Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

Maejo International Journal of Science and Technology

ISSN 1905-7873

Available online at www.mijst.mju.ac.th

Full Paper

Selection of efficient wavelengths in NIR spectrum for determination of dry matter in kiwi fruit

Lü Qiang[1], Tang Mingjie[1], Cai Jianrong[1,*], Lu Huazhu[1] and Sumpun Chaitep[2]

[1] School of Food and Biological Engineering, Jiangsu University, Zhenjiang, Jiangsu, 212013, China

[2] Department of Mechanical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, 50200, Thailand

* Corresponding author, e-mail: lvqiang1111@gmail.com

Received: 19 November 2009 / Accepted: 9 March 2010 / Published: 2 April 2010

Abstract: The feasibility of using efficient wavelengths in the near-infrared (NIR) spectrum for the rapid determination of the dry matter (DM) in kiwi fruit was investigated. Partial least squares (PLS), synergy interval PLS (siPLS) and genetic algorithm siPLS (GA-siPLS) were comparatively performed to calibrate regression models. The number of wavelengths and the number of PLS components were optimised as per the root mean square error of cross-validation (RMSECV) in the calibration set. The performance of the final model was evaluated by the root mean square error of prediction (RMSEP) and the correlation coefficient (r) in the prediction set. Results indicate that the performance of GA-siPLS model is the best one compared to PLS and siPLS models. The optimal model was achieved with r = 0.9020 and RMSEP = 0.5315 in the prediction set. This work shows that it is feasible to determine DM in kiwi fruit using NIR spectroscopy and that GA-siPLS algorithm is most suitable in solving the problem of selection of efficient wavelengths.

Keywords: kiwi fruit, dry matter, NIR spectroscopy, partial least squares (PLS), synergy interval partial least squares (siPLS), genetic algorithm siPLS.

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

Introduction

Kiwi fruit are harvested unripe though physiologically mature but must be left in natural storage to ripen before consumption [1]. Timing of the harvest has a decisive effect on the subsequent postharvest shelf life and fruit quality [2-3]. The dry matter (DM) in kiwi fruit has been proposed as a maturity indicator for the proper time of harvest and also as a predictor of the sensory quality of the fruit once it is ripe [4-6].

Near infrared (NIR) spectroscopy is a fast, accurate and non-destructive technique that can be deployed as a replacement of individuals' labour skills and time-consuming methods. The NIR spectroscopy has been used to grade fruits [7-8], predict fruit maturity [9] and indicate optimal harvesting time [10]. Kiwi fruit are a commodity the sorting of which, based on pre-selected NIR spectral features, can be used to grade them at harvest on the basis of DM. Recent research has established that NIR spectroscopic analysis can be used to assess kiwifruit DM and/or soluble-solid content of the ripe fruit [1, 3, 6, 11-12].

In addition to these, NIR spectral data calibrations have been made with the classical multivariate calibration analysis, e.g. partial least squares (PLS) regression [13-14]. Many spectral pretreatment methods have been developed to reduce the effects of variations in the spectral data that are not related to the chemical variations in the samples [15-16]. These methods generally improve the calibrations. However, they did not take into account that there might be spectral regions that do not contain any information about the chemical variations in the samples [17]. In fact, one of the major tasks in multivariate data analysis is to select appropriate spectral regions in order to achieve the best performance. A number of researchers have constructed PLS models in different spectral regions to quantify ingredient content in kiwi fruit. However, these regions were selected manually [2-3]. Without prior detailed knowledge about NIR spectroscopy, spectral regions selected manually might as well weaken the performance of the calibration model.

According to some other researchers, both theoretical and experimental evidence has been published to the effect that spectral region selection can significantly improve the performance of these calibration techniques [18-19]. It is important to select specific regions that contain much information based on which of the more stable models can be generated with superior interpretability and lower prediction error. Methods [e.g.19] have been recently described in the literature in implementing spectral region selection and PLS used for multivariate calibration in each subset.

A graphically oriented local modelling procedure called interval partial least squares (iPLS) has been presented for use on NIR spectral data. It has been shown that selective optimum interval in the spectral data can yield precision prediction models [19-20]. A method called synergy interval partial least squares (siPLS) has also been proposed to be used to select several interval spectral data which can split the data set into a number of intervals (variable-wise) and to calculate all possible PLS model combinations of two, three or four intervals [17]. Genetic algorithm (GA) has already been used in variable selection problem and seems to be a solution to the multivariate selection of variables [21-22].

This study investigates and compares the results provided by PLS, siPLS and GA-siPLS procedures for NIR quantitative analysis of DM in kiwi fruit. Two specific objectives of this research

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

are: (1) to establish relationships between the NIR measurements and the DM of kiwi fruit based on the new method, and (2) to compare the prediction performance of calibration models at different wavelengths and then find out the optimal wavelengths and develop the best calibration models.

Materials and Methods

Sample preparation

One hundred and twelve "*Zhonghua*" kiwi fruit samples, purchased from a farm in Zhouzhi, Shanxi Province, China, were used in this study. All sizes of the fruit from peewee to jumbo were used. However, the fruit with irregular shape were not incorporated in the data analysis. The fruit were sent to our laboratory in October 2008, then were stored for one month. Experiments were done under controlled condition (20^{o}C). Before being examined by NIR technique, the fruit were acclimatised to equilibrium for 12 h in the controlled condition.

Collection of spectra

The NIR spectra were measured in the reflectance mode using the FT-NIR spectrophotometer (Antaris^{TM} Analyser, Thermo Electron Co., USA) with an integrating sphere. Each spectrum was obtained from an average of 32 scans. The range of spectrum was 10,000-4,000 cm^{-1} and the data were collected in 1.928 cm^{-1} intervals, which resulted in 3,112 variables. Each kiwi fruit was measured three times around equatorial locations. The average of the three spectra, which were measured at the equator of each kiwifruit, was used in the sequence analysis.

Measurement of kiwi fruit reference DM

The fruit DM was determined by cutting two equatorial slices of approximately 3-mm thickness each, and drying them at 65^{o}C to constant weight (approximately 24 h). The fruit DM was calculated from the final dry weight and the initial wet weight of the slices, recorded as a percentage of fresh weight.

Software

All algorithms were implemented in Matlab V7.0 (Mathworks, USA) under Windows XP. Result Software (Antaris System, Thermo Electron Co., USA) was used in NIR spectral data acquisition. The iPLS, siPLS and GAPLS algorithms used in this work were downloaded from http://www. models.kvl.dk/.

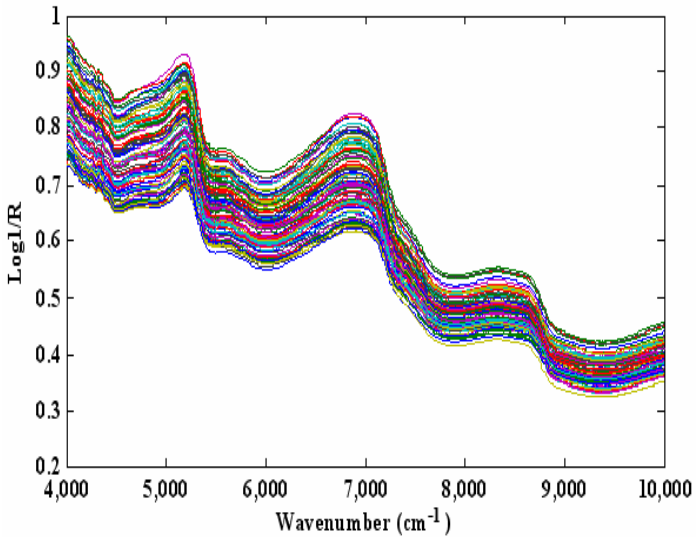
Results and Discussion

Spectral pre-processing

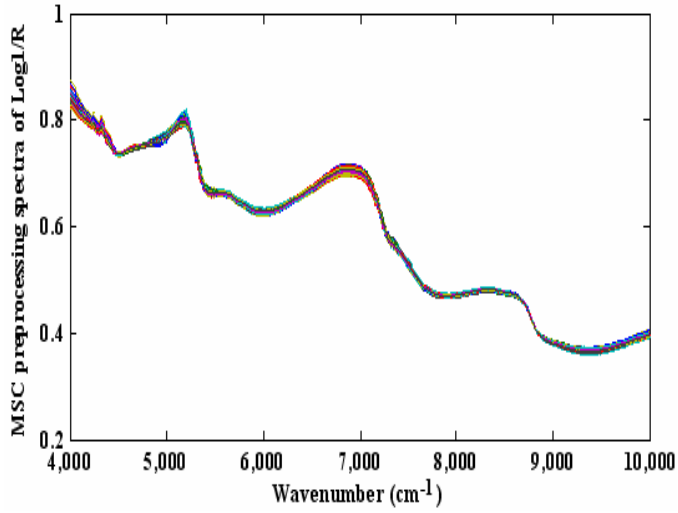
Figure 1(a) presents the raw spectral profile of the kiwi fruit, the raw spectral data being conducted on spectral pre-processing. Each mean spectrum was recorded as log(1/R), where R is the reflectance. In this research, the spectral data were analysed with multiplicative scatter correction (MSC) pre-processing technique because MSC is an important procedure for the correction of

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

scattered light, and the technique is often used to correct for additive and multiplicative effects in the spectra [23]. The spectra after MSC pre-processing are presented in Figure 1(b).



(a)



(b)

Figure 1. Spectra of kiwi fruit obtained from (a) raw data and (b) MSC pre-processed data

Calibration of models

All 112 samples were divided into two subsets. The first one was the calibration set, which was used to build the models, whereas the other was the prediction set, which was used to test the robustness of the established models. To avoid bias in the subset division, it was made by sorting all samples according to their respective y-value (viz. the reference measurement value of dry matter). In order to achieve a 2/1 division of calibration/prediction spectra, one spectrum of every three samples was assigned to the prediction set so that finally the calibration set contained 74 spectra and the remaining 38 spectra constituted the prediction set. Seen from Table 1 is the range of y-value in

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

the calibration set that covers the range in the prediction set. Therefore, the distribution of the samples was appropriate in both the calibration and prediction sets.

Table 1. Reference measurements of DM and sample numbers in calibration and prediction sets

Set	Unit	Number of samples	Mean value	Range	Standard deviation	CV /%
Calibration set	% (g/g)	74	16.1736	13.526-18.757	1.2554	7.7622
Prediction set	% (g/g)	38	16.2237	13.758-18.584	1.2202	7.5210

Note: CV = coefficient of variation

The performance of the final PLS model was evaluated in terms of the root mean square error of cross-validation (RMSECV), the root mean square error of prediction (RMSEP), and the correlation coefficient (r). For RMSECV, a leave-one-sample-out cross-validation was performed: the spectrum of one sample of the training set was deleted from this set and a PLS model was built with the remaining spectra of the calibration set. The left-out sample was predicted with this model and the procedure was repeated by leaving out each of the samples of the calibration set. The RMSECV was calculated by Eq. 1 [24]:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{I_c}(\hat{y}_i - y_i)^2}{I_c - 1}} \quad (1)$$

where $\hat{y}_i$ is the predicted value of the i th observation, y_i the measured value of i th observation and I_c the number of observation in the calibration set. The number of PLS factors included in the model was chosen according to the lowest RMSECV. This procedure was repeated for each of the pre-processed spectra.

For the prediction set, the RMSEP was calculated by Eq. 2 [24]:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I_p}(y_i - \tilde{y}_i)^2}{I_p}} \quad (2)$$

where $\tilde{y}_i$ is the predicted value for sample i of the prediction set, y_i is the measured value for sample i of the prediction set, and I_p is the number of observation in the prediction set. The correlation coefficients (r) between the predicted and measured values were calculated by Eq. 3 [24] for both the calibration and prediction sets:

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

$$r = \sqrt{1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}} \quad (3)$$

where $\hat{y}_i$ and y_i are the predicted and measured values respectively of sample i in calibration or prediction set, $\bar{y}_i$ is the mean of the reference measurement results for all samples in the calibration or prediction set, and n is the number of observation in the calibration or prediction set.

To verify the superior capability of the PLS calibration models based on the selected region by different methods, each calibration model mentioned above was used to predict the calibration data set and the prediction data set. The RMSECV, RMSEP and correlation coefficients of each model for the calibration data set (r_c) and validation data set (r_p) were taken into account.

Results of PLS model

In the application of PLS algorithm, it is generally known that the number of PLS components is a critical parameter in calibrating the model. The optimum number of PLS components is determined by the lowest RMSECV, which is 0.5513 when 12 PLS components are included in the calibration model. Therefore the optimal number of PLS components is 12.

In the optimal model, RMSECV is 0.5513 and correlation coefficient (r) is 0.8913 in calibration set. When the performance of PLS model is evaluated by the samples in the prediction set, RMSEP is 0.5926 and correlation coefficient (r) is 0.8806. Figure 2 represents the scatter plot showing a correlation between reference and NIR-predicted DM in the prediction set by PLS model.

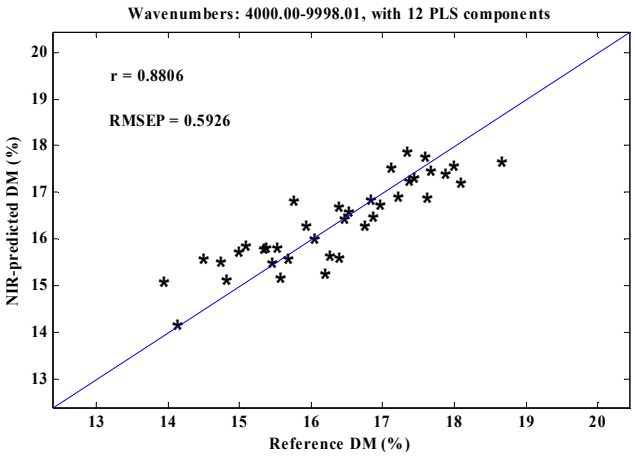


Figure 2. Reference versus NIR-predicted DM by PLS in prediction set

Results of siPLS model

The synergy interval PLS (siPLS) algorithm used here has been developed by Nørgaard et al. [19]. First, the data set is split into a number of intervals (variable-wise). Next, PLS regression models are established for all possible combinations of two, three or four intervals. Thereafter, RMSECV is calculated for every combination of intervals. The combination of intervals with the lowest RMSECV is then chosen.

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

The number of intervals is also optimised according to RMSECV in siPLS model calibration. Table 2 shows the results of siPLS model calibration when splitting the spectra into different numbers of intervals. The optimal siPLS model is obtained with 15 intervals and 10 PLS components, the lowest RMSECV being 0.5139. The optimal combination of intervals selected is 3, 4, 8 and 12. It corresponds to 4,802.04-5,201.14, 5,203.07-5,602.16, 6,807.16-7,204.33 and 8,403.55-8,800.71 cm^{-1} in the spectral regions as shown in Figure 3.

For the optimal model, RMSECV is 0.5139, and correlation coefficient (r) is 0.9062 in the calibration set. When the performance of siPLS model is evaluated by the samples in the prediction set, RMSEP is 0.5710 and correlation coefficient (r) is 0.8903. Figure 4 represents the scatter plot showing a correlation between reference and NIR-predicted DM in the prediction set by siPLS model.

Table 2. Results of siPLS model calibration for different spectral regions

Number of intervals	No.of PLS components	Selected intervals	Calibration set		Prediction set	
			r	RMSECV	r	RMSEP
13	9	[3 7 10 12]	0.9115	0.4992	0.8829	0.5822
14	9	[3 7 11 13]	0.9247	0.4625	0.8862	0.5800
15	**10**	**[3 4 8 12]**	**0.9062**	**0.5139**	**0.8903**	**0.5710**
16	7	[1 9 13]	0.9014	0.5254	0.8348	0.6693
17	13	[2 4 7 13]	0.9056	0.5142	0.8592	0.6213
18	6	[3 9 14]	0.8896	0.5538	0.8352	0.6817
19	8	[3 4 7 15]	0.9283	0.4499	0.8695	0.6086
20	6	[1 11 16]	0.9032	0.5196	0.8548	0.6370

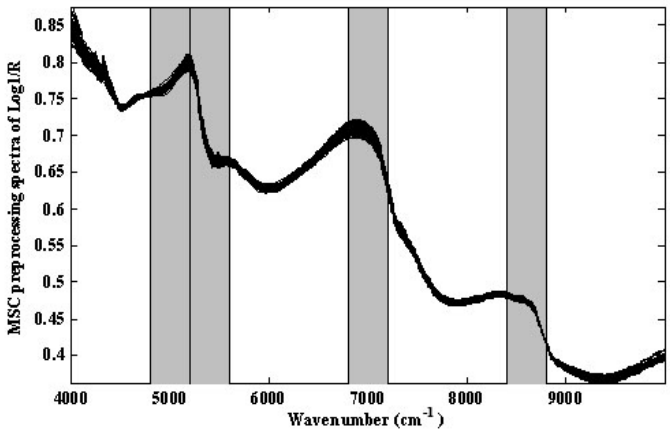


Figure 3. Optimal spectral regions selected by siPLS with wavenumbers of 4,802.04-5,201.14, 5,203.07-5,602.16, 6,807.16-7,204.33 and 8,403.55-8,800.71 cm^{-1}

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

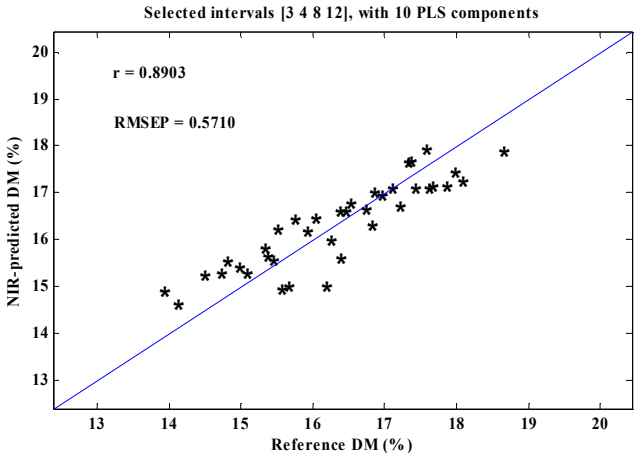


Figure 4. Reference versus NIR-predicted DM by siPLS in prediction set

Results of GA-siPLS model

GA is an optimisation method based on the principles of genetics and natural selection. This algorithm is inspired by the theory of evolution. In a living environment, the 'best' individuals have a greater chance to survive and a greater probability to spread their genomes by reproduction. The mating of two 'good' individuals causes the mixing of their genomes, which may result in a 'better' offspring. The terms 'good', 'better' and 'best' are related to the fitness of the individuals to their environment [21-22].

The number of wavelengths is also similarly optimised by RMSECV using GA in the optimal combination of intervals (4,802.04-5,201.14, 5,203.07-5,602.16, 6,807.16-7,204.33 and 8,403.55-8,800.71 cm^{-1}) selected by siPLS model. The optimal parameters are set as follows: number of generations = 100, population size = 30, mutation probability = 0.1, and recombination probability = 0.8. Figure 5 shows the selected frequency versus DM variable in the first spectral region (4,802.04-5,201.14 cm^{-1}) of the optimal combination of intervals. Wavelength variables are individually added to PLS model in accordance with the selected frequency. The best number of wavelength variables is then identified according to the RMSECV of the model. The optimal GA-siPLS model is obtained with 229 wavelengths and 9 PLS components when the lowest RMSECV is 0.4724.

In the optimal model, RMSECV is 0.4724 and correlation coefficient (r) is 0.9209 in the calibration set. When the performance of GA-siPLS model is evaluated by the samples in the prediction set, RMSEP is 0.5315 and correlation coefficient (r) is 0.9020 in the prediction set. Figure 6 represents the scatter plot showing a correlation between reference and NIR-predicted DM in the prediction set by GA-siPLS model.

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

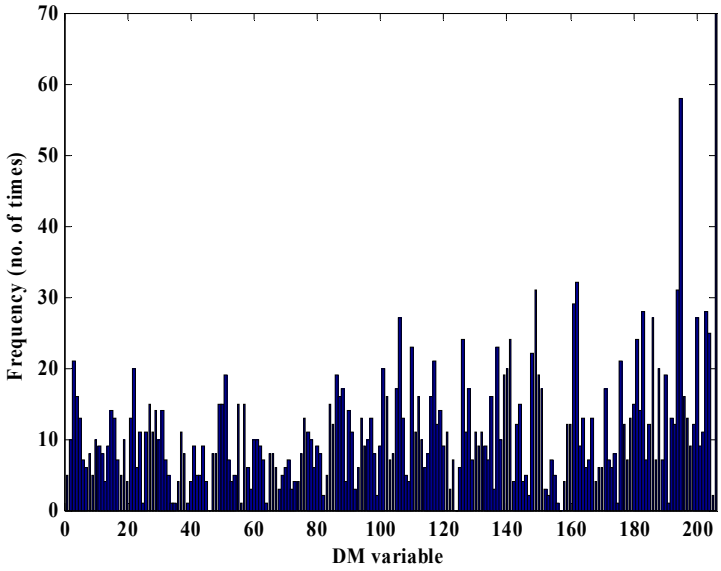


Figure 5. Selected frequency versus DM variable

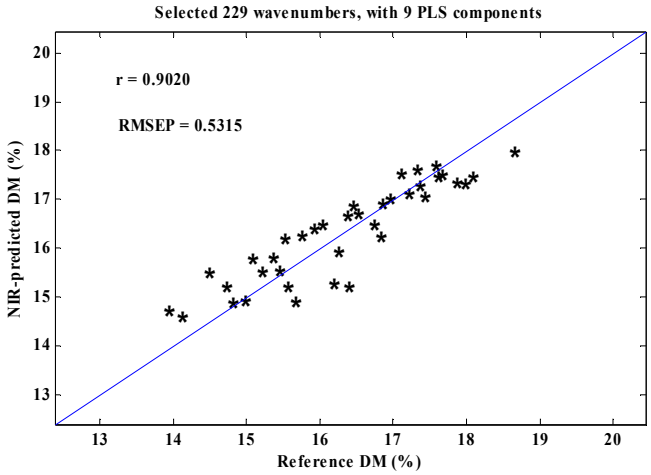


Figure 6. Reference versus NIR-predicted DM by GA-siPLS in prediction set

Table 3 shows results from different PLS models. Comparing among these models, one can see that GA-siPLS seems to be the best one followed by siPLS. Such phenomena can be explained by the following: (1) PLS is performed on full spectral range (4,000.00-10,000 cm^{-1}) to calibrate the global model. Thus, some noisy spectral information has inevitably weakened the performance of the model; (2) siPLS overcomes the disadvantages of PLS since siPLS combines with two, three or four intervals to calibrate the PLS model so as to remove some noisy regions and obtain useful information in the calibrated model; and (3) in contrast with siPLS, GA-siPLS selects interesting variable wavelengths and removes noisier spectral information based on siPLS.

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

Table 3. Results from different PLS models

Model	Number of variables	PLS components	Calibration set		Prediction set	
			r	RMSECV	r	RMSEP
PLS	3112	12	0.8913	0.5513	0.8806	0.5926
siPLS	830	10	0.9062	0.5139	0.8903	0.5710
GA-siPLS	229	9	0.9209	0.4724	0.9020	0.5315

Conclusions

In the present study, it has been demonstrated that NIR spectroscopy is a suitable tool for quantification of dry matter in kiwi fruit with small prediction errors over the entire range studied. Three models were studied. The PLS model was performed on full spectral region (4,000.00-10,000 cm^{-1}, 3112 variables) to calibrate the model. It requires a large number of variables and some noisy spectral information has inevitably reduced the prediction accuracy of the model. The siPLS model was performed on four intervals of the spectral region (4,802.04-5,201.14, 5,203.07-5,602.16, 6,807.16-7,204.33 and 8,403.55-8,800.71 cm^{-1}, 830 variables) to calibrate the model. Some noisy regions were removed so as to reduce variables and improve prediction accuracy. The GA-siPLS model was performed on the most informative wavelengths (4,802.04-5,201.14 cm^{-1}, 229 variables) to calibrate the model. Compared with PLS and siPLS models, the GA-siPLS model requires fewer variables and improves prediction accuracy by removal of more spectral noises.

Acknowledgements

This work was financially supported by the National High Technology Research and Development Program of China (Project No. 2006AA10Z263) and the Key Natural Science Foundation of Jiangsu Province (Grant No. BK2006707-1). We are grateful to the website http://www.models.kvl.dk/ for their generous contribution in permitting the download of software for iPLS, siPLS and GAPLS free of charge.

References

1. V. A. McGlone, R. B. Jordan, R. Seelye and P. J. Martinsen, "Comparing density and NIR methods for measurement of kiwi fruit dry matter and soluble solids content", *Postharvest Biol. Technol.*, **2002**, *26*, 191-198.
2. D. C. Slaughter and C. H. Crisosto, "Nondestructive internal quality assessment of kiwi fruit using near-infrared spectroscopy", *Semin. Food Anal.*, **1998**, *3*, 131-140.
3. V. A. McGlone, C. J. Clark and R. B. Jordan, "Comparing density and VNIR methods for predicting quality parameters of yellow-fleshed kiwi fruit (Actinidia chinensis)", *Postharvest Biol. Technol.*, **2007**, *46*, 1-9.

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

4. J. Burdon, D. McLeod, N. Lallu, J. Gamble, M. Petley and A. Gunson, "Consumer evaluation of "Hayward" kiwi fruit of different at-harvest dry matter contents", *Postharvest Biol. Technol.*, **2004**, *34*, 245-255.
5. R. B. Jordan, E. F. Walton, K. U. Klages and R. J. Seelye, "Postharvest fruit density as an indicator of dry matter and ripened soluble solids of kiwi fruit", *Postharvest Biol. Technol.*, **2000**, *20*, 163-173.
6. V. A. McGlone and S. Kawano, "Firmness, dry-matter and soluble-solids assessment of postharvest kiwi fruit by NIR spectroscopy", *Postharvest Biol. Technol.*, **1998**, *13*, 131-141.
7. I. W. Budiastra, Y. Ikeda and T. Nishizu, "Prediction of individual sugars and malic acid concentrations of apples and mangoes by the developed NIR reflectance system", *J. Jpn. Soc. Agric. Machine.*, **1998**, *60*, 117-127.
8. V. Steinmetz, J. M. Roger, E. Molto and J. Blasco, "On-line fusion of colour camera and spectrophotometer for sugar content prediction of apples", *J. Agric. Eng. Res.*, **1999**, *73*, 207-216.
9. A. Peirs, N. Scheerlinck and B. M. Nicolaï, "Temperature compensation for near infrared reflectance measurement of apple fruit soluble solids contents", *Postharvest Biol. Technol.*, **2003**, *30*, 233-248.
10. B. Park, J. A. Abbott, K. J. Lee, C. H. Choi and K. H. Choi, "Near-infrared diffuse reflectance for quantitative and qualitative measurement of soluble solids and firmness of delicious and dala apples", *Trans. Am. Soc. Agric. Eng.*, **2003**, *46*, 1721-1731.
11. C. J. Clark, V. A. McGlone, H. N. De Silva, M. A. Manning, J. Burdon and A. D. Mowat, "Prediction of storage disorders of kiwi fruit (Actinidia chinensis) based on visible-NIR spectral characteristics at harvest", *Postharvest Biol. Technol.*, **2004**, *32*, 147-158.
12. P. N. Schaare and D. G. Fraser, "Comparison of reflectance, interactance and transmission modes of visible-near infrared spectroscopy for measuring internal properties of kiwi fruit (Actinidia chinensis)", *Postharvest Biol. Technol.*, **2000**, *20*, 175-184.
13. J. Lammertyn, A. Peirs, J. D. Baerdemaeker and B. Nicolaï, "Light penetration properties of NIR radiation in fruit with respect to non-destructive quality assessment", *Postharvest Biol. Technol.*, **2000**, *18*, 121-132.
14. A. Peirs, J. Lammertyn, K. Ooms and B. M. Nicolaï, "Prediction of the optimal picking date of different apple cultivars by means of VIS/NIR-spectroscopy", *Postharvest Biol. Technol.*, **2001**, *21*, 189-199.
15. Q. S. Chen, J. W. Zhao, X. Y. Huang, H. D. Zhang and M. H. Liu, "Simultaneous determination of total polyphenols and caffeine contents of green tea by near-infrared reflectance spectroscopy", *Microchem. J.*, **2006**, *83*, 42-47.
16. Q. S. Chen, J. W. Zhao, H. D. Zhang and X. Y. Wang, "Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration", *Anal. Chim. Acta.*, **2006**, *572*, 77-84.

Maejo Int. J. Sci. Technol. **2010**, *4*(01), 113-124

17. C. Abrahamsson, J. Johansson, A. Sparén and F. Lindgren, "Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets", *Chemom. Intell. Lab. Syst.*, **2003**, *69*, 3-12.
18. O. Kleynen, V. Leemans and M. F. Destain, "Selection of the most efficient wavelength bands for 'Jonagold' apple sorting", *Postharvest Biol. Technol.*, **2003**, *30*, 221-232.
19. L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck and S. B. Engelsen, "Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy", *Appl. Spectrosc.*, **2000**, *54*, 413-419.
20. R. Leardi and L. Nørgaard, "Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions", *J. Chemom.*, **2004**, *18*, 486-497.
21. R. Leardi and A. L. González, "Genetic algorithms applied to feature selection in PLS regression: How and when to use them", *Chemom. Intell. Lab. Syst.*, **1998**, *41*, 195-207.
22. R. Leardi, "Application of genetic algorithm-PLS for feature selection in spectral data sets", *J. Chemom.*, **2000**, *14*, 643-655.
23. X. L. Chu, H. F. Yuan and W. Z. Lu, "Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique", *Prog. Chem.*, **2004**, *4*, 528-542.
24. X. B. Zou, J. W. Zhao and Y. X. Li, "Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of 'Fuji' apple based on BiPLS and FiPLS models", *Vib. Spectrosc.*, **2007**, *44*, 220-227.