PLS-DA Infrared Spectra Model of Citrus Leaves For The Characterization of Citrus Huanglongbing

Dong-mei CHEN, Xue-juan LV, Hua-tang WANG, Jia-li LIU and Xin-nian ZENG*

Guangdong Engineering Research Center for Insect Behavior Regulation, South China Agricultural University, Guangzhou 510642, China

zengxn@scau.edu.cn

*Corresponding author

Keywords: Citrus Huanglongbing, Near-Infrared Spectrum, Principal Component Analysis, Pls-Da.

Abstract. Citrus huanglongbing disease is an important disease of citrus species, which leads to the change of chemical composition in leaves. The infrared technique may be advantageous to characterize the chemical differences between Huanglongbing infected and healthy leaves rapidly and nondestructively. In this study, the near-infrared spectra of citrus leaves were obtained in the field by using MicroNIR 1700 Spectrometer, and then precision detection of citrus HLB-associated bacteria (Candidatus Liberibacter asiaticus, CLas) by RT-qPCR for verification. The method of Soft Independent Modeling of Class Analogy(SIMCA) and Partial Least Squares Discriminant Analysis (PLS-DA) was used to establish the qualitative discriminant models of HLB diagnosis, and using first derivative and Savitzky-Golay for further data processing. The results showed that the correct rejection rate of Principal Component Analyses (PCA) model of CLas-negative leaves was more than 80%, and correct recognition rate of CLas-positive leaves was more than 95% in different citrus orchards from the field. The correct rejection rate of PLS-DA model of CLas-negative leaves was more than 83%, and correct recognition rate of CLas-positive leaves was more than 99% in different citrus orchard from the field. Both models were well distinguished whether the citrus leaves were CLas-negative or CLas-positive(even if the citrus leaves was asymptomatic), which provided a new method for the rapid diagnosis and early warning of HLB in the field.

Introduction

Citrus huanglongbing (HLB) is caused by the phloem-limiting bacteria, *Candidatus* Liberibacter asiaticus (CLas). It is a devastating disease of citrus, also known as a serious threat to the citrus industry in worldwide, that has greatly affected citrus production and resulted in great economic losses [1]. Typical symptoms of CLas-positive citrus trees are yellowing of leaves and shoots, with mottled or blotchy leaves. Fruit from CLas-positive trees are small and malformed, or asymmetric [2]. CLas-positive citrus trees often act as a source of inoculum and cause further spread of the disease in field, for which no cure has been found. Therefore, rapidly identification and removal of CLas-positive trees are useful for reduce the spread of HLB in citrus orchards and economic losses. However, traditional field assessment based on characteristics symptom is difficult since they resemble other diseases (such as stubborn disease and tristeza) and nutritional deficiencies [3-5]. At present, laboratory techniques, such as polymerase chain reaction (PCR) provide accurate detection of HLB [6], but PCR is an expensive and time-consuming process. Therefore, an easy-

to-use, fast, accurate, and inexpensive HLB diagnostic approach is greatly needed, especially for small growers to monitor their orchard and control the spread of the disease.

Near-infrared spectroscopy (NIRS) detects the stretching and bending of CH, NH, and OH functional groups caused by the light absorption of organic molecules from 350nm to 2500nm, and it has the potential to detect the chemical "fingerprint" of a specific citrus leaves [7, 8]. Changes in spectral reflectance can indicate physiological stress in trees that result from the changes in photosynthetic pigments such as chlorophyll, carotenoids and other factors [9, 10]. The spectral reflectance from the tree canopy in the visible and infrared regions of the electromagnetic spectra can be used as an indication of plant stress. Spectroscopy in the range of visible and near infrared has been investigated for disease detection in a great variety of crops since it is a rapid and non-destructive tool that can be used in real-time crop assessment under field conditions [11].

With improvements in spatial, spectral and temporal resolution of remote sensing, multispectral imagery remains advantageous due to its real-time or near real-time imagery for visual assessment [12]. In another spectroscopy study, it was shown that the reflection of dried ground leaves in the mid-infrared band can be used to determine the HLB status of a sample with >95% accuracy [5]. Perez has demonstrated that Raman spectroscopy can be discriminated between orange plants HLB-positive and healthy plants with PCA–LDA analysis. The results of PCA–LDA analysis showed a sensitivity of 86.9%, a specificity of 91.4%, and a precision of 89.2%. They proposed that Raman spectroscopy combined with PCA–LDA could be applied as a rapid pre-diagnostic methodology, which has the advantage of being a non-invasive optical technique. And it is easy to conduct and only requires a very compact set-up, meaning that it can be portable and produces immediate results [13].

Based on previous studies, the present research aims to use the MicroNIR 1700 Spectrometer with chemometrics analysis to establish qualitative discrimination models and find the best way to scan the spectrum, as a rapid and cost-effective way to determine whether plants (both symptomatic and asymptomatic) were CLas-positive or not, as an alternative for rapid detection during the phytosanitary epidemiological surveillance activities, prior to undergoing molecular confirmatory DNA and RT-qPCR analysis.

Materials and Methods

Samples

121 CLas-negative citrus leaves (including symptomatic or asymptomatic) and 176 CLas-positive samples are used to establish models. All citrus leaves were picked from Citrus reticulate Blanco orchards of Guangzhou Fuhe (23°44′N, 113°69′E), and each tree was 6-years-old and about 250-300 cm tall. Each sample used in the experiment must be tested by RT-qPCR to determine whether it carries the CLas or not, the RT-qPCR components and conditions were those described by Li et al, and the primers HLBasf/HLBasr and the probe HLBp, which were designed to amplify the 16S rDNA gene of CLas were used [14].The threshold cycle (C_t) from the PCR analysis indicates the presence or absence of CLas bacteria infected in leaves. C_t>32 is considered as CLas-negative, while C_t<30 is considered CLas-positive.

Data Collection

The near-infrared absorption spectra of samples were collected by MicroNIR 1700

Spectrometer (JDSU Co., Ltd, American) with a steel plate as scanning background, 0.5s scan speed of each sample in the wave number range of 910-1650nm. The steel plate was specially designed which has low spectral absorption, suitable reflectivity and refractive index. In this study, 6 different scanning methods were used to collect spectrum to find the most suitable method for leaves scanning of citrus leaves, they are front blade side, back blade side, front blade midrib, back blade midrib, front blade edge, back blade edge, respectively. Each scanning method was repeated 3 times, and the near-infrared raw spectra of different citrus leaves are shown in Figure1.

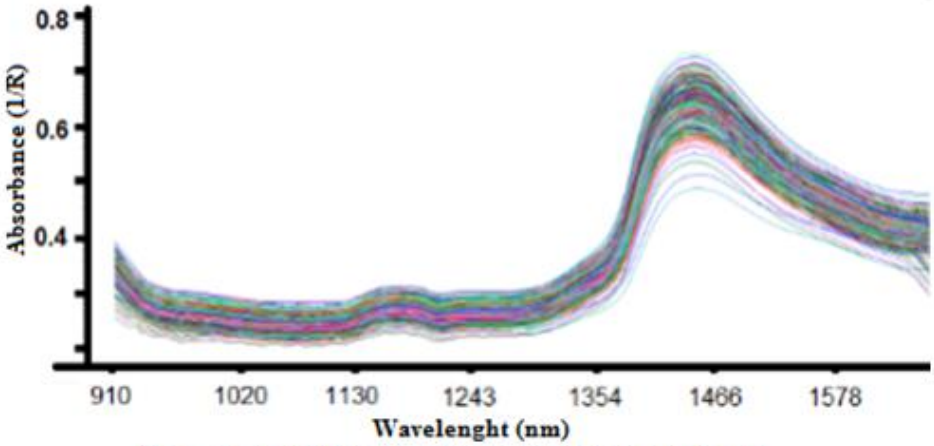

Fig. 1. The near-infrared raw spectra of different citrus leaves

Data Analysis

Preprocessing of The Spectra. The calibration and validation spectral datasets collected from samples were preprocessed before further analysis. The spectra of all samples were exported to The unscrambler software and divided per class samples into two groups, taking every alternate sample as calibration (training) set and remaining as validation (testing) set [15]. There have 101 calibration samples and 20 validation samples of CLas-negative samples, at the same time, CLas-positive samples were divided into 133 calibration samples and 43 validation samples.

The preprocessed (normalized) spectral reflectance data were used to calculate the first derivatives based on Savitzky-Golay filtering. This correction will bring the spectra to a common baseline by pulling out changes in the Y-axis and increases the signal-to-noise ratio [16, 17]. In Savitzky-Golay filtering, an unweighted linear least-square fit based on polynomial equation is used to calculate the filter coefficients. Spectral regions dominated by noise were excluded. The Savitzky-Golay filtering performs data smoothing, in addition to calculating the derivatives. In the meantime, the number of principal components (PCs) is an important parameter that greatly affects discrimination results of NIRS models. Previous practices indicated that more complicated samples were more principal components needed [18], the model have maximum accuracy when meet suitable principal component number [18, 19]. Under-fitting will occur when the number of principal components too low [20]. Usually, the symptoms of HLB were often considered as a complicated samples, which often confusion with the symptoms of nutrient deficiency. Therefore, 10, 15, 20 were used as the principal component number after repeated screening, and compared the prediction accuracy of the models in the experiment. Another factor that affects the quality of the model is smoothing point [14]. The selection of smoothing points is an essential step in Savitzky-Golay filtering. In the process of model building, the smoothing points also affected accuracy of the the models were found. At last, models

when the smoothing number was 3, 5, 7 and 9 under different PCs were established respectively.

Different datasets were generated to determine the most suitable method for detecting CLas-positive samples. The datasets used were summarized as follows: raw data, a first derivatives and 3 smoothing point in Savitzky-Golay filtering (SG-1st 3 smoothing point) dataset, a first derivatives and 5 smoothing point in Savitzky-Golay filtering (SG-1st 5 smoothing point) dataset, a first derivatives and 7 smoothing point in Savitzky-Golay filtering (SG-1st 7 smoothing point) dataset, a first derivatives and 9 smoothing point in Savitzky-Golay filtering (SG-1st 9 smoothing point) dataset, derived from preprocessed raw data. And every dataset saved with different number of PCs (10, 15, 20), respectively.

Chemometric Analysis. The initial examination of the average spectra showed considerable overlap between different species. In this study, the calibration samples was developed using principal component analyses (PCA) and partial least squares discriminant analysis (PLS-DA) including CLas-negative and CLas-positive using the leave-one-out cross validation samples.

PCA (mean centred) full cross validation option and constant weight to all variables was performed to investigate clustering of samples in different groups [21]. In this study, PCA was then performed using the complete spectra (including all wave number ranges and subsets of these ranges), before and after pre-processing of spectral data. PLS-DA is performed using an exclusive binary coding. The predicted origins seldom lead to a binary result not exactly equal 0 or 1 but to a result near 0 or 1, which is justified by the natural variability of the sample constituents. PLS-DA allows modeling several response variables (Y). In calibration datasets, the Y value of CLas-negative samples is “0”, while the Y value of CLas-positive is “1” in this study. For PLS-DA, the values of root mean error of calibration (RMSEC), coefficient of determination (R^2) and root mean square error of prediction (RMSEP) were used to identify a better calibration and validation. The RMSEC value is used as an indication of the uncertainty in the calibration model, while the R^2 value and RMSEP were determined for the validation data sets in prediction samples [22]. In order to establish a more practical model, R^2, RMESP and SEP were used in our PLS-DA model.

The formula of R^2 as in Eq. (1):

$$R^2 = 1 - \sum_{i-1}^{n} \frac{(\widehat{y_i} - y_i)^2}{\sum_{i-1}^{n}(\widehat{y_i} - \bar{y})^2} \tag{1}$$

where n is the number of spectra for the calibration, y_i is the quantity of lipid present in the mixture corresponding to the spectrum i, $\widehat{y_i}$ is the quantity of lipid estimated by the model with spectrum i and $\bar{y}$ is the average of all reference measurements values in the calibration set.

The formula of RMESP as in Eq. (2):

$$\mathrm{RMESP} = \sqrt{\frac{\sum_{i-1}^{n}\left(y_{pi} - \widehat{y_{p1}}\right)^2}{n_p}} \tag{2}$$

where n_p is the number of spectra in the prediction set, y_{pi} is the quantity of lipid measured with another technique for spectrum i, $\hat{y}_{p1}$ is the estimated amount of lipids y the model using spectrum i [22-25].

Classification of Samples

Soft Independent Modeling of Class Analogy (SIMCA), another multivariate classification model based on collection of PCA models, was used to predict class memberships using validation samples at 5% confidence interval level. SIMCA class models were interpreted based on class projections, misclassifications, discriminating power, and interclass distances. In SIMCA-based classification, as a part of model development, the PC scores were generated for each class based on the variation in each class, rather than utilizing the overall variation in the data. The models residuals were then used for classification of unknown samples [2].

Different PCA models were established by different type of samples and parameters, the calibration datasets of CLas-negative samples and CLas-positive samples were respectively established their PCA models. Validation datasets of CLas-negative samples and CLas-positive samples were put into its corresponding PCA models, counted results to determine the best model and best preprocess parameters, and the most suitable method for the spectra scanning of citrus leaves. Different from SIMCA, one PLS-DA model with two types of calibration datasets but different parameters was just need established. Then validation datasets were put into PLS-DA model, and counted results to determine the best model and the best preprocess parameters.

Field Applicability Verification

Two different citrus orchards samples were applied to verify the model applicability of the field, and each orchard by five point random sampling method for sampling. East, West, South, north and Middle, each azimuth represents a point. Collecting 33 samples in each of azimuth, each of the citrus orchards were sampled 165, and two citrus orchards were sampled 330. The results of RT-qPCR showed that 297 were CLas-positive, while 33 were CLas-negative. After preprocessing the NIRS of these samples, then input PCA model and PLS-DA model for classification, respectively.

Statistical Analyses

The data were analyzed using SPSS software version 17.0 (SPSS Inc., Nie et al, Chicago, IL, USA), and the Unscrambler 9.8 software. In this study, correct recognition rate and correct rejection rate were used to evaluate the predictive ability of the PCA model. The higher the correct recognition rate and the correct rejection rate, the better the prediction ability of the model.

Correct recognition rate= $N_r / N_1 \times 100\%$

Correct rejection rate= $N_{re} / N_2 \times 100\%$

Nr-correctly identify the sample number of the sample itself; N_1-the number of samples from the same source; N_{re}-correctly reject the number of samples from other sources; N_2-the number of samples from other sources.

For PLS-DA model, the R^2, RMSEP and SEP were used to identify its quantitative. R^2 closer to the 1, the relationship between the predictive value and the real value of the correction model is better. The lower and closer of the RMSEP and SEP value, indicating the better the performance of the model is and the higher the prediction accuracy [21, 24, 27].

Results and Discussion

Classification Results of PCA Model

As shown in Figure 2, the green area represented the CLas-negative samples, red area represented CLas-positive samples. The visual of PCA model can distinguish between CLas-negative and CLas-positive samples. Description of the above two types of calibration datasets has the basic ability of discriminant analysis, which can verify the validation samples, and then calculate the specific classification accuracy.

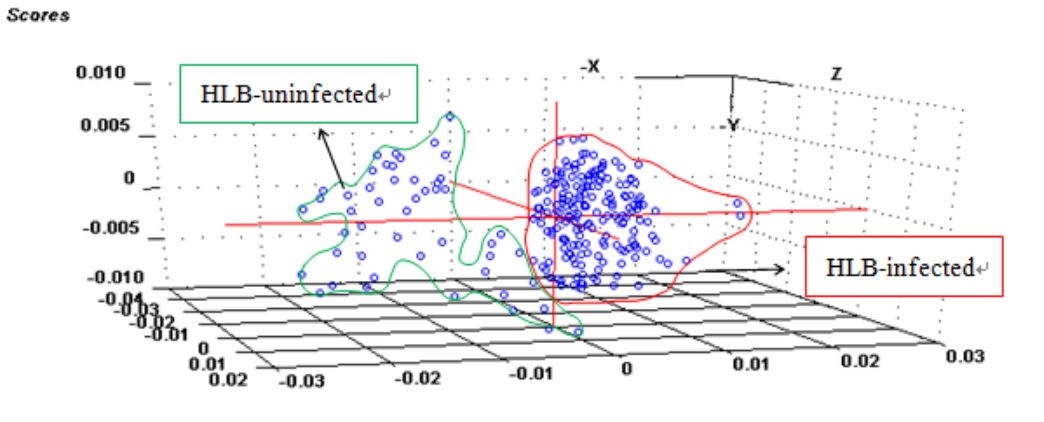

Fig. 2. The 3D models of calibration datasets

The classification results of PCA models under different scanning methods and different spectral pretreatment methods were shown in Table 1. After SG first derivative processing, in addition to the back blade edge method, the maximum correct recognition rate of the other scanning methods were all more than 79%. Among them, the back blade midrib scan method for spectra data by SG-1st 3 smoothing point, when PCs value was 10, 15 and 20, the correct recognition rate was high up to 96.83%. The model classification accuracy of all the parameters in the back blade midrib scan mode was above 87%, and the overall classification accuracy was more than 93% after SG first derivative preprocessing. At the same time, although the front blade edge method for SG-1st 5 smoothing point, when PCs value was 10, the correct recognition rate was also 96.83%. However, the classification accuracy will reduce when the PCs was changed, indicated that the change of the PCs has great influence on the classification ability of the model, and the stability of the PCA model under this parameter was not good. The classification accuracy of the back blade midrib scanning method was the highest. The results showed that the difference PCs were not affected the model's classification ability, so the stability of the PCA model was the best one with respect to the other scanning method and preprocess parameters.

Related research also showed that the CLas was first through the petiole infect midrib, then harm leaves and finally influence of diseased trees overall development, and under-fitting will occur when the number of PCs too low and the number of PCs too high will lead to over-fitting. Therefore, the back blade midrib scanning method was the optimal and the SG-1st 3 smoothing point and15 PCs was the best preprocess parameter.

Table 1. The results of PCA models with different scanning and prepossessing methods

Prepossessing parameter			Correct recognition rate (%)					
			Front blade side	Back blade side	Front blade midrib	Back blade midrib	Front blade edge	Back blade edge
SG-1st derivative	Smoothing[b]=3	PCs[a] =20	68.25	69.84	65.08	96.83	93.65	9.52
		PCs =15	69.84	68.25	65.08	96.83	93.65	9.53
		PCs =10	71.43	66.67	65.08	96.83	93.65	9.52
	Smoothing=5	PCs =20	74.60	88.89	79.37	95.24	95.24	25.40
		PCs =15	74.60	88.89	79.37	95.24	95.24	25.40
		PCs =10	74.60	88.89	79.37	95.24	96.83	25.40
	Smoothing=7	PCs =20	76.19	88.89	82.54	95.24	95.24	30.16
		PCs =15	76.19	87.30	82.54	95.24	95.24	30.16
		PCs =10	76.19	90.48	82.54	95.24	95.24	30.16
	Smoothing=9	PCs =20	79.37	92.06	77.78	93.65	95.24	34.92
		PCs =15	79.37	90.48	77.78	93.65	95.24	34.92
		PCs =10	79.37	90.48	77.78	93.65	95.24	33.33
Raw data		PCs =20	63.49	87.30	73.02	87.30	63.49	47.62
		PCs =15	63.49	87.30	73.02	87.30	63.49	47.62
		PCs =10	63.49	87.30	73.02	87.30	63.49	47.62

Note: Table notes.
PCs[a] means the number of principal components of PCA model.
Smoothing[b] means the number of smoothing points.

Classification Results of PLS-DA Model

The spectral data of the optimal scanning method (back blade midrib scan method) was obtained in 3. 1, then PLS-DA analysis was carried out, and the model was optimized to determine the qualitative discriminant model of PLS-DA. After processing the raw data by PLS-DA method, the CLas-negative samples and CLas-positive samples were correctly classified by the models (Fig. 3). Because the R^2 value closer to the 1, and the lower the RMSEP value, the closer the RMSEP and SEP value, indicating the better the performance of the model is and the higher the prediction accuracy. After the SG-1st7 smoothing point preprocess, the R^2=0.970317, RMSEP=0.080198, SEP=0.074656, the validation datasets classification accuracy up to 100%, compared with other groups, the result was more satisfied with the model evaluation criteria. So the best preprocess parameter for optimal PLS-DA model was the SG-1st7 smoothing point and 15 PCs.

Table 2. The results of back blade midrib scan with different prepossessing parameter (PLS-DA).

Prepossessing parameter		R^2	RMSEP	SEP
SG 1stderivative	Smoothing[a]=3	0.931313	0.121996	0.079807
	Smoothing=5	0.940347	0.113691	0.110176
	Smoothing=7	0.970317	0.080198	0.074656
	Smoothing=9	0.967320	0.084150	0.080602
Raw data		0.964341	0.087901	0.074843

Note: Smoothing[a] means the number of smoothing points

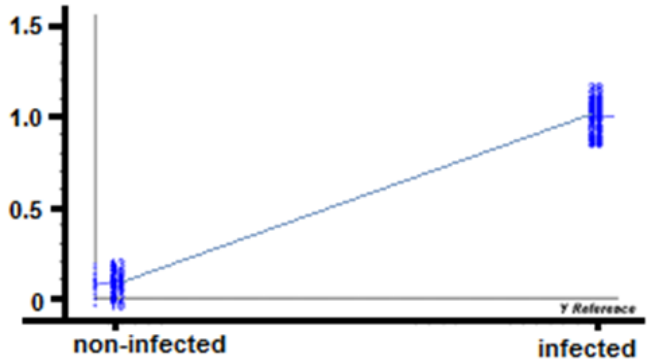

Fig. 3.Prediction map of PLS-DA model under the best parameters of back blade midrib scan

The Applicability of the Model in the Field Verification

The samples of two different citrus orchards were preprocessed according to the optimum parameters for PCA and PLS-DA model, and then carried on the classification analysis (Table 3). The classification results of PCA and PLS-DA model showed that the correct recognition rate was all above 95%, the highest reached 100%, meanwhile the correct rejection rate was all above 80%, and the maximum is 86.67%. The high classification ability of the two models was proved. Whether the correct rejection rate or correct recognition rate, the correct verification rate of the two orchards used PLS-DA model were all higher than that of the PCA model, indicating that the discriminant analysis ability of PLS-DA model was better than that of the PCA model.

At the same time, comparing the classification results of two models and the actual results of RT-qPCR detection, in order to determine the applicability of the two models, the chi-square test were carried out to classification results of the field and the actual results of RT-qPCR detection. The results displayed that there no significant difference between the results of RT-qPCR detection and both PCA and PLS-DA models classification ($p>0.05$). It suggested that the samples could be correctly classified by PCA and PLS-DA models in field.

Table 3. Verification results of different citrus orchards

Samples	Correct rejection rate (%)		Correct recognition rate (%)	
	PLS-DA	PCA	PLS-DA	PCA
No.1 orchard	86.67	80.00	99.33	95.33
No.2 orchard	83.33	83.33	100.00	97.28

Conclusions

Rapid diagnosis of HLB in field is the precondition for HLB control. The application of MicroNIR 1700 Spectrometer in conjunction with chemometric analysis (SIMCA and PLS-DA) of spectral data to predict the presence HLB of citrus leaves was investigation in this study. The results indicated that statistical classifier models such as PCA and PLS-DA could distinguish between CLas-negative and CLas-positive leaves with high classification accuracies of greater than 80%, maximum up to 100%. Comparison of multiple experimental results, we found the classification accuracy and discriminant analysis ability of PLS-DA model was better than that of the PCA model, and the most suitable method of near infrared spectrum scanning method was also found(the back blade midrib scan method). It indicated that MicroNIR 1700 Spectrometer has potential to detected HLB with optimal model and the best

preprocess parameter for data, even if the citrus leaves was asymptomatic CLas-positive.

However, the evaluate of applicability of MicroNIR 1700 Spectrometer in predicting different commercial citrus species that easily infected with HLB disease, and the quantity of CLas in each disease samples with quantitative models were also necessary. Some of these aspects would be involved in the future studies.

Acknowledgments

The present research is financially co-supported by projects of Guangdong Science and Technology Plan (2014A040401072) and Guangdong Engineering Research Center for Insect Behavior Regulation (2015B090903076).

References

[1] K. R. Chung, and R. H. Brlansky, Citrus diseases exotic to Florida: Huanglongbing (citrus greening), Plant Pathology Department Fact Sheet pp-210, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida, http://edis. ifas. ufl. edu/pp133 (2005).

[2] S. Sankaran, A. Mishra, J. M. Maja, and R. Ehsani, Visible-near infrared spectroscopy for detection of Huanglongbing in citrus orchards,*Comput Electron Agr*77, 127-134 (2011).

[3] S. E. Halbert, K. L., and Manjunath, Asian citrus psyllids (Sternorrhyncha: Psyllidae) and greening disease of citrus: Aliterature review and assessment of risk in Florida, *Florida Entomological Society* 3, 330-353 (2004).

[4] S. A. Hawkins, B. Park, G. H. Poole, T. R. Gottwald, W. R. Windham, J. Albano, and K. C. Lawrence, Comparison of FTIR Spectra between Huanglongbing (Citrus Greening) and Other Citrus Maladies, *J Agr Food Chem* 58, 6007-6010 (2010).

[5] W. Li, J. A. Abad, R. D. French-Monar, J. Rascoe, A. Wen, N. C. Gudmestad, G. A. Secor, I. Lee, Y. Duan, and L. Levy, Multiplex real-time PCR for detection, identification and quantification of 'Candidatus Liberibacter solanacearum' in potato plants with zebra chip, *J Microbiol Meth* 78, 59-65 (2009).

[6] J. S. Shenk, J. J. Workman, and M. O. Westerhaus, Application of NIR spectroscopy to agricultural products, *Practical Spectroscopy Series* 27, 419-474 (2001).

[7] C. Pasquini, Near infrared spectroscopy: fundamentals, practical aspects and analytical applications, *J Brazil Chem Soc*14, 198-219 (2003).

[8] J. H. Everitt, and D. E. Escobar, The status of video systems for remote sensing applications, in *Proc.of 12th Biennial Workshop on Color Photography and Videography in the Plant Sciences and Related Field*, pp. 6-29(1989).

[9] J. H. Everitt, D. E. Escobar, I. Cavazos, J. R. Noriega, and M. R. Davis, A three-camera multispectral digital video imaging system, *Remote Sens Environ* 54, 333-337 (1995).

[10] S. Sankaran, A. Mishra, R. Ehsani, and C. Davis, A review of advanced techniques for detecting plant diseases, *Comput Electron Agr* 72, 1-13 (2010).

[11] S. Sankaran, J. Maja, S. Buchanon, and R. Ehsani, Huanglongbing (Citrus Greening) Detection Using Visible, Near Infrared and Thermal Imaging Techniques, *Sensors* 13, 2117-2130 (2013).

[12] M. R. V. Perez, M. G. G. Mendoza, M. G. R. Elias, F. J. Gonzalez, H. R. N. Contreras, and C. C. Servin, Raman Spectroscopy an Option for the Early Detection of Citrus Huanglongbing, *Appl Spectrosc* (2016).

[13] W. Li, J. S. Hartung, and L. Levy, Quantitative real-time PCR for detection and identification of Candidatus Liberibacter species associated with citrus huanglongbing, *J Microbiol Meth* 66, 104-115 (2006).

[14] L. Zhang, X. Zhang, L. Ni, Z. Xue, X. Gu, and S. Huang, Rapid identification of adulterated cow milk by non-linear pattern recognition methods based on near infrared spectroscopy, *Food Chem* 145, 342-348 (2014).

[15] B. Schrader, Infrared and Raman spectroscopy: methods and applications (John Wiley & Sons, 2008).

[16] A. Savitzky, and M. J. Golay, Smoothing and differentiation of data by simplified least squares procedures, *Anal Chem 36*, 1627-1639 (1964).

[17] P. Ciosek, Z. Brzozka, W. Wroblewski, E. Martinelli, C. Dinatale, and A. Damico, Direct and two-stage data analysis procedures based on PCA, PLS-DA and ANN for ISE-based electronic tongue—Effect of supervised feature extraction, *Talanta* 67, 590-596 (2005).

[18] B. Worley, S. Halouska, and R. Powers, Utilities for quantifying separation in PCA/PLS-DA scores plots, *Anal Biochem* 433, 102-104 (2013).

[19] P. Jaiswal, S. N. Jha, A. Borah, A. Gautam, M. K. Grewal, and G. Jindal, Detection and quantification of soymilk in cow–buffalo milk using Attenuated Total Reflectance Fourier Transform Infrared spectroscopy (ATR–FTIR), *Food Chem* 168, 41-47 (2015).

[20] Q. Chen, P. Jiang, and J. Zhao, Measurement of total flavone content in snow lotus (Saussurea involucrate) using near infrared spectroscopy combined with interval PLS and genetic algorithm, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 76, 50-55 (2010).

[21] G. Erik, and R. Jean-Marie, Subtraction of atmospheric water contribution in Fourier transform infrared spectroscopy of biological membranes and proteins, *Spectrochimica Acta Part A: Molecular Spectroscopy* 50, 2137-2144 (1994).

[22] Y. Tominaga, Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN, *Chemometr Intell Lab* 49, 105-115 (1999).

[23] Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, and M. Hanpin, Variables selection methods in near-infrared spectroscopy, *Anal Chimacta* 667, 14-32 (2010).

[24] O. Anjos, M. G. Campos, P. C. Ruiz, and P. Antunes, Application of FTIR-ATR spectroscopy to the quantification of sugar in honey, *Food Chem* 169, 218-223 (2015).

[25] A. Derenne, O. Vandersleyen, and E. Goormaghtigh, Lipid quantification method using FTIR spectroscopy applied on cancer cell extracts, *Biochimica et*

Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1841, 1200-1209 (2014).