International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 2, Issue 6 (Nov-Dec 2014), PP. 69-72

NEAR-INFRARED SPECTROSCOPIC MODELING OPTIMIZATION FOR QUANTITATIVE DETERMINATION OF SUGAR BRIX IN SUGARCANE INITIAL-PRESSURE JUICE

Hua-Zhou Chen*, Jiang-Bei Wen, Jie-Chao Chen, Ling-Hui Li, Ya-Juan Huo
College of Science, Guilin University of Technology, Guilin, 541004, China
*huazhouchen@163.com

Abstract— Initial-pressed juice is an important intermediate product in cane sugar industry, and sugar brix is a key indicator for evaluating sugar quality. Real-time evaluation of sugar quality requires determining the content of sugar brix in all steps of the cane sugar process. Near-infrared (NIR) spectroscopy is simple, rapid and non-destructive technologies on the analysis of material contents. In this study, the chemometric algorithm of parameter-combined tuning of Savitzky-Golay (SG) smoother and Partial Least Squares (PLS) regression was utilized for NIR analysis of sugar brix contents in sugarcane initial-pressure juice. The algorithms of combined optimization of SG smoother and PLS regression was achieved and the calibration models were optimally established by screening the expanded 540 SG smoothing modes and the 1-30 latent valuables (LV). The optimized models have high predictive accuracy. These results confirm that the combined optimization of SG smoothing modes and PLS LVs is effective in the quantitative determination of sugar brix contents in sugarcane initial-pressure juice, and that the NIR spectroscopic technology with its chemometric algorithms have the potential in the analysis of cane sugar intermediates.

Key Words— Initial-Pressure Juice, Sugar Brix, Near-Infrared Spectroscopy, Partial Least Squares regression, Savitzky-Golay smoother.

I. INTRODUCTION

Near-infrared (NIR) spectroscopy is a well-performed technology on analysis of both the structure of matter and material contents. It has the advantages of simple, rapid, non-destructive and reagent-free measurement, multi-component simultaneous determination, etc. With the development of modern science and computation, NIR spectroscopic analytical technology is widely applied to many fields, such as agriculture, food, environment, biomedicine [1-5]. In recent years, there are preliminary studies on the application of NIR to cane sugar industry [6-8]. In the progress of cane sugar, initial-pressed juice is an intermediate product of much importance, and sugar brix is an important indicator for evaluating sugar quality. Real-time evaluation of sugar quality requires determining the content of sugar brix in all steps of cane sugar process. Conventional chemical analytical methods in laboratory cannot achieve the fast (or online) determinations and this has been a long-standing problem to be solved in cane sugar industry.

Online fast detection of cane sugar intermediates is expected to be achieved by using NIR spectroscopic technology. NIR spectroscopy with its chemometric methods owns the ability to output the perspective detection results in just a few minutes [9-10]. Common chemometric methods include Classical Linear Regression, Multiple Linear Regression, Principal Component Analysis, Partial Least Squares and etc. Partial Least Squares (PLS) is a common chemometric analytical algorithm integrating principal component analysis and multivariate linear regression. It can effectively eliminate spectral collinearity by creating comprehensive latent valuables. The number of latent valuables (LV) is the key parameter of PLS, to reduce spectral noises and extracting responsive information [11-14]. Model predictive results will be reduced if LV is out of the suitable range, too small or too large. Thus a reasonable LV should be selected by taking it as a tunable parameter.

As the NIR spectroscopy is a reflection of the comprehensive information of all chemical components in the analytes, and, indispensably, including some kinds of noises generated in the detecting process. Therefore it is necessary to study the chemometric methods of data pretreatment to reduce spectral noises [15-16]. Savitzky-Golay (SG) smoother is a famous and widely used method for spectral data pretreatment [17-19]. Its major steps contain smoothing and differential, in which the smoothing mode is quite important for model improvement. There are many smoothing modes, determined by the three parameters of Order of Differential (OD), Degree of Polynomial (DP) and Number of Points (NP), and a specific smoothing mode outputs a corresponding calculation equation with its specific coefficients. Thus it is necessary to select a suitable SG smoothing mode and the best way to find it out is to screen it by tuning the three parameters combined with the optimization of PLS latent valuables.

The aim of this research is to quantitatively determine the sugar brix content in sugarcane initial-pressure juice by using the NIR spectral responses. Savitzky-Golay smoother is employed as the method for data pretreatment and PLS is utilized as the algorithm for establishing calibration models. Model improvement is achieved by the combined optimization of PLS LV and the parameters of SG smoother. For a wide-range optimization, we try to expand the tuning range of the three SG smoothing parameters, and then establish NIR calibration models with SG smoothing pretreatment and PLS regressions. The reasonable smoothing modes and the optimal PLS LV are simultaneously determined according to the model predictive results, in the combined computational algorithm platform. The selected pretreatment and modeling methods are examined by the prediction sample set, to have the potential of NIR modeling enhancement.

II. EXPERIMENT AND METHODS

A. Materials and Instruments

Eighty-six samples of sugarcane initial-pressure juice were collected. The sugar brix content of each sample was measured using the traditional chemical methods and the measured values were used as the modeling reference values for NIR quantitative analysis. The sugar brix content ranges from 19.0 to 22.2 (%Bx).

We detected the NIR absorption spectra of initial-pressure juice samples by using Foss Rapid Liquid grating spectrometer

with a 1-mm pathlength quartz cuvette. The scanning range is set as 800-2500 nm. Because the rotation of the cuvette cell can effectively reduce the unevenness, and multiple scans can effectively reduce the influence of background noise, we designed to measure the spectra while the cuvette cell is rotating. The temperature was controlled at 25±1°C and the relative humidity was at 46±1% RH throughout the spectral scanning process.

B. Partitioning of Calibration and Prediction Samples

NIR spectroscopic analysis requires partitioning the samples into calibration set and prediction set. Calibration samples are used for model establishment and prediction samples for model evaluation. A suitable partition will lead to perspective modeling results. The method of sample partitioning based on joint x-y distances (SPXY) is common used for sample partitioning in the spectroscopic field [20-21]. SPXY process is the performance of Kernard-Stone (KS) algorithm on both the spectral data and the chemical data. The classic KS algorithm is aimed at selecting a representative subset from the sample pool (*N* samples). In order to ensure a uniform distribution of such a subset along the *x* matrix (spectral response), KS follows a stepwise procedure in which new selections are taken in regions of the space far from the samples already selected. For this purpose, the algorithm employs the Euclidean distances $d_x(j, k)$ between the *x*-vectors of each pair (j, k) of samples calculated as

$$d_x(j,k) = \sqrt{\sum_{p=1}^{P}(x_j(p) - x_k(p))^2}, \quad j,k \in [1, N]. \tag{1}$$

For spectral data, $x_j(p)$ and $x_k(p)$ are the instrumental responses at the p-th wavelength for samples j and k, respectively. P denotes the number of wavelengths in the spectra.

The algorithm selects the sample that exhibits the largest minimum distance with respect to any sample already selected. The proposal of the present paper consists of augmenting the distance defined in Eq. (1) with a distance in the dependent variable *y* vector (the contents of the target component in each sample) for the parameter under consideration. Such a distance $d_y(j, k)$ can be calculated for each pair of samples j and k as

$$d_y(j,k) = \sqrt{(y_j - y_k)^2} = |y_j - y_k|, \quad j,k \in [1, N]. \tag{2}$$

In order to assign equal importance to the distribution of the samples in the *x* matrix and *y* vector, distances $d_x(j, k)$ and $d_y(j, k)$ are divided by their maximum values in the data set. In this manner, a normalized *xy* distance is calculated as

$$d_{xy}(j,k) = \frac{d_x(j,k)}{\max\limits_{j,k\in[1,N]} d_x(j,k)} + \frac{d_y(j,k)}{\max\limits_{j,k\in[1,N]} d_y(j,k)}, \quad j,k \in [1, N]. \tag{3}$$

With a stepwise selection procedure similar to the KS algorithm, SPXY can be applied with $d_{xy}(j, k)$ instead of $d_x(j, k)$ alone.

C. The Extended Savitzky-Golay Smoother

SG smoother is a famous and widely-used pretreatment method to eliminate spectral noise. SG smoothing parameters include Order of Differential (OD), Degree of Polynomial (DP) and Number of Points (NP). For convenience, we denoted that the original spectral smoothing is 0th order differential. And NP is usually an odd number, denoted as NP=2m+1. It means that 2m+1 consecutive spectral data as a window, the spectral data in the window were fitted by using polynomial function whose independent variable was the serial number i of the spectral data, (i=0, ±1, ±2… ±m), and the polynomial coefficients were determined. Then the smoothing value and each order derivative value at the center point (i=0) of the window were calculated by using the determined polynomial coefficients. By moving the window in the whole spectral collecting region, the SG smoothed spectra and SG derivative spectra were obtained.

According to the above method, the smoothing value and each order derivative value at the center point of the window can be expressed as a linear combination of the measured spectral data in the window. The coefficients of the linear combination (i.e. smoothing coefficients) were uniquely determined by number of smoothing points (i.e. the number of points in the window), degree of polynomial, and order of derivatives. In Savitzky and Golay's paper [19], it was set that OD=0, 1, 2, 3, 4, 5, DP＝2, 3, 4, 5, and NP=5, 7… 25 (odd numbers). Different combinations of parameters correspond to different smoothing modes, and further correspond to different smoothing coefficient sets. There were a total of 117 smoothing modes (i.e. 117 sets of smoothing coefficients). The appropriate smoothing mode can be selected according to different analytes.

However, for the spectroscopic analysis of cane sugar intermediates (e.g. initial-pressed juice and clarified juice), if the interval between spectral points was very small and number of points was small, then the window was narrow and the information in the window for smoothing was not sufficient. In this case, it was difficult to get satisfying smoothing effect. Hence, it was very necessary to expand the range of NP. In this paper, NP was expanded to 5, 7… 81 (odd), DP was expanded to 2, 3, 4, 5, 6, and the corresponding sets of new smoothing coefficients were calculated, so that a total of 540 smoothing modes were obtained including the original 117 modes, which is a SG smoothing group with a wider application scope.

D. Modeling Indicators

The model evaluation indicators mainly include correlation coefficient of predication (Rp), root mean squared error of predication (RMSEP) and the relative RMSEP (RRMSEP), which are calculated by the Eqs. (4-6),

$$\mathrm{Rp} = \frac{\sum_{i=1}^{M}(C_{ip} - C_{mp})(C'_{ip} - C'_{mp})}{\sqrt{\sum_{i=1}^{M}(C_{ip} - C_{mp})\sum_{i=1}^{M}(C'_{ip} - C'_{mp})}}, \tag{4}$$

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{M}(C'_{ip} - C_{ip})^2}{M-1}}, \tag{5}$$

$$\mathrm{RRMSEP} = \frac{\mathrm{RMSEP}}{C_{mp}} \times 100\%, \tag{6}$$

where C'_{ip} and C_{ip} were predictive value and chemical values of the sample i in the prediction set, C'_{mp} and C_{mp} were the mean predicted value and mean chemical value of all samples in the prediction set, and M was the sample number in the prediction set.

The value of Rp is in coherent with RMSEP, usually that a higher Rp corresponds to a lower RMSEP. And, RRMSEP is always proportional to RMSEP. Thus, we take Rp and RMSEP as the main indicators for model optimization.

III. Results and Discussions

The NIR spectra of 86 sugarcane initial-pressure juice were showed in Figure 1. The spectral responses contain the absorption information of many hydrocarbon groups in the initial-pressure juice samples, such as sucrose, organic acids, amino acids and etc. As is showed in Figure 1, the NIR spectra of initial-pressure juice reflect severe spectral overlap, and the absorption was weak. The absorbance is quite strong around 1460 nm and around 1940 nm because of water molecules. In order to reduce the interference of the water molecules, it is necessary to use SG smoother to deal with data pretreatment in spectral modeling.

Table 1. Mean value and standard derivation of the measured content of sugar brix for calibration/prediction samples

(unit: %Bx)	Mean value	Standard deviation
Calibration	20.80	0.63
Prediction	20.39	0.57

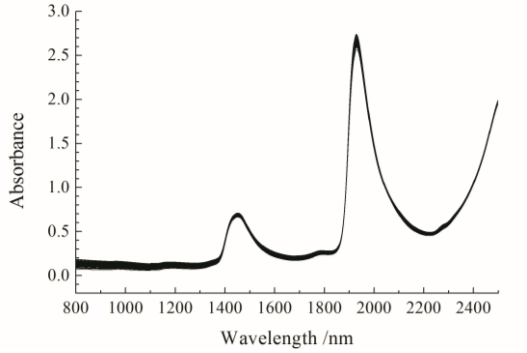

Figure 1. NIR spectra of 86 initial-pressure juice samples

The partitioning of calibration samples and prediction samples has to be finished before model establishment. Using the SPXY method, we have the 86 initial-pressure juice samples divided into 56 samples (for calibration) and the other 30 (for prediction). The mean value and the standard derivation of the measured values of sugar brix content for all calibration/prediction samples were showed in Table 1.

The computational algorithm platform was built up for establishing NIR quantitative analytical models, by combining the 540 kinds of SG smoothing modes and the PLS LV tuning and optimizing, where the PLS LV was set changing from 1 to 30. The optimized model parameters were selected according to the model predictive results.

The optimal models corresponding to each order of differential, with its parameters and predictive results, were showed in Table 2. The non-smoothed full-range PLS modeling results was also listed in Table 2 for comparison. According to the maximum Rp (or minimum RMSEP), the optimal NIR model output the predictive results of Rp=0.954, RMSEP=0.762%Bx and RRMSEP=3.7%, with the best smoothing mode of OD=2, DP=4 (or 5) and NP=55, and with the optimal LV=12.

Table 2. The optimal models corresponding to each order of differential, with its parameters and predictive results

	DP	NP	LV	Rp	RMSEP (%Bx)	RRMSEP
Non-smoothed	—	—	11	0.915	1.193	5.8%
0th order	4, 5	53	9	0.945	0.932	4.6%
1st order	3, 4	77	10	0.949	0.832	4.1%
2nd order	**4, 5**	**55**	**12**	**0.954**	**0.762**	**3.7%**
3rd order	3, 4	77	11	0.953	0.776	3.8%
4th order	2, 3	71	9	0.944	0.917	4.5%
5th order	5, 6	39	11	0.900	1.212	5.9%

It can be concluded from Table 2 that the model predictive results have high accuracy at each smoothing order of differential, regardless for the NIR data of initial-pressure juice when integrating the optimization of PLS models combined with SG smoother. And the modeling results are significantly superior in the PLS models with SG smoothing than without SG smoothing. The global optimal SG smoothing mode was 2nd order of differential and 4th (or 5th) degree of polynomial, and the optimal number of points was obviously larger than 25. These results indicated (1) the samples partitioning method of SPXY lead to well-done calibration models; (2) the DP, NP of SG smoother and the LV of PLS always altered corresponding to the varied OD; (3) the NP of SG smoother is necessary to be expanded to the range of larger than 25; and (4) the spectroscopic analytical chemometric parameters are somewhat similar for sugarcane initial-pressure juice.

To view insight the pretreatment effect of SG smoother, we sketch the figures showing the RMSEP corresponding to each order of differential and each number of points, optimized from different DP of SG smoother and different LV of PLS (see Figure 2). Figure 2 demonstrated that a lower-than-25 NP of SG cannot reach the minimum values of RMSEP and the expansion of NP would output the much optimal results.On another aspect, the influence of LV of PLS on modeling effect was also investigated. Figure 3 showed the RMSEP values corresponding to the varied LV of PLS, optimized by SG smoothing mode, with 2nd order of differential and 4th or 5th degree of polynomial. The figure confirmed that the optimized LVs were larger than 10.

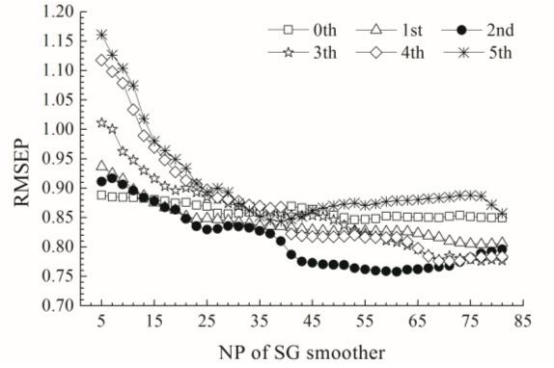

Figure 2. RMSEP corresponding to each order of differential and each number of points for initial-pressure juice (optimized from different DP of SG smoother and different LV of PLS)

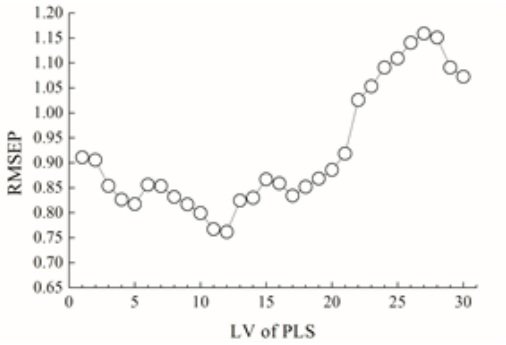

Figure 3. RMSEP values corresponding to the varied LV of PLS for initial-pressure juice (optimized by SG smoothing mode, with 2nd order of differential and 4th or 5th degree of polynomial)

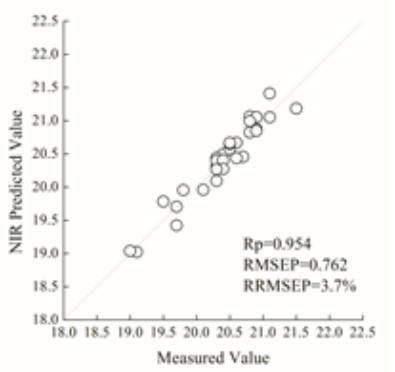

Figure 4. The comparative relationship between the NIR predicted values and the chemical measured values of prediction samples

IV. Conclusions

The chemometric algorithm combined parameter-tuning of SG smoother and PLS regression was utilized for NIR spectroscopic analysis of sugar brix contents in sugarcane initial-pressure juice, to establish and screen for the optimized calibration model. SPXY method was used smoothly and seemed much effective in the partition of calibration and prediction samples. The algorithms of combined optimization of SG smoother and PLS regression was achieved and the calibration models were optimally established by screening the expanded 540 SG smoothing modes and the 1-30 LVs. The combined optimized calibration models have high predictive accuracy, and the optimized modeling results were quite appreciated. These results confirm that the expansion of SG parameters is quite necessary, and the combined optimization of SG smoothing modes and PLS LVs is an important method for the quantitative determination of sugar brix contents in sugarcane initial-pressure juice. Our conclusions demonstrated that the NIR spectroscopic technology with its chemometric algorithms has the potential in the analysis of cane sugar intermediates. This rapid, non-destructive and reagent-free technology has practical meanings and is perspective in the online detection for cane sugar industry.

References

[1] D.A. Burns and E.W. Ciurczak, Handbook of near-infrared analysis, 3rd ed., CSC Press LLC, Boca Raton, 2006.
[2] W.Z. Lu, Modern near infrared spectroscopy analytical technology, 2nd ed., China petrochemical press, Beijing, 2007.
[3] D.I. Givens and E.R. Deaville, "The current and future role of near infrared reflectance spectroscopy in animal nutrition: a review," Aust. J. Agric. Res., 1999, vol. 50, pp. 1131-1145.
[4] A. Alishahi, H. Farahmand, N. Prieto and D. Cozzolino, "Identification of transgenic foods using NIR spectroscopy: a review," Spectrochim Acta A., 2010, vol. 75, pp. 1-7.
[5] A. Saleem, C. Canal, D.A. Hutchins and *et al.*, "Techniques for quantifying chemicals concealed behind clothing using near infrared spectroscopy," Anal. Methods. 2011, vol. 3, pp. 2298-2306.
[6] N. Berding and G.A. Brotherton, "Near Infrared Reflectance Spectroscopy for Analysis of Sugarcane from Clonal Evaluation Trials: I. Fibrated Cane," Crop Science, 1991, vol. 31, pp. l017-l023.
[7] N. Berding and G.A. Brotherton, Near Infrared Reflectance Spectroscopy for Analysis of Sugarcane from Clonal Evaluation Trials: II. Expressed Juice," Crop Science, 1991, vol. 31, pp. l024-l028.
[8] J.H. Meyer, "Near infrared spectroscopy (NIRS) research in the South African sugar industry," International Sugar Journal (Cane Sugar Ed.), 1998, vol. 100, pp. 279-286.
[9] P. Williams and K. Norris, Near-infrared Technology in the Agricultural and Food Industries (2nd ed.). the American Association of Cereal Chemists, Inc. St. Paul, Minnesota, 2001.
[10] X.L. Chu, Y.P. Xu and W.Z. Lu, "Research and Application Progress of Chemometrics Methods in Near Infrared Spectroscopic Analysis," Chinese Journal of Analytical Chemistry, 2008, vol. 36, pp. 702-709.
[11] M. Daszykowski, M.S. Wrobel, H. Czarnik-Matusewicz and B. Walczak, "Near-infrared reflectance spectroscopy and multivariate calibration techniques applied to modelling the crude protein, fibre and fat content in rapeseed meal," Analyst, 2008, vol. 113, pp. 1523-1531.
[12] B. Igne, J.B. Reeves, G. McCarty and *et al*., "Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils," J. Near Infrared Spec. 2010, vol. 18, pp. 167-176.
[13] S. Kasemsumran, Y.P. Du, K. Maruo and *et al*., "Improvement of partial least squares models for in vitro and in vivo glucose quantifications by using near-infrared spectroscopy and searching combination moving window partial least squares," Chemometr. Intell. Lab., 2006, vol. 82, pp. 97-103.
[14] H.Z. Chen, G.Q. Tang, Q.Q. Song and W. Ai, "Combination of Modified Optical Path Length Estimation and Correction and Moving Window Partial Least Squares to Waveband Selection for the Fourier Transform Near-Infrared (FT-NIR) Determination of Pectin in Shaddock Peel," Anal. Lett. 2013, vol. 46, pp. 2060-2074.
[15] 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," Progress in Chemistry, 2004, vol. 16, pp. 528-542.
[16] K. Wang, G.Y. Chi, R. Lau and T. Chen, "Multivariate Calibration of Near Infrared Spectroscopy in the Presence of Light Scattering Effect: A Comparative Study," Anal. Lett. 2011, vol. 4, pp. 824-836.
[17] A. Savitzky and M.J.E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," Analytical Chemistry, 1964, vol. 36, pp. 1627-1637.
[18] A. Rinnan, F. Vandenberg and S.B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," Trac-Trends in Analytical Chemistry, 2009, vol. 28, pp. 1201-1222.
[19] H.Z. Chen, Q.Q. Song, G.Q. Tang, and *et al.*, "The Combined Optimization of Savitzky-Golay Smoothing and Multiplicative Scatter Correction for FT-NIR PLS Models," ISRN Spectroscopy, Volume 2013, Article ID 642190, 2013.
[20] R.W. Kennard and L.A. Stone, "Computer Aided Design of Experiments," Technometrics, 1969, vol. 11, pp. 137-148.
[21] R.K.H. Galvao, M.C.U. Araujo, G.E. Jose and *et al.*, "A method for calibration and validation subset partitioning," Talanta, 2005, vol. 67, pp. 736-740.