IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 25, 2019, accepted November 28, 2019, date of publication December 10, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2958841

Comparative Study of Two Different Strategies for Determination of Soluble Solids Content of Apples From Multiple Geographical Regions by Using FT-NIR Spectroscopy

GUANGZHAO TIAN, XIAONA LI, BAOHUA ZHANG, JUN ZHOU, AND BAOXING GU
College of Engineering, Nanjing Agricultural University, Nanjing 210031, China

Corresponding author: Baohua Zhang (bhzhang@njau.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Project 31901415, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20180515, and in part by the Key Research and Development Program of Jiangsu under Project BE2017370.

ABSTRACT Apple is one of the most popular fresh fruits with an extensive scope of regions owing to its nutrition and sweet in flavor. There is a large difference in the composition of the fruits growing in varying regions because of the variation in the growing regions, such as temperature, soil nutrients, etc. As a result, it is of significance to decrease the impact of region variability on the measurement of soluble solids content (SSC) in apples. To lessen the impact of region variability and enhance the predictive ability of on the model, our manuscript compared the performance of the two multi-region prediction models for the estimation of SSC in apples from multiple geographical regions. One multi-region prediction model was developed by merging SSC values and spectral data of all samples from multiple regions. The other multi-region prediction model was built for the determination of SSC in combination with region discriminant, model search strategy, and single-region models. Support vector machine (SVM) was applied to establish the model for discriminating the apples from multiple geographical regions. It was found that the region discriminant model achieved great results, with the classification accuracy of 99.52%. By comparing and analyzing the two multi-region prediction models, the optimal multi-region prediction model was obtained. Finally, to decrease the irrelevant spectral information and reduce the computational cost, the multi-region SSC prediction model was optimized in combination with various spectral preprocessing methods (multiple scatter correction (MSC), standard normal variate (SNV), and first derivative (FD) correction) and variable selection methods (Monte Carlo uninformative variables elimination (MC-UVE), competitive adaptive reweighted sampling (CARS), and random frog (RF)). The overall results denoted that it was more accurate to estimate SSC in apples from the different geographical regions by using the multi-region models based on the region discriminant model in combination with SNV preprocessing algorithm and MC-UVE variable selection algorithm, and the prediction accuracy preceded the single-region models.

INDEX TERMS FT-NIR spectroscopy, apple, SSC prediction, region discriminant, multi-region model measurement.

I. INTRODUCTION

Apple is one of the most popular fresh fruits for the consumers owing to its nutrition and sweet in flavor. Soluble solids content (SSC) is a major internal parameter that affects the flavor, postharvest storage requirements, and harvest time of apples [1]. In consequence, the development of a reliable and fast SSC estimation approach is of great significance to satisfy the growing market demands for high-quality fruit [2]. In the past, a variety of standard analytical approaches such as high-performance liquid chromatography [3], and gas chromatography [4], [5] have been employed to evaluate the quality of fruit. However, most of them exhibit great reliability and high accuracy but have some certain limitations,

The associate editor coordinating the review of this manuscript and approving it for publication was Liandong Zhu.

namely, time-consuming, expensive, and destructive. NIR spectroscopy has been considered as a replacement of the traditional destructive analytical methods during the past decades since it is nondestructive, fast, accurate, and economically reasonable [6].

The recorded NIR spectra comprise both chemical and physical information of the irradiated samples [7]. In combination with a suitable predictive model, Flourier transformation NIR (FT-NIR) spectroscopy has been demonstrated to be a fast and accurate analytical technique which is utilized for quantifying SSC in the assessment of fruits and vegetables, such as melon [8], apple [9], [10], and peach [11]. Giovanelli *et al.* [12] probed into the feasibility of NIR spectroscopy in optimizing post-harvest management and following fruit quality changes during storage. It was found that the average correct classification of validation set was higher than 93% and that of classification set nears 100%, confirming that NIR spectroscopy was a valid technology to measure internal quality parameters of apples. More applications of FT-NIR spectroscopy applied for the quality assessment of fruits and vegetables were researched by Alamar *et al.* [13] and Wang and Han [14].

Apple is an extensively cultivated fruit with an extensive scope of regions. Because of large-scale planting areas and changes in the growing environment (i.e., lighting effect, temperature, soil nutrient, rainfall), the composition of the fruits produced in diverse regions varies greatly [15]. Zhang *et al.* [1] gave a detailed summary of the impact of physical and biological variability which includes region variability and compensation methods for eliminating the effects in fruit and vegetable quality non-invasive assessment by applying imaging and NIR spectroscopy technology. In addition, Alamar *et al.* [13] researched the influence of region variability on the NIR spectroscopy assessment of SSC values. Functional analysis of variance was applied to interpret the variance in the spectra with regard to biological variability. It was concluded that the impact of region variability on the NIR spectra was great of significance. Therefore, it is necessary to decrease the effect of region variability on the quality measurement of apples.

Nevertheless, there were few reports that using FT-NIR spectroscopy to decrease the impacts of geographical region variability on the assessment of SSC values. Nowadays, two main approaches to determine the SSC in apples from multiple regions were reported. One approach was to construct the prediction model by merging the spectral data and SSC values of all studied samples [16]. For the second approach, the prediction model was established based on the region discriminant and the single-region model [7], [17]. Consequently, in order to lessen the impact of region variability on the SSC assessment model, our manuscript compared the performance of the two multi-region prediction models mentioned above on the determination of SSC values in apples from multiple geographical regions. Then, in an attempt to enhance the predictive ability, the multi-region model optimization was performed by combining with the spectral preprocessing algorithms and variable selection algorithms.

II. MATERIALS AND METHODS

A. SAMPLES PREPARATION

In this research, the apple samples were collected from 'Fuji' apple commercial orchards in Akesu, Qixia, and Yichuan of China. 208 samples in total with no damages were washed, dried, numbered, and then stored in the laboratory (20° temperature and 60% relative humidity) for 24h. All samples from each region (76 samples came from Akesu, 72 samples came from Qixia, and 60 samples came from Yichuan) were marked around the equator and scanned by the FT-NIR spectrometer. In order to avoid errors, three replicate measurements were made on all the apple samples that are utilized for this experiment, and the mean of three measurements was utilized for the subsequent analysis.

B. SPECTRA ACQUISITION

The frame diagram of the FT-NIR spectral data acquisition system for samples was depicted in Fig. 1a. The spectral data of apple samples were collected in the reflectance mode by an Antaris II FI-NIR spectrometer (Thermo Electron Co., USA) which equipped with an InGaAs detector with high sensitivity, an integrating sphere, and a tungsten lamp (20W) [18]. The detector covered the spectral range from 10,000 to 4000cm^{-1} (1000-2500nm), and the spectral resolution of this spectrometer is 1.928cm^{-1}. The intact apples were put into fruit holder of FT-NIR spectrometerone by one, finally getting 3112 spectral variables per sample. At the equidistant position around the equator of each apple, three reflection spectra were measured, and the averaged spectrum of each apple was utilized as the raw spectrum of the sample for analysis.

C. SSC MEASUREMENT

After the spectral data were acquired, the SSC values in apples were measured immediately by a conventional destructive method. At the same points of the spectra measurement, three slices of fresh with peel were cut from the equidistant points at the equator of each apple, respectively. Then, the juice from each apple was pressed and dripped onto a temperature-compensated refractometer to measure the actual SSC values. By averaging from the soluble solids at the three different positions of apple, the value of SSC in each sample was obtained as a reference for the FT-NIR spectral analysis approach.

D. DATA ANALYSIS METHODS

1) SPECTRAL DATASET DIVISION AND SPECTRAL PREPROCESSING

All the calculations in this paper were performed by self-editing programs in MATLAB R2018b under Windows 10 with 3.4 GHz CPU and 16GB memory. Prior to the step of building the calibration model, the linear correlations

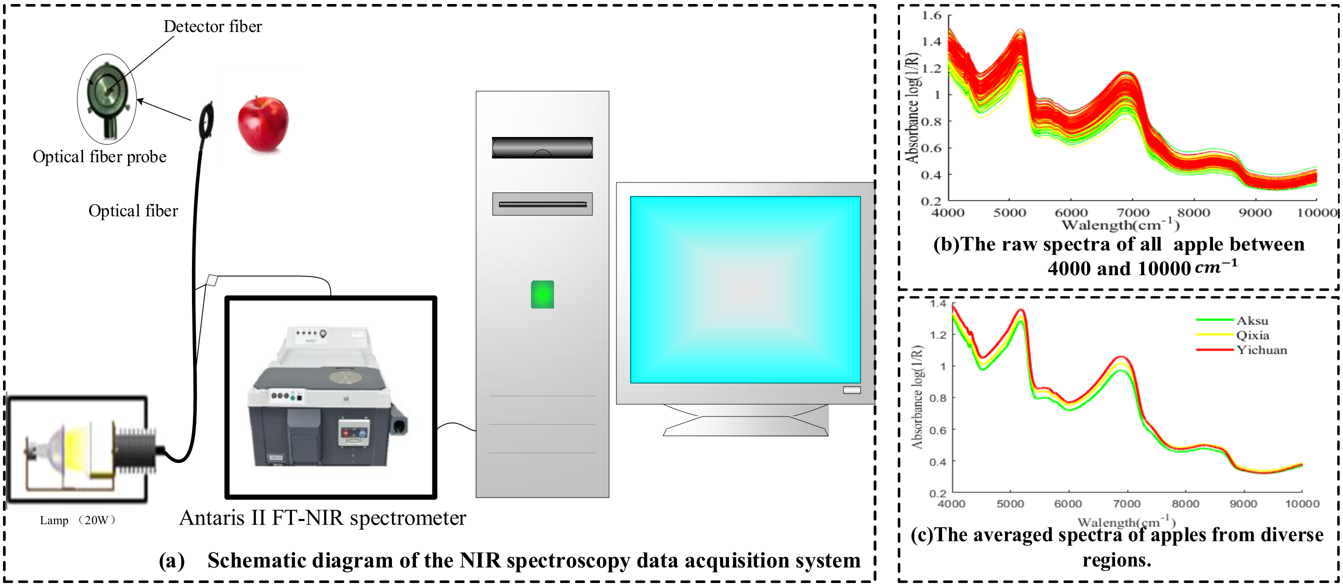

Detector fiber
Optical fiber probe
Optical fiber
Lamp （20W）
Antaris II FT-NIR spectrometer
(a) Schematic diagram of the NIR spectroscopy data acquisition system
Absorbance log(1/R)
Walength(cm⁻¹)
(b)The raw spectra of all apple between 4000 and 10000 cm⁻¹
Aksu
Qixia
Yichuan
(c)The averaged spectra of apples from diverse regions.

FIGURE 1. The frame diagram of the FT-NIR spectral data acquisition system for samples and the spectral data of apples between 4000 and 10000cm^{-1}.

between FT-NIR spectra and SSC measurements were acquired by converting the reflectance spectra (R) to absorbance value (log (1/R)) [19]. For the purpose of getting a more stable and predictable regression model, the apple samples were segmented into two subsets on the basis of the Kennard-Stone (KS) algorithm. The first one was the training sets, which were applied to construct the calibration models, while the other was the test sets, which were applied to detect the accuracy of the developed models. In this study, the KS algorithm was used to divide the samples from every region into the training set and test set according to a ratio of 3:1, respectively.

The entire data processing generally includes the following several procedures: spectral preprocessing, variable selection, model calibration, and model evaluation [20]. The spectral data needed to be preprocessed to eliminate multiplicative and additive impacts in the spectra and improve the subsequent multivariate analysis. In our study, we applied and compared several commonly used spectral preprocessing approaches, which included multiple scatter correction (MSC), standard normal variate (SNV), and first derivative (FD) correction. The detailed description of these data preprocessing approaches could be searched in the literature of [21] and [22].

2) VARIABLE SELECTION

To decrease the calculative burden of spectra data, enhance the efficiency of the detection, and predigest the model for estimating SSC values of apples, variable selection is a crucial and inevitable procedure to select the optimal variables [20]. In this research, four wavelength selection algorithms were employed to extract effective variables with the highest predictive capacity, including Monte Carlo uninformative variables elimination (MC-UVE), competitive adaptive reweighted sampling (CARS), and random frog (RF). A detailed description of these variable selection methods can be found in the literature [7].

3) PARTIAL LEAST SQUARE (PLS) REGRESSION

PLS regression is an extensively applied chemometric approach for developing calibration models in NIR spectral analysis [23]. It has the advantage to solve the situation when the input variables contain noise and are highly correlated, but also applicable when the matrix of predictors has more wavelengths than that of observations [24]. In recent decades, the PLS algorithm was welcome to develop many calibration models for fruits and vegetables in present chemometric analysis, and a lot of applications are reported on apples [9]. The principle of the PLS algorithm is to search a set of latent variables (LVs) by projecting the X variables and the Y variables into a new latent space under the constraint of maximization of covariance between inputs and outputs [25]. In this manuscript, we applied the PLS regression algorithm to establish the calibration models to determine the SSC values of apples. When establishing the PLS model, it is of significance to utilize cross-validation of the calibration sets to determine the optimal number of LVs. The optimal number of LVs was determined by performing 10-fold cross-validation of calibration sets until the root mean square error of cross-validation (RMSECV) attained the minimum [7].

In terms of the correlation coefficient of calibration (R_c) and the root mean square error of calibration (RMSEC) between the predicted values and the measured values in the training set, the performance of the calibration model was evaluated [15]. Likewise, the correlation coefficient of prediction (R_p) and the root mean square error of prediction (RMSEP) are applied to evaluate the prediction model. RMSEC,RMSEP, R_c, and R_p are defined in the following equations:

$$\text{RMSEC} = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n_c}(\hat{y}_i - y_i)^2} \tag{1}$$

$$\text{RMSEP} = \sqrt{\frac{1}{n_p}\sum_{i=1}^{n_p}(\hat{y}_i - y_i)^2} \tag{2}$$

$$R_c = \sqrt{\frac{\sum_{i=1}^{n_c}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_c}(\hat{y}_i - y_{cm})^2}} \tag{3}$$

$$R_p = \sqrt{\frac{\sum_{i=1}^{n_p}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_p}(\hat{y}_i - y_{pm})^2}} \tag{4}$$

where, $\hat{y}_i$ is the predicted value of the ith observation, y_i is the measured value of the ith observation, y_{cm} or y_{pm} is the mean value of the calibration or prediction set. n_c and n_p are separately the number of observations in the training and test set. Generally, a great model needs to possess high R_c and R_p values and low RMSEC and RMSEP values [7].

E. REGION DISCRIMINANT WITH SUPPORT VECTOR MACHINE (SVM)

SVM is a statistical learning model on the basis of structural risk minimization which analyzes data applied to perform non-probabilistic binary linear classification or multivariate function assessment [26]. Although initially designed for binary classification, the basic SVM algorithm can be extended to the multi-class discrimination task [27]. Nowadays, SVM has been extensively utilized for supervised pattern recognition. Compared with other machine learning algorithms, this algorithm develops a model with fewer training samples, thus overcoming the local minimum required for the neural network. The reader can refer to the tutorials and mathematical explanations about SVM in detail [28]. In order to avoid over-fitting, the complexity (c value) of the model is determined by a penalty error function. There are three diverse kernel functions applied in establishing SVM models, including radial basis function (RBF), polynomial, linear kernel functions. In this study, SVM with the polynomial as the kernel function was selected for classifying samples in diverse geographical regions. Furthermore, we use a grid-search and cross-validation procedure to get the best kernel function parameters (g) and varying penalty parameters (c) to achieve the highest recognition rate. All the SVM computations were performed by using LIBSVM (version 3. 24) package in MATLAB R2018b.

For the purpose of obtaining an SVM classification model with good performance, we evaluated the SVM model based on the statistical parameters of the correct classification rate (CCR) [27]. CCR is defined in the following equations:

$$\text{CCR} = \frac{N_{Right}}{N_{all}} \tag{5}$$

where, N_{all} and N_{Right} refer to the total number of samples in that class and the numbers of samples that are correctly classified, respectively. For instance, in terms of samples from Aksu, N_{all} is the number of samples from Aksu in total, and N_{Right} refers to the number of samples from Aksu when they actually belong to samples from Aksu.

F. DEVELOPMENT OF TWO TYPES OF MULTI-REGION SSC PREDICTION MODEL

To examine the influence of the geographical region on the FT-NIR spectral analysis of SSC in apples, the PLS algorithm was utilized to establish the multi-region model of SSC. In this study, two different strategies were proposed to quantitatively determinate the SSC in 'Fuji' apples from multiple geographical regions to decrease the effect of region variability. In terms of the first strategy, the other multi-region SSC prediction model (multi-region prediction model_1) was developed by merging SSC values and spectral data of all studied samples from diverse regions. That was to say, SSC values and spectral data of all studied apple samples were used as input for constructing a multi-region model.

For the second strategy, one multi-region SSC prediction model (multi-region prediction model_2) was established for determining SSC values in 'Fuji' apples on the basis of region discriminant model, model search strategy, and multivariate regression analysis. For the single-region models, they were constructed based on varying data (spectral and SSC of apple samples coming from Aksu, Qixia, and Yichuan). The multi-region SSC prediction was performed in the following procedures: (1) separately establishing three specific single-region prediction models for estimating the SSC values of apple samples from Aksu, Qixia, and Yichuan; (2) discriminating and identifying the geographical region of the unknown apple samples through region discriminant model; (3) developing the multi-region SSC prediction model_2 by PLS algorithm in combination with the results of the region discriminant, model search strategy, and the single-region SSC prediction models.

To obtain an SSC prediction model with a more stable and predictive property, the two multi-region SSC prediction models were employed and compared to separately determine SSC in apples from multiple geographical regions. The optimal multi-region SSC prediction model was obtained by comparing and analyzing the two multi-region SSC prediction models. Then, in order to enhance the predictive ability, the multi-region prediction model optimization was performed by combining with the spectral preprocessing algorithms and variable selection algorithms. Fig. 2 showed the flowchart for the determination of SSC values in apples from the multiple geographical regions based on two multi-region models.

III. RESULTS AND DISCUSSION

A. FT-NIR SPECTRAL ANALYSIS AND SAMPLE DIVISION

In this study, the FT-NIR spectra were collected from apple samples within the range from 4000 to 10000cm^{-1}, finally

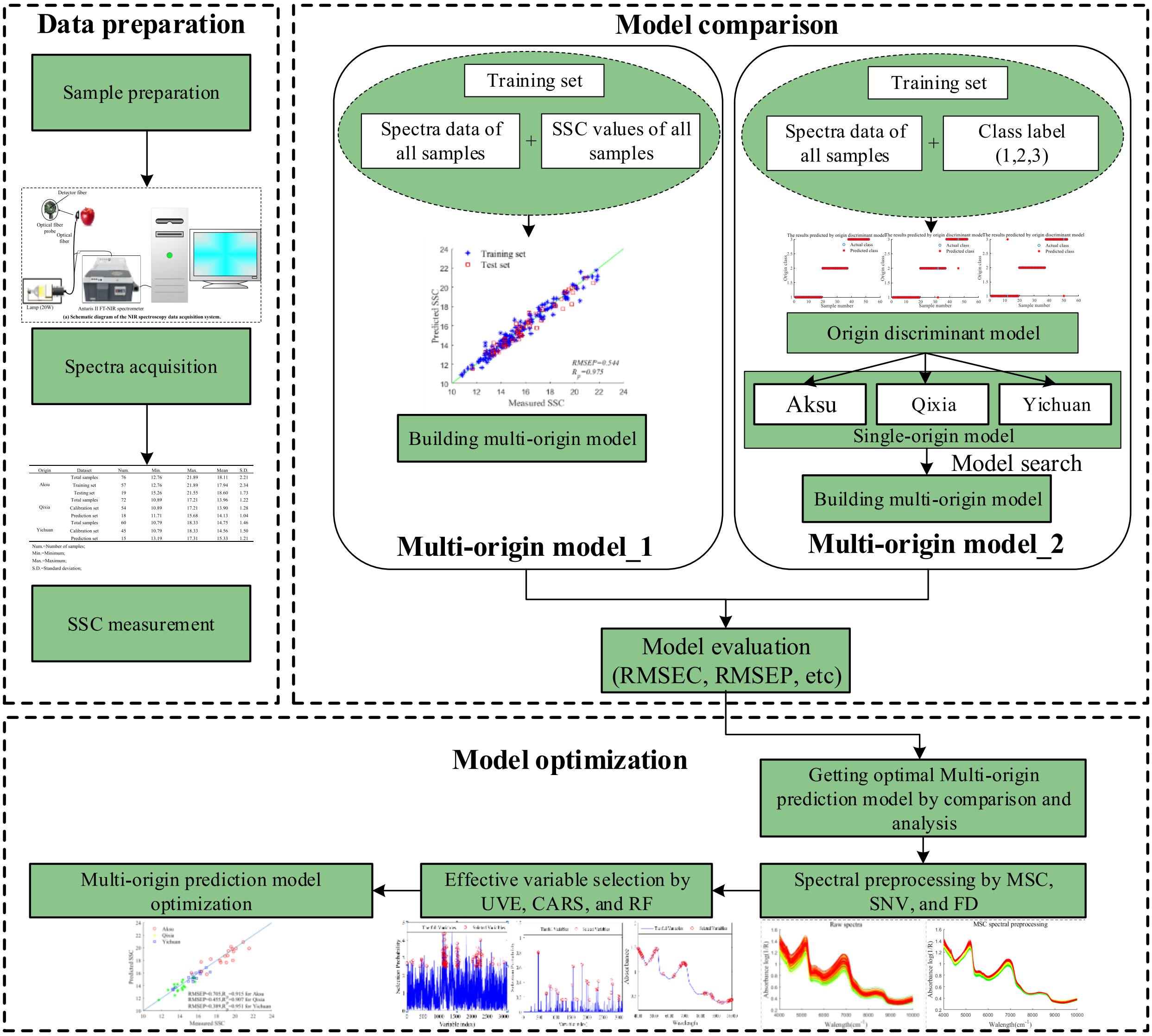

Data preparation
Sample preparation
Spectra acquisition
SSC measurement
Model comparison
Training set
Spectra data of all samples
+
SSC values of all samples
Building multi-origin model
Multi-origin model_1
Training set
Spectra data of all samples
+
Class label (1,2,3)
Origin discriminant model
Aksu
Qixia
Yichuan
Single-origin model
Model search
Building multi-origin model
Multi-origin model_2
Model evaluation (RMSEC, RMSEP, etc)
Model optimization
Getting optimal Multi-origin prediction model by comparison and analysis
Spectral preprocessing by MSC, SNV, and FD
Effective variable selection by UVE, CARS, and RF
Multi-origin prediction model optimization
Measured SSC
Predicted SSC
Training set
Test set
RMSEP=0.544
Raw spectra
MSC spectral preprocessing

FIGURE 2. The flowchart for the determination of SSC values in apples from the multiple geographical regions based on two multi-region models.

getting 3112 data points per spectrum. The original FT-NIR absorbance spectra of 208 apple samples from three geographical regions are shown in Fig. 1b. In this figure, the spectra curves of 'Fuji' apples coming from Aksu, Qixia, and Yichuan were depicted separately by the green, yellow, and red lines. It could be observed that there are two strong absorption peaks around 5195cm^{-1} and 6917cm^{-1}. The absorption peak around 5195cm^{-1} was related with the first overtone of O-H stretching, while the absorption peak around 6917cm^{-1} was attributed to the combination of the band of the second overtone of O-H stretching [16]. The graphic also revealed that there were some overlapping and crossovers among these spectra, however, the apple samples from diverse geographical regions had similar spectral trends. Fig. 1c showed the averaged absorbance spectra of apple samples which came from three regions. As shown in Fig. 1c, a variation of the spectral intensity of samples in varying regions existed. This variability might be influenced by the geographical region which included the factor of light effects, nutrition, soil characteristics, as well as weather conditions.

According to the KS algorithm, the apple samples from three regions of Aksu, Qixia, and Yichuan were respectively partitioned into the training sets and test sets with a proportion of 3:1. The reference measurement results of SSC in the training sample sets and test sample sets from three diverse regions were pooled to create reference data for the multi-region SSC model. The mean and standard deviation values of samples from each region were depicted in Table 1. For the apple samples from each region, there is a relatively large SSC change scope with 12.76-21.89°Brix for Aksu apples, 10.89-17.21°Brix for Qixia apples, and 10.79-18.33°Brix for Yichuan apples, respectively. In addition, the ranges of SSC

TABLE 1. The mean and standard deviation values of SSC measured in samples from three regions.

Region	Dataset	Num.	Min.(°Brix)	Max. (°Brix)	Mean(°Brix)	S.D. (°Brix)
Aksu	Total samples	76	12.76	21.89	18.11	2.21
	Training set	57	12.76	21.89	17.94	2.34
	Test set	19	15.26	21.55	18.60	1.73
Qixia	Total samples	72	10.89	17.21	13.96	1.22
	Training set	54	10.89	17.21	13.90	1.28
	Test set	18	11.71	15.68	14.13	1.04
Yichuan	Total samples	60	10.79	18.33	14.75	1.46
	Training set	45	10.79	18.33	14.56	1.50
	Test set	15	13.19	17.31	15.33	1.21

Num.=Number of samples;
Min.=Minimum;
Max.=Maximum;
S.D.=Standard deviation;

TABLE 2. Results of region discriminant models by SVM algorithm using different kernel functions.

Kernel function	SVs	CCR in the training set and test set		Overall accuracy(%)
		Training set (%)	Test set (%)	
Linear	45	99.36	100	99.52
Polynomial (degree=3)	34	100.00	96.15	99.04
RBF	81	93.59	96.15	94.23

value in training sets covered the ranges of in the test sets, which was helpful for establishing a good calibration model.

B. RESULTS OF REGION DISCRIMINANT USING SVM

SVM was utilized to build models for discrimination of 'Fuji' apples from diverse geographical regions. To estimate properly the predictive capacity of the region discriminant model established, the data obtained from 208 apple samples were sorted into a training set and a test set. The training set consisted of 156 samples (Aksu: 57; Qixia: 54; Yichuan: 45), while the test set was composed of the remaining 52 samples (Aksu: 19; Qixia: 18; Yichuan: 15). To discriminant the diverse geographical regions of the samples, this study applied the SVM algorithm to construct a region discriminant model by using the spectral data as inputs. In the SVM model, the X variables were related to the spectral data and the Y variables were associated with class labels of the region. In terms of Y variables, three geographical regions of samples were labeled, where 1, 2, and 3 separately stand for Aksu, Qixia, and Yichuan.

The results of region discriminant models by the SVM algorithm using different kernel functions were presented in Table 2 and Fig. 3. After the comparison of the overall accuracy of the region discriminant models by the SVM algorithm using different kernel functions, the best kernel function was found to be the Linear with the support vectors (SVs). The kernel function parameter of c was 256 and the kernel function parameter of g was 0.011. The excellent results were acquired, and the classification accuracy for training and test set were 99.36% and 100%, respectively. When the best region discriminant model was applied to recognize the region of all studied apple samples, the overall accuracy of 99.52% was obtained. In Fig. 3, the region discriminant model was carried out to predict the region of samples in the test set. In this figure, the little blue circle stood for the actual region variety, while the region predicted by the region discriminant model was described with the red asterisk. As can be seen in Fig. 3, the region discriminant model performed an excellent classification rate. Thus, this region discriminant model was identified to be appropriate to build the multi-region to use for the establishment of multi-region prediction model_2. Then, according to the result of the region discriminant model, the corresponding single-region SSC prediction model was selected. In this way, the multi-region prediction model_2 was built.

C. COMPARISON ANALYSIS OF TWO MULTI-REGION SSC PREDICTION MODELS

The performances of different multi-region models and different single-region models for assessing SSC in apple

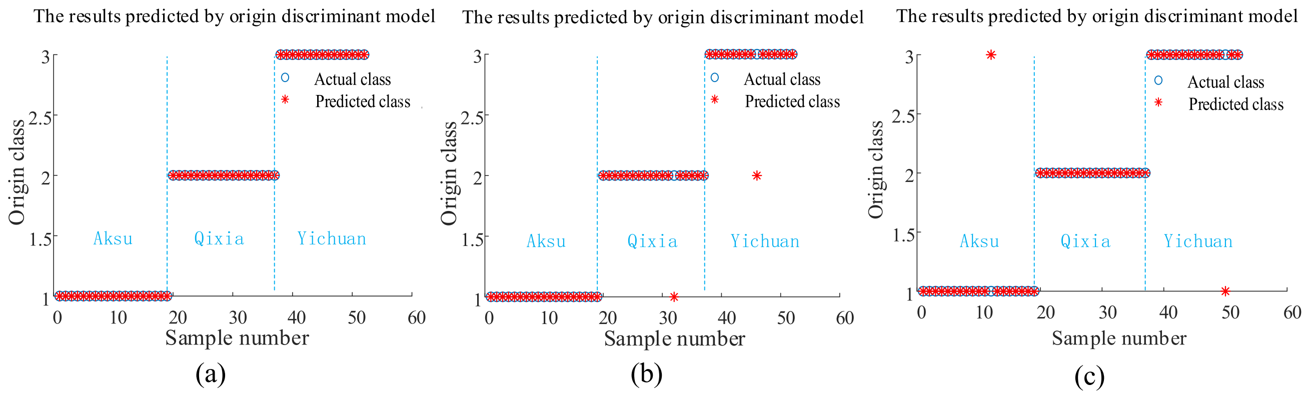


FIGURE 3. The results of region discriminant models by the SVM algorithm based on different kernel function (a) Linear kernel function (b) Polynomial kernel function (c) RBF kernel function.

TABLE 3. The training and test results determined for SSC by different multi-region models and different single-region models.

Prediction Model	LVs	Training set		Test sets					
				Aksu		Qixia		Yichuan	
		RMSEC (°Brix)	R_c	RMSEP (°Brix)	R_p	RMSEP (°Brix)	R_p	RMSEP (°Brix)	R_p
Single-region model (Aksu)	10	0.607	0.965	0.753	0.898	0.957	0.819	0.962	0.902
Single-region model (Qixia)	7	0.459	0.932	1.504	0.869	0.612	0.799	1.735	0.606
Single-region model (Yichuan)	14	0.352	0.995	1.442	0.892	0.710	0.933	0.495	0.950
Multi-region model_1	14	0.489	0.981	0.705	0.915	0.455	0.907	0.389	0.951
Multi-region model_2	10/7/14	-	-	0.753	0.898	0.612	0.799	0.495	0.950

samples from varying geographical regions were illustrated in Table 3. For the purpose to enhance the prediction accuracy and avoid the over-fitting of the data, the optimal LVs were determined according to the minimum RMSECV by applying cross-validation. As depicted in Table 3, single-region prediction models were established separately by using the training set of per region (Aksu, Qixia, and Yichuan), and then they were validated by employing all the test sets of the diverse regions. For the single-region prediction models (Aksu model, Qixia model, and Yichuan model), they had the lowest RMSECV (0.810, 0.589, and 0.401°Brix) for test sets when the number of LVs were 10, 7, and 14, respectively. It was found that the single-region SSC prediction model obtained satisfying prediction results if both the training set of a single-region model and the test set used to be predicted came from a similar geographical region. However, if the single-region prediction model was applied to assess the apple samples from other regions, the results of this model did not possess reliable prediction accuracy (higher RMSEP and lower R_p). For example, when the single-region model built with the training set from Qixia was used to predicting the SSC values of samples from Qixia, the RMSEP and R_p values were 0.612 °Brix and 0.799, respectively, while the RMSEP and R_p values were separately 1.504, 1.735 °Brixand 0.869, 0.606 when this model was utilized to estimate the SSC values of apple samples from Aksu and Yichuan. Obviously, the prediction results of the single-region model of Qixia was used to predict the samples from Qixia were far great than those of this model used to assess the samples from Aksu and Yichuan. The results indicated that the single-region prediction model was sensitive to the variability of the geographical region and not dependable enough for practice application.

In an attempt to correct for the effect of region variation, the multi-region prediction model_1 based on a training set containing samples data from three diverse regions was established. As illustrated in Table 3, the RMSEP and R_p values of the multi-region prediction model_1 for predicting the SSC values of samples from three geographical region were 0.705, 0.455, 0.389 °Brix and 0.915, 0.907, 0.951. Compared with the single-region model, it seemed that the results of the multi-region prediction model_1 developed on the basis of all samples data from three regions obtain were worse than the results of the single-region prediction model established by using the samples from the same region. But in fact, the multi-region prediction model_1 generated more accurate results for all the samples from the varying regions, which denoted that the variability of the sample region had a little impact on the prediction accuracy of the multi-region prediction model_1. Therefore, building a multi-region prediction model by merging all the sample data was helpful and effective to decrease the effect of region variability and enhance the stability and robustness of the model for SSC measurement.

As discussed above (Table 2), we could accurately search the corresponding single-region prediction model through the model search strategy based on the best result obtained by the region discriminant model (overall accuracy of 99.52%). In combination the results of region discriminant, model search strategy, and single-region model, the multi-region prediction model_2 was developed. The results of the multi-region prediction model_2 to predict the SSC values of apples from three varying regions were also shown in Table 3. Compared with the single-region models, the multi-region model_2 obtained better prediction results for the samples from diverse regions, with the corresponding RMSEP and R_p values were 0.753, 0.616, 0.624°Brix and 0.898, 0.796, 0.895, respectively. According to the principle of establishing multi-region model_2, we could conclude that the prediction accuracy of multi-region model_2 would be infinitely close to the single-region model which used to prediction the sample from the same region when the correct discriminant rate of region discriminant model approached 100%.

Although the targets of the multi-region prediction model_1 and multi-region prediction model_2 were both to minimize the effect of geographical region variability, those two models were suitable for different application circumstances. By incorporating a great amount of sample data per region, the multi-region prediction model_1 could reduce the impact of geographical region variability on the accuracy of SSC measurement. Whereas, when the model was utilized to determine the SSC values of apples that were not included in the training set, the performance of the multi-region prediction model_1 still got poor results, just like single-region models. In addition, the prediction accuracy of the multi-region model_1 would drop as the variety of the geographical regions increases. On the contrary, although the variety of geographical regions were in escalation, the multi-region prediction model_2 would not be affected due to the existence of the region discriminant model. However, if there was not a sufficient number of sample data per region, the prediction accuracy of the multi-region model_2 was not particularly satisfying when compared with the multi-region prediction model_1. Just like the result in this manuscript, the prediction accuracy of the multi-region prediction model_1 was superior to that of the multi-region prediction model_2, and it had higher R_p and lower RMSEP.

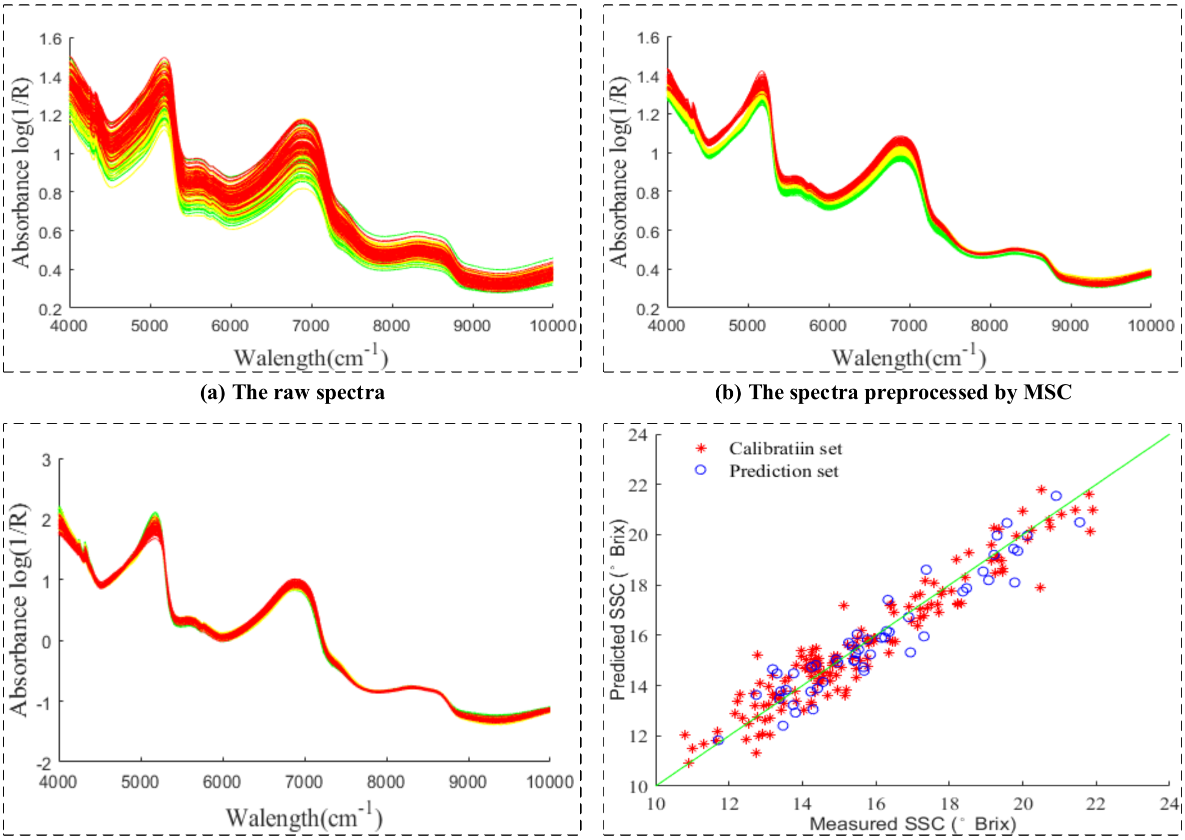
D. OPTIMIZATION OF THE MULTI-REGION MODEL

As discussed above, the multi-region prediction model_1 and the multi-region prediction model_2 had their own merits when applied to estimate the SSC values of apples from varying regions. When the amount of the sample data per region was not sufficient, the multi-region prediction model_2 could not acquire excellent analysis results because the model contained insufficient information. If the variety of the geographical regions was too much, the multi-region prediction model_1 would get poor results. According to Table 3, all the multi-region models with full spectral obtained great results in the SSC measurement of apples. Nevertheless, some spectral improvement in multivariate data analysis was still required to decrease the irrelevant spectral information and reduce the computational cost, achieving more accurate results.

In this manuscript, the raw spectral data were separately preprocessed with MSC (Fig. 4b), FD correction, and SNV (Fig. 4c) algorithms before constructing the calibration model. The effects of these preprocessing approaches were evaluated based on the PLS calibration model for SSC. In Table 4,the accuracy of the multi-region model_1 and multi-region model_2 based on different spectral preprocessing methods for determining SSC values in apples from Aksu, Qixia, and Yichuan were compared and analyzed, respectively. As shown in Table 4, the model based on the spectra selected by SNV algorithm had the best prediction accuracy, and RMSEP and R_p of multi-region model_1 (Fig. 4d) and multi-region model_2 was surrounded by a red outline. It was concluded that SNV preprocessing algorithm showed good optimization results and effectively removed the slop and baseline effects. Therefore, the subsequent computation of this manuscript was based on spectral data which preprocessed by the SNV method.

As mentioned above, the SSC prediction models based on the full spectra were time-consuming for spectral calculation and analysis, and thus variable selection algorithms area significant procedure to decrease the calculative burden of spectra data.

Variable selection algorithms including MC-UVE, CARS, and RF were applied to select the most crucial wavelengths based on the samples with full spectral data. During the process of MC-UVE variable selection, 156 samples (from Aksu, Qixia, and Yichuan) in the training set were chosen to acquire the variable selection result. Effective wavelengths can be selected by evaluating the stability values of each wavelength. Variables with stability values above the threshold (0.75) were regarded as the informative variables. Fig. 5a described the stability of each wavelength and the red circles in this figure stood for the spectral variables selected for



(a) The raw spectra

(b) The spectra preprocessed by MSC

(c) The spectra preprocessed by SNV

(d)The scatter plots of measured verse predicted SSC based on the SNV spectra

FIGURE 4. The raw spectra (a) preprocessed by MSC (b) and SNV (c) and the result of multi-region model_1 based on the SNV spectral preprocessing method (d).

TABLE 4. The results determined for SSC by two multi-region models based on different spectral preprocessing methods.

Model	Preprocessing Methods	Aksu		Qixia		Yichuan	
		RMSEP(°Brix)	R_p	RMSEP(°Brix)	R_p	RMSEP(°Brix)	R_p
Multi-region model_1	Raw spectra	0.705	0.915	0.455	0.907	0.389	0.951
	MSC	0.750	0.914	0.406	0.933	0.392	0.964
	SNV	0.703	0.917	0.407	0.934	0.337	0.965
	FD	1.104	0.812	0.595	0.834	0.651	0.863
Multi-region model_2	Raw spectra	0.753	0.898	0.612	0.799	0.495	0.950
	MSC	0.675	0.920	0.495	0.885	0.476	0.951
	SNV	0.665	0.924	0.493	0.878	0.470	0.953
	FD	0.950	0.855	0.712	0.844	1.051	0.715

building the calibration model based on MC-UVE variable selection. The lowest RMSEP values and the highest R_p values were obtained for the calibration model when the variable number is 100. For the purpose of eliminating non-informative variables, the CARS variable selection approach was also able to be used for the calibration model to select the most effective variables. After the calculation, the number of sampling runs was 30, 40 sampling variables number to be selected from the full spectra as the key wavelengths of SSC measurement were determined by 10-fold cross-validation. And when the number of sampling runs was equal to 30, the lowest RMSECV of 0.582 (°Brix) was obtained. The variable selection process of CARS for the prediction of SSC values in 156 samples from three regions and the distribution of effective variables selected from the full spectra by CARS were described in Fig. 5c. Furthermore, RF was also

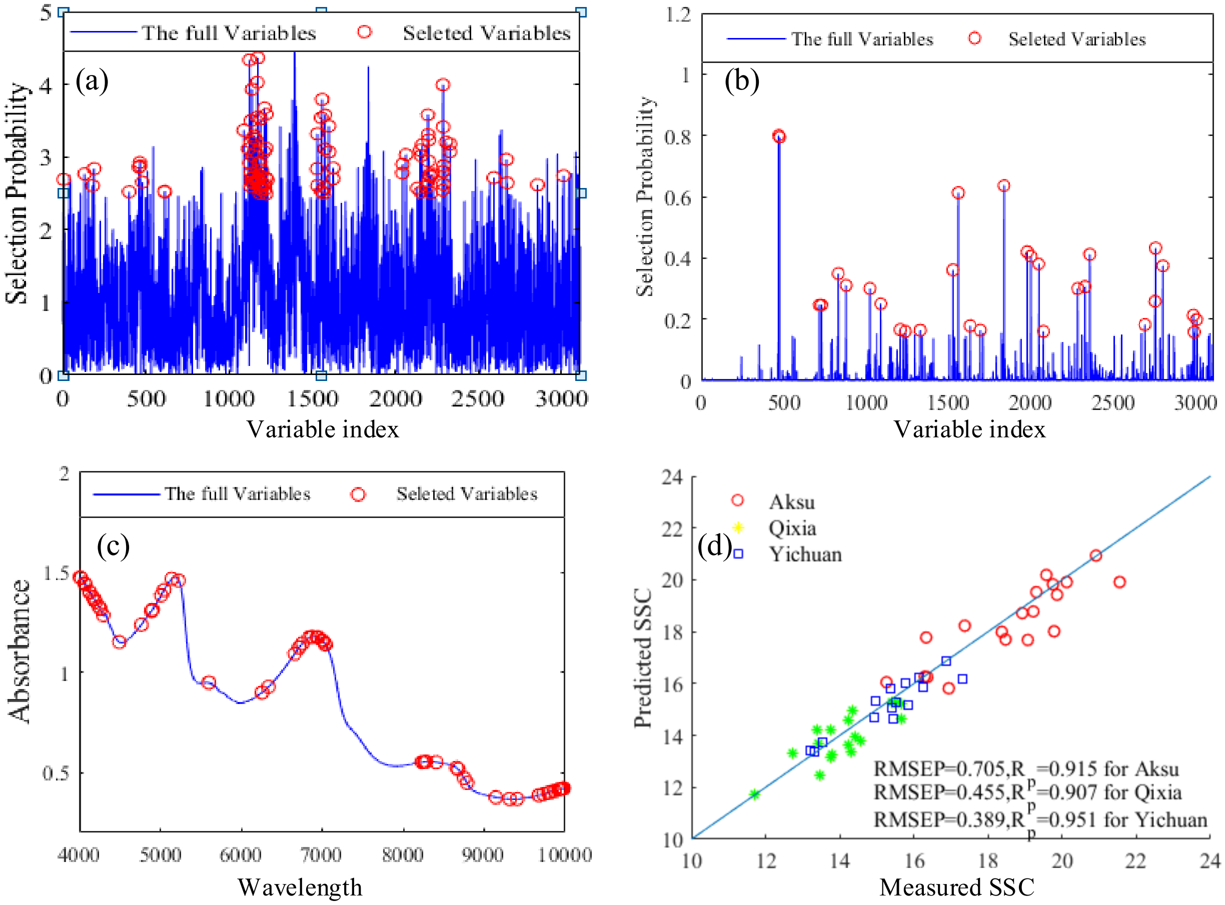


FIGURE 5. The variable selected from the full spectra by the MC-UVE (a), RF (b), and CARS algorithms and the prediction results of optimal multi-region model based on SVN and MC-UVE algorithm.

TABLE 5. The results determined for SSC by two multi-region models based on different variable selection methods.

Model	Variable selection methods	Variable number	Aksu		Qixia		Yichuan	
			RMSEP	R_p	RMSEP	R_p	RMSEP	R_p
Multi-region model_1	Full spectra	3112	0.705	0.915	0.455	0.907	0.389	0.951
	MC-UVE	100	0.527	0.952	0.490	0.934	0.481	0.934
	RF	30	0.596	0.937	0.519	0.897	0.408	0.955
	CARS	40	0.663	0.920	0.570	0.894	0.395	0.934
Multi-region model_2	Full spectra	3112	0.753	0.898	0.612	0.799	0.495	0.950
	MC-UVE	100	0.521	0.920	0.412	0.885	0.476	0.951
	RF	30	0.620	0.936	0.441	0.900	0.570	0.942
	CARS	40	0.555	0.945	0.438	0.913	0.529	0.934

carried out to select key spectral variables. Fig. 5b displayed the selection probability of wavelength determined by RF algorithms. The large the selection probability was, the more crucial the corresponding wavelength was. The cutoff threshold of the selection probability of each variability was set to be 0.15. Thus, 30 crucial variables were selected as the inputs for the development of the calibration model. As shown in Table 5 and Fig. 5d, the multi-region model_1 and multi-region model_2 based on the variables selected by MC-UVE, CARS, and RF performed better than those models on the basis of the full spectra. The results showed that MC-UVE, CARS, and RF were both effective approaches to remove non-informative wavelengths and enhance the accuracy of the calibration model. In conclusion, the multi-region model_1 and multi-region model_2 could achieve the best results when two models were carried out to prediction SSC values of apples from Aksu, Qixia, and Yichuan by applying the SNV preprocessing algorithm and MC-UVE variable selection algorithm, with RMSEP values of 0.521, 0.412, 0.476 (°Brix) and R_p values of 0.920, 0.885, 0.951.

IV. CONCLUSION

In this study, the impact of geographical region variability on the FT-NIR spectral analysis of SSC values in apples was investigated. As shown in Table 3, the prediction performance of the single-region models would be excellent if the test samples and the training samples came from the same geographical region. However, the single-region model achieved a poor result when it was used to estimate the SSC values of apples from other regions. These results denoted that the influence of region variability on the performance of SSC in apples existed. In an attempt to decrease the effect of region variability, two different strategies were proposed to quantitatively determinate the SSC in 'Fuji' apples from multiple geographical regions. For the first strategies, the multi-region prediction model_1 was developed by merging SSC values and spectral data of all samples from multiple regions. For the second strategy, the multi-region prediction model_2 was built for the determination of SSC by combining with region discriminant, model search strategy, and single-region models. SVM was applied to establish the model for discriminating the apple samples from diverse geographical regions. It was found that the region discriminant model achieved great results, with the classification accuracy of 99.52%. According to the result of the region discriminant model, we could accurately search the corresponding single-region prediction model through the model search strategy.

Then, two multi-region prediction models were compared and analyzed to determine SSC values in apples from diverse regions. It was found that both of the two multi-region prediction models demonstrated region performance. By incorporating a great amount of sample data per region, the impact of geographical region variability on the spectroscopy accuracy for SSC measurement could be decreased by applying the multi-region prediction model_1. Whereas, when the model was utilized to predict the SSC values in apples that were not included in the training set, the performance of the multi-region prediction model_1 still got poor results, just like single-region models. On the contrary, owing to the existence of the region discriminant model, the multi-region prediction model_2 was not sensitive to the variability of the region. Nevertheless, the multi-region prediction model_2 cannot acquire excellent analysis results when the amount of the sample data per region is not sufficient. Finally, the multi-region SSC prediction model was optimized in combination with the spectral preprocessing methods and variable selection methods. The overall results (shown in Table 5) denoted that it was feasible to accurately determine SSC in apples from the different geographical regions using the multi-region model_2 in combination with SNV preprocessing method and MC-UVE variable selection method, and the prediction accuracy was superior to the single-region models.

REFERENCES

[1] L. Zhang, B. Zhang, J. Zhou, B. Gu, and G. Tian, "Uninformative biological variability elimination in apple soluble solids content inspection by using Fourier transform near-infrared spectroscopy combined with multivariate analysis and wavelength selection algorithm," *J. Anal. Methods Chem.*, vol. 2017, pp. 1–9, Oct. 2017.

[2] S. Fan, C. Li, W. Huang, and L. Chen, "Detection of blueberry internal bruising over time using NIR hyperspectral reflectance imaging with optimum wavelengths," *Postharvest Biol. Technol.*, vol. 134, pp. 55–66, Dec. 2017.

[3] P. D. Drogoudi, Z. Michailidis, and G. Pantelidis, "Peel and flesh antioxidant content and harvest quality characteristics of seven apple cultivars," *Scientia Horticulturae*, vol. 115, no. 2, pp. 149–153, 2008.

[4] T. Lavilla, J. Puy, M. L. López, I. Recasens, and M. Vendrell, "Relationships between volatile production, fruit quality, and sensory evaluation in Granny Smith apples stored in different controlled-atmosphere treatments by means of multivariate analysis," *J. Agricult. Food Chem.*, vol. 47, no. 9, pp. 3791–3803, 1999.

[5] H. Yamada, H. Ohmura, C. Arai, and M. Terui, "Effect of preharvest fruit temperature on ripening, sugars, and watercore occurrence in apples," *J. Amer. Soc. Horticultural Sci. Amer. Soc. Horticultural Sci.*, vol. 119, no. 6, pp. 1208–1214, 1994.

[6] Y. Xia, Y. Xu, J. Li, C. Zhang, and S. Fan, "Recent advances in emerging techniques for non-destructive detection of seed viability: A review," *Artif. Intell. Agricult.*, vol. 1, pp. 35–47, Mar. 2019.

[7] X. Li, J. Huang, Y. Xiong, J. Zhou, X. Tan, and B. Zhang, "Determination of soluble solid content in multi-origin 'Fuji' apples by using FT-NIR spectroscopy and an origin discriminant strategy," *Comput. Electron. Agricult.*, vol. 155, pp. 23–31, Dec. 2018.

[8] J. Guthrie, C. Liebenberg, and K. B. Walsh, "NIR model development and robustness in prediction of melon fruit total soluble solids," *Austral. J. Agricult. Res.*, vol. 57, no. 4, pp. 411–418, 2006.

[9] X. Zou, J. Zhao, X. Huang, and L. I. Yanxiao, "Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of 'Fuji' apple based on different PLS models," *Chemometrics Intell. Lab. Syst.*, vol. 87, no. 1, pp. 43–51, 2007.

[10] A. Peirs, N. Scheerlinck, K. Touchant, and B. M. Nicolaï, "PH-postharvest technology: Comparison of Fourier transform and dispersive near-infrared reflectance spectroscopy for apple quality measurements," *Biosyst. Eng.*, vol. 81, no. 3, pp. 305–311, 2002.

[11] Y. Ying, Y. Liu, J. Wang, X. Fu, and Y. Li, "Fourier transform near-infrared determination of total soluble solids and available acid in intact peaches," *Trans. ASAE*, vol. 48, no. 1, pp. 229–234, 2005.

[12] G. Giovanelli, N. Sinelli, R. Beghi, R. Guidetti, and E. Casiraghi, "NIR spectroscopy for the optimization of postharvest apple management," *Postharvest Biol. Technol.*, vol. 87, pp. 13–20, Jan. 2014.

[13] M. C. Alamar, E. Bobelyn, J. Lammertyn, B. M. Nicolaï, and E. Moltó, "Calibration transfer between NIR diode array and FT-NIR spectrophotometers for measuring the soluble solids contents of apple," *Postharvest Biol. Technol.*, vol. 45, no. 1, pp. 38–45, 2007.

[14] J.-H. Wang and D.-H. Han, "Analysis of near infrared spectra of apple SSC by genetic algorithm optimization," *GuangpuXueyuGuangpu Fen Xi*, vol. 28, no. 10, pp. 2308–2311, 2008.

[15] L. Qiang, Q. Liao, Y. Liu, and Y. Lan, "Feasibility of SSC prediction for navel orange based on origin recognition using NIR spectroscopy," *Intell. Automat. Soft Comput.*, vol. 21, no. 3, pp. 305–317, 2015.

[16] S. Fan, W. Huang, Z. Guo, B. Zhang, and C. Zhao, "Prediction of soluble solids content and firmness of pears using hyperspectral reflectance imaging," *Food Anal. Methods*, vol. 8, no. 8, pp. 1936–1946, 2015.

[17] Y. Bai, Y. Xiong, J. Huang, J. Zhou, and B. Zhang, "Accurate prediction of soluble solid content of apples from multiple geographical regions by combining deep learning with spectral fingerprint features," *Postharvest Biol. Technol.*, vol. 156, Oct. 2019, Art. no. 110943.

[18] D. Ireri, E. Belal, C. Okinda, N. Makange, and C. Ji, "A computer vision system for defect discrimination and grading in tomatoes using machine learning and image processing," *Artif. Intell. Agricult.*, vol. 2, pp. 28–37, Jun. 2019.

[19] Y. Liu, W. Liu, X. Sun, R. Gao, Y. Pan, and A. Ouyang, "Potable NIR spectroscopy predicting soluble solids content of pears based on LEDs," *J. Phys., Conf. Ser.*, vol. 277, no. 1, p. 12026, 2011.

[20] D. Liu, D.-W. Sun, and X.-A. Zeng, ''Recent advances in wavelength selection techniques for hyperspectral image processing in the food industry,'' *Food Bioprocess Technol.*, vol. 7, no. 2, pp. 307–323, 2014.

[21] D. Wu and D.-W. Sun, ''Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review-Part I: Fundamentals,'' *Innov. Food Sci. Emerg. Technol.*, vol. 19, pp. 1–14, Jul. 2013.

[22] H. Wang, J. Peng, C. Xie, Y. Bao, and Y. He, ''Fruit quality evaluation using spectroscopy technology: A review,'' *Sensors*, vol. 15, no. 5, pp. 11889–11927, 2015.

[23] A. Peirs, N. Scheerlinck, and B. M. Nicolaï, ''Temperature compensation for near infrared reflectance measurement of apple fruit soluble solids contents,'' *Postharvest Biol. Technol.*, vol. 30, no. 3, pp. 233–248, 2003.

[24] F. Mendoza, R. Lu, D. Ariana, H. Cen, and B. Bailey, ''Integrated spectral and image analysis of hyperspectral scattering data for prediction of apple fruit firmness and soluble solids content,'' *Postharvest Biol. Technol.*, vol. 62, no. 2, pp. 149–160, 2011.

[25] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O. L. García-Navarrete, and J. Blasco, ''Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment,'' *Food Bioprocess Technol.*, vol. 5, no. 4, pp. 1121–1142, 2012.

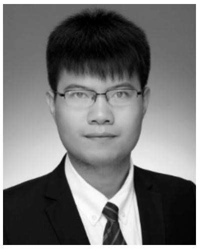
[26] X. Feng, Y. Zhao, C. Zhang, P. Cheng, and Y. He, ''Discrimination of transgenic maize kernel using NIR hyperspectral imaging and multivariate data analysis,'' *Sensors*, vol. 17, no. 8, p. 1894, 2017.

[27] H. Nouri-Ahmadabadi, M. Omid, S. S. Mohtasebi, and M. S. Firouz, ''Design, development and evaluation of an online grading system for peeled pistachios equipped with machine vision technology and support vector machine,'' *Inf. Process. Agricult.*, vol. 4, no. 4, pp. 333-341, 2017.

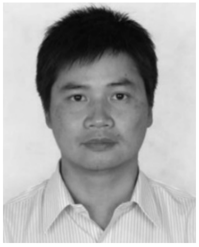
[28] Q. Chen, J. Zhao, and H. Lin, ''Study on discrimination of roast green tea (Camellia sinensis L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition,'' *SpectrochimicaActa A Mol. Biomol. Spectrosc.*, vol. 72, no. 4, pp. 845–850, 2009.

GUANGZHAO TIAN received the B.S. degree in agricultural electrification and automation and the master's and Ph.D. degrees in agricultural engineering from Nanjing Agricultural University. He is currently a Lecturer with the College of Engineering, Nanjing Agricultural University (NJAU). He has published six SCI and EI cited articles in international journals. His research interests include agriculture robots, artificial intelligence, computer vision, and pattern recognition.

XIAONA LI received the B.S. degree in engineering from Nanjing Agricultural University, in 2019. Her research interests are machine learning, pattern recognition, and spectral analysis and modeling.



BAOHUA ZHANG received the B.S. degree in mechanical and electronic engineering from Northwest A&F University and the Ph.D. degree in mechanical engineering from Shanghai Jiao Tong University. He is currently an Associate Professor with the College of Engineering, Nanjing Agricultural University (NJAU). He is also the Founder and the Managing Editor of the international journal, *Artificial Intelligence in Agriculture*. He has published more than 40 SCI cited articles, and ten EI cited articles in national and international journals. His research interests include harvesting robots, robot vision, robotic grasping, spectral analysis and modeling, robotic systems and their applications in agriculture, food, and bio-system engineering.



JUN ZHOU received the Ph.D. degree in mechanical engineering from Nanjing Agricultural University, in 2005. He is one of the experts in agricultural robots in China and a Full Professor in Nanjing Agricultural University (NJAU). He is also the Director of the Department of Science and Technology, NJAU. He has published more than 50 SCI/EI cited articles in national and international journals about agricultural robots and vision based control problems in agriculture in the past 10 years. His research interests include agricultural robot control, robotic grasping, object recognition and location, mapping, and guidance.

BAOXING GU received the B.S. degree in mechanical design and manufacturing and automation and the master's and Ph.D. degrees in agricultural engineering from Nanjing Agricultural University. He is currently a Lecturer with the College of Engineering, Nanjing Agricultural University (NJAU). He has published five SCI and EI cited articles in international journals. His research interests include agriculture robots, artificial intelligence, and intelligent agricultural equipment.