Ultrasonic concentration measurement of citrus pectin aqueous solutions using PC and PLS regression

Meng Ruifeng, Zhong Jianjun, Zhang Lifen, Ye Xingqian, Liu Donghong*

(*School of Biosystems Engineering and Food Science*, *Zhejiang University*, *Hangzhou* 310058, *China*)

Abstract: This work demonstrated the use of multivariate statistical techniques called principal component (PC) and partial least squares (PLS) to extract the acoustic features of citrus pectin water solution. The concentration of citrus pectin water solution was predicted by PC and PLS regression method using the spectra of ultrasound pulse echoes travelling through mixtures. The values of root mean square error of validation (RMSEV) were 0.0675 g/100 g and 0.0662 g/100 g for PC and PLS regression model, respectively. Since the response variable was taken into account, PLS regression model was more accurate than PC regression model. Also, a method for temperature compensation was proposed to correct the impact of temperature variation on analyzed data. The proposed methods for pectin concentration measurement are easily adaptable to similar applications using existing hardware.

Keywords: Partial Least Square Regression, Principal Component Regression, concentration measurement, acoustic velocity

DOI: 10.3965/j.ijabe.20120502.010

Citation: Meng R F, Zhong J J, Zhang L F, Ye X Q, Liu D H. Ultrasonic concentration measurement of citrus pectin aqueous solutions using PC and PLS regression. Int J Agric & Biol Eng, 2012; 5(2): 76－81.

1 Introduction

Ultrasonic is an old sensing technique, while it gained increasing attentions in food and agriculture in recent years. The power ultrasound (frequency: 16-100 kHz, intensity: > 1 W/cm^2) have been extensively researched in the field of pectin extraction from fruits (Ultrasound Assisted Extraction)[1-4] as an assistant technology to increase the yield, while low intensity ultrasound (frequency: > 1 MHz, intensity: < 1 W/cm^2) can be used to measure concentration of solution non-invasively, non-destructively and fast.

Parameter of concentration of citrus pectin solution can be used to monitor state and quantities of industrial process. The modern industrial process control requirements including robust, accurate, non-invasive, continuously measuring, safe and low maintenance are well matched when low intensity ultrasound is applied. It is used successfully for non-invasive detection (process control) and for characterizing physicochemical properties of food materials (product assessment or control)[4-9], using the three basic sound parameters (velocity c, attenuation α, and impedance Z). The acoustic velocity c is the prevalent parameter used in practice now by correlating itself with process-charactering parameter, but the amplitude information of ultrasound signal is often ignored (the amplitude information is also related to mechanical properties of the propagation liquid). While the spectral composition, including amplitude and phase information, is related to mechanical properties of the propagation medium at all frequencies. However, acoustic measuring system is a complicated system which results in unmanageable or inaccurate physical models. One solution to this problem is statistical modeling. This means finding a connection among some responses Y that are not directly measurable by studying some directly measurable

Received date: 2012-02-02 **Accepted date:** 2012-06-28

Biographies: Meng Ruifeng, PhD candidate, Email: mrfnmg@yahoo.cn; **Ye Xingqian**, Professor, Email: psu@zju.edu.cn.

***Corresponding author: Liu Donghong,** PhD, Professor, majored in food processing technology and equipment. School of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China. Email: dhliu@zju.edu.cn

descriptor variables matrix X, using Multiple Linear Regression (MLR), such as Principal Components Regression (PCR) and Partial Least Square Regression (PLSR), so that

$$Y = X \cdot A \quad (1)$$

This study demonstrated how to use PCR and PLSR to model the variations in spectra of ultrasonic pulses transmitted through citrus pectin water solutions with a conventional transducer, which designed as it was applied in contemporary process instrumentation. Though the specific frequency changes may not be linear with various concentrations, multivariate analysis of the results can allow analytical quantification[10]. The PCR and PLSR deliver the coefficients A.

2 Materials and methods

2.1 Measurement apparatus

Measurements of citrus pectin concentration were made using the transmission mode ultrasound configuration depicted in Figure 1. A Pulser-Receiver CTS-8077PR (Guangdong Goworld Co. Ltd. China) was used to produce 100 ns width, -25 V square electric pulses at 100 Hz repetition rate. A repetition rate of 100 Hz was used to ensure that any reflections had been completely attenuated by the media before the subsequent impulse was generated. The electric pulses were converted into pressure pulses by a 5 MHz ultrasound transmitter with 6 mm in diameter (transducer part number: TOFD-5MHz-6mm, made by Guangdong Goworld Co. Ltd. China). The wave train resulting from a single impulse reverberated back and forth within a sample cell. Transmitted pulses were received by another 5 MHz transducer opposite to the transmitter, and amplified 10 dB by the Pulser-Receiver. Overlapping bandwidth allows better frequency coverage and an increased sensitivity. A 16-bit Data Acquisition Unit (ADLINK, PCI-9846H/512) sampled the amplified incoming signal at sample rate 40 MHz as voltage signals V (t), and transferred them to a PC via PCI bus for data processing later. The length of each sampled signal was 8000. In order to eliminate the influence of temperature variation, the measuring cell in this work, with 8 mm inner diameter and 3 mm thick glass walls, was placed in a water bath for temperature consistent, such as at (20±0.05)℃.

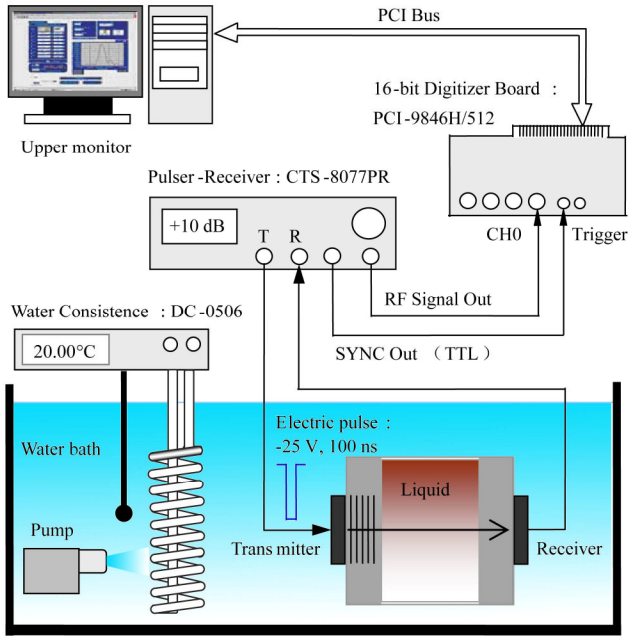


Figure 1 Schematic diagram of an ultrasound system used for concentration measurements of aqueous solution

2.2 Calculation of regression coefficients

The columns of descriptor matrix X are linear-dependent. Hence, ordinary least square regression won't work here. The most common remedy in this case is PCR[11]. The principal components matrix X_S of the descriptor matrix X can be calculated by function $X_S = XW$, where W denotes the weight matrix which is composed of eigenvectors of covariance matrix $C_{XX} = X^T X$. Since $X_S^T X^S = diag(\lambda)$, λ is the corresponding vector of eigenvectors of C_{XX}, the corresponding PCR estimate is:

$$\hat{A}_{PCR} = W diag^{-1}(\lambda) W^T X^T Y \quad (2)$$

The PCR estimate is optimal when it can be guaranteed that the most significant principal components explain the major part of variance occurring in Y. Otherwise the cross-covariance, $C_{YX} = (Y^T X)^T (Y^T X) = X^T Y Y^T X$, should be considered. This is efficiently done by PLS[12].

The PLS iteratively projects the descriptor matrix X onto an orthogonal basis, T (scores matrix) spanned by the PLS components. At each iteration step, one PLS component in terms of a new right-hand side column of T is computed as a product of the most significant

eigenvector of C_{YX} and the error between X and its projection onto T of the previous step. This difference equals X initially. The subsequent cross-covariance matrix is computed using the error of X and its projection onto T of the current step (deflation). Deflation assures the orthogonality of T as well as a decreasing error while iterating. The respective coefficient is $R=W(P^TW)^{-1}$, where the columns of W hold the computed dominant eigenvectors of C_{YX}. Weights R directly relates to X through function $T=XR$. $P=X^TT(T^TT)^{-1}$ yields the transposed coefficients of the projection of X onto T. P is called loadings matrix. The PLS estimate, given by:

$$\hat{A}_{PLS}=R(T^TT)^{-1}R^TX^TY \quad (3)$$

is optimized in a sense that it best explains the cross-covariance of X and Y.

The codes for calibration were programmed and carried out using MATLAB (Version 7 Release 14, The MathWorks, Inc., USA).

2.3 Experiment description

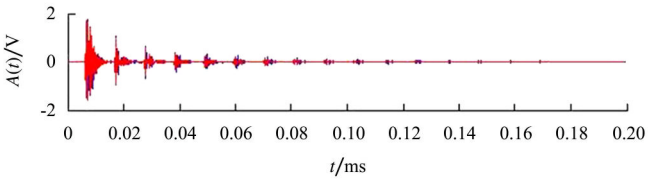
A representative experiment was conducted to estimate the concentration of citrus pectin purchased from Sigma Co. (St. Louis, MO, USA) in water solution. There were 11 measured concentrations covering a range of 0-3% by weight. The intervals between each concentration were 0.3%. Each concentration was measured 5 times. Hence, N=55 ultrasound pulse echo signals were recoded in all. For calibration and validation, the data set were split. The validation data set used 20 ultrasound pulse echoes of concentration 0.3%, 0.9%, 2.1% and 2.7%. Other 35 ultrasound pulse echoes were used as calibration data set.

2.4 Input, output and pre-processing of regression models

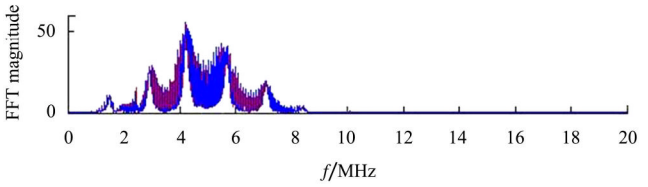
The response variable was the concentrations (by weight) of citrus pectin in calibration data set. The values were stored in a column vector $Y=\{c_n\}_1^{35}$, $Y\in\Re^{35\times1}$. The descriptor variables of the experiments were the spectra of the ultrasound pulses sampled at the receiving transducer at a constant temperature. The n^{th} sampled pulses $P(t)_n$ were transformed using the Fast Fourier Transform (FFT) to examine the frequency content, giving the spectral representation $P(f)_n$ in the terms of magnitude and phase. Once magnitude $P=\{|P(f)_n|\}$ and unwrapped phase $\mathbf{\Phi}=\{\arg(P(f)_n)\}$ were calculated, they would be stored as rows of the descriptor matrix of calibration samples $X=[P\vdots\Phi]$.

Received signals (3% citrus pectin solution and pure water) and their FFT were showed in Figure 2. They are significantly different, especially for spectral information. The signal-to-noise ratio (SNR) is sufficiently high from 1 to 9 MHz. Hence, the information in that bandwidth was selected in descriptor matrix X, $X\in\Re^{35\times3200}$. Every line contains the 1600 FFT amplitude and 1600 FFT phase data for one sample.

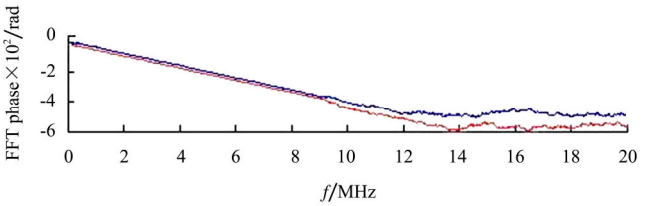
Finally, the columns of X and Y were scaled to unit variance, and subtracting their mean values, which was called data normalization. In this way, no column was given any greater significance than others.



a. Shape of a recorded ultrasound pulse echoes in time domain



b. Magnitude spectrum of echoes



c. Unwrapped phase spectrum

Figure 2 Typical received signal and it's FFT. The blue line represents pure water; and red line 3% citrus pectin solution. The relevant information is located within a frequency range from 1 MHz to 9 MHz.

2.5 Estimation of model size and accuracy

Calculation of regression coefficients (matrix $\hat{A}$) was carried out following Equations (2) and (3). Finding the most reliable model order (number of PC taken for the estimation of the parameter matrix $\hat{A}$) causes the major

problem in terms of accuracy and stability of the calculated regression model. One possible criterion is choosing the model size by the minimum predictive error (for example *RMSE*: root mean square error). The formal description of *RMSE* is shown in Equation (4) determined over the most commonly used cross-validation, where n is the number of samples, y_i is the reference value, and $\hat{y}_i$ is the corresponding estimated value using model.

$$RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ n} \tag{4}$$

This method includes the risk of achieving an over-fit or under-fit model size. A possibility to overcome over- or under-fitted models is to apply calculation of the predictive error by using an external validation test set not taken into account while calibrating. To choose the right test set depends on two main points: the remaining calibration set should still cover the whole region of interest, and the more representative the test set is, the more it resembles the whole data field of interest. In this contribution, the root mean square error of validation over an external validation set is used for model order prediction (*RMSEV*).

Besides *RMSEV*, the ratio performance deviation (*RPD*) is also used for model validation, which is defined as below:

$$RPD = SD / RMSEV \tag{5}$$

Parameter *RPD* is the standard deviation (*SD*) of the reference concentrations of validation samples. If *RPD* is larger than 3, the model is suitable for prediction[13].

3 Results and discussion

Using functions (2) and (3), regression coefficients matrix $\hat{A}_{PCR}$ and $\hat{A}_{PLSR}$ were calculated. Using different components, regression model have different accuracy as shown in Table 1.

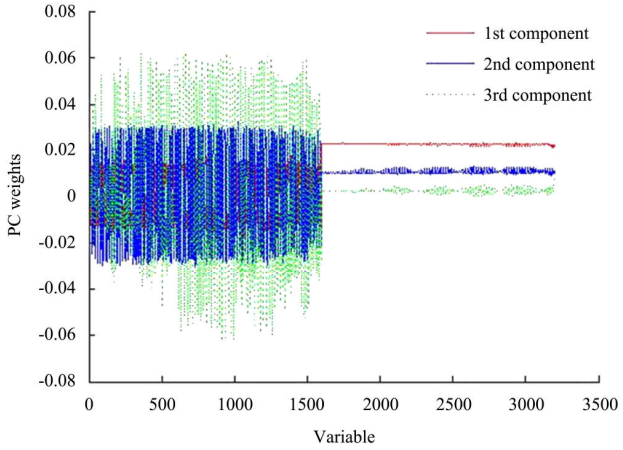
Table 1 Comparison between different types of citrus pectin solution concentration models developed by PC and PLS regression

PCs	PCR				PLSR			
	R_{cal}^2	R_{val}^2	*RPD*	*RMSEV*/%	R_{cal}^2	R_{val}^2	*RPD*	*RMSEV*/%
1	0.2816	0.1364	1.076	0.8752	0.9598	0.9316	3.821	0.2464
2	0.9994	0.9949	13.93	0.0676	0.9994	0.9951	14.22	0.0662
3	0.9994	0.9949	13.94	0.0675	0.9999	0.9949	13.94	0.0675
4	0.9998	0.9948	13.86	0.0679	0.9999	0.9948	13.80	0.0682
5	0.9999	0.9946	13.63	0.0691	1.0000	0.9947	13.75	0.0685

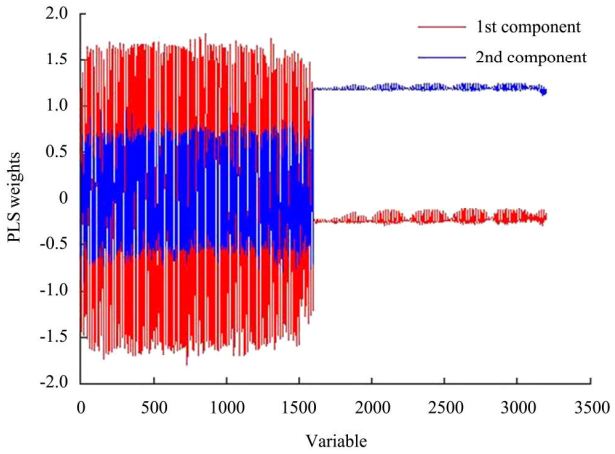
It was shown that the parameter *RMSEV* (summarized in Table 1) for the PLS model was almost the same small with that for the PCR model. The best PLSR model reached the smallest *RMSEV* just using two PLS components, and the best PCR model just using three principal components. The smallest RMSEV was 0.0662% for PLSR model and 0.0675% for PCR model, and the parameter *RPD* was larger than 3. So the best developed PCR and PLSR models were suitable for prediction, and these models were more accurate than sound velocity model (results were not shown here). While, including more components in regression model could introduce errors both for PLSR and PCR, as the latter components contain noise. Results showed that the ultrasonic sensor array used in this study was able to detect citrus pectin content in aqueous solutions accurately.

As described is section 2.2, principal components and PLS components are the products of the descriptor matrix $X \in \Re^{35\times3200}$ and weights matrix. Figure 3 showed the weights of components. Based on the structure of descriptor matrix $X \in \Re^{35\times3200}$, the frontal 1600 predictor variables were the magnitude information of echo signal, while latter 1600 predictor variables were the phase information. The principal components and PLS components reasonably included phase (determined by acoustic velocity) and magnitude (determined by acoustic attenuation and impedance) information of spectral of ultrasound signal simultaneously. The PC and PLS components retrieved information of wave shape interrelating with concentration, and only relative changes of ultrasound signal due to changes in citrus

pectin concentration were taken into account.



a. Principal components



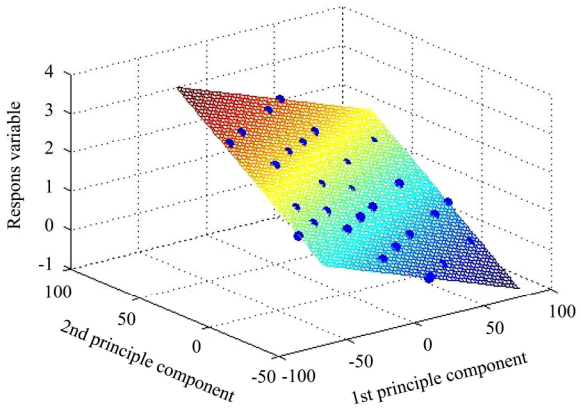
b. PLS components

Figure 3 Plots of the weights of the 3200 variable in each

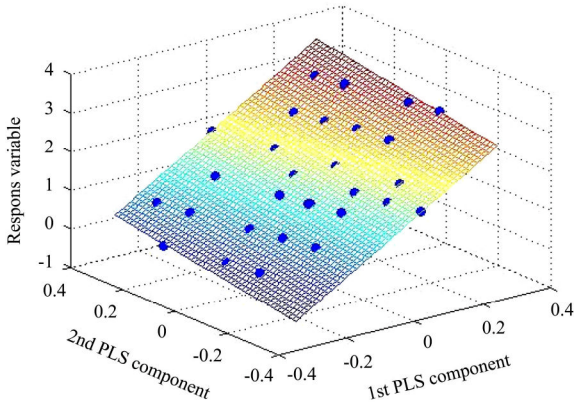
Since both attenuation and speed of sound are highly temperature dependent, it is important to maintain the temperature constant during measurements. Another way to avoid misinterpretations of changes in pulse spectral information from temperature dependent attenuation and sound velocity is to include the measured temperatures as a column of the descriptor matrix X.

Linear relationships between response variable $Y=\{c_n\}_1^{35}$ and principal components and PLS components were clearly illustrated in Figure 4. Calibration dada set included 35 sample points, which were shown as points in Figure 4. Some points are under the regression plane so that they are invisible.

Figure 5 showed the PC and PLS regression results, using 3 and 2 components respectively. The estimated concentrations of 20 validation samples were very close to the reference values both for PCR and PLSR. All dada points were near to a 45° line.

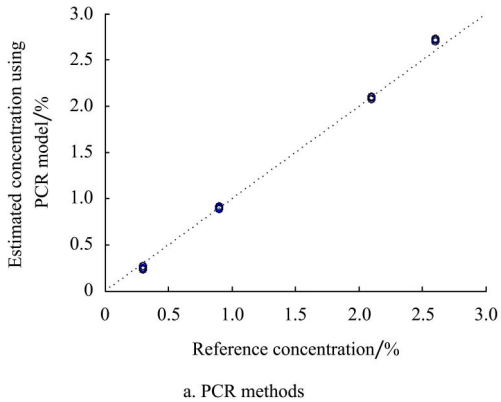


a

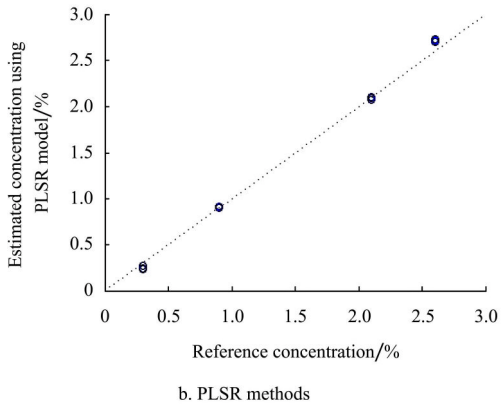


b

Figure 4 Linear relationships between components and response variable $Y=\{c_n\}_1^{35}$



a. PCR methods



b. PLSR methods

Figure 5 Estimated concentrations and reference values using the best PCR and PLSR methods

The above results of this work showed the possibility to calculate the citrus pectin concentrations in water solutions by only using ultrasonic sensors accurately, which made this quite noteworthy. The presented system showed big advantages in service and maintenance due to completely contactless investigations of the solution of interest. Compared with other possibilities for offline analysis, the present method is advantageous by means of sample preparation.

4 Conclusions

This paper demonstrated how PC and PLS regression can be used to accurately estimate the concentration of citrus pectin in an aqueous solution using frequency spectra of ultrasound pulses. The PC and PLS regression model can retrieve information of wave shape interrelating with concentration. This prevents estimation of the absolute attenuation, impedance and velocity of measured medium, and only relative changes, because changes in citrus pectin concentration are observed. Since a single vector multiplication is required to estimate the concentration from the descriptors, this ultrasonic method is an on-line, rapid, non-contract and accuracy technology.

Acknowledgments

This work is supported by the National Scientific and Technological supporting Program (2008BAD91B00), NSFC (30972282) and the National High Technology Research and Development Program ("863" Program) (2007AA091802), in China.

[References]

[1] Yan Y L, Yu C H, Chen J, Li X X, Wang W, Li S Q. Ultrasonic-assisted extraction optimized by response surface methodology, chemical composition and antioxidant activity of polysaccharides from Tremella mesenterica. Carbohydrate Polymers, 2011; 83(1): 217-224.

[2] Hou X J, Chen W. Optimization of extraction process of crude polysaccharides from wild edible BaChu mushroom by response surface methodology. Carbohydrate Polymers, 2008; 72 (1): 67-74.

[3] Vilkhu K, Mawson R, Simons L, Bates, D. Applications and opportunities for ultrasound assisted extraction in the food industry-A review. Innovative Food Science and Emerging Technologies, 2008; 9(2): 161-169.

[4] Xia T, Shi S Q, Wan X C. Impact of ultrasonic-assisted extraction on the chemical and sensory quality of tea infusion. Journal of Food Engineering, 2006; 74(4): 557–560.

[5] Hauptmann P, Hoppe N, Püttmer A. Application of ultrasonic sensors in the process industry. Measurement Science and Technology, 2002; 13(8): R73-R83.

[6] Resa P, Elvira L, Espinosa F M. Concentration control in alcoholic fermentation process from ultrasonic velocity measurements. Food Research International, 2004; 37(6): 587-594.

[7] Schäfer R, Carlson J E, Hauptmann P. Ultrasonic concentration measurement of aqueous solutions using PLS regression. Ultrasonics, 2006; 44: e947-e950.

[8] Kaatze U, Eggers F, Lautscham K. Ultrasonic velocity measurements in liquids with high resolution-techniques, selected applications and perspectives. Measurement Science and Technology, 2008; 19(6): 1-21.

[9] Krause D, Schöck D, Hussein M A, Becker T. Ultrasonic characterization of aqueous solutions with varying sugar and ethanol content using multivariate regression methods. Journal of Chemometrics, 2011; 25(4): 216-223.

[10] Schäfer R, Hauptmann P. Statistical modelling of utrasonic sensors in process industries-new prospects for conventional devices. Measurement Science and Technology, 2007; 18(5): 1627-1636

[11] Joliffe I T. Principle Component Analysis. 2nd ed. New York: Springer. 2002; pp. 167-198.

[12] Wold S, Sjöström M, Eriksson L. PLS regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 2001; 58(2): 109-130.

[13] Williams P C. Implementation of Near-Infrared Technology. In: Williams P and Norris K, Editors, Near-Infrared Technology in the Agricultural and Food Industries. 2nd ed., St. Paul, MN. American Association of Cereal Chemists, 2001; pp. 145-169.