PREPROCESSING EFFECTS ON ON-LINE SSC MEASUREMENT OF FUJI APPLE BY NIR SPECTROSCOPY

D. S. Ryu[1], S. H. Noh[1], I. G. Hwang[2]

[1]School of Biological Resources and Materials Engineering
College of Agriculture and Life Sciences, Seoul National University
Suwon, Kyonggi-Do 441-744, Korea
[2]R & D Center, Tong Yang Moolsan Co. Ltd.
Yongin, Kyonggi-Do 440-870, Korea
E-mail: noh@snu.ac.kr

ABSTRACT

The aims of this research were to investigate the preprocessing effect of spectrum data on prediction performance and to develop a robust model to predict SSC in intact apple. Spectrum data of 320 Fuji apples were measured with the on-line transmittance measurement system at the wavelength range of 550~1100nm. Preprocess methods adopted for the tests were Savitzky Golay, MSC, SNV, first derivative and OSC. Several combinations of those methods were applied to the raw spectrum data set to investigate the relative effect of each method on the performance of the calibration model. PLS method was used to regress the preprocessed data set and the SSCs of samples, and the cross-validation was to select the optimal number of PLS factors.

Smoothing and scattering corection were essential in increasing the prediction performance of PLS regression model and the OSC contributed to reduction of the number of PLS factors. The first derivative resulted in unfavorable effect on the prediction performance. MSC and SNV showed similar effect. A robust calibration model could be developed by the preprocessing combination of Savitzky Golay smoothing, MSC and OSC, which resulted in SEP= 0.507, bias=0.032 and R^2=0.8823.

Key Word: VIS/NIR Spectroscopy, Transmittance, Spectrum, Preprocess, PLS Model, Soluble Solid Contents, Fuji Apple

INTRODUCTION

Appearance quality such as color, shape and defects has been the most important factor in consumer's choice of fruits before nondestructive evaluation technology of internal quality such as sugar, acid and firmness was not developed. With the development of new technology, consumer's choice is focused on the internal quality.

Many researchers have studied to develop on-line fruit sorting system by using VIS/NIR spectroscopic method that could evaluate internal quality. In developing such an on-line sorter, two aspects should be considered such as the hardware design composed of light source, detector, interface, etc. and the software involving signal correction, spectrum analysis, calibration, etc. Particularly, in case of transmittance spectrum data

which are measured with the sample at on-line state, not only the signal intensity itself is very weak, but also they contains relatively large noise and variations duo to the difference in physical characteristics of each sample such as size, firmness, texture, color, etc.

In chemometrics, various preprocessing methods and regression models have been introduced for successful analysis of NIR spectrum data set. Examples are moving average, Hanning window, polynomial fitting, wavelet, Fourier transform etc. for smoothing the spectrum data, MSC, SNV etc. for scattering correction, derivatives for extraction of significant spectral information, OSC for reduction of variation, etc. MLR, PCR, PLS, PLS-neural net, etc. are examples of regression model.

The objectives of this study are to investigate the effect of preprocessing methods on the prediction of SSC in intact apple and to develop a robust calibration model with transmittance spectrum data set measured at on-line state.

MATERIALS AND METHODS

Materials and measurement of spectrum data

VIS/NIR transmittance spectra of 320 Fuji apples that harvested at two different locations (Kimchun and Yesan) in 1999 were measured at speed of two apples per second with the on-line sugar grading system developed by Hwang (2000). Spectral range was 550~1050nm and transmission energy spectra were measured at an interval of 1.8nm.

All spectrum data were divided randomly into 2 groups of 200 data set for calibration and 120 for validation. Sample juice was extracted from apple slices and SSC (Soluble Solid Contents) was measured by digital refractometer (DBX-55, ATAGO, Japan). Statistics of sample SSCs are shown in Table 1.

Table 1. Statistics of samples

Data	Internal qualities	Range	Statistics		
			Mean	Std.	Var.
Fuji apple	SSC(%Brix)	8~14.9	12.4	1.39	1.94

Preprocessing of spectrum data

Variations in the spectrum data are caused not only by chemical constituents but also by physical characteristics, measurement conditions (integration and accumulation time), environmental condition (sample temperature), etc. Especially, more significant variations are observed from the transmittance spectrum data than from the reflectance data because of difference in the pathlength of the transmitted ray from sample to sample. S/N ratios are also relatively low duo to the weak transmission energy. These act on the spectrum data as the systematic variation. Preprocessing of the spectrum data set is the first step in developing a robust regression model to predict internal quality.

In this study Savitzky-Golay for smoothing, MSC and SNV for scattering correction,

the first derivatives for extraction of significant spectral information and OSC for reduction of variation were adopted among the several preprocessing methods. PLS regression method was used to correlate the preprocessed data set to the SSCs of intact Fuji apples. All algorithms were coded by Matlab (ver. 5.31, MathWorks, USA) and PLS_Toolbox (ver. 2.0, Eigenvector Research, Inc., USA).

(1) Smoothing and derivative

In Savitzky-Golay algorithm, the polynomial order and smoothing points were fixed at the first order and 5 points (9nm), respectively, based on the preliminary test. Gap size of the fist derivative was also fixed at 5 points (9nm).

(2) MSC (Multiplicative Scattering Correction)

This method assumes that the wavelength dependency of the light scattering is different from that of the constituent absorption. Main idea of this method is to remove the effects of scattering by linearlizing each spectrum to an ideal spectrum of the samples. However, because ideal spectrum does not exist, mean spectrum of training data set is used instead. Mean (Eq. 1) and MSC corrected spectrum (Eq. 3) were computed as follows.

$$\overline{X} = \frac{\sum_{i=1}^{m} X_{ij}}{m} \tag{1}$$

$$X_i = a_i \overline{X} + b_i \tag{2}$$

$$X_{iMSC} = (X_i - b_i)/a_i \tag{3}$$

where X: spectrum matrix with m samples and n wavelength,
m: No. of samples i=1,2,3,⋯, m
n: No. of wavelength j=1,2,3, ,⋯, n .

(3) SNV (Standard Normal Variate)

SNV also reduces the scattering effect like MSC, but the calculation is quite different. Ideal spectrum is not necessary in this method. The scattering is eliminated by normalizing each spectrum with the standard deviation of the spectrum across the entire spectral range.

$$\overline{X}_i = \frac{\sum_{j=1}^{n} X_{ij}}{n} \tag{4}$$

$$X_{iSNV} = \frac{(X_i - \overline{X}_i)}{\sqrt{\sum (X_{ij} - \overline{X}_i)^2 / (n-1)}} \tag{5}$$

(4) OSC (Orthogonal Signal Correction)

OSC was suggested by Wold et. al (1998) to eliminate the variation of spectrum that is not related to sample constituents. From the NIPALS algorithm of PLS, spectrum matrix X can be decomposed into score (T) and loading vector (P) as Eq. (6).

$$X = TP' + E \tag{6}$$

where E is residual matrix.

To orthogonalize the score vector and constituent vector (Y), the orthogonal score vector is calculated by Eq. (7).

$$T^* = (1 - Y(Y'Y)^{-1}Y')T \tag{7}$$

Using this new score vector, the reconstructed spectrum (T^*P') can be obtained. The OSC spectrum unrelated to constituents can be calculated as Eq. (8).

$$X_{OSC} = X - T^* \times P' \tag{8}$$

(5) Combinations of the preprocessing methods

To investigate the effects of preprocessing on the prediction of regression model, various combinations of preprocessing methods were applied to the spectrum data set as shown in Table 2.

Table 2. Combinations of the preprocess methods

Notation	Combinations methods
N	No preprocessing
M	MSC
S	SNV
O	OSC
M1	1st derivative+MSC
S1	1st derivative+SNV
O1	1st derivative+OSC
MS	MSC+SNV
MS1	1st derivative+MSC+SNV
MO	MSC+OSC
MO1	1st derivative+MSC+OSC
SO	SNV+OSC
SO1	1st derivative+SNV+OSC

Modeling

PLS (Partial Least Square) regression method was used to develop the calibration model. Optimal number of score factors of PLS model was determined by cross-validation method (leave-one-out).

The robustness of the calibration model was evaluated on the basis of the standard error of calibration (SEC), standard error of prediction (SEP), the determination coefficient (R^2) between the predicted and measured values, and bias, which were defined by the following.

$$SEC = \sqrt{\frac{\sum_{i=1}^{m_c}(\bar{y}_i - y_i)}{m_c}} \tag{9}$$

$$bias = \frac{\sum_{i=1}^{m_P}(\bar{y}_i - y_i)}{m_P} \tag{10}$$

$$SEP = \sqrt{\frac{\sum_{i=1}^{m_P}(\bar{y}_i - y_i - bias)^2}{m_p - 1}} \tag{11}$$

where mc, mp: number of calibration data set and validation data set,
$\bar{y}, y$: predicted and measured values of constituents.

RESULTS AND DISCUSSION

Change in spectrum patterns by preprocessing

Seven spectrum data measured from samples having different SSCs were selected and treated with different preprocessing methods. According to Beer-Lambert's law, some linearity between the concentration and the spectral data should be observed from the raw spectrum data (Fig. 1(a)) at certain wavelength band at least, but it was not. Fig. 1(b) and (c) present the results preprocessed with MSC and SNV, respectively. From those treatments it was observed that the range of spectral intensity in y-axis was narrowed and the order of graphs were rearranged in favor of improving the correlation between the spectrum intensity and concentration level of SSC in certain wavelength band. In both cases, the preprocessed results were similar to each other. OSC also resulted in changing the order of graphs into the order of concentration level of SSC throughout the whole wavelength range (Fig. 1(d)).

Effect of preprocessing on correlation coefficient spectra

PLS model is based on the assumption that linearity exists between the dependent and independent variables. Therefore, it is important to improve correlation between spectrum values at each wavelength and the constituent values (%Brix) by preprocessing. Fig. 3 shows correlation spectra that were obtained from the spectrum data preprocessed

by various methods. It is noted that MSC is the most effective in improving the correlation. SNV brought about similar effect to MSC except for certain range (880~1050nm) where the transmission intensity is relatively weak. The effect of OSC was mild but appeared over the entire wavelength range.

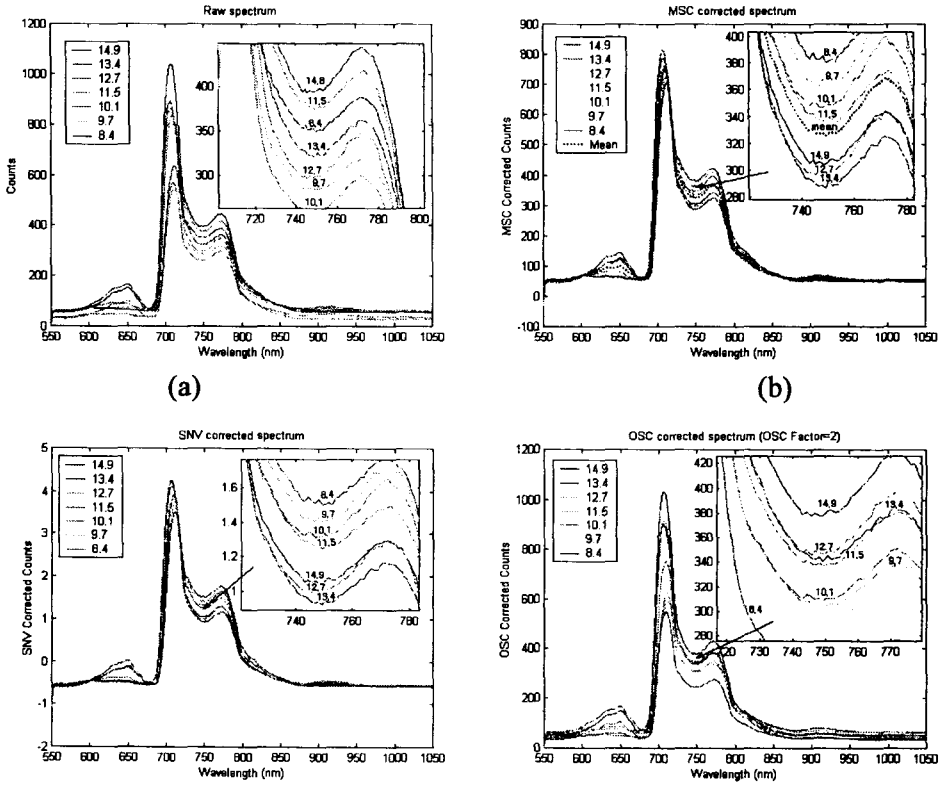


Fig. 1 The raw spectrum(a) and changes in spectrum patterns by MSC(b), SNV(c) and OSC(d).

Effect of preprocessing on robustness of calibration model

Effects of the various preprocessing methods on robustness of calibration models were investigated by comparing SECs, SEPs, R^2s and number of factors. It is known that SEC, SEP and their difference and bias should be as small as possible and R^2s should be as large as possible for a model to be robust. Table 3 and Fig. 3 present the results from various combinations of preprocessing. It is noted that scattering correction by MSC or SNV is quite necessary for improving the robustness of calibration model, and MSC is more effective than the SNV.

Comparisons of the models preprocessed by MSC and MSC+OSC (SNV and SNV+OSC) are indicating that OSC contributes to decreasing the number of PLS factors. The first derivative gave unfavorable effect on the performances of models in general.

However, it is noted that number of PLS factors of the models preprocessed by MSC+OSC+1st derivative or SNV+OSC+1st derivative are less than 3, while SEC, SEP and R^2 are not much sacrificed as compared with those by MSC+OSC or by SNV+OSC. Further examination on the robustness of these models is recommended.

The model preprocessed by MSC+OSC was resulted in the best, indicating SEP=0.507%Brix, R^2=0.8823, bias=-0.0327 and number of factor=11. The OSC was done twice. Fig. 4(a) and (b) shows the relation between the measured and the predicted without any preprocessing and the relation between the measured and the predicted by the best model, respectively.

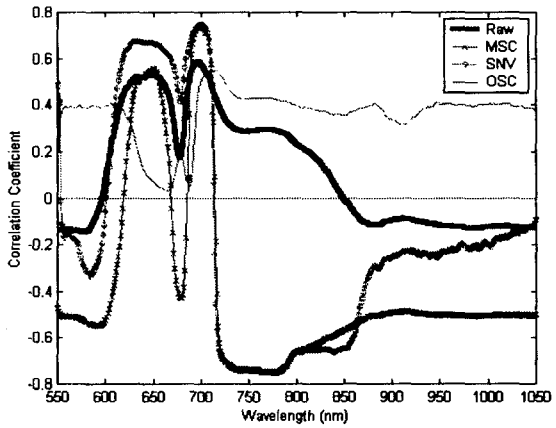


Fig. 2 Effect of MSC, SNV and OSC on correlation coefficient spectra between SSC and transmitted energy

Table 3. Effect of preprocessing on the prediction performance of the SSC model

Preprocessing Combinations	No. of PLS Factor	Calibration		Validation		
		R^2	SEC	R^2	SEP	Bias
N	15	0.7688	0.7832	0.6272	0.9482	0.0985
S	12	0.8933	0.4946	0.8789	0.5145	-0.0229
S1	10	0.9571	0.3195	0.8142	0.6407	-0.0510
M	12	0.9086	0.4690	0.8823	0.5071	-0.0319
M1	7	0.9089	0.4621	0.8194	0.6282	-0.0430
O	10	0.7893	0.7082	0.7063	0.8058	-0.0155
O1	7	0.8219	0.6582	0.4073	1.2649	0.1134
MS	12	0.8983	0.4946	0.8789	0.5145	-0.0229
MS1	10	0.9571	0.3195	0.8142	0.6407	-0.0510
MO	**11**	**0.9086**	**0.4678**	**0.8823**	**0.5071**	**-0.0327**
MO1	2	0.9119	0.4486	0.8200	0.6269	-0.0418
SO	10	0.8963	0.4969	0.8727	0.5271	-0.0454
SO1	3	0.9233	0.4196	0.8236	0.6205	-0.0588

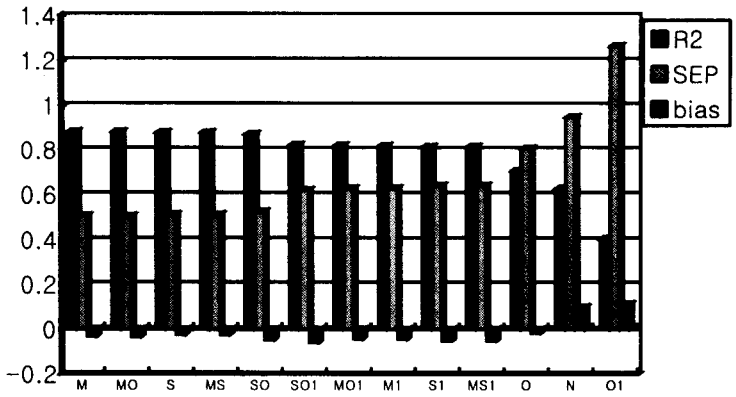
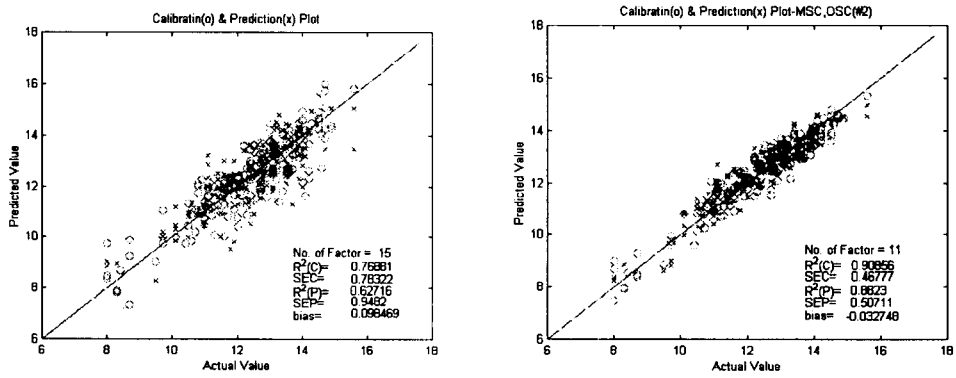


Fig. 3 Comparison of validation results



(a) (b)

Fig. 4 Calibration and validation results by no preprocessing (a) and those by preprocessing of MSC+OSC(b)

CONCLUSION

This study was accomplished to investigate the effect of preprocessing the VIS/NIR transmittance spectrum data on the prediction of SSC of intact apple. Even though many preprocessing methods are reported to reduce the noise and the systematic variation in the transmittance spectrum data set, there is no rule of thumbs in adopting the preprocessing method. Various preprocessing methods were applied to raw spectrum data set and their effects on the performance of calibration model.

The correlation spectra that were obtained from the spectrum data preprocessed by various methods presented that MSC was the most effective in improving the correlation between the SSCs and the spectral data in each wavelength. SNV brought about similar effect to MSC except for certain range (880~1050nm) where the transmission intensity is relatively weak. OSC increased the correlation over the entire wavelength range but the

effect was relatively mild.

Scattering correction by MSC or SNV was quite necessary for improving the robustness of calibration model, and MSC is more effective than the SNV. OSC contributes to decreasing the number of PLS factors. The first derivative gave unfavorable effect on the performances of models in general but number of PLS factors of the models preprocessed by MSC+OSC+1st derivative and SNV+OSC+1st derivative were less than 3, while SEC, SEP and R2 are not much sacrificed. Further examination of the effectiveness of OSC on robustness of calibration model is recommended.

The model preprocessed by MSC+OSC was resulted in the best, indicating SEP=0.507%Brix, R2=0.8823, bias=-0.0327 and number of factor=11. Finally, it is concluded that SSCs of the intact apples could be predicted by the transmittance spectrum data that were measured at the wavelength range of 550~1050 nm on on-line state

REFFERENCES

1. Hwang, I. G., S. H. Noh. 1999. Preliminary study for development of an algorithm for on-lie sugar content of intact fruits using NIR spectroscopy. Abstract book of the 9th International Conference on Near-Infrared Spectroscopy. Towards the 3rd millennium:3-26

2. Hwang, I.G. 2000. Development of on-line apple(Fuji) sorting system by the soluble solid and acid contents using VIS/NIR spectroscopy. Ph.D. thesis of Seoul National University.

3. Wold, S., H. Antti, F. Lindgren, J. Ohman. 1998. Orthogonal signal correction of near infrared spectra. Chemometrics and Intelligence Laboratory System 44:175-185

4. Sjoblom, J., O. Svensson, M. Josefeson. 1998. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. Chemometrics and Intelligence Laboratory System 44:229-244

5. Wise, B. M. 1998. PLS_Toolbox manual. Eigenvector Research, Inc.

6. 1998. Statistics Toolbox User's Guide. MathWorks, Inc.