SVM models, as shown in Table 4 and Table 5.

As shown in Table 4, in accordance with the modeling assessment principles, the comparison between PLSR models, which were built with the characteristic wavelengths extracted by 6 dimension-reducing methods as the input, and FS-PLSR (Full-spectrum-PLS), the quantity of wavelength variables showed significant reduction, and improvement of model stability and precision to different extents. In addition, Rc, RMSEC, Rp and RMSEP are all better than those of FS-PLSR. By comparing models of UVE-PLSR, CARS-PLSR, RC-PLSR, SPA-PLSR, UVE-SPA-PLSR, and UVE-CARS-PLSR, RC-PLSR retained 12 variables and showed poorer precision; CARS-PLSR and UVE-CARS-PLSR outperformed the rest, and the comparison between them two showed that their Rc, RMSEC, Rp and RMSEP are very close. When the variable quantity of UVE-CARS-PLSR is 10, Rc of UVE-CARS-PLSR is 0.8995, and Rp is 0.8579, both values are closer to 1; while RMSEC is 0.8897, RMSEP is 0.9059 and RPD is 1.8766, indicating that the models have good calibration and prediction performance, and the preferably selected 10 characteristic wavelengths may effectively reduce the dimensions of original spectral data.
Table 5

| Modeling method | Extraction method | Variable number | \([\gamma, \sigma^2]\) | Calibration set | Prediction set | RPD |
|-----------------|-------------------|-----------------|-------------------|-----------------|-----------------|-----|
| FS              |                   | 2051            | 1.78×10^2, 1.88×10^3 | 0.8558, 0.8464 | 0.8014, 0.9625 | 1.7662 |
| UVE             |                   | 94              | 1.32×10^2, 1.87×10^3 | 0.8762, 0.8914 | 0.8521, 0.9276 | 1.8327 |
| CARS            |                   | 19              | 6.68×10^2, 2.19×10^3 | 0.9025, 0.8614 | 0.8743, 0.9225 | 1.8428 |
| LS-SVM          | RC                | 12              | 2.43×10^3, 1.87×10^3 | 0.8644, 0.8999 | 0.7961, 0.9380 | 1.8124 |
|                 | SPA               | 4               | 6.66×10^2, 2.17×10^3 | 0.8352, 0.9828 | 0.7632, 0.9883 | 1.7201 |
|                 | UVE-SPA           | 5               | 8.28×10^2, 1.95×10^3 | 0.8211, 0.9576 | 0.7825, 0.9812 | 1.7326 |
|                 | UVE-CARS          | 10              | 4.86×10^3, 8.45×10^2 | 0.9097, 0.8528 | 0.8766, 0.9116 | 1.8649 |
|                 |                   |                 |                   | Rc = 0.8521, RMSEC = 0.8528, RMSEP = 0.9116, RPD = 1.8648 |

Table 5 illustrates the performance of the LS-SVM models in calibration and prediction. Compared with the establishment PLSR model, the established LS-SVM model also showed more satisfactory results. and FS-LS-SVM model coefficients of calibration set, prediction set and root mean square errors are: \(R_c = 0.8558, R_p = 0.8014, \text{RMSEC} = 0.8464, \text{RMSEP} = 0.9625, \text{RPD}=1.7662\). In SPA-LS-SVM model, the \(R_p\) of 0.7632 was the lowest and \(\text{RMSEP}\) of 0.9883 was the highest may be because variables with important information were eliminated by SPA, and UVE-SPA-LS-SVM model did so too. Although RC-LS-SVM model the number of variables was reduced to 12, that showed poorer precision. In addition, as shown in Table 5, UVE-CARS-LS-SVM model had better prediction performance with higher \(R_P\) of 0.8766, lower \(\text{RMSEP}\) of 0.9116 and higher \(\text{RPD}\) of 1.8648 than UVE-LS-SVM model (with \(R_p=0.8521, \text{RMSEP}=0.9276, \text{RPD}=1.8327\)). For CARS-LS-SVM model and UVE-CARS-LS-SVM model, that comparison between them two showed that their \(R_c, \text{RMSEC}, R_p \) and \(\text{RMSEP}\) are very close. However, fewer variables (only 10 variables) were used in UVE-CARS-LS-SVM model. \(R_c\) of UVE-CARS-LS-SVM is 0.9097, and \(R_p\) is 0.8766, both values are closer to 1; while \(\text{RMSEC}\) is 0.8528 and \(\text{RMSEP}\) is 0.9116. Therefore, in accordance with the modeling assessment principles, among all LS-SVM models, UVE-CARS-LS-SVM model was the best for effectively predicting.

RESULTS

As shown in Table 4 and Table 5, PLSR and LS-SVM models of CH SSC content based on full spectra and different wavelengths, all models can achieve the effective prediction. It can be observed that the optimal linear PLSR models (UVE-CARS-PLSR) has slightly similar prediction ability compared with the optimal LS-SVM models (UVE-CARS-LS-SVM), that their \(R_c, \text{RMSEC}, R_p \) and \(\text{RMSEP}\) are very close. The results indicated that UVE-CARS has the potential to select Vis/NIR spectroscopy effective wavelengths. However, UVE-CARS-PLSR model had better prediction performance with higher \(\text{RPD}\) of 1.8766 than UVE-CARS-LS-SVM model (with \(\text{RPD}=1.8649\)).

Fig.12 shows the scatter plots of measured and predicted SSC results of UVE-CARS-PLSR built with the preferred characteristic wavelengths with the use of UVE-CARS dimension-reducing algorithm.
CH SSC prediction models were established with MSC-UVE-CARS and PLSR in the research, and 10 characteristic wavelengths were extracted, obtaining $R_c$, $R_p$, RMSEC, and RMSEP, which are 0.8995, 0.8579, 0.8897, and 0.9059, and RPD is 1.8766. As for the PLSR model on FS, its $R_c$, $R_p$, RMSEC, and RMSEP are 0.7406, 0.7302, 0.9276 and 1.3398, and RPD is 1.2688. It means that MSC-UVE-CARS-PLSR reflects the characteristic spectra absorption of CH SSC more directly than FS-PLSR does. For example, the Ar et al., (2019) used MSC and PLS to establish a prediction model for the SSC in persimmons. According to the analysis, $R_c$ and RMSEC are 0.86 and 1.4866; RMSEP is 1.4663, RPD is 1.79 and CV is 9.84 when the calibration model factor quantity is 17. By comparison, the prediction models established by the author have better stability and precision, which means that different dimension-reducing methods used for original spectra to extract valid variables would improve the stability and prediction precision of models.

CONCLUSIONS

In this study, with cerasus humilis “Nongda 5” as the research object, the spectral info in between 400-2450nm were collected and reflection spectra were extracted to establish the PLSR and LS-SVM of CH spectral info and SSC, achieving the prediction of CH SSC. The main conclusion has been drawn as follows:

(1) The original spectra of CH samples were preprocessed with 6 methods and MSC was proved to be the best preprocessing method which improved the PLSR modeling performance. $R_c$ is 0.8511, $R_p$ is 0.7939, RMSEC is 0.8607 and RMSEP is 1.2125;

(2) Based on the MSC preprocessing method, UVE, CARS, RC, SPA, UVE-SPA and UVE-CARS methods were adopted to extract the characteristic wavelengths, and the numbers of preferred characteristic wavelengths were 94, 19, 12, 4, 5, and 10;

(3) Based on the full spectrum data, FS, UVE, CARS, RC, SPA, UVE-SPA and UVE-CARS were respectively adopted to extract the characteristic wavelengths to establish the linear model (PLSR) and non-linear model (LS-SVM). By comparison, the optimal model is proved to be UVE-CARS-PLSR, and its coefficients of calibration set, prediction set and root mean square errors are: $R_c = 0.8995$, $R_p = 0.8579$, RMSEC = 0.8897, RMSEP = 0.9059, RPD = 1.8766. Extracting characteristic wavelengths with UVE-CARS may cut down the redundant data of the original spectra, reduce the wavelength dimensions, and retain valid info, which can provide references and theoretical basis for subsequent portable detectors and online sorting detection research.

However, it should be noted that samples with consistent size, shape and have no damage were used in this study. In practice, however, size, shape and damage degree of samples is different. Thus, size, shape and damage degree parameter should be considered for establishment of models in future studies. In addition, establishment of various models based on different growing sites, more cultivars and storage days for developing the more accurate and robust prediction models, to improve the universality of the model are necessary.

ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (31271973); Key Research and Development Program of Shanxi Province, China (201703D221029-1)

REFERENCES

[1] Ar N. H., Purwanto Y. A., Budiastra I. W., (2019), Prediction of soluble solid content, vitamin C, total acid and firmness in astringent persimmon (Diospyros kaki L.) cv. Rendeu using NIR spectroscopy, Materials Science and Engineering, Vol.2019, Issue 557, pp.12-18, Amsterdam/Netherlands;

[2] Bao Y., Liu F., Kong W., et al., (2014), Measurement of soluble solid contents and pH of white vinegars using VIS/NIR spectroscopy and least squares support vector machine, Food and Bioprocess Technology, Vol.7, Issue 1, pp.54-61, New York/United States;

[3] Coen T., Saeys W., Ramon H., et al., (2006), Optimizing the tuning parameters of least squares support vector machines regression for NIR spectra, Journal of Chemometrics, Vol.20, Issue 5, pp.184-192, London/U.K.;
[4] Escribano S., BiasiW. V., LerudR., et al., (2017), Non-destructive prediction of soluble solids and dry matter content using NIR spectroscopy and its relationship with sensory quality in sweet cherries, Postharvest Biology and Technology, Vol.2017, Issue 128, pp.112-120, Amsterdam/Netherlands;

[5] Galvão R. K. H., Araújo M. C. U., José G. E., et al., (2005), A method for calibration and validation subset partitioning, Talanta, Vol.2005, Issue 67, pp.736-740, Amsterdam/Netherlands;

[6] Guo Zhiming, Huang Wenqian, Peng Yankun., et al., (2016), Colour compensation and comparison of shortwave near infrared and long wave near infrared spectroscopy for determination of soluble solids content of Fuji apple, Postharvest Biology and Technology, Vol.2016, Issue 115, pp.81-90, Amsterdam/Netherlands;

[7] Gao S., Wang Q., Li Q., et al., (2019), Non-destructive detection of vitamin c, sugar content and total acidity of red globe grape based on near infrared spectroscopy, Chinese Journal of Analytical Chemistry, Vol.47, Issue 6, pp.941-949, Beijing/China;

[8] Jiang S., Sun J., Xin Z., et al., (2016), Visualizing distribution of pesticide residues in mulberry leaves using NIR hyperspectral imaging, Journal of Food Process Engineering, Vol.40, Issue 4, pp.1-6, New York/United States;

[9] Liu D., Zhang S., Wang B., et al., (2015), Detection of hawthorn fruit defects using hyperspectral imaging, Spectroscopy and Spectral Analysis, Vol.35, Issue 11, pp.3167-3171, Beijing/China;

[10] Maniwar P., Nakano K., Boonyakiat D., et al., (2014), The use of visible and near infrared spectroscopy for evaluating passion fruit postharvest quality, Journal of Food Engineering, Vol.143, Issue 2, pp.33-43, Los Angeles/California;

[11] Maniwa P., Nakano K., Boonyakiat D., et al., (2014), The use of visible and near infrared spectroscopy for evaluating passion fruit postharvest quality, Journal of Food Engineering, Vol.2014, Issue 143, pp.33-43, Los Angeles/California;

[12] Parpinello G. P., Nunziatini G., Rombolà A. D., et al., (2013), Relationship between sensory and NIR spectroscopy in consumer preference of table grape (cv Italia), Postharvest biology and technology, Vol.2013, Issue 83, pp.47-53, Amsterdam/Netherlands;

[13] Purwanto Y. A., Sari H. P., Budiastra I. W., (2015), Effects of preprocessing techniques in developing a calibration model for soluble solid and acidity in 'Gedong Gincu' mango using NIR spectroscopy, International Journal of Engineering and Technology, Vol.2015, Issue 7, pp.1921-1927, Singapore/Singapore;

[14] Sun T., Mo X., Liu M., (2018), Effect of pericarp on prediction accuracy of soluble solid content in navel oranges by visible/near infrared spectroscopy, Spectroscopy and Spectral Analysis, Vol.38, Issue 5, pp.1406-1411, Beijing/China;

[15] Tamaki Y., Mazza G., (2011), Rapid determination of carbohydrates, ash, and extractives contents of straw using attenuated total reflectance Fourier transform mid-infrared spectroscopy, Journal of Agricultural and Food Chemistry, Vol.59, Issue 12, pp.6346-6352, Washington/United States;

[16] Yu J., Wang H., Sun X., et al., (2017), Parameter optimization in soluble solid content prediction of entire bunches of grape based on near infrared spectroscopic technique, Journal of Food Measurement and Characterization, Vol.11, Issue 4, pp.1676-1680, New York/United States;

[17] Zhang P., Li J., Meng X., et al., (2011), Research on nondestructive measurement of soluble tannin content of astringent persimmon using visible and near infrared diffuse reflection spectroscopy, Spectroscopy and Spectral Analysis, Vol.31, Issue 4, pp.951-954, Beijing/China;

[18] Zhang S., Zhang H., Zhao Y., et al., (2012), Comparison of modeling methods of fresh jujube soluble solids measurement by NIR spectroscopy, Transactions of the Chinese Society for Agricultural Machinery, Vol.43, Issue 3, pp.108-112, Beijing/China.