International Food Research Journal 23(Suppl): S231-S236 (December 2016)
Journal homepage: *http://www.ifrj.upm.edu.my*

Influence of different pre-processing methods in predicting sugarcane quality from near-infrared (NIR) spectral data

[1]Lazim, S. S. R. M., [1*]Nawi, N. M., [2]Chen, G., [2]Jensen, T. and [1]Rasli, A. M. M.

[1]*Department of Biological and Agricultural Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400 Selangor, Malaysia*
[2]*Faculty of Health, Engineering and Sciences, University of Southern Queensland, Toowoomba, QLD 4350, Australia*

Article history

Received: 25 September 2016
Received in revised form: 15 October 2016
Accepted:17 October 2016

Keywords

Pre-processing
Sugarcane
Spectral data
Chemometric
Spectroradiometer

Abstract

The influence of different data pre-processing methods (smoothing by moving average (MA), multiplicative scatter correction (MSC), Savitzky-Golay (SG), standard normal variate (SNV) and mean normalization (MN) on the prediction of sugar content from sugarcane samples was investigated. The performance of these pre-processing methods was evaluated using spectral data collected from 292 sugarcane internode samples using a visible-shortwave near infrared spectroradiometer (VNIRS). Partial least square (PLS) method was applied to develop both calibration and prediction models for the samples. If no pre-processing method was applied, the coefficient of determination (R^2) values for both reflectance and absorbance data were 0.81 and 0.86 respectively. The highest prediction accuracy values were obtained when the data was treated with MSC method, where the R^2 values for reflectance and absorbance being 0.85 and 0.87, respectively. From this study, it was concluded that pre-processing can improve the model performances where MSC method was found to give the highest prediction accuracy value.

Introduction

In recent years, rapid development of near infrared spectroscopic (NIRS) techniques combined with multivariate analysis has enabled the technologies to be applied in sugarcane industries especially to predict quality level of the crop. Many studies have reported the application of spectroscopic methods to predict sugar content of sugarcane (Madsen *et al*., 2003; Mehrotra and Siesler, 2003; Taira *et al*., 2010; Nawi *et al*., 2012, 2013; Nawi, Chen and Jensen, 2013). The application of spectroscopic method however requires the multivariate analysis to extract useful data from spectral data. In this process, data pre-processing method is a critical task in the knowledge discovery process to ensure a robust calibration and prediction models can be developed.

For any spectroscopic measurements, a large amount of spectral data collected from NIRS instruments usually contains a lot of useful analytical and background information such as light scattering, path length variations and random noise as well as sample information (Blanco and Illarroya, 2002). This problem is more obvious when the spectral data was collected from solid samples. Since the robustness of the calibration and prediction models is the primary requirement for spectroscopic measurements, removing unwanted background information and noise are very essential.

The spectral data of solid sugarcane samples are influenced by their physical properties with scattering phenomena (which is wavelength-dependent and non-linear) is the most common factor for causing error in absorbance values. In order to obtain reliable, accurate and stable calibration models, it is compulsory to pre-process spectral data before modelling (Cen and He, 2007). Spectral pre-processing techniques are required to remove any irrelevant information including noise, uncertainties, variability, interactions and unrecognized features. Spectral pre-processing method should be used to minimize the influences of irrelevant information into spectra in order to be able to develop more simple and robust models (Blanco and Villarroya, 2002). Pre-processing of spectral data is a key part of spectral analysis used to improve the quality and accuracy of the regression models (Wu *et al*., 2008). Thus, the goal of this study was to investigate the influence of different spectral pre-processing methods on partial least square (PLS) model performance for both reflectance and absorbance data.

*Corresponding author.
Email: *nazmimat@upm.edu.my*

Materials and Methods

Crop samples

A total of 22 sugarcane stalks consist of 292 internode samples were collected from the research plot belongs to the Bureau of Sugar Experimental Station (BSES), Bundaberg, Queensland. The stalks belong to commercial variety trials representing three different maturity stages, namely early maturing (Q155), mid-maturing (Q208) and late-maturing (Q190) crops. The Brix obtained from these three varieties ranged from 7.6 to 22.2°Brix. The stalk samples were harvested after eight months of planting. The leafy part of each stalk sample was removed. Then, the stalks were cut on the node portion into an individual internode using a cutter. Each internode sample was cut into four sections of approximately the same length, representing the node and internode areas (Figure 1). The detailed information about the samples preparation and their characteristic has been reported by Nawi, Chen and Jensen (2013).

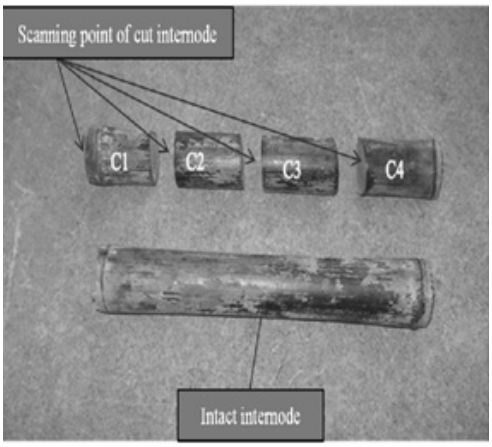

Figure 1. Intact internode vs cut internode with scanning positions (Nawi, Chen and Jensen, 2013).

Instrumentation and spectral measurement

The spectral data reflected from the cross-sectional surface of the cut internode was collected using a handheld visible/shortwave (325–1075 nm) near infrared spectroradiometer (Vis/SW-NIRS; FieldSpec HandHeld and FieldSpec Pro FR, from Analytical Spectral Devices (ASD), Inc., Boulder, CO, USA. The measurement was undertaken using the 25° field-of-view (FOV) of the spectroradiometer. The equipment was set to record the average of 20 scans for each spectrum. Relative reflectance spectra were calculated by dividing the reflectance of the internode samples with the reflectance from the white reference panel.

The spectral data was collected inside a measurement box (900 mm × 600 mm × 450 mm). The box was constructed to eliminate the influence of ambient light on the spectral measurement and to ensure a consistent distance and measurement angle between the probe and samples (Nawi *et al*., 2013). Two halogen lamps (Lowell Pro-Lamp 14.5 V tungsten bulb, Ushio Lighting, Inc., Japan) were placed at a distance of 800 mm above the sample at the angle of 45° to illuminate the samples.

All spectral data were stored in a computer and processed using the RS3 software for Windows (Analytical Spectral Devices, Boulder, CO, USA) designed with a graphical user interface. The reflectance spectra were transformed into ASCII format using the ASD ViewSpecPro software (Analytical Spectral Devices, Boulder, CO, USA). Then, the reflectance data (R) were transformed into absorbance data (A). In order to avoid a low signal-to-noise ratio, only the wavelength regions between 400 and 1000 nm were used for the calculations.

°Brix measurement

After the spectral measurement, each cut section was squeezed using a clamp to extract the juice samples. The juice from all cut sections of the same internode were collected and mixed in a container, shaken and poured onto a refractometer to measure the °Brix value. The °Brix values were measured using a hand-held °Brix refractometer (Model: RHB-32ATC, from Huake Instrument Co., Ltd, Baoan, Shenzhen, China; the °Brix range is 0–32% with automatic temperature compensation). The refractometer was cleaned after each measurement to avoid cross contamination.

Spectral data pre-processing

The spectral data of solid samples is normally influenced by the skin roughness of the samples which can cause some problems in assessing their internal quality attributes. Furthermore, the spectral data normally contains background information such as light scattering, path length variations and random noise as well as sample information. In order to obtain reliable, accurate and stable calibration models, it is essential to pre-process spectral data before modeling (Cen and He, 2007). Nicolaï *et al*. (2007) divided the pre-processing methods into four categories namely smoothing, standardization, normalization and differentiation. Smoothing techniques have been proposed to remove random noise from spectral data and to optimize the signal-to-noise ratio (Cen and He, 2007). The most common smoothing techniques are moving average and the Savitzky–Golay algorithm (Næs *et al*., 2004).

A standardization technique is used to divide the spectrum at every wavelength by the standard deviation

of the spectrum at a certain wavelength. Typically, variances of all wavelengths are standardized to one, which results in an equal influence of the variables in the model (Næs *et al*., 2004). A normalization technique is applied to compensate for additive (baseline shift) and multiplicative (tilt) effects in the spectral data, which are induced by physical effects such as the non-uniform scattering throughout the spectrum as the degree of scattering is dependent on the wavelength of the radiation, the particle size and the refractive index. Multiple scatter correction (MSC) and standard normal variate correction (SNV) are the most popular normalisation techniques. Differentiation methods include the first and second derivative are employed to remove background and increase spectral resolution (Cen and He, 2007).

Before calibration, the spectral data were pre-processed for optimal performance. The effect of several pre-processing techniques on the performance of PLS models investigated in this study included smoothing by moving average with three segments (MA3) and nine segments (MA9) (Wu *et al*., 2008), multiplicative scatter correction (MSC), Savitzky-Golay first derivative (SG1), Savitzky-Golay second derivative (SG2) (Swierenga *et al*., 1999), standard normal variate (SNV), mean normalization (MN) (Griffiths, 1995) and combinations of them. For comparison purposes, the raw spectral data without any pre-processing method was also analyzed. The performance of the models developed using different pre-processing methods were compared with one another based on R^2 and root means square error of predictions (RMSEP) values. The pre-processing processes were implemented using the Unscrambler, V 9.6 software (Camo Process AS, Oslo, Norway).

Development of calibration and validation models

Prior to the development of a calibration model, principal component analysis (PCA) was applied to extract useful information from the spectra, decrease the noise and determine the optimum number of latent variables (Wu *et al*., 2008). PCA is a well-known chemometrics method used to search for directions of maximum variability in sample grouping and using them as new axes called principle components (PCs) that can be used as new variables, instead of the original data, in the following calculations (Blanco and Illarroya, 2002). In this study, 10 PCs were used in all pre-processing methods to ensure the models comparable to each other.

Partial least square (PLS) method was used to simultaneously consider the variable matrix Y (sugarcane °Brix) and the variable matrix X (spectral data). In the development of the PLS model, full cross validation (leave-one-out) was used to evaluate the quality and prevent over fitting of the calibration model (Arana *et al*., 2005). In this paper, both PCA and PLS modelling were run using the Unscrambler V 9.6 software.

External validation was used in this study to check the performance of the PLS models. The samples in the external validation set had not been used for the calibration development. Before calibration, samples were divided into two sets; 75% of the samples were used in the calibration model while 25% of the samples were used in the validation model. Samples for validation were selected by taking one of every four samples from the entire sample set, taking care to ensure that each set included samples that covered the entire range of °Brix values.

Results and Discussion

Samples characteristics and spectral overview

A summary of statistical characteristics for calibration and prediction data sets of the internode samples is shown in Table 1. The calibration and prediction data sets show similar means, ranges and standard deviations, indicating that the selection of samples for each data set was appropriate. A relatively wide range of °Brix values was obtained due to the inclusion of three different varieties with different stage of maturity. The range of °Brix values for internode samples from the top to the bottom of the Q155, Q208 and Q190 varieties were 7.6 to 22.2, 8 to 21.4 and 8 to 21, respectively.

Table 1 Summary of statistical characteristics of internode samples

Model	No of sample	Min	Max	Mean	Standard Deviation
Calibration	220	7.5	22.2	17.86	3.04
Prediction	72	8.2	22	17.83	2.93

The typical raw absorbance and reflectance spectra (before any pre-processing method applied) of three internode samples having low (14.2°Brix), medium (18°Brix) and high (22°Brix) values, as measured by the vis/SW-NIR spectroradiometer, are shown in Figure 2(a) and 2(b) respectively. In both figures, no obvious difference could be seen in the shape of the spectra for different °Brix values. However, gaps could be clearly observed among these three spectra in the region of 700 to 1000 nm. These spectral patterns were due to different samples molecular vibration at different sugar concentration.

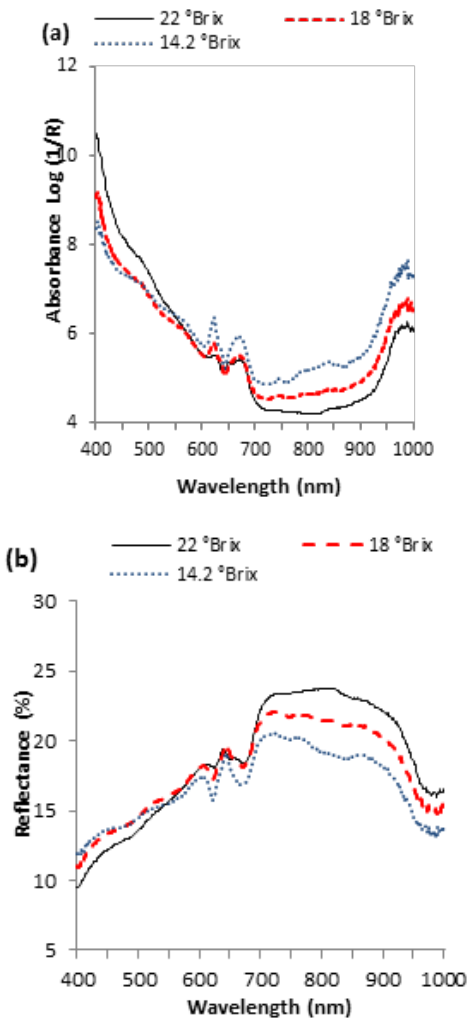

Figure 2. Typical raw spectral curve of sugarcane internode at different °Brix values: (a) Absorbance spectra (b) Reflectance spectra

Effects of different pre-processing methods on prediction accuracy

The influence of different spectra pre-processing methods on PLS model performance for both reflectance and absorbance data collected from internode samples are shown in Table 2. For comparison purposes, the PLS model performance for spectral data without any pre-processing method (raw spectral data) is also shown.

From Table 2, it can be seen that if no pre-processing method was applied, the R^2 values for both reflectance and absorbance data were 0.81 and 0.86 respectively. Absorbance data has indeed performed better than reflectance data. The highest R^2 value was obtained when the data was treated with MSC, where the R^2 values for reflectance and absorbance being 0.85 and 0.87, respectively. The RMSEP values for reflectance and absorbance were 1.54 and 1.45°Brix, respectively. The MSC technique is the most popular normalization technique offered by most chemometrics software packages (Næs *et al*., 2004). MSC compensates for additive (baseline shift) and multiplicative (tilt) effects in the spectral data, which are induced by physical effects such as the non-uniform scattering throughout the spectrum because the degree of scattering is dependent on the wavelength of the radiation, the particle size and the refractive index. In contrast, the prediction models show lower accuracy when treated with SG2, or a combination of these methods because SG2 is normally used to remove slope of the spectral data (Swierenga *et al*., 1999). This observation is in good agreement with that reported by Montalvo *et al*. (1994).

Table 2 The effect of different pre-processing methods on the PLS models performance.

Pre-processing method	Reflectance				Absorbance			
	Calibration		Prediction		Calibration		Prediction	
	R^2	RMSEC	R^2	RMSEP	R^2	RMSEC	R^2	RMSEP
Raw	0.81	1.79	0.81	1.72	0.83	1.70	0.86	1.52
MSC	0.88	1.44	0.85	1.54	0.87	1.49	0.89	1.45
SNV	0.87	1.48	0.84	1.59	0.85	1.60	0.86	1.49
SG1	0.92	1.22	0.80	1.78	0.91	1.24	0.76	1.90
SG2	0.75	2.20	0.39	2.84	0.70	2.15	0.69	2.09
MN	0.86	1.58	0.82	1.68	0.84	1.66	0.86	1.49
MA (3)	0.80	1.87	0.79	1.77	0.81	1.78	0.86	1.53
MA (9)	0.77	1.94	0.77	1.84	0.79	1.86	0.85	1.60
MSC + SNV	0.87	1.48	0.84	1.59	0.85	1.60	0.86	1.49
MA(3) + SG2	0.91	1.24	0.81	1.75	0.73	2.08	0.49	2.60
SG2 + MSC	0.92	1.20	0.78	1.83	0.49	2.65	0.46	2.58
MSC + SNV + SG2	0.93	1.15	0.79	1.80	0.76	1.99	0.39	2.88

* n for calibration model=220; n for prediction model=72.

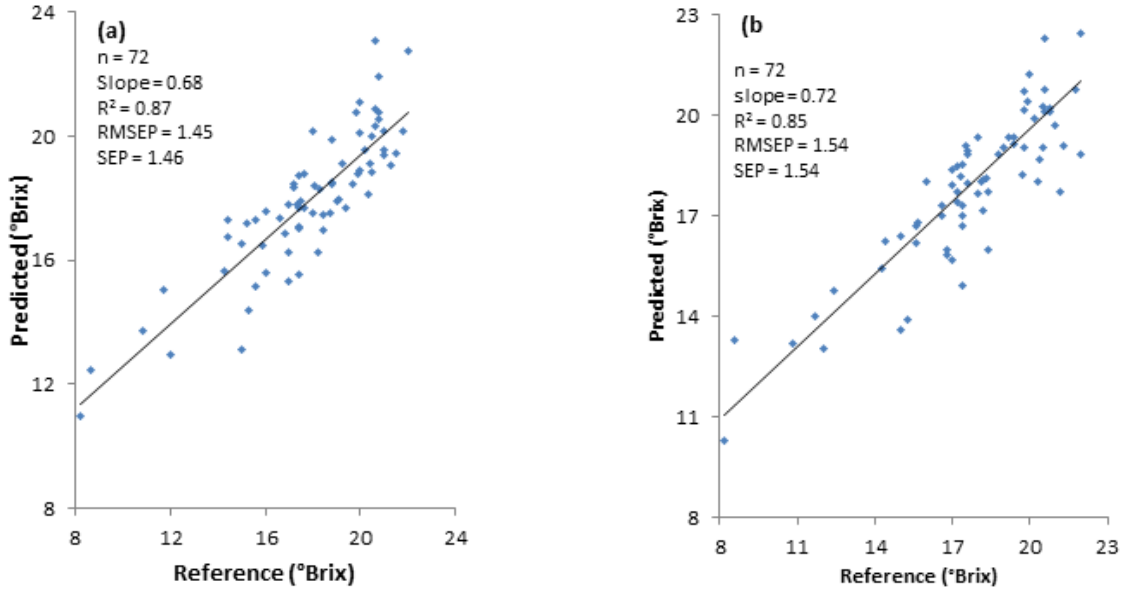

Figure 3. Scatter plots of reference versus predicted °Brix for prediction model: (a) absorbance spectra (b) reflectance spectra

Reflectance data versus absorbance data

The prediction models for both absorbance and reflectance data were presented as the scatter plots as shown in Figure 3(a) and 3(b), respectively. The R^2 for the models developed from absorbance and reflectance data were 0.87 and 0.85, respectively. Absorbance data performed better than reflectance data because the absorbance spectra contains the chemical components information such as sugar, while the reflectance spectra contain the information of the chemical components as well as scattering properties of the tissues (Nicolai *et al*., 2008). Since the spectral measurement was done on the flesh of the internode samples, the scattering problem was minimised. Overall, the prediction accuracy obtained from both reflectance and absorbance data could be considered as good, noting the heterogeneous nature of the stalk samples. The results indicated that the VNIRS and PLS models could provide a satisfactory method for predicting °Brix from stalk samples.

Conclusion

This study has demonstrated that vis/SW-NIR spectroscopy could be applied to predict sugarcane °Brix from internode samples. This study has also shown that the pre-processing could improve the performance of the calibration and prediction models. It was also found that absorbance spectra gave higher prediction accuracy compared to reflectance spectra. For raw spectral data, the R^2 of prediction models for reflectance and absorbance data were 0.81 and 0.86 respectively. However, after the data was treated with MSC method, the R^2 of prediction models for reflectance and absorbance data were improved up to 0.85 and 0.87 respectively. Overall, it can be concluded that right pre-processing method can improve the performance of both calibration and prediction models for spectroscopic analysis.

Acknowledgements

The authors acknowledge the financial support provided by Ministry of Education Malaysia, Universiti Putra Malaysia and National Center for Engineering in Agriculture (NCEA), Toowoomba, Australia. The authors also thank BSES Limited, Bundaberg, for providing samples and equipment.

References

Arana, I., Jarén, C. and Arazuri, S. 2005. Maturity, variety and origin determination in white grapes (Vitisvinifera L.) using NIRS. Journal of Near Infrared Spectroscopy 13: 349–357.

Blanco, M. and Villarroya, I. 2002. NIR spectroscopy: a rapid response analytical tool. Trends in Analytical Chemistry 21(4): 240-250.

Cen, H. and He, Y. 2007. Theory and application of near infrared reflectance spectroscopy in determination of food quality. Trends in Food Science and Technology 18: 72-83.

Griffiths, P. R. 1995. Letter: practical consequences of math pre-treatment of near infrared reflectance data: log (1/R) versus F(R). Journal of Near Infrared Spectroscopy 3: 60-62.

Madsen, L. R., White, B. E. and Rein, P. W. 2003. Evaluation of a near infrared spectrometer for the direct analysis of sugar cane. Journal of American Society of Sugar Cane Technologists 23: 80-92.

Mehrotra, R. and Siesler, H. W. 2003. Application of mid infrared/near infrared spectroscopy in sugar industry. Applied Spectroscopy Reviews 38: 307–354.

Montalvo, Jr. J. G., Buco, S. E. and Ramey, Jr. H. H. 1994. Studies to measure cotton fiber length, strength, micronaire, and color by vis/NIR reflectance spectroscopy', Part II: Principal components regression', Journal of Near Infrared Spectroscopy 2(4): 185-198.

Næs, T., Isaksson, T., Fearn, T. and Davies, T. 2004. A user-friendly guide to multivariate calibration and classification. Charlton, Chichester, UK: NIR Publications.

Nawi, N. M., Chen, G. and Jensen, T. 2013. Visible and shortwave near infrared spectroscopy for predicting sugar content of sugarcane based on a cross-sectional scanning method. Journal of Near Infrared Spectroscopy 21: 289-297.

Nawi, N. M., Chen, G., Jensen, T. and Mehdizadeh, S. A. 2013. Prediction and classification of sugar content of sugarcane based on skin scanning using visible and shortwave near infrared. Biosystems Engineering 115: 154-161.

Nawi, N. M., Jensen, T. and Chen, G. 2012. The application of spectroscopic methods to predict sugarcane quality based on stalk cross-sectional scanning. Journal of American Society of Sugar Cane Technologists 32: 16-27.

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I. and Lammertyn, J. 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. Postharvest Biology and Technology 46: 99-118.

Nicolaï, B. M., Verlinden, B. E., Desmet, M., Saevels, S., Saeys, W. and Theron, K. 2008. Time-resolved and continuous wave NIR reflectance spectroscopy to predict soluble solids content and firmness of pear. Postharvest Biology and Technology 47: 68-74.

Swierenga, H., Weijer, A. D. P., Wijk, R. J. V. and Buydens, L. M. C. 1999. Strategy for constructing robust multivariate calibration models. Chemometrics and Intelligent Laboratory Systems 49: 1-17.

Taira, E., Ueno, M. and Kawamitsu, Y. 2010. Automated quality evaluation system for net and gross sugarcane samples using near infrared spectroscopy. Journal of Near Infrared Spectroscopy 18: 209-215.

Wu, D., Feng, L., Zhang, C. and He, Y. 2008. Early detection of Botrytis Cinerea on eggplant leaves based on visible and near-infrared spectroscopy. Transactions of the American Society of Agricultural and Biological Engineers (ASABE) 51(3): 1133-1139.