MULTIVARIATE ANALYZING AND ARTIFICIAL NEURAL NETWORKS FOR PREDICTION OF PROTEIN CONTENT IN WINTER WHEAT USING SPECTRAL CHARACTERISTICS

Ass. Prpf. Dr. Rasooli Sharabiani V.[1], PhD. Stu. Soltani A.[1], Prof. Dr. Noguchi N.[2]

Department of Biosystem Engineering, Faculty of Agriculture and Natural Resources, University of Mohaghrgh Ardabili, Ardabil, IRAN[1]
Laboratory of Vehicle Robotics, Graduate School of Agricultural Engineering, Hokkaido University, Sapporo, JAPAN[2]
Email: vrasooli@uma.ac.ir, arazsoltani@yahoo.com, noguchi@cen.agr.hokudai.ac.jp,

Abstract: *This study aimed to predict the protein content(PC) and canopy spectra in winter wheat were measured based on field test. Key spectral bands were chosen by principal component analysis (PCA) method, and the predicted models were built by Partial Least Squares Regression (PLSR) and Artificial Neural Network (ANN). The performance of the feed forward and cascade forward ANNs was compared with those of PLS regression models using root mean square error (RMSE) and the correlation coefficient (R^2). The finest consequence by CFBP was related to topology of 8-8-1 with Levenberg-Marquardt (LM) algorithm, threshold function of TANSIG-TANSIG-PURELIN and the initial strategy. This arrangement output was RMSE=0.0289 and R2=0.9881 at 14 epochs. The consequences of estimate for correlation values of PLSR model was 0.9783. The results of prediction for the two models were in order of ANN > PLSR with correlation values of 0.9881 and 0.9783, respectively. Therefore, NIRS shows the potential for predicting protein content with accuracies suitable for process control.*

KEYWORDS: PROTEIN CONTENT, MULTIVARIATE ANALYZIN, NEURAL NETWORKS, WINTER WHEAT

1. Introduction

The protein content in winter wheat is essential for safety of society [1,23]. Precise and suitable prediction of the PC of winter wheat can aid agriculturists create correct choices about manure application, winter wheat diversity choice, and crop assortment [2, 24]. There are typical calculation techniques for protein parameter [2]. These techniques are usually hard and lengthy, and cannot be applied to quickly calculate quality features. NIRS method is to acquisition correlation between spectral features by creating the model between the reference values and spectral data [20]. As a non-destructive technique that protects the sample's integrity and provides the calculation of a great number of characteristics applying an only sample also reduces the price of typical examinations [3], it has been a replacement to typical chemical methods and extensively applied in food investigations [4,5]. In the quantitative analysis, choice of modeling techniques performs a vital role as diverse arranger will generate diverse results when they act on the similar datasets. PLSR is a usual technique that has some relation to PCA, put in spectroscopy [6], and has as well as generated considerable consequences. N. Shetty et al. [7] applied NIRS mixed with PLSR technique to estimate amount of fructan in vegetables, and acquired a model with extremely estimation capability, root mean square error of prediction (RMSEP) and R_2 were 1.13 and 0.93 respectively. Millar established NIRS standardizations using NIR spectra from U.K. and French winter wheat and displayed possibilities for guessing protein and moisture amount, water absorption, and flour color, but had negligible consequences when trying to guess loaf volume and crumb grain score [8]. Y. Liu et al. [9] demonstrated PLSR to be a high potent instrument for the calculation of soluble solid content of orange, the technique terminated in correlation coefficient, RMSEP, average difference between predicted and measured values of 0.90, 0.68° Brix and 0.16° Brix, respectively. Sissons et al [10] applied NIR spectra from grain to guess seed, flour, and dough features. Their consequences displayed possibilities for classification samples into little, average, and great groups for check weight, thousand kernel weights, semolina yield, semolina yellow color, semolina browning, grain hardness, and cooked pasta firmness. To evaluate the procedures presented in this paper, the relationship between protein chemical content and spectral data in winter wheat samples was studied using the PLS multivariable regression modeling. For this goal, firstly, outlier samples were recognized and removed in terms of spectral data using the PCA technique. standard normal variate (SNV) were used to correct both multiplicative and additive effects of the spectra. To rise the spectral resolution, first and second derivatives of the spectra (D1, D2) based on the Savitzky–Golay algorithm with five smoothing points and polynomial order of 2, were also performed. For prediction of PC, nonlinear models are extra appropriate because of suppleness and nonlinear behavior of natural crops. Furthermore, numerous making procedures include with fluctuations in procedure circumstances and depend on the ability and knowledge of operators. So, artificial neural network (ANN) models recently have obtained motion for modeling and control processes. ANN models are known as fine tools for dynamic modeling since they do not need parameters of physical models [11]. Such models are capable to learn the explanation of problems from a set of experiential data, and to use compound systems with nonlinearities and interplay among decision variables [12]. ANN models can be categorized into two groups: supervised networks and unsupervised networks. Supervised networks require a training algorithm and a training data anthology to regulate the connection weights, while unsupervised networks can adjust weights by themselves to attain the essential consequences without using any training algorithm. Supervised networks are mainly applied for grouping, estimate, and function approximation processes. Unsupervised networks are appropriate for clustering of patterns.

2. Materials and Methods

2.1. Experimental field

In order to obtain actual data, an experimental winter wheat field was cultivated with a conventional variety in three consecutive years. A Dimensions of this field were 40 m x 120 m. The field was divided into 8 areas (blocks). Reflectance data by using a Spectroradiometer (Field Spec 3) during of growth season, and protein content of wheat grain was measured after harvesting from target points as reference area which was randomly set in the field.

2.2. Spectroradiometer

In this study a high-resolution spectroradiometer, FieldSpec®3 (FS3) (Analytical Spectral Devices, Inc., USA) was used to take the spectral data. FS3 is a hyper spectral sensor designed to collect data on solar radiance, irradiance, and reflectance. It can measure the spectral reflectivity within the range of 350 nm to 2500 nm with sampling intervals of 1.4 nm and 2 nm in the ranges of 350-1050 nm and 1000-2500 nm, respectively (2150 wavelengths in total).

2.3. Protein Determination

The protein content was measured by Kjeldahl technique according to the relevant standard of GB/T 5009.5-1985. The protein for 40 rows of winter wheat samples which ranged from 9.4395% to 16.9796% had representative significance for this study.

2.4. Data analysis

Before creating the grouping models, outliers which having a harmful effect on modelling [13] were discovered and deleted. for this purpose, the spectral variation of all the samples was evaluated using PCA and the outlier samples were calculated as the points outside the normal range of variability in the PCA scores plot [21]. After eliminating the outliers (4 samples), all the 36 remnant samples were prepared in order of their protein content. Then, 25% of the samples was selected for the external validation set, and 75% of the samples were selected for the calibration set. After key spectral bands were chosen by the PCA technique, the anticipated models were constructed by PLSR model. For PLSR model, spectra were imported into Unscramble x10 software (CAMO, Oslo, Norway). The accuracy of regression model was evaluated by root mean square error and correlation coefficient (R^2). R^2 changes from 0 to 1, and the closer it is to 1, the superior the model influence, while the model's performance will be better when the RMSE closer to 0. These indices were applied to assess the ability of the model in predicting samples, intended as the reliability of the quality parameters estimation [22].

2.5. Artificial neural networks modeling

The topology of the developed ANN is shown in Fig. 1. The ANN model was implemented using neural network toolbox of Matlab software. The network was simulated based on a multi-layer feed-forward and cascade forward algorithm. The variables of wavelengths were contemplated as the inputs (first four principal components scores were put as input variables), whereas one parameter of protein was applied as the outputs. The number of neurons in the primary and second hidden layers varied from 2 to 15. Each data set was divided into two groups, consisting of 70 % for training and 30 % for test. In this paper, the total data of wavelengths from 350-1350 nm for artificial neural networks were used.

In this study, to gain the best network, several numbers of multilayer feed-forward back propagation (FFBP) and cascade forward back propagation (CFBP) were made and tested with different number of hidden layers (1 and 2) and neurons. Two types of multilayer perceptron (MLP), namely feed forward back propagation (FFBP) and cascade forward back propagation (CFBP) were applied to prediction the winter wheat's protein. Two learning algorithms of BFGS Quasi-Newton and Levenberg-Marquardt (LM) were also applied for finalization of ANN models. FFBP and CFBP include of one input layer, one or two hidden layers and one output layer [19]. For training this structure, the back propagation (BP) algorithm was used. In the informing process, the weight coefficients were restructured with learning rules. During network training, calculations were managed from input toward output layers of the network and error values were propagated to previous layers. CFBP is similar to FFBP in weights updating, but the important diversity between these networks is in the relationship between neurons [15]. Three transfer functions, defined in the next equations, were engaged to reach the enhanced network arrangement [16, 17]:

$$Y_j = X_j \quad \text{(PURELIN)} \quad (1)$$

$$Y_j = \frac{2}{(1+\exp(2X_j))-1} \quad \text{(TANSIG)} \quad (2)$$

$$Y_j = \frac{1}{1+\exp(-X_j)} \quad \text{(LOGSIG)} \quad (3)$$

where X_j is computed as follow:

$$X_j = \sum_{i=1}^{m} W_{ij} * Y_i + b_j \quad (4)$$

where m is the number of output layer neurons, W_{ij} is the weight of between ith and jth layers, Y_i is the jth output neuron, X_j is the jth input neuron, b_j is the bias of jth neuron for FFBP and CFBP networks.

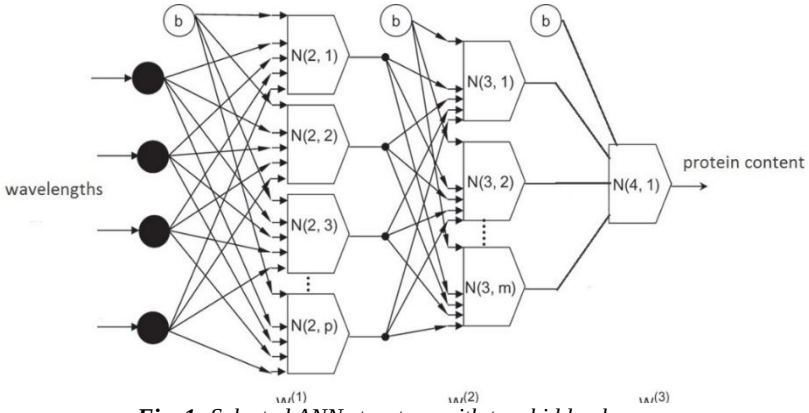

Fig. 1: *Selected ANN structure with two hidden layers.*

3. Result and Discussion

3.1. PLSR model results

The canopy in winter wheat showed a diverse raw reflectance in the 520–680 nm and 720–1000 nm spectral regions (Fig. 2). For a crop canopy, reflectance is little amid the 480- and 680- nm state because of the powerful absorption by chlorophylls and other pigments, but is high in the near infrared state because of the microcellular constructions in leaf material and canopy constructions [18]. In this paper, our consequences presented that there were two clear band ranges amid 520–680 nm and 720–1000 nm detected (Fig. 2). Alike consequences were informed in earlier papers [19]. Since unnecessary spectral bands and noise interference influence accuracy of regression model, principal component analysis (PCA) technique was applied to decrease main input factors dimension. The number of variables can be decreased by eliminating the lower-level components without any important damage of information included in the raw data set by PCA. After principal components were chosen by PCA, the anticipated model was constructed by partial least squares regression (PLSR) technique. For this purpose, various preprocessing methods include standard normal variate (SNV), first and second derivatives of the spectra (D1, D2) based on the Savitzky–Golay algorithm and their combinations were used. In order to investigate the effect of these different methods on the accuracy of the developed models, a multivariate modeling for spectra was performed without any pre-processing. After development of PLS models, the evaluation of the models was done by a full cross-validation method.

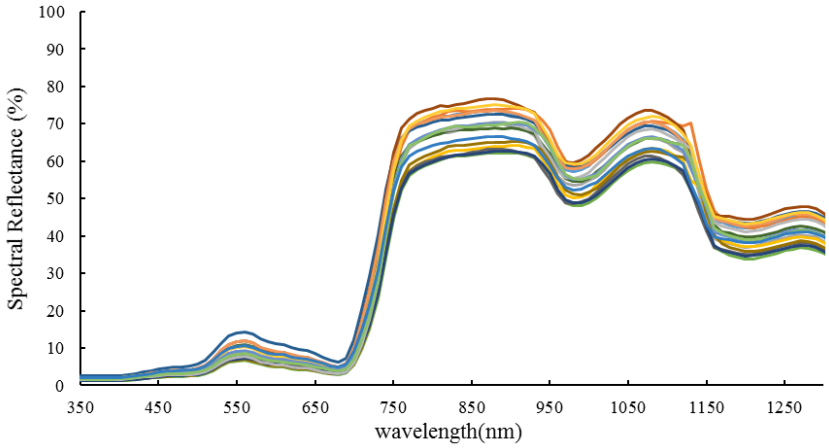

Fig. 2: *Spectral reflectance of canopy in winter wheat changing with wavelength*

In order to determine the best predictive model of protein, regression coefficients and standard error predictions of models were calculated and compared. In order to more study the characteristics of uncertain samples and specify whether they were outlier samples, the effect of each of them on the efficiency of the model were investigated. For this purpose, initially, the result of model removing any one uncertain sample with that of no removing were compared (Table 1). The amounts of R_2 were 0.9773, 0.9746, 0.9808, and 0.9828 respectively after removing No. 7, No. 15, No. 24 and No. 39 respectively, each was higher than the R_2 (0.9735) which flowed from the state of no removing; moreover, contrasted to the RMSE (0.2384) attained in the case of no-removing, the RMSE amounts decreased to 0.2194, 0.2223, 0.2010, 0.1899, respectively.

Table 1 *The effects of No. 7, No. 15, No. 24 and No. 39 samples on models.*

Deleted samples	R^2	RMSE
No deleting	0.9735	0.2384
No. 7	0.9773	0.2194
No. 15	0.9746	0.2223
No. 24	0.9808	0.2010
No. 39	0.9828	0.1899
No. 24, No. 39	0.9877	0.1590
No. 15, No. 39	0.9835	0.1856
No. 15, No. 24	0.9827	0.1905
No. 7, No. 39	0.9857	0.1726
No. 7, No. 24	0.9838	0.1845
No. 7, No. 15	0.9782	0.2155
No. 7, No. 15, No. 39	0.9861	0.1704
No. 7, No. 15, No. 24	0.9854	0.1746
No. 7, No. 24, No. 39	0.9896	0.1458
No. 15, No. 24, No. 39	0.9889	0.1501
No. 7, No. 15, No. 24, No. 39	**0.9907**	**0.1374**

As said before, S-G, SNV, D_1, D_2, and different compositions of them were used to process the main data and the assessment elements of relative PLSR models under diverse pre-processing in this paper are shown in Table 2. The PLSR models for PC designation displayed in Table 3 had R_2 values 0.9587–0.9783 and the confine of RMSE values were 0.4333 to 0.1457. Investigating the consequences attained from all the pre-processing procedures, the usage of the SG+D1+SNV formed outstanding consequences where the RMSE value reduced importantly in contrast with the direct regression model on raw spectra, which decreased from 0.2670 to 0.1457. In the similar instance, the R^2 was 0.9783, where the number of PLSR components was five. Fig. 3 displays the investigational data of PC against the anticipated values of PC.

Table 2: *Prediction results for protein content in wheat using different pre-processing methods.*

Pre-treatments	R^2	RMSE	The number of components
Raw	0.9724	0.2670	7
D_1	0.9688	0.3185	5
D_2	0.9664	0.3488	5
SNV	0.9589	0.4304	6
S-G	0.9723	0.2676	7
D_1+SNV	0.9782	0.1486	5
D_2+SNV	0.9726	0.2637	4
S-G+SNV	0.9587	0.4333	6
S-G+D_1	0.9694	0.3120	5
S-G+D_2	0.9704	0.2957	8
S-G+D_1+SNV	**0.9783**	**0.1457**	**5**
S-G+D_2+SNV	0.9750	0.2206	4

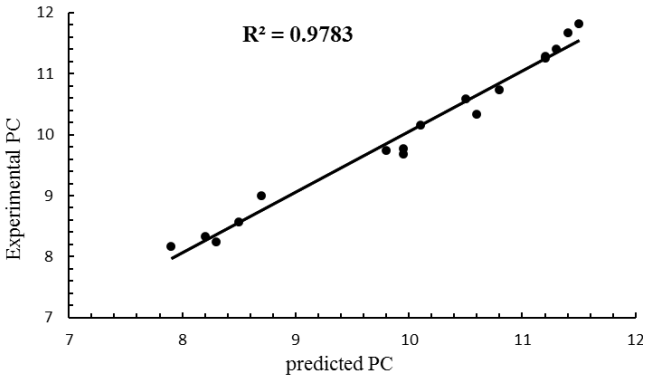

Fig. 3: *Predicted values of PC against experimental values for testing data set.*

3.2. Artificial Neural Network Results

In the training process, the data were applied to calculate the best number of hidden layers and neurons per hidden layer to supply the finest consequences. ANN structures were tested by 1 and 2 hidden layers with 2–15 neurons per each hidden layer. The models error was calculated after each simulation. To study the influence of diverse threshold functions on FFBP and CFBP outputs, two strategies of similar and various threshold functions for all of the layers were used. The predicted PC has shown in Table 3. Both strategies, also learning algorithms of LM and BFGS, were applied for training of FFBP and CFBP networks. Some topologies were chosen as the finest consequences from each network, threshold function and training algorithms. Consequences of model's performance indices (R^2 and RMSE) for all randomly chosen datasets based on testing are shown in Table 3.

Table 3: *Best selected topologies including training algorithm, different layers and neurons for FFBP and CFBP for PC.*

Network	Training algorithm	Threshold function	Number of layers and neurons	RMSE	R^2	Epoch
FFBP	**LM**	TANSIG-TANSIG-TANSIG	5-5-1	0.0385	0.9785	10
		LOGSIG-TANSIG-TANSIG	**10-5-1**	**0.0298**	**0.9879**	**9**
	BFG	TANSIG-TANSIG-TANSIG	5-5-1	0.0374	0.9810	10
		TANSIG-TANSIG-PURLIN	8-8-1	0.0342	0.9838	10
CFBP	**LM**	TANSIG-TANSIG-TANSIG	10-10-1	0.0361	0.9828	12
		TANSIG-TANSIG-PURLIN	**8-8-1**	**0.0289**	**0.9881**	**14**
	BFG	TANSIG-TANSIG-TANSIG	5-5-1	0.0319	0.9874	10
		TANSIG-LOGSIG-PURLIN	15-10-1	0.0322	0.9862	12

The finest consequences attained by feed forward back propagation (FFBP) in prediction of PC was related to 10-5-1 topology and LOGSIG-TANSIG-TANSIG threshold function with LM algorithm in the initial strategy. This structure produced RMSE=0.0298 and R_2=0.9879 with 9 epochs. The finest consequence for the next strategy of FFBP with BFG algorithm was assigned to 8-8-1 topology and threshold functions of TANSIG-TANSIG-PURELIN. This structure created RMSE=0.0342 and R^2=0.9838 for PC. The finest consequence by CFBP was related to topology of 8-8-1 with LM algorithm, threshold function of TANSIG-TANSIG-PURELIN and the initial strategy. This arrangement output was RMSE=0.0289 and R^2=0.9881 at 14 epochs. CFBP with the second strategy, BFG algorithm, topology of 5-5-1 and threshold functions of TANSIG-TANSIG-TANSIG introduced the output of RMSE=0.0319 and R^2=0.9874. Fig. 4 displays the investigational data of PC against the anticipated values PC.

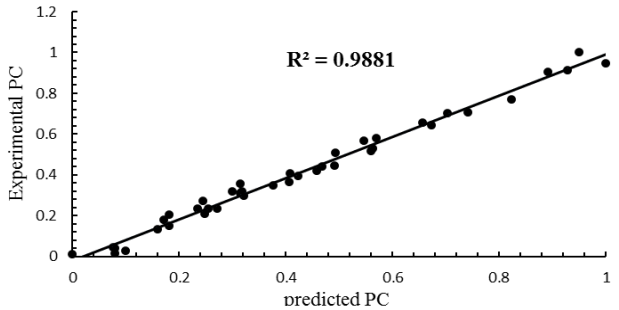

Fig. 4: *Predicted values of PC using artificial neural networks against experimental values for testing data set.*

4.Conclusion

In this paper, numerous ANN models and PLSR models were established to predict protein content of winter wheat grain. The consequences showed that the advanced models were extra appropriate. The consequences showed the capability of advanced ANN models and PLSR models to calculate and estimate the protein content in winter wheat. The best result for PLSR model was obtained using the SG+D1+SNV pre-processing method where the RMSE value reduced importantly in contrast with the direct regression model on raw spectra, which decreased from 0.2670 to 0.1457. In the similar instance, the R_2 was 0.9783, where the number of PLSR components was five. Also eliminating 4 variables gave us the best result. The finest consequences attained by feed forward back propagation (FFBP) in prediction of PC was related to 10-5-1 topology and LOGSIG-TANSIG-TANSIG threshold function with LM algorithm in the initial strategy. This structure produced RMSE=0.0298 and R2=0.9879 with 9 epochs. The finest consequence by CFBP was related to topology of 8-8-1 with LM algorithm, threshold function of TANSIG-TANSIG-PURELIN and the initial strategy. This arrangement output was RMSE=0.0289 and R2=0.9881 at 14 epochs. According to the results, it was found that using different models of artificial neural network provides better results in predicting the amount of winter wheat protein in comparison with partial least squares regression model. Generally, this study has importantly enhanced the predictive efficiency of the model by enhancing data analysis techniques, which enhanced a non-damaging, rapid, and accurate technique for online calculation of PC in winter wheat.

5. Reference

[1]. Tang Y.L., Huang J.F., Wang R.C. (2004): Study on estimating the contents of crude protein and crude starch in rice panicle and paddy by hyperspectral. Scientia Agricultura Sinica, 37: 1282–1287.

[2]. Diker K., Bausch W.C. (2003): Potential use of nitrogen reflectance index to estimate plant parameters and yield of maize. Biosystems Engineering, 85: 437–447.

[3] H.L. Galasso, M.D. Callier, D. Bastianelli, J.P. Blancheton, C. Aliaume, the potential of near infrared spectroscopy (NIRS) to measure the chemical composition of aquaculture solid waste, Aquaculture 476 (2017) 134–140.

[4] L.S. Magwaza, S.I.M. Naidoo, S.M. Laurie, M.D. Laing, H. Shimelis, Development of NIRS models for rapid quantification of protein content in sweetpotato [Ipomoea batatas (L.) LAM.], LWT, Food Science and Technology 72 (2016) 63–70.

[5] T.B. Bagchi, S. Sharma, K. Chattopadhyay, Development of NIRS models to predict protein and amylose content of brown rice and proximate compositions of rice bran, Food Chem. 191 (2016) 21–27.

[6] S.A. Moreira, J. Sarraguça, D.F. Saraiva, R. Carvalho, J.A. Lopes, Optimization of NIR spectroscopy based PLSR models for critical properties of vegetable oils used in biodiesel production, Fuel 150 (2015) 697–704.
[7] N. Shetty, R. Gislum, Quantification of fructan concentration in grasses using NIR spectroscopy and PLSR, Field Crop Res 120 (2011) 31–37.
[8] Millar, S. J. 2003. The development of near-infrared (NIR) spectroscopy calibrations for the prediction of wheat and flour quality. The Home-Grown Cereals Authority Project Report No. 310. HGCA: London.
[9] Y. Liu, X. Sun, A. Ouyang, Nondestructive measurement of soluble solid content of navel orange fruit by visible–NIR spectrometric technique with PLSR and PCABPNN, LWT Food Sci. Technol. 43 (2010) 602–660.
[10]. Sissons, M., Osborne, B., and Sissons, S. 2006. Application of near infrared reflectance spectroscopy to a durum wheat breeding programme. J. Near Infrared Spectrosc. 14:17-25.
[11] Amiri Chayjan R, Kaveh M. Physical parameters and kinetic modeling of fix and fluid bed drying of terebinth seeds. J Food Process Preserv. 2014;38:1307–20.
[12] Yousefi G, Emam-Djomeh Z, Omid M, Askari GR. Prediction of physicochemical properties of raspberry dried by microwave-assisted fluidized bed dryer using artificial neural network. Drying Technol. 2014; 32:4–12.
[13]. Heise HM, Winzen R. 2006. Chemometrics in near-infrared spectroscopy. In: Siesler HW, Ozaki Y, Kawata S, Heise HM, editors. Near-infrared spectroscopy: principles, instruments, applications. Germany: Wiley-VCH; p. 125–162.
[14] Demuth H, Beale M, Hagan M. Neural network toolbox 5. Natick, MA, USA: The MathWorks; 2007.
[15] Amiri Chayjan R, Salari K, Barikloo H. Modelling moisture diffusivity of pomegranate seed cultivars under fixed, semi fluidized and fluidized bed using mathematical and neural network methods. Acta Sci Polym Technol Aliment. 2012;11(2):137–49.
[16] Aghajani N, Kashaninejad M, Dehghani AA, Garmakhany AD. Comparison between artificial neural networks and mathematical models for moisture ratio estimation in two varieties of green malt. Qual Assur Saf Crop Food. 2012; 4:93–101.
[17] Kaveh M, Amiri Chayjan R. Mathematical and neural network modelling of terebinth fruit under fluidized bed drying. Res Agr Eng. 2015;61(2):55–65.
[18]. Thomas J.R., Oerther G.F. (1972): Estimating nitrogen content of sweet pepper leaves by reflectance measurements. Agronomy Journal, 64: 11–13.
[19]. Li Y., Zhu Y., Tian Y., Yao X., Zhou C., Cao W. (2006): Quantitative relationship between leaf nitrogen accumulation and canopy reflectance spectral. Scientia Agricultura Sinica, 32: 203–209.
[20] L.S. Magwaza, S.I.M. Naidoo, S.M. Laurie, M.D. Laing, H. Shimelis, Development of NIRS models for rapid quantification of protein content in sweetpotato [Ipomoea batatas (L.) LAM.], LWT, Food Science and Technology 72 (2016) 63–70.
[21] M. Zhang, S. Zhang, J. Iqbal, Key wavelengths selection from near infrared spectra using Monte Carlo sampling–recursive partial least squares, Chemom. Intell. Lab. Syst. 128 (2013) 17–24.
[22] K. Haddad, A. Rahman, M.A. Zaman, S. Shrestha, Applicability of Monte Carlo cross validation technique for model development and validation using generalized least squares regression, J. Hydrol. 482 (2013) 119–128.
[23]. Vali RASOOLI SHARABIAN, Noboru NOGUCHI and Kazunobu ISHI, Optimal Vegetation Indices for Winter Wheat Growth Status Based on Multi-Spectral Reflectance, Environ. Control Biol., 51 (3), 105112, 2013
[24]. Vali Rasooli Sharabian, Noboru Noguchi, Kazunobu Ishi, Significant wavelengths for prediction of winter wheat growth status and grain yield using multivariate analysis, Engineering in Agriculture, Environment and Food 7 (2014) 14-21.