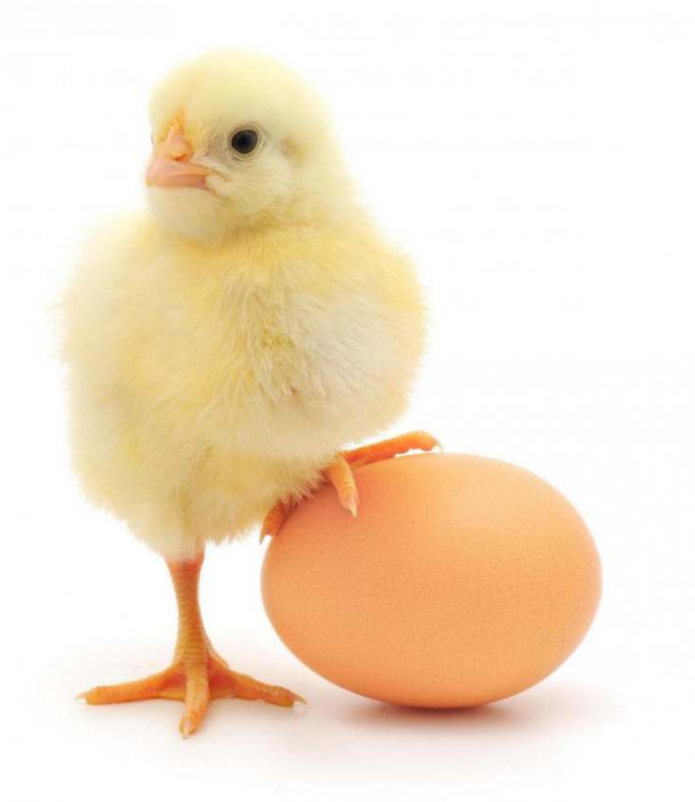
ISSN 1682-8356
ansinet.org/ijps



INTERNATIONAL JOURNAL OF POULTRY SCIENCE

ANSInet
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan
Mob: +92 300 3008585, Fax: +92 41 8815544
E-mail: editorijps@gmail.com

International Journal of Poultry Science 8 (2): 170-178, 2009
ISSN 1682-8356

Combined Maximum R^2 and Partial Least Squares Method for Wavelengths Selection and Analysis of Spectroscopic Data

N. Abdel-Nour, M. Ngadi, S. Prasher and Y. Karimi
Department of Bioresource Engineering, McGill University, Macdonald Campus, 21111 Lakeshore Road,
Ste-Anne-de-Bellevue, Quebec, H9X 3V9, Canada

Abstract: The selection of wavelengths in multivariate analysis is of utmost importance in order to build a strong and robust predictive model. The aim of this research was to investigate the feasibility of an automated selection of sets of relevant wavelengths in Visible/Near Infra-Red (VIS/NIR) spectroscopy by combining Maximum R^2 (MAXR) method with Partial Least Squares (PLS) regression (MAXR-PLS) to build a PLS predictive model. The data used to test this method was derived from the determination of albumen pH and Haugh Unit (HU) as tools for testing the egg quality. For this purpose, 360 eggs were stored during 16 days under a temperature of 18°C and a relative humidity of 55%. For each egg, the VIS/NIR transmission spectra and the two most widely used methods for the assessment of egg quality namely the HU and the albumen pH were performed. A PLS model was built using the full spectra and compared with the models built by selected wavelengths using MAXR-PLS method. Using the mentioned method, the correlation coefficients between the measured and predicted values were up to 95% and the Root Mean Square Error for Cross-validation (RMSECV) were 0.05 and 5.05 for pH and HU, respectively. In addition, this method reduces the complexity of the models by reducing the Latent Variables (LV). Despite the complexity of the spectral data, the Maximum R^2 method leads to a robust predictive model that uses the informative wavelengths.

Key words: Spectroscopy, maximum R^2, wavelengths selection, egg quality, partial least squares

INTRODUCTION

VIS/NIR spectroscopy is a powerful, fast and non-destructive technique that is increasingly used for measuring a large number of chemical and physical properties of agricultural products. Successful applications have been reported for the determination of soluble solids in cherry and apricot (Carlini *et al.*, 2000), dry matter in onions (Birth *et al.*, 1985) and egg freshness (Kemps *et al.*, 2006). However extracting the useful information from spectral data is the fundamental challenge. Therefore, there is a need for automated techniques to extract relevant information. The Partial Least Squares (PLS) method is often used in spectroscopy to analyze spectral data containing overlapping absorption peaks, interference effects from diffuse light scatter and noise from the hardware used to collect the data (Osborne *et al.*, 1997). The choice of wavelengths should be adequate to build a strong model using PLS having good predictive ability and excluding those wavelengths that are irrelevant for the model.

The choice of the wavelengths is required to establish a calibration model with minimum errors in prediction. The benefits gained from wavelength selection are the stability of the model to the collinearity in multivariate spectra as well as the interpretability of the relationship between the sample composition and the model (Jiang *et al.*, 2002). Bangalore *et al.* (1996) studied the feasibility of coupling genetic algorithm method for the selection of wavelengths and partial least squares regression for analysing spectral data. They found that the results obtained after selection were better than those obtained with no spectral range selection. Du *et al.* (2004) used the changeable size window partial least squares and searching combination moving window partial least squares and they found that the combination of these two methods improved the prediction ability of the PLS model. Todeschini *et al.* (1999) proposed the use of Kohonen Artificial Neural Networks (K-ANN) for selecting a set of wavelengths. The results have shown that the predicting ability of the PLS model was improved. Perez-Mendoza *et al.* (2003) used PLS beta coefficient for the selection of important wavelengths to build a PLS model for the classification of flours with and without insect fragment. Ventura *et al.* (1998) used the Multiple Linear Regression (MLR) procedure to select the best wavelengths for determination of soluble solids in apple.

The aim of this research was to develop a method with the capacity to select the informative and relevant wavelengths. With this method we attempted to combine the maximum R^2 (MAXR) method with PLS regression (MAXR-PLS). The specific objectives of this study were to select sets of wavelengths and eliminate the uninformative wavelengths, improve the predictive ability of the PLS model by using the optimal set of wavelengths and decrease the complexity of the model by decreasing the Latent Variables (LV) used in the model.

Theory and algorithm

Partial Least Squares (PLS): PLS is a quantitative spectral decomposition technique that is advantageous as it performs the decomposition on both spectral and concentration data. Two sets are generated when the calibration spectral data are processed using PLS method. These are a set of spectral loadings corresponding to the common variation in spectral data and a set of spectral weights corresponding to the changes in spectra due to the differences in concentration. The method assigns a set of scores for spectral data and for concentration data. As a result, the spectra containing higher concentration of the components are weighed more than those with low concentration (Thermo Galactic, 2003).

Having 2 data sets (blocks of variables): N-dimensional and M-dimensional space of variables, the PLS models the relations between these two blocks of variables. After observing n data samples from each block of variables, PLS decomposes the (n x N) matrix of zero-mean variables X and the (n x M) matrix of zero-mean variables Y into the form in Equations (1) and (2):

$$X = TP^t + E \qquad (1)$$

$$Y = UQ^t + F \qquad (2)$$

Where T and U are (n x p) matrices of the p extracted latent variables, P is the (N x p) matrix, Q is the (M x p) matrix. Q and P represent the matrices of loadings. E is the (n x N) matrix and F is the (n x M) matrix where E and F represent the matrices of residuals (Rosipal and Krämer, 2006).

Validation of the PLS calibrated model is performed by leave-one-out cross-validation technique where the same spectra used as training set are predicted back against the same model. This means that with m numbers of calibration samples, the model is built with m-1 samples and the mth sample is predicted as unknown sample. The selection of Latent Variables (LV) number when building the model is of paramount importance: more variables selected cause an overfitting, while less variables causes underfitting. From the statistical point of view, the number of samples must be equal or more than five times the number of LVs. The Prediction Residual Error Sum of Squares (PRESS) is a commonly used method for the selection of the LV's number. It is plotted against the number of factors. The plot falls down to a minimum corresponding to the best number of LVs on PLS calibration model. It is calculated using the Equation (3):

$$\mathrm{PRESS}(r) = \sum_{i=1}^{n} (\hat{y}_i - y_i) \qquad (3)$$

Where $\hat{y}_i$ is the prediction of the concentration of interest in calibration sample, y_i is the measured value in calibration sample and PRESS (r) is the PRESS value obtained with r factors.

Multiplicative Scatter Correction (MSC): In order to remove background and noise from NIR spectra, MSC has been proposed as a pre-treatment technique before the PLS calculation. The MSC technique is a transformation that allows the removal of amplification and offset effect from the spectra. In principle, this estimation should be applied only on the part of the spectrum that is influenced by light scattering. In practice the whole spectrum is sometimes used. For the MSC an ideal or reference spectrum is required. As an estimate to that ideal, the average of the calibration set can be used. Each spectrum is regressed in the set-mean spectrum; the effect of scatter is responsible for variations along a straight line whereas deviations from the line correspond to the absorption by the sample components (Blanco *et al.*, 1997). The main multiple linear regression is cited in Equation (4):

$$X_{ik} = a_i + b_i r_k + e_{ik} \qquad (4)$$

Where a_i represents the "common shift" and is related to proportional additive effect, b_i represents the "common amplification" and is related to multiplicative effect, e_{ik} represents the residuals and is related to the chemical information.

The corrected spectrum is calculated by using Equation (5):

$$X_{ij}(\mathrm{MSC}) = \frac{X_{ij} - a}{b} \qquad j = 1, 2, \ldots\ldots.. p \qquad (5)$$

Maximum R^2 (MAXR): MAXR technique is used in order to choose sets of wavelengths which contain information about the variables. Goel (2003) used the MAXR criterion with PROC REG procedure of SAS software to choose the best model for estimations of various crop's biophysical parameters. The MAXR technique finds the best one-variable, best two-variable and so on which produces the model with highest R^2 value. It begins by finding the best one-variable model. A second variable is added to the model in a manner to increase the R^2 values and to build the two-variable model. The best two-variable model is chosen by removing one variable in the model and replacing it one by one, with each of the other variables until it produces the best R^2. In the same way, it builds the best three-variable and four-variable model and so on. This process is called "compare and switch".

Root mean square error based on cross-validation: The Root Mean Square Prediction Error Based on Cross-validation (RMSECV) is an estimate of the standard error in prediction and is calculated by using Equation (6):

$$\mathrm{RMSECV} = [1/N\sum_{1}^{N}(\hat{y}_i - y_i)^2]^{1/2} \qquad (6)$$

The RMSECV is used, in addition to R^2, to compare the predictive models of the different sets of wavelengths. The RMSECV indicates the mean difference between the measured and predicted values (Kobayshi and Salam, 2000). The predictive ability of the model will be better with smaller RMSECV values: the smaller the RMSECV value for a model the stronger the prediction ability of the model will be (Todeschini *et al.*, 1999).

MATERIALS AND METHODS

Visible/Near-infrared spectroscopic data: VIS/NIR transmission data of 360 intact white-shell leghorn eggs were obtained using a spectroradiometer (FieldSpec® Pro, Analytical Spectral Devices, Boulder, CO, USA) in 2151 wavebands. The spectroradiometer measures transmittance at wavelengths from 350-2500 nm with 1 nm increment. Four halogen lamps 500 watts each were used as a light source. A white chamber was placed between the light and the samples to avoid heating up the samples and to uniformly distribute the light. A pure white standard was used for calibration. Each egg was scanned 3 times and averaged. After the transmission spectra were determined, the eggs were weighed (±0.01mg) and broken. The albumen height was measured using a mounted digital Vernier Caliper (Marathon Watch Company Ltd., On, Ca). These measurements allowed determination of the Haugh Unit (Haugh, 1937) using Equation (7):

$$\mathrm{HU} = 100 \log_{10} (h-1.7w^{0.37}+7.6) \qquad (7)$$

Where HU is the haugh unit, h is the observed height of albumen in millimetres and *w* is the weight of the egg in grams.

After separating the albumen from the yolk, the albumen pH was determined using pH-meter (Accumet Basic AB15 pHmeter, Fisher Scientific, USA).

Method description: The flowchart that described the method used for selecting the wavelengths is shown in Fig. 1. In the beginning, full MSC is applied to the whole spectra and then the MAXR technique was used to choose the appropriate sets of wavelengths. After that, a PLS model was built using the different sets of wavelengths. Finally the models were tested to determine their robustness and good predictive ability.

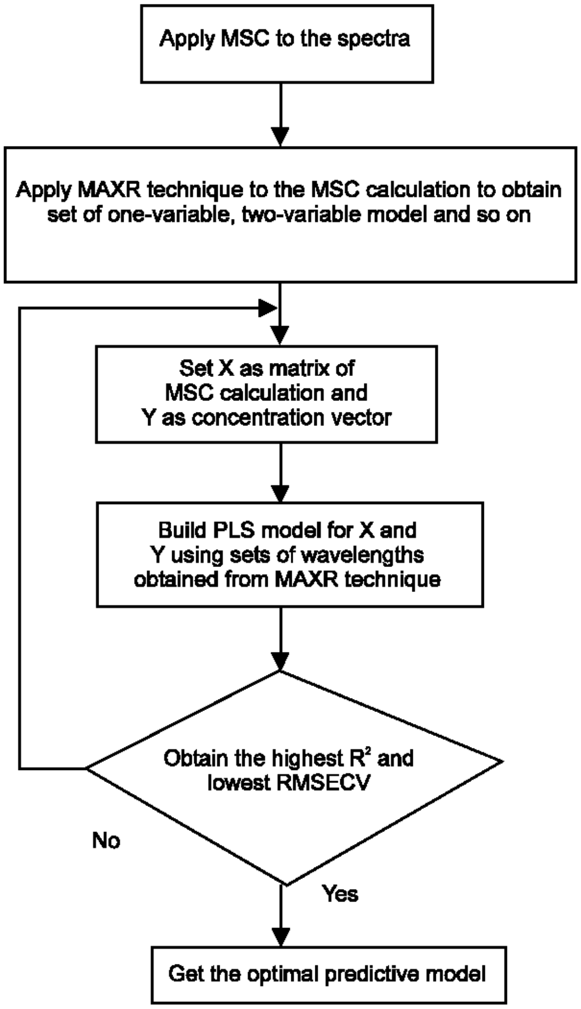

Apply MSC to the spectra
Apply MAXR technique to the MSC calculation to obtain set of one-variable, two-variable model and so on
Set X as matrix of MSC calculation and Y as concentration vector
Build PLS model for X and Y using sets of wavelengths obtained from MAXR technique
Obtain the highest R² and lowest RMSECV
No
Yes
Get the optimal predictive model

Fig. 1: Flowchart for execution of wavelengths selection

Data processing: VIS/NIR Spectral data were analysed using SAS® (version 9.1, 2003) statistical software package and The Unscrambler® (version 9.7, 2007).

Firstly, full MSC was performed using The Unscrambler program. MAXR technique in SAS program was used to identify the most informative wavelengths. After choosing the sets of relevant wavelengths, the VIS/NIR spectral data were linked to HU and albumen pH by a Partial Least Squares, type1 (PLS1) regression model. Multivariate analysis of the samples was performed in The Unscrambler program. The models obtained were validated by using the leave-one-out cross-validation method. In the leave-one-out method, all data are used to build the model except the one used to test this model. The slope and the intercept of regression lines were calculated and compared with their ideal values of 1 and 0, respectively using a statistical program called IRENE® (version Beta 1.00, 2003) (Fila *et al.*, 2003).

RESULTS AND DISCUSSION

The transmittance measurements from 350-400 as well as from 1751-2500 nm were not included in the analysis because of extreme noise. Sets of wavelengths containing less than 8 wavelengths gave low coefficient correlations and R^2 as well as higher RMSECV. Therefore, the PLS models was judged not to have a strong predictive ability. Sets of wavelengths containing more than 10 wavelengths did not improve the PLS model. Thus, following results focused on the sets containing 8, 9, 10 and full spectra 1350 wavelengths.

Figures 2 and 3 show the plot of predicted versus measured for albumen pH and HU, respectively with different number of wavelengths chosen using MAXR-PLS method and the full spectra. The parameters *r*, *b* and *a* are the correlation coefficient, slope and intercept, respectively obtained by least-squares regression. Figure 2 shows that there are 2 groups of scatter due to the difference of temperature between day 0 and the other days. It also shows that the best result obtained for *r*, *a* and *b* was by using 10 wavelengths in building the PLS model. The results obtained with 10 wavelengths and full spectra for building the model are the same in terms of the coefficient correlation and the slope but using the relevant wavelengths improves the intercept. As a result, the PLS could be built with 10 wavelengths instead of the full spectra with improved robustness.

As a comparison between the sets of wavelengths and the full spectra, Fig. 3 shows that by using MAXR-PLS method, the 3 parameters *r*, *a* and *b* have numerically improved. On the other hand, it can be seen that the best correlation coefficient is obtained by using 10 wavelengths whereas the best intercept is obtained by using 9 wavelengths. Addition of wavelengths to the model constituted with 10 wavelengths for the prediction of albumen pH and HU did not improve these three parameters. The reason for this is that the presence of uninformative wavelengths decreases the predictive ability. As a result, the MAXR-PLS method improved the PLS model built with selected wavelengths for the prediction of HU and decreased the complexity of the model for prediction of albumen pH.

For both the HU and the albumen pH for all the sets of wavelengths used and after tested in IRENE program using Least Squares method, the results showed that there was no statistical significance between the value of slope and intercept obtained from the predictive model and their ideal (0 and 1 for the intercept and slope, respectively).

The number of wavelengths used in MAXR, PLS and MAXR-PLS methods is shown in Tables 1 and 2. Every set of wavelength chosen used MAXR-PLS method is used to build PLS models with the calibration set and then the validation set to test the performance of these models. The calculated Root Mean Squared Error of Cross Validation (RMSECV) and the R^2 are also listed.

Table 1: Prediction results for albumen pH obtained with MAXR, PLS and MAXR-PLS methods

Method	Number of wavelength	Latent Variable number	R^2	RMSECV
MAXR	8	-	0.84	0.07
	9	-	0.88	0.06
	10	-	0.89	0.06
PLS	All spectra (1350)	7	0.90	0.06
MAXR	8	6	0.84	0.08
-PLS	9	6	0.89	0.07
	10	7	0.90	0.06

Table 2: Prediction results for HU obtained with MAXR, PLS and MAXR-PLS methods

Method	Number of wavelength	Latent Variable number	R^2	RMSECV
MAXR	8	-	0.78	5.17
	9	-	0.78	5.10
	10	-	0.79	5.06
PLS	All spectra (1350)	6	0.74	5.51
MAXR	8	6	0.78	5.14
-PLS	9	6	0.78	5.11
	10	7	0.79	5.05

From these 2 Tables, it can be observed that when the number of wavelengths increases from 8-10, the R^2 increases whereas the RMSECV decreases. The number of LVs in PLS and MAXR-PLS methods increases with an increasing the number of wavelengths. These results are similar to those obtained by Todeschini *et al.* (1999) who found that when the number of wavelengths decreased, the dimension of the model decreased but the predictive ability was still high. The minimum value of RMSECV and the maximum value of R^2 for the albumen pH and HU are obtained using 10 wavelengths with 7 and 6 LVs, respectively in the MAXR-PLS methods. These results are in contradiction to Kemps *et al.* (2006) who found that the relevant information concerning egg freshness (albumen pH and HU) are detected in 6 wavelengths. Table 1 shows that the model used to predict the pH built from PLS regression had better R^2 and RMSECV than the model built using MAXR method, Whereas, Table 2 shows the opposite for HU. For the reasons cited before, the multivariate analysis is used often in spectroscopy due to its ability to overcome the collinearity problems. Combining PLS and MAXR methods could be useful. Therefore, the results obtained from these 2 Tables show that the PLS model can be built with selective wavelengths using MAXR method and not necessarily with all the spectra. In addition, the robustness and the predictive capability of the model can be improved with selective wavelengths. Table 2 shows that constructing a predictive model for HU with all spectra required 6 LVs and the RMSECV value was 5.51 whereas using the MAXR-PLS method required 7 LVs and the RMSECV value was 5.05. Therefore, the inclusion of uninformative wavelengths can lead to collinearity and the PLS model becomes

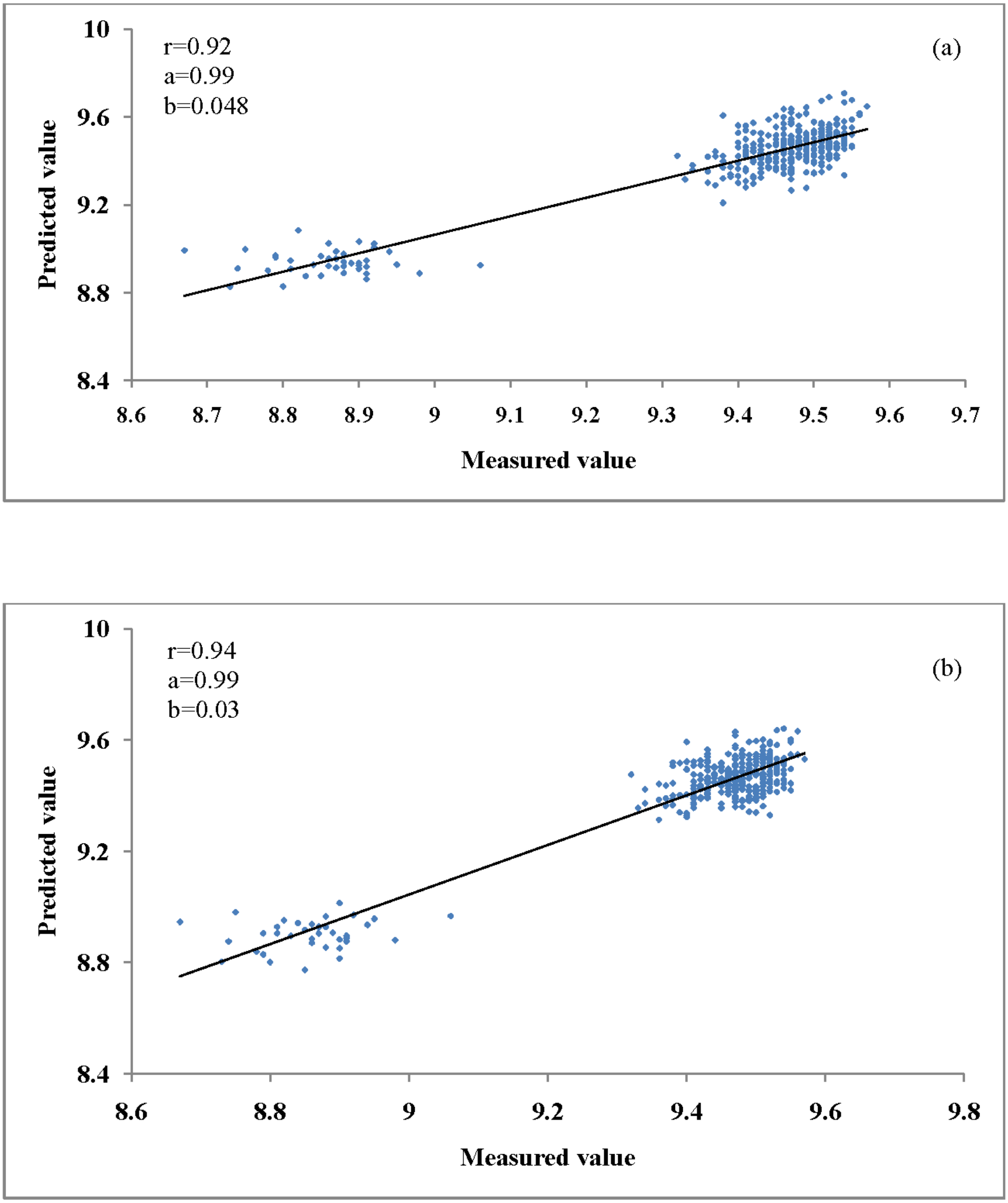
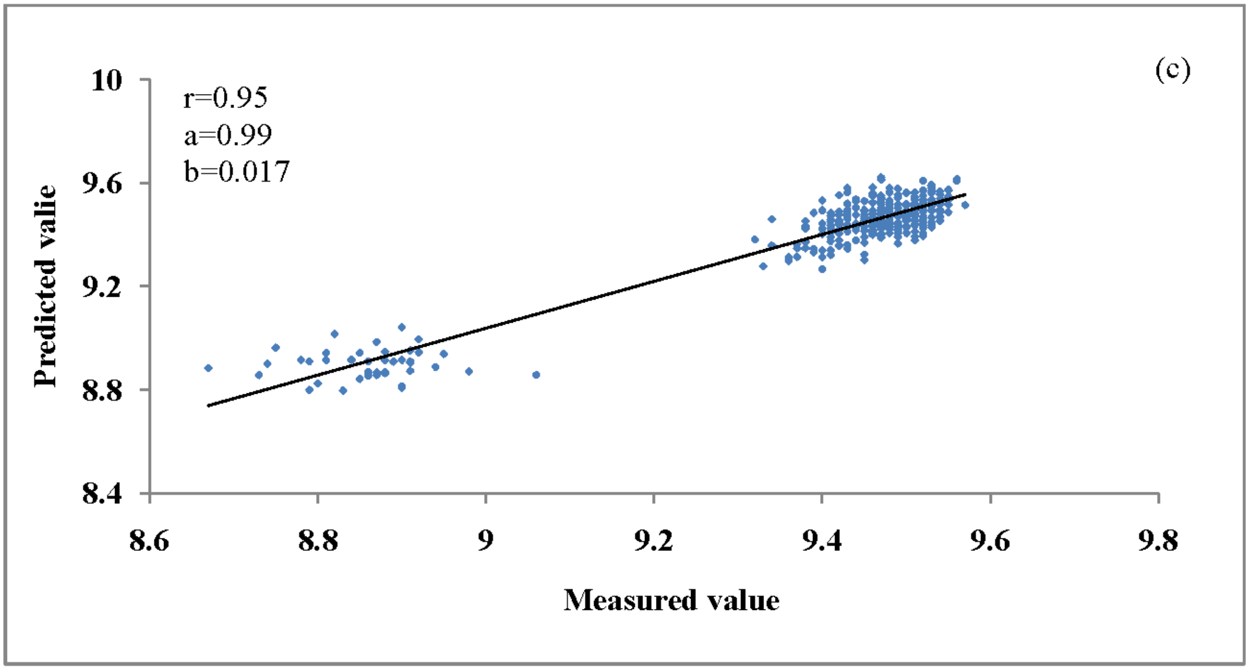


Fig. 2: Relationship between measured and predicted albumen pH obtained with (a) 8, (b) 9, (c) 10 wavelengths and (d) full spectra



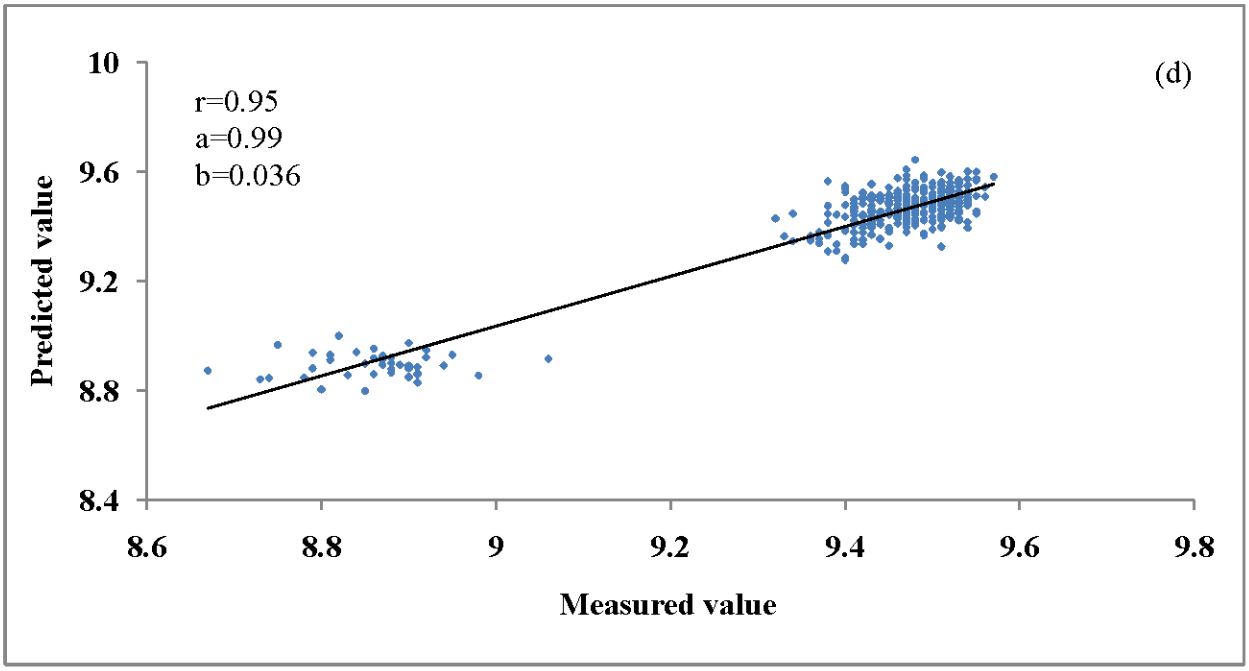


Fig. 3: (cont.) Relationship between measured and predicted albumen pH obtained with (a) 8, (b) 9, (c) 10 wavelengths and (d) full spectra

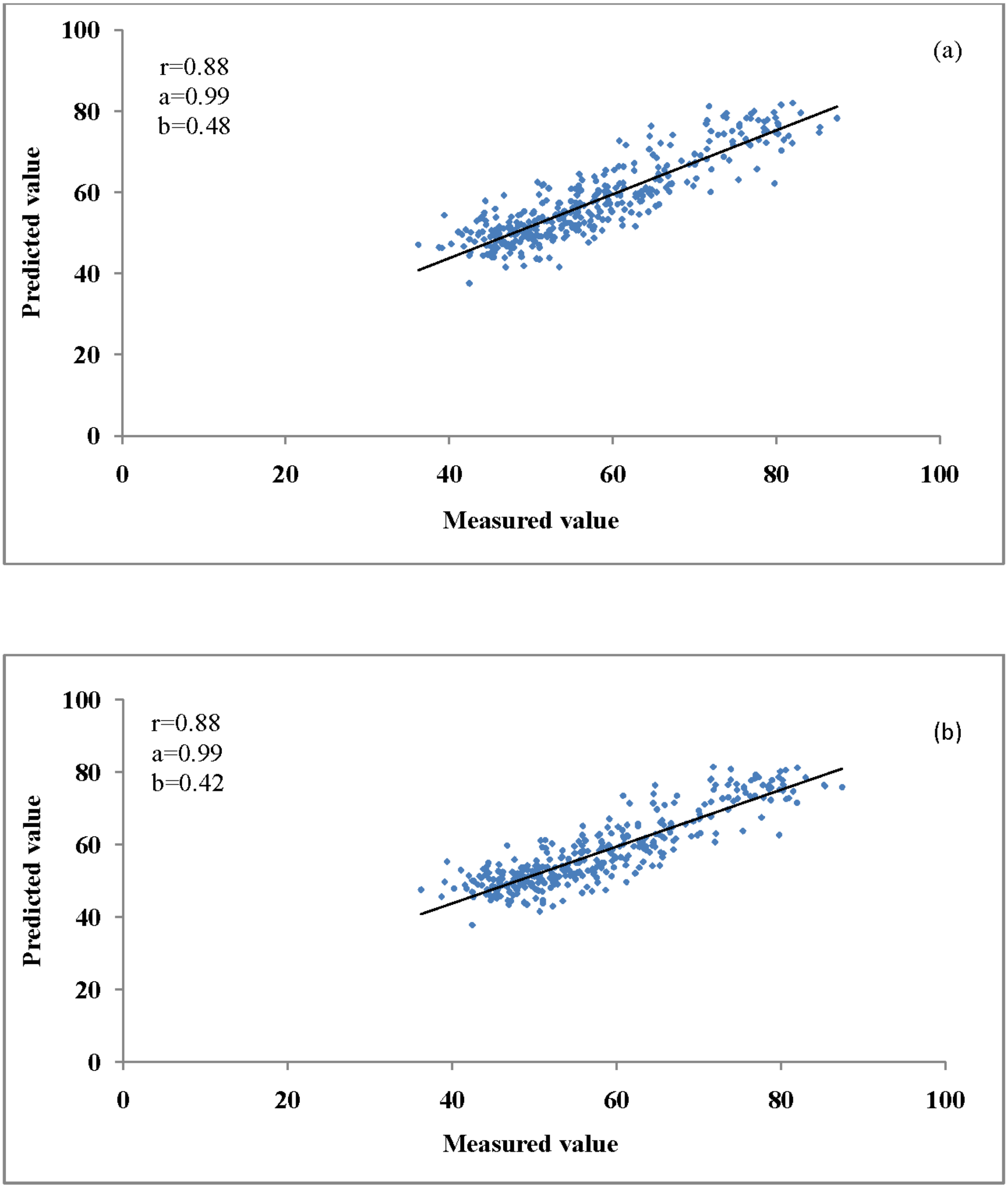


Fig. 4: Relationship between measured and predicted HU obtained with (a) 8, (b) 9, (c) 10 wavelengths and (d) full spectra

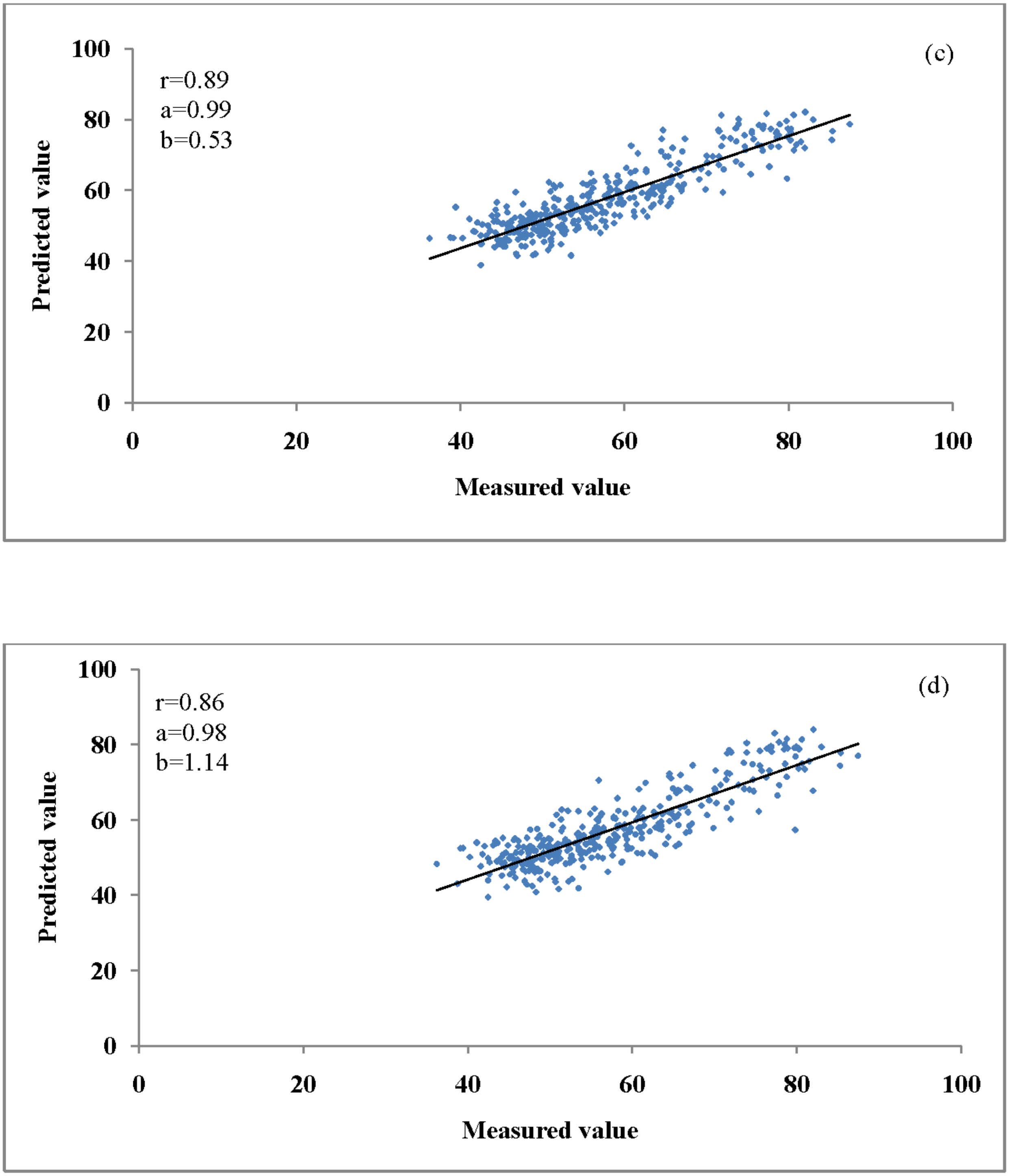


Fig. 5: (cont.) Relationship between measured and predicted HU obtained with (a) 8, (b) 9, (c) 10 wavelengths and (d) full spectra

less stable than the model based on informative wavelengths. It is evident that the whole original spectral matrix does not allow satisfactory predictions for samples. Thus, the use of PLS method alone with all spectral range, the VIS/NIR spectroscopy cannot be used to accurately classify the HU for eggs. The proposed MAXR-PLS method selected the most informative wavelengths and provided a robust predictive PLS model with only 10 wavelengths.

Conclusion: In this article, the usefulness of a new method named as MAXR-PLS for the building of a PLS predictive model was investigated. The results presented above demonstrated that this method is a good tool for this purpose. For the HU and albumen pH, the results showed an improvement in prediction with 10 wavelengths compared to those obtained with the full spectrum. The RMSECV for the model built with 10 wavelengths and the whole spectra for the prediction of albumen pH were 5.05 and 5.51, respectively. The R^2 of the model predicting the HU for the selected wavelengths and the full spectra were 0.89 and 0.86 respectively as well as there was a numerically improvement of the correlation coefficient, slope and intercept. By eliminating the less informative wavelengths and selecting the relevant ones using this method, the PLS models can improve the predictive ability, decrease the model complexity and reduce the time of analysis.

ACKNOWLEDGEMENT

The authors will like to acknowledge the Poultry Industry of Canada for funding this study.

REFERENCES

Bangalore, A.S., R.E. Shaffer and G.W. Small, 1996. Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy. Anal. Chem., 68: 4200-4212.

Birth, G.S., G.G. Dull, W.T. Renfore and S.J. Kays, 1985. Nondestructive spectrophotometric determination of dry matter in onions. J. Am. Soc. Hort. Sci., 110: 297-303.

Blanco, M., J. Coello, H. Iturriaga, S. Maspoch and C. De La Pezula, 1997. Effect of data preprocessing methods in near-infrared diffuse reflectance spectroscopy for the determination of the active compound in a pharmaceutical preparation. Appl. Spectrosc., 51: 240-246.

Carlini, P., R. Massantini and F. Mencarelli, 2000. Vis-NIR measurement of soluble solids in cherry and apricot by PLS regression and wavelength selection. J. Agric. Food Chem., 48: 5236-5242.

Du, Y.P., Y.Z. Liang, J.H. Jiang, R.J. Berry and Y. Ozaki, 2004. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. Anal. Chim. Acta., 501: 183-191.

Fila, G., A. Bellochi, M. Acutis and M. Donatelli, 2003. IRENE: a software to evaluate model performance. Europ. J. Agronomy, 18: 269-372.

Goel, P.K., 2003. Hyper-Spectral remote sensing for weed and nitrogen stress detection. Ph.D. Thesis. McGill University. Montreal. QC., pp: 238-256.

Haugh, R.R., 1937. The Haugh unit for measuring egg quality. US Egg Poult. Mag., 43: 552-555, 572-573.

IRENE, 2003. Integrated Resources for Evaluating Numerical Estimates. Ver. beta 1.00. Bologna, Italy.

Jiang, J.H., R.J. Berry, H.W. Siesler and Y. Ozaki, 2002. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data. Anal. Chem., 74: 3555-3565.

Kemps, B.J., F.R. Bamelis, B. De Katelaere, K. Mertens, K. Tona, E.M. Decuypere and J.G. De Baerdemaeker, 2006. Visible transmission spectroscopy for the assessment of egg freshness. J. Sci. Food Agric., 86: 1399-1406.

Kobayshi, K. and M.U. Salam, 2000. Comparing simulated and measured values using mean square deviation and its components. Agron. J., 92: 345-352.

Osborne, S.D., R.B. Jordan and R. Künnemeyer, 1997. Method of wavelength selection for partial least squares. Analyst., 122: 1531-1537.

Perez-Mendoza, J., J.E. Throne, F.E. Dowell and J.E. Baler, 2003. Detection of insect fragments in wheat flour by near-infrared spectroscopy. J. Stored Prod. Res., 39: 305-312.

Rosipal, R. and N. Krämer, 2006. Overview and recent advances in partial least squares. SLSFS 2005, LNCS 3940, pp: 34-51.

SAS, 2003. SAS User's Guide: statistics. Ver 8.2. Cary, N.C.: SAS Institute, Inc.

Thermo Galactic, 2003. PLS Plus IQ^{TM}. User's Guide. Salem, NH, USA.

The Unscrambler, 2007. The Unscrambler User's Guide: ver. 9.7. Woodbridge, N.J., USA: CAMO Software AS.

Todeschini, R., D. Galvagni, J.L. Vílchez, M. Del Olmo and N. Navas, 1999. Kohonen artificial neural networks as a tool for wavelength selection in multicomponent spectrofluorometric PLS modelling: application to phenol, *o*-cresol, *m*-cresol and *p*-cresol mixtures. TrAC, 18: 93-98.

Ventura, M., A. De Jager, H. De putter and F.P.M.M. Roelofs, 1998. Non-destructive determination of soluble solids in apple fruit by near infrared spectroscopy. Postharvest Biol. Technol., 14: 21-28.