Journal of Scientific & Industrial Research
Vol. 79, February 2020, pp.148–152

Performance Evaluation of Chemometric Prediction Models—Key Components of Wheat Grain

A Anne Frank Joe[1]*, A Gopal[2] and R Pandian[1]

[1]Sathyabama Institute of Science and Technology, Tamil Nadu
[2]CSIR-CEERI, Taramani, Chennai 113, India

Received 3 October 2018; revised 18 August 2019; accepted 10 September 2019

The present study was aimed to evaluate the accuracy of using near-infrared spectroscopy (NIRS) for predicting protein, moisture, starch and ash content values of wheat. The physiochemical properties of wheat were predicted using twelve prediction models of preprocessing coupled with regression tools. The performance measure of SVM aided with extended multiplicative scatter correction gave confident prediction results of protein, moisture, ash and starch content with R^2 values of 0.989, 0.987, 0.976, 0.998 and RMSECV values of 0.263, 0.285793, 0.369 and 0.03 respectively. These results indicate the practical applicability of NIRS in wheat grain quality profiling.

Keywords: Wheat, Support vector machine, Quality parameters, Near Infrared Spectrometer

Introduction

Wheat crop adapts to changes in climatic conditions and thus is a preferred cereal crop[1]. Wheat and wheat based products are used in everyday consumables and also in the preparation of alcoholic drinks, cattle feed to name a few. The success of final wheat product has a direct dependability on the physiochemical properties of the grain[2]. Traditionally, quality tests for wheat grain and flour have been undertaken by lengthy chemical and rheological methods, which are labor intensive, time-consuming, and requires experienced technicians. NIR technology ascertains speedy, non destructive, qualitative and quantitative analysis to facilitate the millers/buyers and sustain complete confidence in the grain they choose to buy.

NIR Instrumentation

The near infrared spectroscopy is an instrumental technique based on the absorption rate of the infrared radiation reflected or transmitted by the sample, due to the vibration of hydrogen bonds (CH, NH, OH). The wheat grains are of complex composition with numerous hydrogen bonds whose vibration contributes to overlapped spectral bands. Assignment of a specific spectral wavelength to a chemical constituent is not simple through direct relationship which is where chemometrics comes to the rescue[3]. NIRS is a secondary analytical method which needs to be backed up with reference data. Prediction ability is based on the relationship modeled between the measured spectral data and the constituent of interest. Chemometric techniques, such as Partial least squares (PLS) regression, Support vector machine(SVM), Multiple linear regression(MLR) and other such calibration model building methods can be used to identify and quantify the key wheat constituents. The NIR instrument used to collect spectral data of whole grain samples was FOSS XDSTM NIR Analyzer, which has a wide wavelength range (450–2,498 nm) including the visible spectrum. The characteristic absorption peaks were marked at 990nm, 1440nm, 1940nm, 2100nm, 2336nm and absorption bands from 2168 to 2180 nm and 973 to 1020 nm in the NIR spectrum based on intensive literature survey and prior work[4]. The amount of near infrared absorption is related to the presence of a few key components in wheat grains[5]. The aim of this study was to apply chemometric models on NIR spectroscopic data to predict the quantity of protein, moisture, starch and ash which are a few vital features that depicts a characteristic profile of wheat[6]. The quantitative analysis of these parameters can further be used to segregate different varieties of wheat.

Wheat Quality Analysis

Moisture content is a key stricture of any grain[7] which affects its sale weight, quality and ability to be

*Author for Correspondence
E-mail: annefrankjoe@gmail.com

stored long-term. The ability of wheat flour to be processed into different food merchandise is largely determined by the protein content[8]. Mature wheat grains contain 8% to 20% proteins. About 75–85% of the total grain proteins is composed of glutenins and gliadins. In common wheat, these parameters contributes to the elasticity during dough formation[9]. Starch content in wheat grain contains 66–76% carbohydrates[10]. It is not merely an important part of our nutrition but also dictates the textural properties of many food products, wherein baked items can be mentioned in specific[11]. The measure of ash content relates to the grade of flour[12]. Ash content indicates milling performance by indirectly revealing the amount of bran contamination in flour. Ash in flour can affect color, imparting a darker color to finished products.

Materials and methods

20 varieties of wheat grain (Triticum aestivum) cultivars like pbw17, ankurkedar , pbw502, sonalika, to name a few that were grown in different locations of India for a period of two years were collected and stored. Severely wrinkled, discolored, and split grains were manually removed from the sample. Five sample sets from each variety were chosen as carefully selected and representative units, which have in turn been quality profiled using analytical reference standards. A total of 100 spectra were collected from the sample set. The procedure flow chart for the measurement of key constituents of wheat grain is illustrated in Figure1. Prediction models were developed for each quality parameter using PLS, MLR and SNV. Statistics used to determine the ability of NIR to predict specific parameters were the R^2; coefficient of determination[13] and RMSECV; the Root mean squared error of cross-validation which is a statistical measure of the proximity of the data to the fitted regression line.

Partial least square Regression (PLS)

After the first step of preprocessing the spectral signal X, the newly built test data matrix was employed in the PLS calculation using the expression[14]

$$y = Xb + e = \sum_{i=1}^{A}(t_i\, p_i^T) * b + e \qquad \dots (1)$$

Where y is the predicted response, b is the regression coefficient, t_i and p_i are the scores and loadings vectors used for training the model, A is the number of Latent variables which indicates the length of test data and e is the residual vector.

Multiple Linear Regression (MLR)

Linear regression with more than one variable is believed to enhance the estimation of new predictor variable. For the model performance with multiple variables, the cost function to fit the best straight line in the data was calculated from the expression

$$J(\theta_0, \theta_1, \dots., \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \qquad \dots (2)$$

Where m is the amount of features, h is predictive value and y is real world value.

The algorithms that minimized the cost function were gradient descent and normal equations[15].

Support Vector Machine (SVM):

The basic idea behind SVM is to construct a hyper plane that lies in close proximity to as many of the data points as possible[16]. The hyperplane (decision boundary) can be expressed as

$$wT\ x + b = 0 \qquad \dots (3)$$

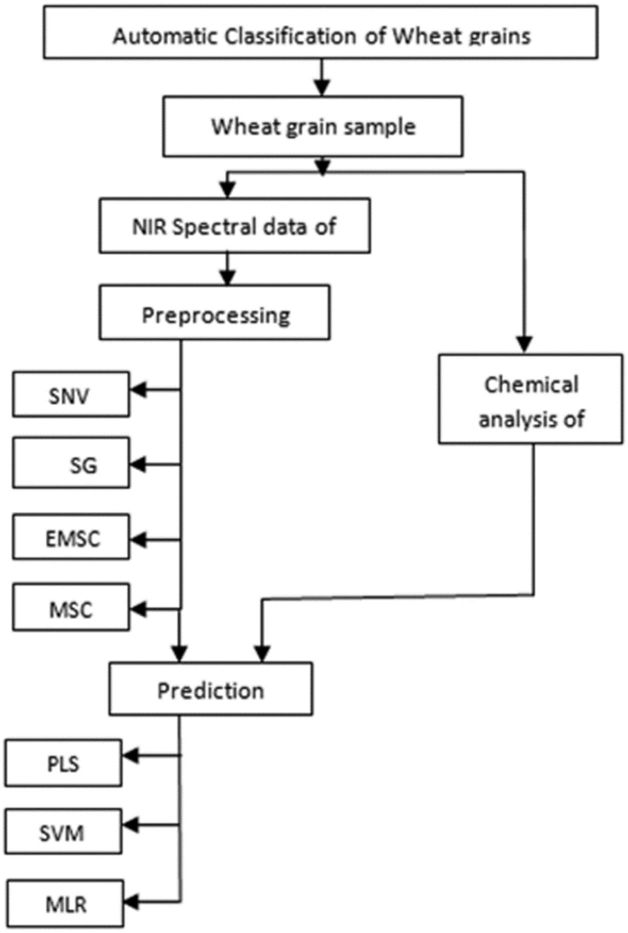

Fig. 1 — The flow diagram of calibration model development

Table1 — The validation result of the wheat grain variety: UP 262 under 12 prediction model combinations of preprocessing and regression model

Prediction model	RMSECV of				R^2 of			
	Moisture	Protein	Starch	Ash	Moisture	Protein	Starch	Ash
SG 1st der with SVM	0.258	0.244	0.369	0.032	0.989	0.989	0.976	0.999
SG 1st der with MLR	0.524	0.302	0.402	0.097	0.961	0.986	0.977	0.998
SG 1st der with PLS	0.299	0.303	0.393	0.097	0.988	0.985	0.978	0.998
SNV with SVM	0.229	0.243	0.367	0.030	0.989	0.989	0.976	0.999
SNV with MLR	1.302	0.511	0.369	0.096	0.781	0.963	0.980	0.998
SNV with PLS	0.290	0.414	0.404	0.095	0.987	0.977	0.978	0.998
MSC with SVM	0.260	0.265	0.369	0.029	0.989	0.988	0.976	0.999
MSC with MLR	0.531	0.318	0.455	0.096	0.961	0.982	0.967	0.998
MSC with PLS	0.336	0.347	0.426	0.096	0.982	0.977	0.969	0.998
EMSC with SVM	0.263	0.286	0.369	0.030	0.989	0.987	0.975	0.999
EMSC with MLR	0.341	0.315	0.402	0.097	0.977	0.985	0.976	0.998
EMSC with PLS	0.289	0.298	0.405	0.097	0.988	0.987	0.977	0.999

Table 2 — RMSECV and R^2 of the different prediction models for the four constituents under consideration namely Moisture, Protein, Starch and Ash

Preprocess + Prediction	→	SG + SVM	SG + MLR	SG + PLS	SNV + SVM	SNV+ MLR	SNV + PLS	MSC + SVM	MSC + MLR	MSC+ PLS	EMSC + SVM	EMSC + MLR	EMSC + PLS
Constituents ↓	Chemical result	Predicted result											
Protein	13	12.85	12.87	12.89	12.61	12.59	12.65	12.85	12.93	12.99	12.62	12.73	12.8
Carbohydrate	70.18	71.05	70.79	70.84	71.05	70.72	70.68	71.06	71.05	71.05	71.05	70.9	70.87
Moisture	10.4	10.28	9.76	10.39	10.3	10.54	10.47	10.27	9.72	10.2	10.27	10.23	10.33
Ash	1.67	1.59	1.6	1.59	1.63	1.6	1.59	1.58	1.6	1.59	1.56	1.6	1.59

Where x is the input vector, w and b are coefficients. Wheat spectrum is not a perfectly linear separable data, thus mapping it into a hyperplane was formulated as

$$\min_{w,b,\varepsilon} \frac{w^T w}{2} + C\sum_i \varepsilon_i, \qquad \ldots (4)$$

While yi(wT xi + b)≥1-ε_i and $\varepsilon_i \geq 0$ determines the trade-off between the flatness of the function and the tolerable signal error, ε - the slack variables introduced as $\varepsilon_i \geq 0$ for each xi. The prediction results produced with SVM on NIR signals were greatly reassuring.

Results and Discussion

Savitzky-Golay first-derivative (SG), Standard normal transformation (SNV), Multiplicative scattering correction (MSC), extended multiplicative signal correction (EMSC) were the different preprocessing methods applied on the raw spectrum. These preprocesses were combined with different prediction models like SVM, PLS and MLR. Accuracy, defined as the conformance of a measured value and the true value, in the case of the NIRS method, it is expressed as a measure of nearness of the NIRS predicted value with the reference (wet chemistry method). SG, SNV, MSC and EMSC were the different preprocessing methods applied on the spectrum. These preprocesses were combined with different prediction models like SVM, PLS and MLR. Twelve different models from these combinations were validated. The validation result of the wheat grain variety: UP 262 under all the twelve prediction model combinations is displayed in the Table 1. 60 spectral signals of the calibration model were worked with all the 12 prediction models and the Model performance was defined in terms of the multivariate coefficient of determination (R^2), the standard error of prediction between the measured and modeled values as listed in Table 2. All the preprocessing and regression model combinations performed confidently well with the RMSECV values in the range of 0.091 to 1.3 and R^2 coefficient values in the range of 0.78 to 0.998. EMSC provided the best prediction ability while SVM displayed an improved performance on nonlinear data, thus an overall enhanced prediction capability was observed when EMSC was combined with SVM. The success is presumably due to the ability of spectral modeling to separate chemical light- absorbance and physical light scatter effects

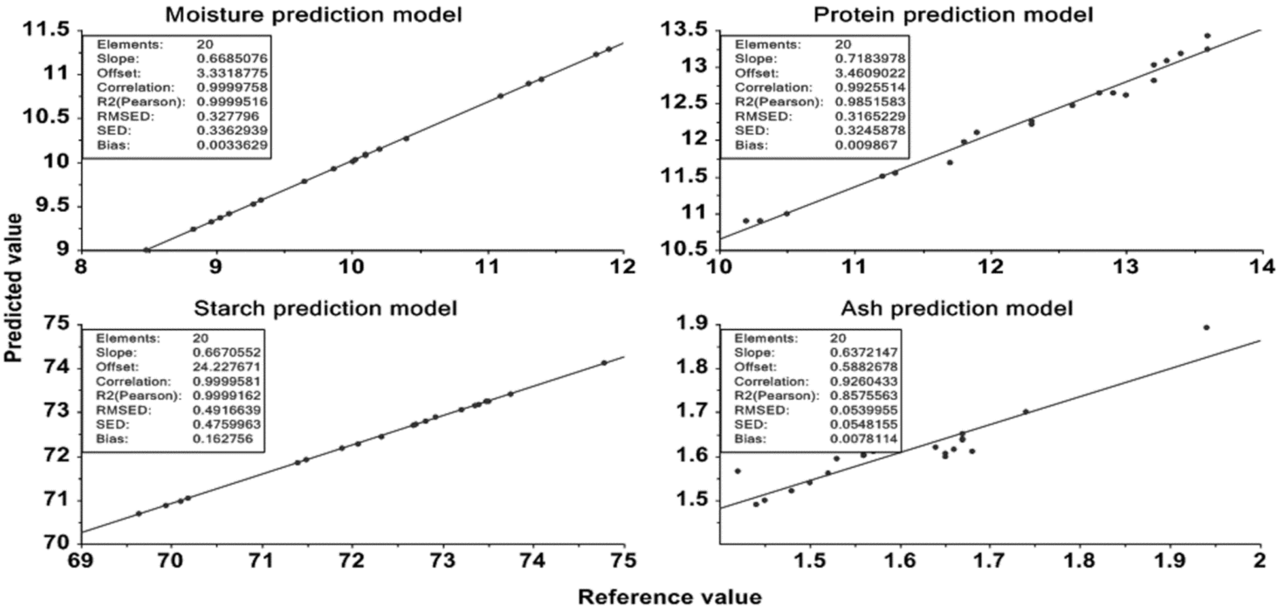

Fig. 2 — The prediction of moisture, protein, starch and ash content using SVM and EMSC.

effectively. Figure 2 shows the plot of reference value against the predicted value of all the 20 wheat varieties.

Conclusion

Chemometric methods were applied in the current study to extract information from the spectra for the total Moisture, Protein, Starch and Ash by analyzing the influence of absorption at particular wavelength of the wheat grain samples. Protein rich varieties are more suitable for baking products, high desirable constituents of wheat grain can be chosen as breeding cultivars to produce higher nutritional values; such are the outcomes of the quantitative analysis of the grain composition. Calibration development and evaluation were performed using MLR, PLS and SVM regression techniques wherein the R^2 and RMSECV values can be fine tuned with a larger sample number. Also many more constituents of the wheat grain can be predicted to provide a complete profiling of the wheat sample. Through the convincing results obtained from the quantitative analysis the corresponding spectral wavelength fingerprinting can be a confidently marked for future work.

References

1 García-Molina M D, García-Olmo J & Barro F, Effective Identification of Low-Gliadin Wheat Lines by Near Infrared Spectroscopy (NIRS): Implications for the Development and Analysis of Foodstuffs Suitable for Celiac Patients, *PloS one*, ***11***(2016), e0152292.

2 Caporaso N, Whitworth M B & Fisk I D, Near-Infrared spectroscopy and hyper spectral imaging for non-destructive quality assessment of cereal grains, *Appl Spectrosc Rev*, ***53*** (2018) 667-687.

3 Pojić M, Mastilović J & Majcen N, The application of near infrared spectroscopy in wheat quality control, in *Infrared spectroscopy-Life and biomedical sciences*, Intech Open, 2012, 167-184.

4 Joe A A F & Gopal A, Identification of spectral regions of the key components in the near infrared spectrum of wheat grain, in *IEEE International Conference on Circuit, Power and Computing Technologies, April* 2017,1-5.

5 Ye D, Sun L, Zou B, Tan W, Zhang D & Che W, Extraction of characteristic spectral bands of wet gluten in wheat based on NIR, in *IEEE 2nd International Conference on Cloud Computing and Internet of Things*, 2016, 105-110.

6 Prajapati Uma, Asrey Ram, Arora A, K Singh A and Hasan M, Differential Effects of Abscisic Acid and Fluridone on Postharvest Quality Parameters of Detached Capsicum (Capsicum Annuum L.) Fruits, *J Sci Ind Res*, **78** (2019) 242-247.

7 Caporaso N, Whitworth M B, Grebby S & Fisk I D, Rapid prediction of single green coffee bean moisture and lipid content by hyper spectral imaging, *J Food Eng*, **227** (2018) 18-29.

8 Ingle P D, Christian R, Purohit P, Zarraga V, Handley E, Freel K & Abdo S, Determination of protein content by NIR spectroscopy in protein powder mix products, *J AOAC Int*, ***99*** (2016) 360-363.

9 Žilić S, Barać M, Pešić M, Dodig D & Ignjatović-Micić D, Characterization of proteins from grain of different bread and durum wheat genotypes, *Int J Mol Sci*, ***12*** (2011) 5878-5894.

10 Bala S, Asthir B & Bains N S, Heat and drought stress responses alter grain characteristics by impeding starch precursors of wheat, *IJEB*, **56** (2018) 565-572.

11 Koehler P, Wieser H, Handbook on Sourdough Biotechnology, in *Chemistry of Cereal Grains*, edited by

Gobbetti M & Gänzle M (Springer, Boston, MA) 2013, 11-45.

12 Dong X & Sun X, A case study of characteristic bands selection in near-infrared spectroscopy: nondestructive detection of ash and moisture in wheat flour, *J Food Meas Char,* ***7*** (2013) 141-148.

13 Bai X, Zhou T, Lai T & Zhang H. Optimization of the Microwave-Assisted Extraction of Poly phenols from Red Pitaya Peel using Response Surface Methodology, *J Sci Ind Res*, **77** (2018) 419-424.

14 Deng B C, Yun Y H, Liang Y Z, Cao D S, Xu Q S, Yi L Z & Huang X, A new strategy to prevent over-fitting in partial least squares models based on model population analysis, *Anal Chim Acta*, ***880*** (2015) 32-41.

15 Sharma N, Dwivedi A, Singh A & Srivastava A K, Quantitative structure activity relationship study based molecular modeling of 4-aminoquinazoline derivatives for Aurora kinase inhibition, *Indian J. Chem., Sect B,* **57B** (2018), 1421-1429.

16 Liu T A, Zhai X Y, Zhang Q, Lv W & Lu W C, Using NIR with Support Vector Regression to Predict the Crude Protein of Alfalfa, in *IEEE Int Conf Inf Syst Artif Intell*, June 2016, 415-418.