Effective Variables Selection in Eggs Freshness Graphically Oriented Local Multivariate Analysis using NIR Spectroscopy

Hao Lin*, Jiewen Zhao, Li Sun, Xia kun Bi, Jianrong Cai
School of Food & Biological Engineering, Jiangsu University
Zhenjiang, P.R. China
linhaolt794@163.com

Abstract— **In multi-component spectral analysis, informative variables selection is important to get satisfied performance. The present research intends to establish relationship between eggs freshness and NIR spectroscopy, and build a compact and robust calibration model. Graphically-oriented local multivariate calibration modeling procedures were used comparatively to select efficient spectral variables in comparison to the full-spectrum model. Three kinds of methods, which were spectral interval selection, effective coefficient variables selection and genetic algorithm, were used for variable selection. Successive projections algorithm (SPA) showed its superior ability in reducing the complexity of model building. A satisfactory result was achieved while only 8 variables were used. Meanwhile, the optimal performance was obtained with genetic algorithm synergy interval partial least-square (GA-siPLS) by using 9 PCs and 42 variables selected, which resulted in root mean square error value of prediction (RMSEP) value of 3.29. This work indicates that it is feasible to identify egg freshness using NIR spectroscopy combined with graphically-oriented local multivariate analysis, and using variables methods is important to reduce he complexity of model building with fewer spectral variables and improve performance of calibration model.**

Keywords-Variable selection; Multivariate calibration; Near infrared spectroscopy; Egg freshness

I. INTRODUCTION

Freshness makes a major contribution to the internal quality of eggs. Development of a reliable and non-invasive method for determining freshness of eggs is critical to this industry.

Reduction of eggs freshness can be explained in terms of changes in ovomucin–lysozyme interaction, in disulfide bonds of ovomucin or in carbohydrate moieties of ovomucin which are mainly involved in the egg white thinning [1]. Recent studies show that it is possible to assess the freshness of eggs with Near Infrared Spectroscopy (NIR) technique. Berardinelli et al. performed a predictive model of thick albumen height[2]. Giunchi et al. reported non-destructive freshness assessment of shell eggs by means of FT-NIR and partial least square regression (PLS)[3]. A predictive model of Haugh unit was obtained with determination coefficient R2 of 0.676. Zhao et al. (2010) employed support vector data description (SVDD) to solve the problem with imbalance training samples in egg freshness measurement by NIR spectroscopy, the identification rates of fresh eggs and un-fresh eggs were both 93.3% in prediction set[4]. Lin et al. (2011) employed NIR spectroscopy combined with multivariate calibration model qualitative and quantitative analyze freshness of eggs[5]. Nicolas et al. (2011) predicted the indexes of HU, albumen pH, and number of storage days related to egg freshness, using VIS and NIR ranging from 411 to 1,729 nm[6].

For these works mentioned above, NIR spectral data analyses were usually implemented by classical multivariate calibration models. They did not take into account the selection of spectral variables. However, not all spectral variables or regions are equally important for modeling. Some spectral regions may contain little information about freshness of egg samples, and some noise regions may even be harmful to the performance of model. Further more, variables selection will do great help to simplify the process of calibrating models, and build a robust model for the measurement of egg freshness. In recent years, the significance of using variable selection methods in multivariate analysis has been demonstrated [7-8].

The present work intends to establish relationship between egg freshness and NIR spectroscopy, and build a compact and robust calibration model. In order to reduce the complexity of model building, several graphically-oriented local multivariate calibration models (i.e. i-PLS, si-PLS, GA-PLS, GA-siPLS, MW-PLS, SPA-PLS, ICA) were used comparatively to select efficient spectral variables relating to the information of eggs freshness. Spectral variables selection is helpful to design of a simpler NIR instrument for real time measurement. PLS calibration model was employed to evaluate the results of spectral variables selection. The performance of the final model was evaluated according to the values of root mean square error of prediction (RMSEP) and the correlation coefficient (R) in prediction set.

II. Materials and Methods

A. Sample preparation

A total of 154 eggs collected from market used in this study, and all of them have not yet been graded. In order to obtain un-fresh egg samples, these eggs were storied at 25℃ for 3 days. At the end of storage, the NIR spectra of all eggs were collected. After FT-NIR acquisition, destructive methods were used to obtain the Haugh units (HU) of eggs (standard method for egg freshness measurement) (Haugh 1937). In the United States egg grades, AA grade eggs score 72 HU or higher; A grade, 60 – 72 HU, and B grade, lower than 60 HU [9]. Eggs with HU score below 60 are considered to be un-fresh. Haugh unit is calculated as follows:

$$HU = 100\lg\left(H + 7.57 - 1.7m^{0.37}\right) \quad (1)$$

where h is the thick albumen height, m is egg mass.

B. NIR spectroscopy measurements

The NIR spectra were recorded in the reflectance mode by using an Antaris II Near-infrared spectrophotometer (Thermo Electron Co., USA) with a fiber optic sampling probe. The fiber probe was placed directly in contact with eggshell to illuminate the sample and collect the diffused light. NIR spectra were recorded from the equatorial axis of the eggshell, because the internal composition changes are more remarkable in this region than other places. In order to avoid possible effects due to differences in internal composition, the diffused reflectance spectra were obtained by averaging triplicate measurements carried out round the equatorial axis of eggshell. Each spectrum was the average of 32 scanning spectra. The NIR spectra for egg samples were acquired in the region between 10000 and 4000cm-1, and the data were measured in 3.856 cm-1 interval, which resulted in 1557 variables. The temperature and the related humidity were kept around 25℃ and 70% in the laboratory, respectively.

C. Spectral data processing

NIR spectra are easily affected by physical properties of the analyzed products and other interferences. Thus, it is necessary to perform mathematical pre-treatments to reduce the systematic noise, and enhance the contribution of the chemical composition. standard normal transformation (SNV) is a mathematical transformation method used to remove slope variation and correct scatter effects in spectra. Each spectrum is corrected individually by first centering the spectral values, and then the centered spectrum is scaled by the standard deviation calculated from the individual spectral values. In this work, SNV，as a mathematical transformation method (Font et al. 2005)[10], was applied to process full NIR spectra. SNV preprocessed spectra are presented in Fig. 1.

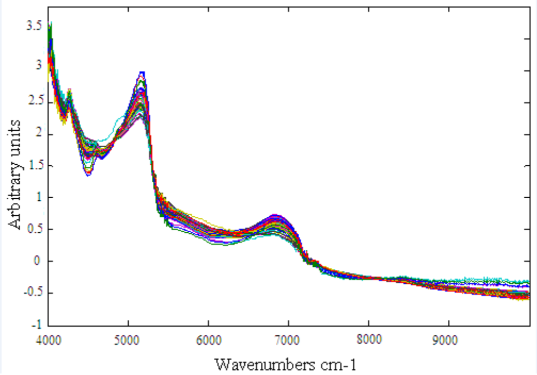

Figure 1. Spectra of egg samples after SNV preprocessing

D. Multivariate calibrations

In order to build a compact and robust calibration model, several graphically-oriented local multivariate calibration models (i.e. i-PLS, si-PLS, GA-PLS, GA-siPLS, MW-PLS, SPA-PLS, ICA) were used comparatively to select effective spectral variables relating to egg freshness. The results of spectral variables selection were evaluated by PLS calibration model. RMSECV was used for evaluation the error of each calibration model [11]. To obtain the RMSECV value, a leave-one-sample-out cross-validation is performed: the spectrum of one sample in calibration set is left out from this set, and a model is built with the remaining spectra in this set. The RMSECV is then calculated with following relation:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{\setminus i} - y_i)^2}{n}} \quad (2)$$

where n is the number of samples in the calibration set,   is the reference measurement result for sample , and   is the estimated result for sample   when the model is constructed with sample   removed. The number of factors including in the model is chosen according to the lowest RMSECV. This procedure is repeated for each of the preprocessed spectra.

i-PLS. The development of spectral variables selection was first accomplished by i-PLS. The interval PLS (i-PLS) algorithm was developed by Norgaard et al. (2000) [12]. Its principle is to split the spectra into some small equidistant regions, and then to develop PLS regression models for each of the sub-intervals. Thereafter, the value of RMSECV is calculated for every sub-interval . The region with the lowest value of RMSECV is chosen.

si-PLS. The algorithm of si-PLS was also used for spectral variables selection. The basic principle of si-PLS is as following. First, it splits the spectra into some small equidistant regions. And next, PLS regression models are developed in all possible combinations of two, three or four intervals. The combination of intervals with the lowest RMSECV is chosen (Thomas et al. 2008) [13].

Genetic algorithms. GA is an adaptive heuristic search algorithm which can be applied for the variable selection. In GA-PLS algorithm, the number of the genes at each chromosome is equal to the number of the samples [14]. The population of the first generation is selected randomly. A gene is given the value of one, if its corresponding variable is included in the subset; otherwise, it is given the value of zero. The selection of variables is based on the lowest RMSECV values of PLS models. In GA-PLS model, genetic algorithm is used to select effective spectral variables, and the performance of the model is evaluated by PLS calibration. The aim of genetic algorithm in this work was to reduce the value of RMSECV in variables selection after running 100 generations for each iteration process. So the fitness function (F) was minimization and it was computed using the equation as follows:

$$F = \frac{1}{1 + RMSECV} \quad (3)$$

where RMSECV is predicted from each chromosome. The best elite chromosomes of size 'N' having with the lowest RMSECV value after running 100 generations for each iteration process.

GA-siPLS. The GA-siPLS algorithm introduced in this study is the combination of GA-PLS and si-PLS algorithm [14]. The algorithm synergy interval partial least square (si-PLS) is firstly attempted to select several NIR spectral regions and then genetic algorithm is employed to choose fewer effective variables from the spectral regions selected by si-PLS. The final spectral variables are selected by the GA-siPLS model that gives the best performance with respect to lowest RMSECV value.

Moving windows method. MW-PLS is another wavelength interval selection method used for spectra analysis in this study. Briefly, MW-PLS develops PLS calibration models with various principal components in every window that moves over the whole spectral region, and then calculates the sums of squared residuals (SSR) for each subset. Finally, it locates informative spectral intervals that have the least model complexity and the lowest sum of residuals. MW-PLS provides a viable approach to eliminate the extra variability generated by non-composition-related factors [15]. A salient advantage of MW-PLS is that the calibration model is very stable against the interference from non-composition related factors. In practice, MW-PLS locates more than one region in vibration spectra because of the existence of many informative spectral bands.

Successive projections algorithm. SPA employs simple projection operations in a vector space to obtain subsets of variables with minimal collinearity. It is a forward variable selection algorithm for multivariate calibration [16]. The principle of it is that the new variable selected is the one among all the remaining variables, which has the maximum projection value on the orthogonal sub-space of the previous selected variable. ***Independent component analysis***. The goal of ICA is to find a proper linear representation of non-Gaussian vectors so that the estimated vectors are as independent as possible, and the mixed signals can be denoted by the linear combinations of these independent components, ICs [17]. The problem setting about ICA is as follows. Assuming that there is an L-dimensional zero-mean non-Gaussian source vector s= $[s_1.....s_L]^T$, such components are mutually independent, and an observed data vector x= $[x_1.....x_N]^T$ is composed of linear combinations of sources , such that

$$x = As \quad (4)$$

where A is a full rank N×L matrix with L⩽N. The goal of ICA is to find such a linear mapping that each component of an estimate u of the source vector

$$u = Wx = WAs \quad (5)$$

is as independent as possible. The original sources are exactly recovered when is a pseudo-inverse of up to some scale changes and permutations . According to the weights of wavenumbers by each IC, the wavenumbers corresponding to the highest weights were selected as effective spectral variables.

Calibration models. All samples were divided into two subsets. One of the subsets called calibration set was used to build model, and the other one called prediction set was used to test the robustness of model. Total numbers of 100 samples were randomly selected as calibration set, and the remaining 54 samples constituted the prediction set. As shown in Table 1, the range of y-value in calibration set covered the range in prediction set. Therefore the distribution of the samples was appropriate. The number of PLS factors and spectral variables were optimized according to the RMSECV value in calibration set. The performance of the final model was evaluated according to the values of RMSEP and the correlation coefficient (R) in prediction set.

TABLE 1 REFERENCE MEASUREMENT OF HAUGH UNITS OF EGGS IN CALIBRATION AND PREDICTION SETS

Subsets	Units	S.N[a]	Range	S.D[b]
Calibration set	g/g	100	73.906- 43.834	6.098
Prediction set	g/g	54	70.821- 46.556	5.519

SOFTWARE. All algorithms were implemented in Matlab V7.1 (Mathworks, USA) under Windows XP. Result Software (Antaris II System, Thermo Electron Co., USA) was used in NIR spectral data acquisition.

III. RESULT AND DISCUSSION

A. *i-PLS model*

The selection of spectral variables was first accomplished by i-PLS. The spectra were divided into 20 equidistant sub-intervals, because more dividing did not improve the results. A calibration model based on PLS was developed for each sub-interval. The selected variables were evaluated according to the lowest value of RMSECV. The optimal i-PLS model was obtained with interval number 14 selected, where the lowest RMSECV was 3.35. The best interval was number 14, corresponding to wavenumbers in the range 8291.73-8715.89 cm^{-1}. The full spectrum RMSECV is 3.36 while interval 14 is 3.35. So, the variable selection of i-PLS is not significant to achieve a better result, but it is useful to reduce the complexity of model building.

B. si-PLS model

The selection of spectral variables was secondly accomplished by si-PLS. The spectra were divided into 30 equidistant subintervals, because the usage of more than this number did not improve the results from previously attempts. A calibration model based on PLS was developed for each one, using different numbers of PLS components, as required. For comparison of these models in relation to the global model which used the whole spectrum, the value of RMSECV was used. The optimal si-PLS model was obtained when the spectra were split into 30 intervals and 4 intervals were selected, where the lowest RMSECV was 3.183. The combined intervals selected by si-PLS were corresponding to the wavenumbers of 4605.392-4802.048, 6008.976-6205.632, 6811.024-7007.68, 7412.56-7609.215 cm^{-1} respectively.

B. GA-PLS model

The algorithm of GA-PLS was used for variable selection from 1557 spectral variables. The crossover rate was 0.8; the mutation rate was 0.01; the termination condition was running 100 generations for each iteration process. As the population of the first generation was selected randomly, it would subtly affect the performance and the selection of frequencies. So, the algorithm was implemented for 10 times to remove the effects of population initialized. According to results of experiments, the effect of random population initialized was very subtle, and selected spectral variables were similar within 10 times implementation. The optimal model was obtained when 98 variables and 7 PLS components were used, the lowest RMSECV was 3.2781. In GA-PLS, some variables were mostly used in calibrating GA-PLS model, and some variables were totally not used. So, it is possible to build robust calibration models using few variables relating to egg freshness.

C. GA-siPLS model

Based on 4 regions of NIR spectra (208 variables) selected in si-PLS model, further more study was carried on by GA algorithm to select effective variables from these regions. The fitness function and other parameters were the same as GA-PLS model for consistent comparison. The algorithm was also implemented for 10 times. When 42 variables and 9 PCs were used, the optimal PLS model could be obtained, with respect to the lowest RMSECV value of 3.2117.

D. MW-PLS

All of the informative regions of egg freshness were obtained from the wavelength region of 4000–10000 cm^{-1}. The windowsize was set to 31 in MW-PLS model. Because egg freshness reflects the information of various chemical compositions, and the amide groups may be separated into several informative regions in their NIR spectra. It is possible to get better PLS models and prediction results by using the combinations of informative regions. Therefore, we collected all promising informative regions and applied MW-PLS to search for their optimized combinations. The best performance was obtained with the combination of 4 spectral regions, where the lowest RMSECV was 3.32. The corresponding wavenumbers were 5673.504-5719.776, 6984.544-7030.816, 7358.576-7389.424, 7412.56-7485.824 cm^{-1}, respectively.

E. SPA

SPA was carried out for selecting effective spectral variables from the full spectra. The maximum number of effective spectral variables selected by SPA was set from 5 to 30 according to previous experience. It was found when 8 variables were selected; the lowest RMSECV value of 3.327 can be obtained, corresponding to the optimal PLS model. The selected spectral variables selected by SPA corresponding wavenumbers were 4682.512, 4813.616, 5341.888, 7987.104, 4007.712, 7200.48, 5318.752, 4092.544 cm-1, respectively.

F. ICA

ICA was also applied to select effective spectral variables from the full spectra. During the process of ICA, the coefficient matrix for each IC and weights matrix for each wavelength were obtained. The effective wavenumbers were selected with the largest absolute weight value of each IC. Fig. 8 shows the weight plots of top three ICs (IC1, IC2, IC3). The largest absolute weight values of top three ICs were corresponding to wavelength at 4967.856, 5415.152, 4632.384 cm^{-1}, respectively. So, the three variables were selected as effective spectral variables from top three ICs. Based on previous experiments, the best performance was obtained when top ten ICs were used. When 9 variables were selected, the optimal PLS model was obtained with the lowest RMSECV value of 3.59.

G. Discussion of spectral variables selection by different methods

PLS is performed on full spectral region (4000-10000cm^{-1}, with 1557 spectral variables) to calibrate global model, some noisy spectral information inevitably weaken the performance of model. So, the performance of PLS model was worse than other calibration models which used spectral variables selection algorithms to remove noisy spectral information. Different graphically-oriented local multivariate calibrations models (i.e. i-PLS, si-PLS, GA-PLS, GA-siPLS, MW-PLS, ICA and SPA) for spectral variables selection were investigated and compared in this work. The results of these calibration models are presented in Table 2. According to investigation of the results from Table 2, these calibration models all resulted in acceptable performance while only fewer spectral variables (usually only several to dozens of variables) were used. The variables selection algorithms were helpful to simplify the process of calibrating models, and build a robust model with fine stability for the measurement of egg freshness.

According the principle of graphically-oriented local multivariate calibrations models, they were divided into three types, which were listed as follows. (1) Interval selection (i.e. i-PLS, si-PLS), which dividing the NIR spectra into several equidistant parts, and then developing the local multivariate calibration models with these selected parts. They reduced the number of variables from thousands or more reduced to tens to hundreds, and the size of selected variables depends on the number of intervals. (2) Variables selection (i.e. MW-PLS, SPA, ICA) based on effective coefficient of spectra. They develop the multivariate calibration models with variables corresponding to effective coefficient of spectra or minimum of collinearity, which quickly reducing the number of variables from thousands or more reduced to tens even to no more than ten. (3) Genetic algorithms, an adaptive heuristic search methods based on fitness function. Variable selection of genetic algorithm has certain randomness, quickly reducing the number of variables from thousands or more reduced to tens to hundreds.

TABLE 2 SPECTRAL VARIABLES OBTAINED BY DIFFERENT METHODS AND THEIR CORRESPONDING PLS PREDICTION RESULTS

Method	Wavenumber s(cm^{-1})	No. V	PCs	RMSEP	R
PLS	4000-10000	1557	7	3.71	0.8007
i-PLS	8291-8715	112	5	3.61	0.8122
Si-PLS	4605-4802 6008- 6205 6811-7007 7412-7609	207	9	3.42	0.8436
GA-PLS	98 variables	98	7	3.667	0.7966
GA-siPLS	42 variables	42	9	3.29	0.8442
MW-PLS	5673-5719 6984-7030 7358-7389 7412.748	51	7	3.64	0.8356
ICA+PLS	4967, 5415, 4632, 5280, 4273, 7235, 4281., 4277, 5276	9	8	3.64	0.8161
SPA+PLS	4682, 4813 5341, 7987 4007, 7200 5318., 4092	8	8	3.56	0.8193

The algorithms of i-PLS, si-PLS aim to select effective spectral interval. They have a consistent concept, that is they both divide the NIR spectra into several equidistant parts, and then develop the local multivariate calibration models with these selected parts . It is possible to get better models comparing to the full-region model, because the selected variables are more related to the information of eggs freshness. i-PLS actually gives an overview of spectral data in selecting the interesting spectral region and removing some noisy regions, but only one selected spectral region usually can not express all useful information about egg freshness. In contrast to i-PLS, si-PLS not only possesses same advantages like i-PLS, but also overcomes the disadvantages of it, because si-PLS combines with two, three or four intervals to calibrate PLS model, so as not to lose much useful information in calibrating model. Si-PLS (i-PLS) algorithms are useful tools for conveniently reducing massive data. However, the selected spectral regions may still contain useless information about analyte.

Variables selection based on effective coefficient of spectra quickly reduced the massive variables to tens even to no more than ten. The goal of MW-PLS is to search for informative regions for the spectral analysis. Informative regions mean that they contain useful information for PLS model building and are helpful to improve the performance of the model . The goal of ICA is to find a proper linear representation of non-Gaussian vectors so that the estimated vectors are as independent as possible, and the mixed signals can be denoted by the linear combinations of these independent components. According to the weights of wave-numbers by each IC, the wave-numbers corresponding to the highest weights are selected as the effective wavelengths. SPA selected several informative spectral variables, which employed projection operations to select variables with the minimum of collinearity [18]. Comparing with the variable selection methods based on the effective coefficient of spectra, SPA employs only a few data reprehensive most of the sample spectrum information, greatly avoided information overlapping.

The goal of MW-PLS is to search for informative regions for the spectral analysis. Informative regions mean that they contain useful information for PLS model building and are helpful to improve the performance of the model . The goal of ICA is to find a proper linear representation of non-Gaussian vectors so that the estimated vectors are as independent as possible, and the mixed signals can be denoted by the linear combinations of these independent components. According to the weights of wave-numbers by each IC, the wave-numbers corresponding to the highest weights are selected as the effective wavelengths. SPA selected several informative spectral variables, which employed projection operations to select variables with the minimum of collinearity. Comparing with the variable selection methods based on the effective coefficient of spectra, SPA employs only a few data reprehensive most of the sample spectrum information, greatly avoided information overlapping.

Genetic algorithms (GA) takes account all spectral variables (1557 variables) in building model, and it is an adaptive heuristic search algorithm applied for the variable selection. However, the variable searching process has a certain degree of blindness. Especially the number of variables to be in the thousands, randomness more obviously presents in the results of variable selection. GA-siPLS is the combination of genetic algorithm and si-PLS, it selected variables within the spectral regions obtained by si-PLS. Therefore, the variable searching processes of GA-siPLS is more purposeful. GA-siPLS selected even fewer spectral variables than si-PLS and GA-PLS, but got better performance.

In conclusion, using variables methods is important to reduce he complexity of model building with fewer spectral variables and improve performance of calibration model. The algorithm of SPA is an effective method for reducing the complexity of model building. A satisfactory result could be achieved (the value of RMSECV and RMSEP were 3.327 and 3.56 respectively) while only 8 variables were used.

Meanwhile, the GA-siPLS got the best performance (lowest value of RMSEP) with RMSEP was 3.29, while with the value of RMSECV was 3.2117. The scatter plot of references measured and NIR spectra predicted by GA-siPLS model in prediction set is shown in Fig.2, and the correlation coefficient was 0.8442.

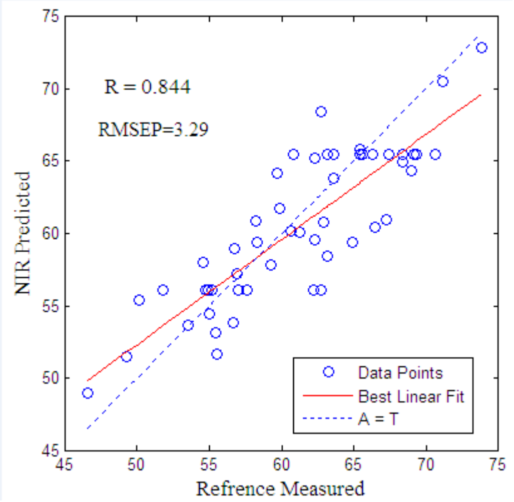

Fig. 2 The scatter plot of references measured and NIR spectra predicted by GA-siPLS model in prediction set

Acknowledgement

The authors express their acknowledge National Natural Science Foundation of PR China (Grant No. 31201451) and open foundation of agricultural product physical processing laboratory (Grant No. JAPP2011-4).

References

[1] E. Li-Chan and S. Nakai, Biochemical basis for the properties of egg white. Poultry Biology, 1989, pp. 21-50.

[2] A. Berardinelli, A. Giunchi, A. Guarnieri, F. Pezzi and L. Ragni, Comparison of linear and nonlinear calibration models based on near infrared spectroscopy data for gasoline properties prediction. Trans. ASAE,2005, pp. 1426-1428.

[3] A. Giunchi, A. Berardinelli, L. Ragni, A. Fabbri and F. A. Silaghi, Nondestructive freshness assessment of shell eggs using FT-NIR spectroscopy, J. Food Eng., 2008, pp.142-148.

[4] J.W. Zhao, H. Lin, Q.S. Chen , X.Y. Huang , Z. B. Sun and F. Zhou, Identification of egg's freshness using NIR and support vector data des-cription J. Food Eng., 2010, pp. 408-414.

[5] H. Lin, J.W. Zhao, L. Sun, Q.S. Chen and F. Zhou, Innov. Food Sci. Emerg. Technol., Freshness measurement of eggs using near infrared spectroscopy and multivariate data analysis 2011, pp.182-186

[6] N. Abdel-Nour, M. Ngadi, S. Prasher and Y. Karimi, Food Bio. Technol.,Prediction of egg freshness and albumen quality using visible/near infrared spectroscopy 2011, pp.713-736

[7] R. Leardi, J. Chemometri., Application of genetic algorithm-PLS for feature selection in spectral data sets, 2000, pp. 643-655.

[8] R. Leardi, M.B. Seasholtz, R.J. Pell, Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data Anal. Chim. Acta., 2002, pp.189-200

[9] United States Department of Agriculture, Agricultural Marketing Service. USDA Egg-Grading Manual. Agricultural Handbook, Washington. 2000,pp.75.

[10] R. Font, M. D. Río-Celestino, E. Cartea, A.D. Haro-Bailón, Phytochemistry, Quantification of glucosinolates in leaves of leaf rape (Brassica napus ssp. pabularia) by near-infrared spectroscopy, 2005, pp.175-185.

[11] L. Munck, J. Pram Nielsen, B. Møller, S. Jacobsen, I. Søndergaard, S.B. Engelsen, L. Nørgaard and R. Bro, Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics, Anal. Chim. Acta., 2001, pp. 171-186.

[12] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck and S.B. Engelsen, Interval Partial Least Squares Regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy , Appl. Spectrosc. 2000, pp. 413-419.

[13] V. Thomas, S. Robert and J. Richard, An optimal alternative to DiPLS_Cluster for unsupervised classification Chemometr. Intell. Lab. Syst. 2008, pp. 8-14.

[14] H. Lin, J.W. Zhao, L. Sun, Q.S. Chen, Z.B. Sun and F. Zhou, Stiffness measurement of eggshell by acoustic resonance and PLS models, J. Food Eng. 2011, pp.351

[15] Y.P. Du, Y. Z. Liang, J. H. Jiang, R.J. Berry and Y. Ozaki,. 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, 2004, pp. 183-191

[16] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemometr. Intell. Lab. Syst., 2001, pp.65-73.

[17] P. R. Oliveira, R. A. F. Romero, Improvements on ICA mixture models for image pre-processing and segmentation. Neurocomputing , 2008, pp.2180-2193.

[18] X. B. Zou, J. W. Zhao, P. Malcolm, H. Mel, H. P. Mao, Variables selection methods in near-infrared spectroscopy Anal. Chim. Acta, 2010, pp. 14-32.