Journal of Innovative Optical Health Sciences
Vol. 7, No. 6 (2014) 1350065 (8 pages)
© The Authors
DOI: 10.1142/S179354581350065X

Online quantitative analysis of soluble solids content in navel oranges using visible-near infrared spectroscopy and variable selection methods

Yande Liu*, Yanrui Zhou and Yuanyuan Pan
Institute of Optics-Mechanics-Electronics Technology
and Application (OMETA), School of Mechanical
and Electronical Engineering
East China Jiaotong University
Nanchang 330013, P. R. China
**jxliuyd@163.com*

Received 6 May 2013
Accepted 8 October 2013
Published 29 January 2014

Variable selection is applied widely for visible-near infrared (Vis-NIR) spectroscopy analysis of internal quality in fruits. Different spectral variable selection methods were compared for online quantitative analysis of soluble solids content (SSC) in navel oranges. Moving window partial least squares (MW-PLS), Monte Carlo uninformative variables elimination (MC-UVE) and wavelet transform (WT) combined with the MC-UVE method were used to select the spectral variables and develop the calibration models of online analysis of SSC in navel oranges. The performances of these methods were compared for modeling the Vis-NIR data sets of navel orange samples. Results show that the WT-MC-UVE methods gave better calibration models with the higher correlation coefficient (r) of 0.89 and lower root mean square error of prediction (RMSEP) of 0.54 at 5 fruits per second. It concluded that Vis-NIR spectroscopy coupled with WT-MC-UVE may be a fast and effective tool for online quantitative analysis of SSC in navel oranges.

Keywords: Vis-NIR spectroscopy; variables selection; soluble solids content; wavelet transform; moving window partial least squares; Monte Carlo uninformative variables elimination.

1. Introduction

Visible-near infrared (Vis-NIR) spectroscopy is a rapid, accurate and nondestructive technique, which is widely used in the analysis of internal quality of fruits. Faster and more accurate analysis of the soluble solids content (SSC) in fruits is very important to meet consumer demand. The very simple technology, which is based on reflectance model measurement, has been available for online fruit grading since 1989. It has been well established for internal quality evaluation of fruits.[1,2]

Different chemometrics methods, which are used to analyze chemistry such as multivariate calibration methods, are applied to develop a quantitative relation between the NIR spectra and the concentrations.[3] Objects and variables quality often affects the quality of a multivariate calibration model. Some broad, weak, nonspecific and overlapping bands are included in NIR spectra[4] and the data matrix of the spectra is so large with thousands of wavelengths and hundreds of samples. Some irrelevant variables for multivariate calibration exist, and will worsen the precision, prediction and efficiency of the calibration models.[5]

Several methods to eliminate the uninformative variables or selection of informative variables for improving the modeling precision, have been reported, such as regression coefficients of partial least squares regression,[6,7] successive projections algorithm (SPA),[8,9] competitive adaptive reweighted sampling (CARS)[10,11] interval partial least squares (iPLS),[12,13] backward interval PLS (Bi-PLS),[14] moving window partial least squares (MW-PLS),[15,16] stepwise regression analysis (SRA),[17] Monte Carlo combined with uninformative variables elimination (MC-UVE)[18,19] and genetic algorithms (GA).[20] MW-PLS and MC-UVC are the commonly used methods for variable selection of NIR spectra. The improved algorithm is used in MW-PLS[16] and a series of PLS models with varying principal components in a window that moves over the whole spectral, and then calculates the sums of squared residuals (SSR) for each subset. It will locate informative spectral intervals that have the least model complexity and the lowest sum of residuals. MC-UVE is a faster computation and modified method of uninformative variables elimination (UVE) for variable selection, which combines with the Monte Carlo (MC) technique and decreases the risk of over-fitting.[21]

The objective of this study was to simplify calibration models with large wavelength number of Vis-NIR data sets for online analysis of SSC in navel oranges. Different variable selection methods were used to improve the accuracy and efficiency of multivariate calibration models. MW-PLS, MC-UVE and wavelet transform (WT) combined with the MC-UVE methods were used to select the spectral variables and develop the calibration models for online analysis of SCC in navel oranges. The performances of these methods were compared for modeling the NIR data sets of navel orange samples. Specific objectives were (1) to establish relationships between online Vis-NIR spectroscopy measurements and the SCC of intact navel oranges; (2) to compare the different select variables methods of MW-PLS, MC-UVE and WT-MC-UVE; (3) to propose practicable and high-efficient modeling methods for online quantitative analysis of SSC in navel oranges.

2. Materials and Methods

2.1. *Navel Orange Samples*

A total of 123 navel orange samples were purchased at a local market in Jiangxi province, and stored in standard refrigeration (2°C). All samples were equilibrated in an experimental environment at about 25°C and 60% relative humidity (RH) for 24 h before Vis-NIR diffuse reflectance spectral measurements were performed. All measurements including spectral collection and SSC measurement were carried out on the same day. In order to compare the performance of different calibration models, samples in the calibration and prediction sets were kept unchanged for all models. By using Kennard–Stone (KS) method, 31 navel oranges were used for prediction set, and the remaining 92 samples were used for calibration set. In order to ensure the adaptability of the calibration model, the samples with the highest and lowest SSC values were put in the calibration set.

The juice of a whole navel orange flesh was extracted by using a manual fruit squeezer (Type reference: MSL-218, Chengdu, China), and 1.0 mL of filtered juice was taken for SSC measurement. The SSC of navel orange samples was determined by digital refractometer (PR-101αCat. No3442, ATGO, Japan). The measurement accuracy was ±0.1 °Brix, and the measurement range was 0 to 45.0 °Brix with automatic temperature compensation. Statistics of SSC in navel oranges of the calibration and prediction sets are summarized in Table 1. The SSC measurements were fairly normally distributed around the mean value (12.34 °Brix), with the standard deviation (SD) of 1.14. In this study, 123 samples were divided into calibration and prediction sets (92:31). The SSC range of the calibration set was from 9.50 to 15.00 °Brix, and from 9.60 to 14.00 °Brix for prediction set.

Table 1. Statistics of SSC in navel oranges used for the calibration and prediction sets.

Dataset	Number	Min (°Birx)	Max (°Birx)	Mean (°Birx)	SD (°Birx)	CV[a] (%)
Total	123	9.50	15.00	12.34	1.14	9.24
Calibration set	92	9.50	15.00	12.33	1.14	9.24
Validation set	31	9.60	14.00	12.35	1.14	9.23

[a]CV (%): coefficient of variation

2.2. *Vis-NIR spectral measurement*

An online Vis-NIRS fruit sorting system (see Fig. 1) was set up based on conventional grading system, and equipped with a fiber spectrometer and a diffuse reflectance configuration. This system consisted of three parts: optical system, control system and fruit transportation and sorting mechanism. The online optical system was constructed as follows: A fiber spectrometer (USB2000+, Ocean optics Inc., USA), a halogen tungsten lamp light source (12 v/100 W) were supported by a constant current power (CCP) for acquiring more stable spectra of navel oranges, an optic fiber, a condenser and a diffuse reflectance configuration (the angle between incident light source and collector was 30°). Diffuse reflection spectra of navel oranges were collected at 5 fruits per second controlled by motor voltage. Wavelength range was selected to be 600–950 nm with an interval of 0.3 nm. A Teflon white panel (6.5-mm thick) was used as the reference standard. Spectrometer parameters settings, spectra data collection and storage were carried out via a self-developed software package based on JAVA language (SUN MicroSystem, USA). The integration time for navel oranges and reference spectra was 25 ms. Navel oranges were randomly placed upon the fruit holder one by one. Spectra of each sample were collected when they got through the optical system. The experiments were repeated five times, and the spectra collected each time were saved in an independent document. The average spectra of each sample were used for modeling. Software of Unscramble V10.0 (CAMO PROCESS AS, OSLO, Norway) and Matlab 7.0 were used for spectral preprocessing and modeling.

2.3. *Variable selection methods and model evaluation*

In our research, the different variable selection tools including MW-PLS, MC-UVE and WT-MC-UVE methods were adopted for elimination of the uninformative variables of fruit spectral data. For MW-PLS method, variables with relative low SSR or high absolute stability were retained. PLS regression models were developed with different number of retained variables. For MC-UVE

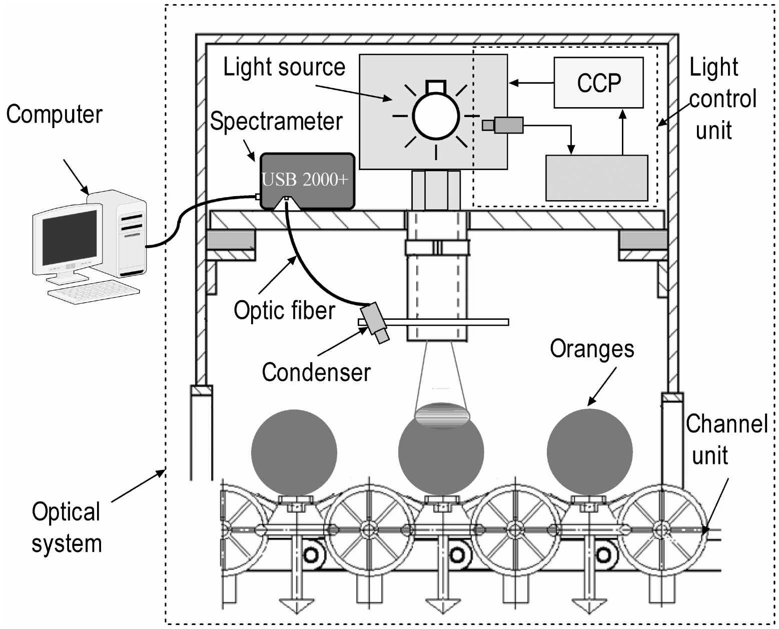


Fig. 1. Diagram of optical system of the online NIRS fruit sorting system.

method, the stability is used to assess the reliability of each variable in the models, and those variables with larger absolute value of the stability are known as informative and used in the modeling. The cutoff value of stability is determined by the root mean square error of prediction (RMSEP). Variables whose stabilities are less than the cutoff value are eliminated by MC-UVE.

The statistic correlation coefficient (r) and RMSEP were used to evaluate the precision of the PLS models with different variable selection methods. The ideal model should have higher value of r and lower value of RMSEP. The numbers of variables were determined when the RMSEP reached the minimum.

3. Results and Discussions

3.1. *Determination of the latent variables (LVs) number and wavelet parameters*

Before variable selection is employed, the number of LVs should be determined which affects both the prediction precision and the calculation efficiency. The optimal number of LVs for PLS regression models were obtained by using leave-one-out cross validation technique and they presented both the relative low value of RMSECV and high value of r. Figure 2 shows the variation of RMSECV and r with different numbers of LVs. In our research, the optimal number of LVs was determined to be 15 in the following calculations.

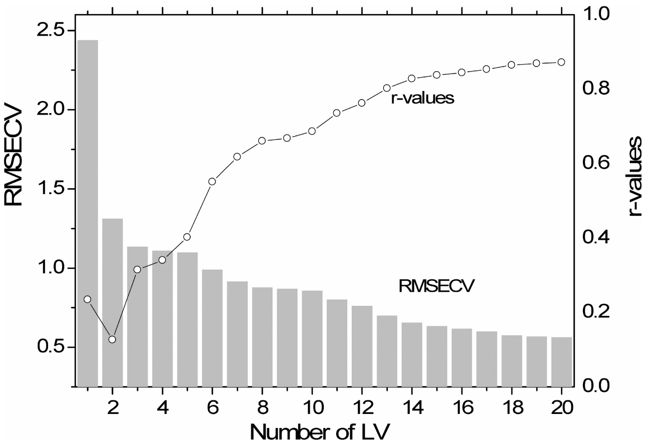


Fig. 2. Variation of RMSECV with the number of factors by PLS method for the raw spectra.

WT has been proved to be a highly efficient tool for analytical data compression and de-noise. To obtain optimum wavelet coefficients construction, two parameters should be optimized: the wavelet filter and decomposition scale based on the optimization results. The RMSEP and r were adopted as criteria to investigate the influence of Daubechies, Symmlet and Coiflet filters for SSC models. By calculation and comparison of calibration models and the results in Shao and Zhuang,[22] the simplest db1 wavelet filter and 5 scale decomposition was adopted in the study.

3.2. *Variables selection based on MW-PLS and MC-UVE methods*

To determine the number of retained variables is the main challenge of variable selection, which decides the accuracy of the calibration model. When the number is too small or too large, the performance of the model may be affected due to the loss of informative variables or the existence of uninformative variables.

For MWPLS method, SSR should be calculated before variable selection in different LVs, information variables usually locate more than one spectral region due to the existence of many spectral absorption bands. It is a useful method in searching information region, and each region obtained by MW-PLS may not always yield the optimum result. The addictive processes were conducted by sorting the SSR of the chosen number of LVs, and the PLS regression models were developed using the retained variables with relatively low SSR. The variation of the RMSEP of the prediction set with different number of retained variables methods of MW-PLS and WT-MWPLS was investigated. Figure 3 shows the variation of RMSEP with the increase of variable numbers for the raw spectra and the wavelet coefficients. Dotted and solid lines in Fig. 3 represent the RMSEP values obtained by PLS regression with the retained variables of raw spectra and WT spectra, respectively. As shown at the beginning, the RMSEP are large, and then with the increase of the number of retained variables, RMSEP decreases sharply. For raw spectra, when the number was 520, the lowest RMSEP was obtained. When the number was bigger than 520, with the increase of the number, RMSEP increased

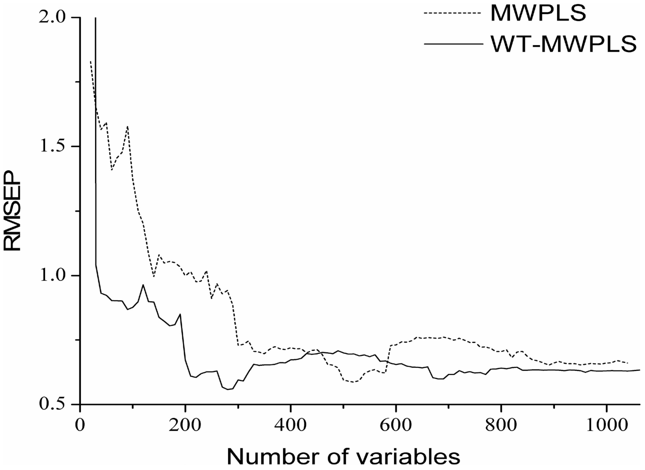
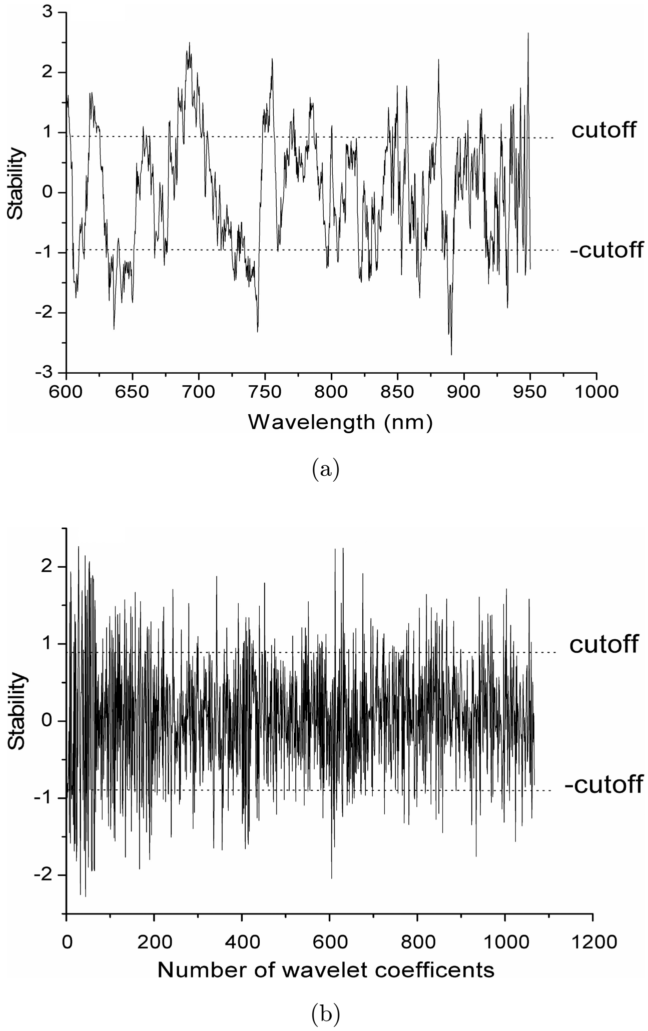


Fig. 3. Variation of RMSEP with the number of retained wavelengths and retained wavelet coefficients selected by MWPLS.

gradually with a little fluctuation. It indicated that, with lesser variables, useful variables were not completely included and led to poor results. Therefore, 520 retained variables were used to further study the raw spectra and WT spectra, 280 variables were selected in our research.

For MC-UVE method, the stabilities of each variable at wavelength range of 600–950 nm of the raw spectra and the wavelet coefficients are shown in Figs. 4(a) and 4(b), respectively. In the two figures, the dotted lines show the cutoff value. Variables whose stability lies within the dotted lines will be eliminated, and the variables whose stability lies out of the dotted lines will be used for PLS calculation. With the comparison of the two figures in Fig. 4, it can be seen that less retained variables were selected when WT was conducted. PLS regression models using the retained variables with relatively high absolute stability were developed. The variation of the RMSEP of the prediction set with different number of retained variables is investigated. Figure 5 shows the variation of RMSEP value with the number of retained variables. The dotted line represents the RMSEP values obtained by PLS regression with the retained variables of raw spectra, and the solid line is obtained with the retained wavelet coefficients. It can be seen that when the number of retained variables are 490 and 180 for the raw spectra and the wavelet coefficients, respectively, the lowest RMSEP are obtained. Therefore, the number of retained variables of 490 was used for the raw spectra for further study and the number of retained variables of 180 was used for the wavelet coefficients.



(a)

(b)

Fig. 4. The stability distribution of each variable: each wavelengths (a) and each retained wavelet coefficient (b) for the prediction of SSC by the MC-UVE method. The two dot lines indicate the lower and upper threshold.

3.3. *Comparison of predicted results by MW-PLS, MC-UVE-PLS and PLS methods*

With the parameters discussed above, PLS models with different inputs (obtained by MWPLS, WT-MWPLS, MC-UVE and WT-MC-UVE methods) were developed. The performances of the models in predicting SSC in navel oranges were compared with the PLS model with raw spectra. Results are given in Table 2. It can be seen that, comparing with the model obtained with raw spectra data,

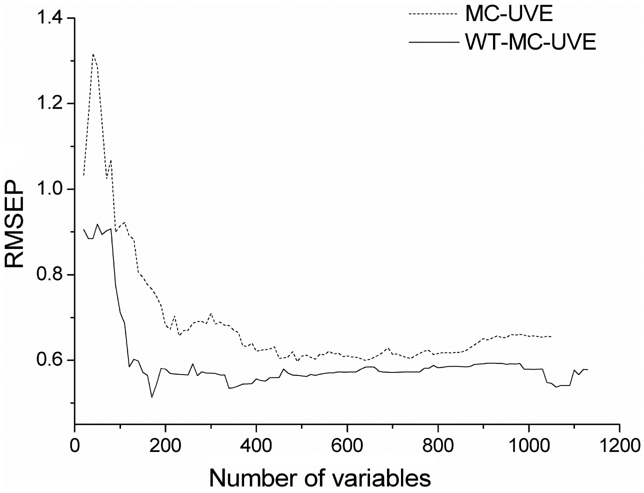


Fig. 5. Variation of RMSEP with the number of retained wavelengths and retained wavelet coefficients selected by MC-UVE.

Table 2. Comparison of the results obtained by PLS regression with different processing methods.

Processing method	No. of variables	r	RMSEP (°Birx)
None	1059	0.84	0.66
MWPLS	520	0.87	0.59
WT- MWPLS	280	0.88	0.57
MC-UVE	490	0.87	0.60
WT-MC-UVE	180	0.89	0.54

better prediction results of SSC were obtained by MW-PLS and MC-UVE treatment methods. Less number of variables was needed, higher value of r and lower value of RMSEP were obtained. When WT was involved to compress the spectra, the models were refined again. Comparing with MW-PLS method, fewer variables were retained by MC-UVE, but the prediction results remain almost the same. The best PLS model of SSC was WT combined with MC-UVE with an r of 0.89, RMSEP of 0.54 °Brix and only 180 variables were required.

As discussed above, WT-MC-UVE resulted in the best r and RMSEP of performance. Therefore, the method was used for all spectra datasets collected in different times.[23] Figure 6 illustrates the predicted result for SSC in navel oranges with WT-MC-UVE method. Vis-NIR spectroscopy successfully predicted the concentration of SSC in navel oranges at the moving situation, as shown in Fig. 6. It illustrates that Vis-NIR analysis of SSC in navel oranges gives r of 0.88, and RMSEP of 0.55 °Birx.

The prediction results for the SSC obtained in this study are superior to those obtained by Cen *et al.*[24] in orange juice with RMSEP of 0.73 °Brix and Miller and Zude-Sasse[25] in Florida citrus with r^2 of 0.67 (coefficient of determination). On the other hand, better results have been obtained by Gomez *et al.*[26] in Satsuma mandarin ($r^2 = 0.88$, RMSEP $= 0.33$ °Brix) and McGlone *et al.*[27] in

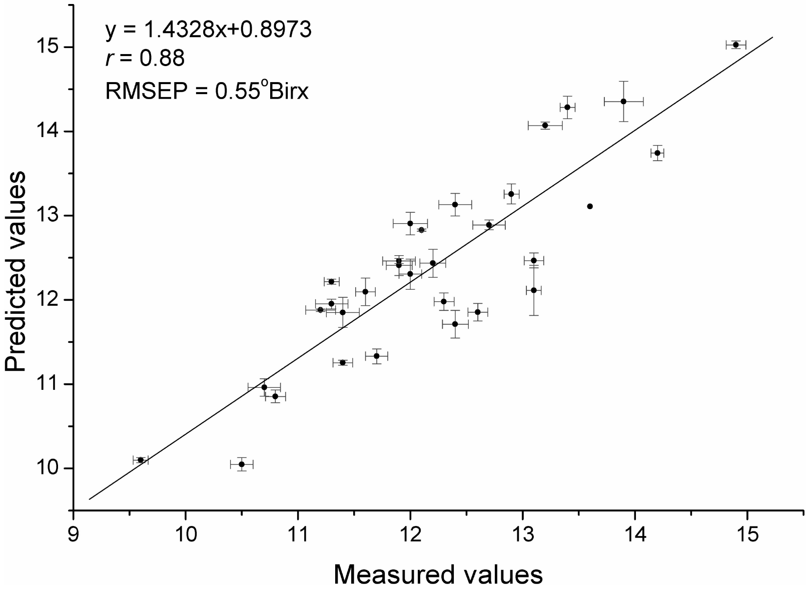


Fig. 6. The predicted results for SSC in navel orange for the prediction set from the PLS models processing with WT-MC-UVE method; The X error bars are the means and the standard errors of the measured values, and the Y error bars are the means and the standard errors of the NIR predictions.

mandarin with r^2 of 0.93 and RMSEP of 0.32 °Brix. This may contribute to the application of transmittance model. In addition, the prediction accuracy in this study is slightly poor in comparison to measurement accuracies achieved on many other fruit types.[28,29] This may contribute to the influence of thicker skin of navel orange than other fruit types. Few studies have reported the ability of an online Vis-NIR technology to determine SSC. Sun *et al.*[30] evaluated the use of Vis-NIR in measuring SSC of intact pears online, and better results were acquired with RMSEP of 0.53 °Brix. However, no reports were found about online SSC measurements of navel oranges.

4. Conclusions

Variable selection methods including MW-PLS and MC-UVE were proposed for simplifying calibration models for online quantitative analysis of SSC by Vis-NIR. It was proved that two methods are efficient, and slightly better results can be obtained compared with full-spectral PLS methods. Furthermore, WT was introduced before variable selection. The results show that WT combined with the MC-UVE method can further simplify modeling process, and enhance the efficiency in building the models, and it is feasible for the online Vis-NIR system for nondestructive detection of SSC in navel oranges.

Acknowledgments

The authors gratefully acknowledge the financial support provided by National Natural Science Foundation of China (60844007, 61178036, 21265006), National Science and Technology Support Plan (2008BAD96B04), Special Science and Technology-Support Program for Foreign Science and Technology Cooperation Plan (2009BHB15200) and Technological expertise and academic leaders training plan of Jiangxi Province (2009DD00700).

References

1. B. M. Nicolaï, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K. Theron, J. Lammertyn, "Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review," *Postharvest Biol. Technol.* **46**(2), 99–118 (2007).
2. L. Xuan, T. Teruo, K. Koki, S. H. Zhang, "Wavelength selection in Vis/NIR spectra for detection of bruises on apples by ROC analysis," *J. Food Eng.* **109**(3), 457–466 (2012).
3. H. Martens, T. Naes, *Multivariate Calibration*, Wiley, Chichester (1989).
4. M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, J. Pagès, "NIR calibration in non-linear systems: Different PLS approaches and artificial neural networks," *Chemometr. Intell. Lab. Syst.* **50**(1), 75–82 (2000).
5. A. Hoskuldsson, "Variable and subset selection in PLS regression," *Chemometr. Intell. Lab. Sys.* **55**(1–2), 23–38 (2001).
6. D. F. Barbin, G. ElMasry, D. W. Suna, P. Allen, "Non-destructive determination of chemical composition in intact and minced pork using near-infrared hyperspectral imaging," *Food Chem.* **138** (2–3), 1162–1171 (2013).
7. H. J. He, D. Wu, D. W. Sun, "Non-destructive and rapid analysis of moisture distribution in farmed Atlantic salmon (Salmo salar) fillets using visible and near-infrared hyperspectral imaging," *Innov. Food Sci. Emerg. Technol.* **18**, 237–245 (2013).
8. M. Kamruzzaman, G. ElMasry, D. W. Sun, P. Allen, "Non-destructive assessment of instrumental and sensory tenderness of lamb meat using NIR hyperspectral imaging," *Food Chem.* **141**(1), 389–396 (2013).
9. D. Wu, D. W. Sun, Y. He, "Application of long-wave near infrared hyperspectral imaging for measurement of color distribution in salmon fillet," *Innov. Food Sci. Emerg. Technol.* **16**, 361–372 (2012).
10. D. Wu, D. W. Sun, "Potential of time series-hyperspectral imaging (TS-HSI) for non-invasive determination of microbial spoilage of salmon flesh," *Talanta* **111**, 39–46 (2013).
11. D. Wu, D. W. Sun, "Application of visible and near infrared hyperspectral imaging for non-invasively measuring distribution of water-holding capacity in salmon flesh," *Talanta* **116**, 266–276 (2013).
12. L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, S. B. Engelsen, "Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy," *Appl. Spectrosc.* **54**(3), 413–419 (2000).
13. H. R. Xu, B. Qi, T. Sun, X. P. Fu, Y. B. Ying, "Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears," *J. Food Eng.* **109**(1), 142–147 (2012).
14. R. Leardi, L. Nørgaard, "Sequential application of backward interval partial least squares and genetic

algorithms for the selection of relevant spectral regions," *J. Chemom.* **18**(11), 486–497 (2004).

15. V. J. Barclay, R. F. Bonner, I. P. Hamilton, "Application of wavelet transforms to experimental spectra: Smoothing, denoising, and data set compression," *Anal. Chem.* **69**(1), 78–90 (1997).
16. J. H. Jiang, R. J. Berry, W. S. Heinz, Y. Ozaki, "Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data," *Anal. Chem.* **74**(14), 3555–3565 (2002).
17. R. F. Kokaly, R. N. Clark, "Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression," *Remote Sens. Environ.* **67**(3), 267–287 (1999).
18. V. Centner, D. L. Massart, O. E. Denoord, S. Dejong, B. M. Vandeginste, Sterna, "Elimination of uninformative variables for multivariate calibration," *Anal. Chem.* **68**(21), 3851–3858 (1996).
19. W. S. Cai, Y. K. Li, X. G. Shao, "A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra," *Chemometr. Intell. Lab. Syst.* **90**(2), 188–194 (2008).
20. H. C. Goicoechea, A. C. Olivieri, "A comparison of orthogonal signal correction and net analyte preprocessing methods, theoretical and experimental study," *Chemometr. Intell. Lab. Syst.* **56**(2), 73–81 (2001).
21. K. Baumann, N. Stiefl, "Validation tools for variable subset regression," *J. Comput. Aided Mol. Des.* **18**, 549–562 (2004).
22. X. G. Shao, Y. Z. Zhuang, "Determination of chlorogenic acid in plant samples by using near-infrared spectrum with wavelet transform preprocessing," *Anal. Sci.* **33**(2), 451–454 (2004).
23. X. Chen, D. Wu, Y. He, "An integration of modified uninformative variable elimination and wavelet packet transform for variable selection," *Spectroscopy* **26**(4), 42–47 (2011).
24. H. Y. Cen, Y. He, M. Huang, "Measurement of soluble solids contents and pH in orange juice using chemometrics and Vis-NIRS," *J. Agric. Food Chem.* **54**(20), 7437–7443 (2006).
25. W. M. Miller, M. Zude-Sasse, "NIR-based sensing to measure soluble solids content of Florida citrus," *Appl. Eng. Agric.* **20**(3), 321–327 (2004).
26. A. H. Gomez, Y. He, A. G. Pereira, "Non-destructive measurement of acidity, soluble solids and firmness of Satsuma mandarin using Vis/NIR-spectroscopy techniques," *J. Food Eng.* **77**(2), 313–319 (2006).
27. V. A. McGlone, D. G. Fraser, R. B. Jordan, R. Künnemeyer, "Internal quality assessment of mandarin fruit by Vis-NIR spectroscopy," *J. Near Infrared Spectrosc.* **11**(5), 323–332 (2003).
28. Y. D. Liu, X. M. Chen, A. G. Ouyang, "Nondestructive determination of pear internal quality indices by visible and near-infrared spectrometry," *LWT - Food Sci. Technol.* **41**(9), 1720–1725 (2008).
29. C. Camps, D. Christen, "Non-destructive assessment of apricot fruit quality by portable visible-near infrared spectroscopy," *Food Sci. Technol.* **42**(6), 1125–1131 (2009).
30. T. Sun, H. J. Lin, H. R. Xu, Y. B. Ying, "Effect of fruit moving speed on predicting soluble solids content of 'Cuiguan' pears (Pomaceae pyrifolia Nakai cv. Cuiguan) using PLS and LS-SVM regression," *Postharvest Biol. Technol.* **51**(1), 86–90 (2009).