BCSIR

Available online at www.banglajol.info

Bangladesh J. Sci. Ind. Res. **51(4)**, 253-260, 2016

BANGLADESH JOURNAL OF SCIENTIFIC AND INDUSTRIAL RESEARCH

E-mail: bjsir07@gmail.com

Simple method for classification of mango varieties on the basis of their physico-chemical properties by chemometric techniques

M. N. Uddin[1]*, R. Ara[2,4], M. Motalab[3], A. N. M Fakhruddin[4] and B. K. Saha[3]

[1]*BCSIR Laboratories Dhaka, Bangladesh Council of Scientific and Industrial Research (BCSIR), Dhaka-1205, Bangladesh*

[2]*Department of Food Engineering and Tea Technology, Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh*

[3]*Institute of Food Science and Technology (IFST), Bangladesh Council of Scientific and Industrial Research (BCSIR), Dhaka-1205, Bangladesh*

[4]*Department of Environment Science, Jahangirnagar University, Dhaka-1342, Bangladesh*

Abstract

Nutritional parameters vary significantly among different varieties of mango. It is therefore necessary to have simple method to classify mango pulps according to their nutritional parameters by effective and economic technique. The present study was carried out to develop a method to classify mango varieties by using two chemometric techniques namely, Artificial Neural Network (ANN) and Linear Discrimination Analysis (LDA). At first, 9 physico-chemical parameters have been chosen from 18 of them by applying Factor Analysis (FA), as quantification of each parameter involves time and cost. Nine varieties of mango available in Bangladesh were studied her. LDA can classify the mango pulps on the basis of their nutritional properties 100 percent accurately, and this rate for ANN is 96.3 percent. Therefore, a method is being proposed where mango can be classified with their 9 physico-chemical parameters into their right varieties by LDA. The proposed chemometric method could be used for regular classification of mango pulps at laboratory and mango product manufacturing industries to improve the quality of mango products.

Keywords : Mango varieties; Artificial neural network; Linear discriminant analysis

Introduction

Like many parts of the world, mangoes (*Mangifera indica* L.) is the most popular fruit in Bangladesh due to its luscious taste, availability and exotic varieties. It is considered as the `king of fruits' here. The more common and widely cultivated varieties of mangoes are Langra, Gopalbhogh, Himsagor, Kheersapat, Tosha, Mallika, Aamrupali, Kohitor, Fazlee, Langra, Ashwina, BARI Aam-1, Kalia, Ranipassand, and many others (Kobra *et al.*, 2012).

Mango is one of the economic agricultural products in Bangladesh. After fulfilling local demand many mangoes and mango products are exported to different parts of the world for their taste and nutritional values. Mango products like mango juice, mango pulp, pickle, etc are normally produced to get the taste of mango round the year. Nutritional values, taste and other characteristics of different varieties of mango available in Bangladesh vary significantly (Shafqat *et al.,* 1992; Jilani *et al,* 2010; Ara e*t al.*, 2014).

Consumers want those varieties of mangoes and their products which are rich in taste and nutritional values. Besides, mango juice and similar product manufacturing industries need to classify their raw pulp to their true varieties on regular basis as they want to enrich the quality of their products with most nutritious and testy varieties of mangoes. However, identifying a mango juice or any other product according to its original varieties is not possible by the existing methods. So, the present study was carried out and proposed a method for classification of mango to the varieties on the basis of some selected physic-chemical parameters and chemometric techniques.

Chemometrics, is the most general sense, is the art of processing data with various numerical techniques in order to extract useful information (Kramer, 1998). One of the most components of chemometrics is classification of individuals to a certain group according to its probability to fall in. In

*Corresponding author e-mail: m2nashir@yahoo.com other words, classification means to assign an individual (sample) to one or more categories based on a set of measurements used to describe or characterize the object itself (Bro, 2013). In analytical chemistry it is very often necessary to classify samples on the basis of its characteristics or components it has. A classification model is used to predict a sample's class based on closest examples. Artificial Neural Network (ANN), Discriminant Analysis (DA), K-nearest neighbour (k-NN) are popularly used for classification in Chemometrics.

Near infrared spectroscopy in combination with chemometric analysis were used in identification of specific mango variety by Shyam *et al.* (2013) with prediction accuracy level of 94 to 99 percent. But, method for classification of mangoes to their varieties on the basis of their nutritional parameters as well as chemometric techniques is yet to be developed.

The objective of the study was to assess the classification performance of ANN and LDA and to compare their efficiencies. Therefore, the study was aimed to develop chemometric method for rapid classification of mango to their true varieties on the basis of their physico-chemical properties by using better performing model.

Materials and methods

Sample collection

Nine popular varieties of mango available in Bangladesh such as, Gopalbhog, Langra, Fazlee, Tosha, Khersapat, Aamrupali, Mollika, Kohitor and Himsagor were collected from local markets of Dhaka City during pick days of their seasons. Nine samples from each variety were taken and thus 81 samples were used finally in the study.

Measuring parameters

The collected mangoes were washed with de-ionized water, pilled and pulps were taken. Nutritional properties of these mango pulps were determined at Fruit Technology Research Laboratory of Institute of Food Science and Technology (IFST) under Bangladesh Council of Scientific and Industrial Research (BCSIR), Dhaka by using suitable instruments for measuring different parameters. Finally, 18 physico-chemical properties of the mango pulp samples were measured, and they are Edible portion, Moisture content, pH, Titratable acidity, Total Soluble Solids (TSS), Total sugar, Reducing sugar, Ash, Vitamin C, Total Protein, Total Fat, Crude fibre, Total energy, Total carbohydrate, Sodium (Na), Potassium (K), Calcium (Ca) and Magnesium (Mg).

Dimension reduction by Factor Analysis (FA)

All parameters are not equally important for classification of mango to their varieties. So dimension of the data could be reduced by selecting comparatively more important parameters of mango pulps. In order to reduce the dimensionality of a dataset, FA is very popular and powerful technique. In most of the scientific research variable are large in number and are correlated. It is very often necessary to reduce the dimensionality of the dataset retaining the variability presented in it as much as possible and with a minimum loss of information (Hair *et al,* 1998). This reduction is achieved by transforming the dataset into a new set of variables-factors, which are orthogonal (non-correlated) and are arranged in decreasing order of importance.

FA can be expressed as

$$F_i = a_1\, x_{ij} + a_2\, x_{2_\mathrm{j}} + \ldots + a_m\, x_m$$

Where F_i= factor, a= loading, x = measured value of variable, i = factor number, j = sample number, m = total number of variables

Before dimension reduction by FA, it is necessary to test sample adequacy by Kaiser-Mayer-Olkin (KMO) test and Sphericity test for assess eligibility of using FA. Then factors are extracted. Next, rotation of the extracted factors are performed to choose comparatively more important variable and thereby dimensionality of dataset is reduced. Finally, these selected variables are used for further application of chemometric techniques.

Artificial neural network (ANN)

ANN is a mathematical representation inspired by the human brain, and has an ability to adapt on the basis of the inflow of new information. Mathematically, ANN is a non-linear optimization tool. The ANN is especially suitable for classification and is widely used in practice. The network consists of one input layer, one or more hidden layers and one output layer, each consisting of several neurons. Each neuron processes its inputs and generates one output value that is transmitted to the neurons in the subsequent layer. Initially, a weighted sum of inputs is calculated at each neuron: the output value of each neuron in the proceeding network layer times the respective weight of the connection with that neuron.

A functional link network introduces a hidden layer of neurons. If model includes estimated weights between the inputs and the hidden layers, and the hidden layers use nonlinear activation function such as the logistic function, the model becomes genuinely nonlinear. The resulting model is called a multilayer perceptron (MLP). MLPs can be used with little knowledge about the form of the relationship between the independent and dependent variable. They are in general purpose, flexible, non-linear models that given enough hidden neurons and enough data, can approximate virtually any function to any desired degree of accuracy (Warren, 1994). One should find the optimal network by trial and error. The most interesting property of a network is its ability to generalize new cases. For this purpose, independent data set is used to test the neural network and check its performance (Anderson *et al.*, 1992; Won *et al.,* 1999). Upon successful completion of the training process, a well-trained neural network is not only capable of computing the expected outputs of any input set of the data used in the training stage, but should also be able to predict, with an acceptable degree of accuracy, the outcome of any unfamiliar set of input located within the range of the training data (Anderson *et al.*, 1994; Nehdi *et al.*, 2011).

Practically, 9 selected physico-chemical parameters were use as input and 9 varieties of mango as outputs in the network. Here we considered hidden layer, and used `Hyperbolic tangent' activation function at hidden layer and `Softmax' at output layer.

Linear Discriminant Analysis (LDA)

Discriminant analysis is a technique for classifying a set of observations into predefined classes. It operates on raw data and the technique constructs a discriminant function for each group (Lattin *et al.*, 2003; Wunderlin *et al.,* 2011). A simple linear discriminant function transforms an original set of measurements on a sample into a single discriminant score (Sanchez *et al.,* 2004). LDA involves the determination of a linear equation that will predict which group the case belongs to. The form of the equation or function is:

$D = v_1 X_1 + v_2 X_2 + ... + v_i X_i + a$

Where D = discriminate function, v = the discriminant coefficient or weight for that variable, X_i = respondent's score

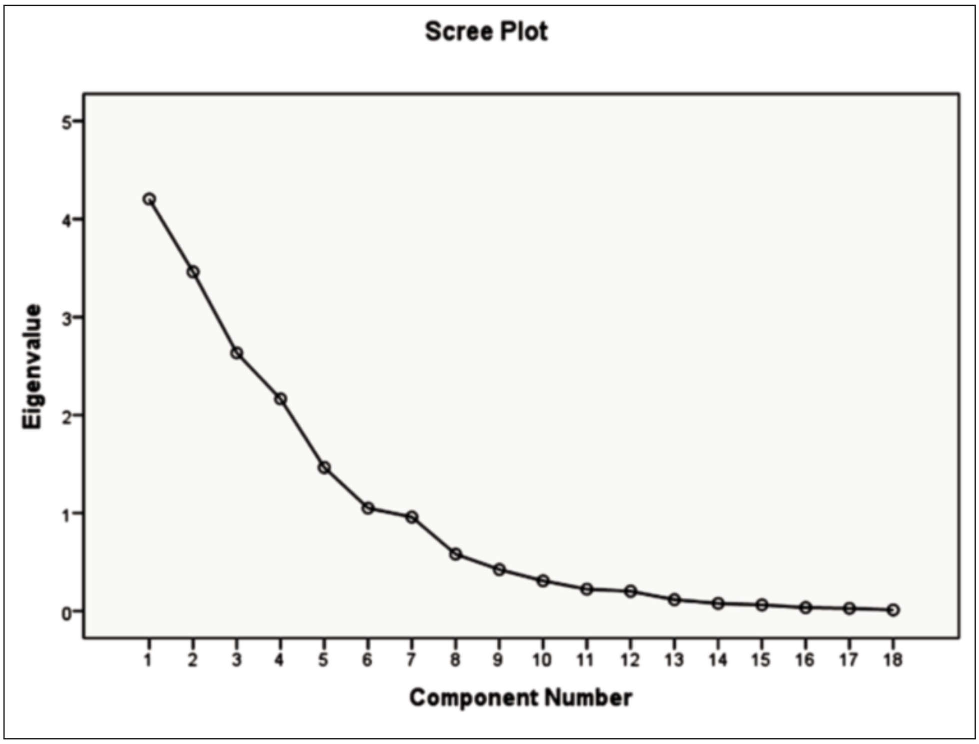

Fig. 1. Scree plot of eigen values of PCs

for that variable, a = constant, i = the number of predictor variables

Computation software

The data were analyzed using SPSS (Statistical Package for Social Sciences), now popularly used in every sector of data analysis of its version 22.0 for dimension reduction of dataset and for computing the ANN and LDA for classification of mango to their respective varieties.

Results and discussion

Kaiser-Meyer-Olkin (KMO) of sampling adequacy is 0.591 and Bartlett's Test of Sphericity significance value is 0.00, so the sample is eligible to use Factor Analysis for significant reduction of physico-chemical parameters.

Total Variance Explained chart in table I shows that eigen values of first six Principal Components (PCs) are greater than 1, and these six components explained 83.21 percent of total variation. Declination trend of eigen values in PCs are shown in the Scree Plot (Fig. 1) where after sixth component the value of eigen values start to be screes.

Table I. Total Variance Explained

Component	Initial Eigen values		
	Total	% of Variance	Cumulative %
1	4.204	23.357	23.357
2	3.461	19.228	42.584
3	2.633	14.629	57.213
4	2.165	12.027	69.240
5	1.465	8.137	77.376
6	1.049	5.830	83.206
7	0.959	5.327	88.533
8	0.580	3.223	91.756
9	0.423	2.352	94.109
10	0.308	1.711	95.820
11	0.223	1.236	97.056
12	0.202	1.121	98.177
13	0.116	0.643	98.820
14	0.077	0.430	99.250
15	0.062	0.346	99.595
16	0.035	0.195	99.790
17	0.027	0.149	99.939
18	0.011	0.061	100.000

Rotated Component Matrix (Table II) shows that Total Protein, Crude Firbre, and Na have significantly positive and Ash has negative contribution to first principal component which express 23.36 percent of the variance (Table I). TSS, Total Fat, Total Engergy and Total Carbohydrate have positive and Moisture has negative significant contribution to second principle component which express 19.23 percent. So if we consider first two principal components, Ash, Total Protein, Crude Fibre, Na, Moisture, TSS, Total Energy and Total carbohydrate are important parameters for further analysis of different varieties of mangos on the basis of their physico-chemical properties. Component Plot in Rotated Space (Fig. 2) presented below depicts the same results where important physico-chemical properties of mangos are shown in solid dots.

Table II. Rotated Component Matrix

Physico-chemical	Component				
property	1	2	3	4	6
Edible portion	-.417	-.240	.284	.479	-.010
Moisture content	-.159	-.801	.030	.065	.056
pH	-.120	.136	-.900	.172	-.066
Titratable acidity	-.151	.026	-.074	.928	.098
TSS	-.243	.618	-.169	.057	.017
Total sugar	-.252	.257	.743	.122	.305
Reducing sugar	.039	-.150	.856	.120	-.113
Ash	-.638	.033	.376	.025	-.382
Vitamin C	.033	.052	.192	-.017	-.095
Total Protein	.878	-.148	-.125	-.242	-.193
Total Fat	-.450	.727	.057	-.009	.157
Crude fibre	.879	.126	.081	.125	-.078
Total energy	-.002	.955	.002	-.027	.123
Total carbohydrate	.121	.939	.000	.005	.111
Na	.914	-.127	.189	.073	.193
K	.260	.141	.022	.008	.893
Ca	.278	-.020	.030	.884	.082
Mg	-.407	.148	.169	.321	.741

Finally, before applying ANN and LDA for classification of mango pulps according to their varieties 9 different physico-chemical parametrs are chosen on the basis of FA which are Ash, Total Protein, Crude Fibre, Na, Moisture, TSS, Total Fat, Total Energy and Total Carbohydrate.

Out of 81 samples, 54 (66.7 percent) were selected as training data and 27 (33.3 percent) as test data randomly to perform artificial neural network.

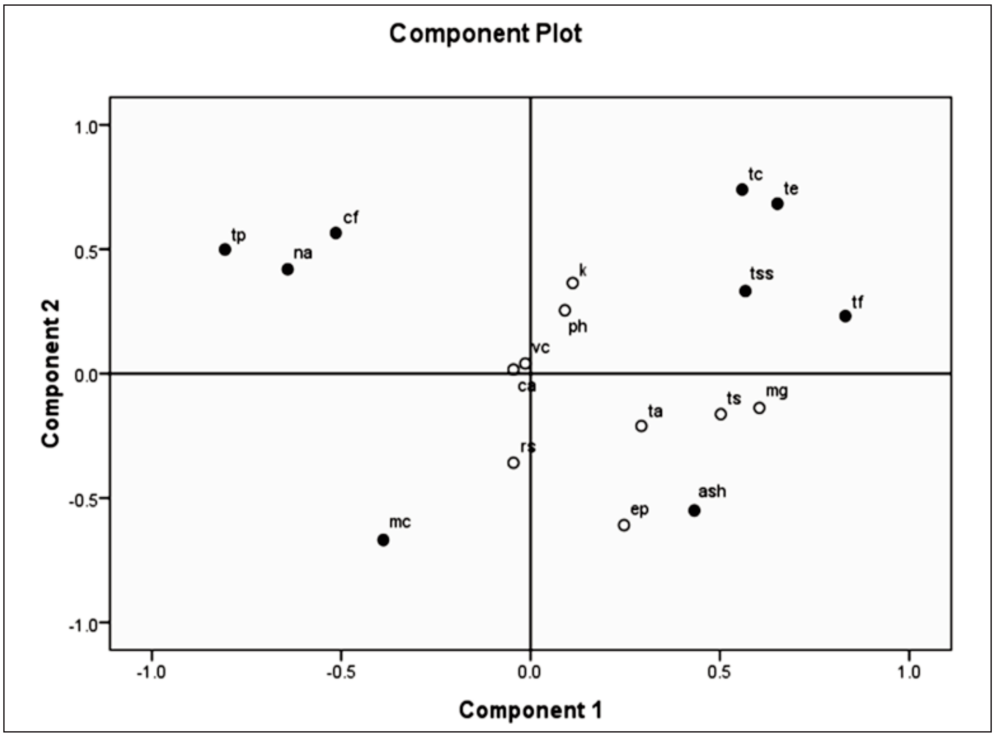

Fig. 2. Component plots in Rotated space of first two PCs

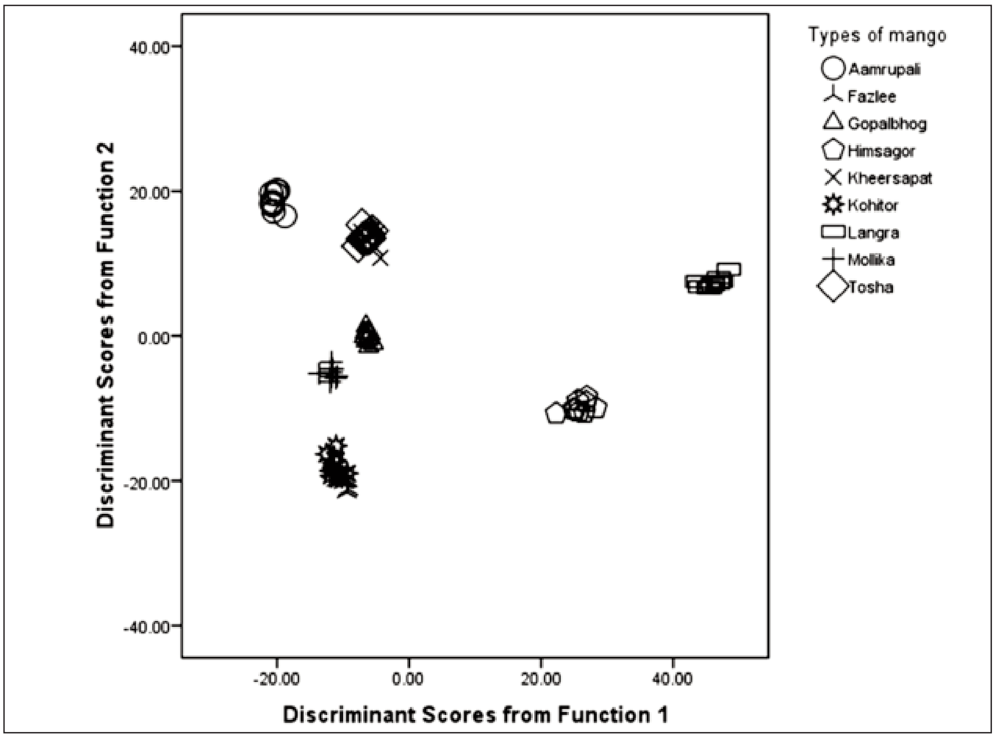

Fig. 3. Score plot of DF1 vs DF2

Table III. Predicted classes and original classes of mango pulp by ANN for test data

	Variety of	Predicted group membership								
	Mango	Gopalbhog	Langra	Fazlee	Tosha	Khersapat	Aamrupali	Mollika	Kohitor	Himsagor
Original group	Gopalbhog	3 100%	0	0	0	0	0	0	0	0
	Langra	0	3 100%	0	0	0	0	0	0	0
	Fazlee	0	0	3 100%	0	0	0	0	0	0
	Tosha	0	0	0	3 100%	0	0	0	0	0
	Khersapat	0	0	0	0	3 100%	0	0	0	0
	Aamrupali	0	0	0	1 33%	0	2 66%	0	0	0
	Mollika	0	0	0	0	0	0	3 100%	0	0
	Kohitor	0	0	0	0	0	0	0	3 100%	0
	Himsagor	0	0	0	0	0	0	0	0	3 100%

Table IV. Independent variable Importance

Physiochemical property	Importance	Normalized Importance (%)
Total energy	0.140	100.0
Crude fibre	0.136	97.1
Total Fat	0.134	95.6
TSS	0.132	93.9
Na	0.114	81.3
Total Protein	0.111	79.0
Ash	0.110	78.0
Total carbohydrate	0.086	61.3
Moisture content	0.036	25.9

Table V. Classification results by LDA

	Variety of	Predicted group membership								
	mango	Gopalbhog	Langra	Fazlee	Tosha	Khersapat	Aamrupali	Mollika	Kohitor	Himsagor
Original group	Gopalbhog	9 (100%)	0	0	0	0	0	0	0	0
	Langra	0	9 (100%)	0	0	0	0	0	0	0
	Fazlee	0	0	9 (100%)	0	0	0	0	0	0
	Tosha	0	0	0	9 (100%)	0	0	0	0	0
	Khersapat	0	0	0	0	9 (100%)	0	0	0	0
	Aamrupali	0	0	0	0	0	9 (100%)	0	0	0
	Mollika	0	0	0	0	0	0	9 (100%)	0	0
	Kohitor	0	0	0	0	0	0	0	9 (100%)	0
	Himsagor	0	0	0	0	0	0	0	0	9 (100%)

ANN can classify mango pulp samples 100 percent correctly to their original call for training data set. For test data, all varieties of mango samples except Amrupali, are correctly classified. In test data one sample of Amrupali, only one is misclassified to Tosha variety. Out of 27 test samples, one is misclassified to a wrong variety. That means, overall incorrect prediction of mango varieties is 3.7 percent. Most important variable for classification of mango varieties by ANN is Total Energy. Subsequent important variables are shown in table IV.

Before applying Linear Discriminant Analyses (LDA), Wilks' Lambda test has been performed first which shows a significance value 0, and that is less than 0.05, means the predictors (physicochemical parameters) significantly discriminate the varieties of mango.

It is evident from the table V that mango pulp samples are classified 100 percent accurately to their original varieties by Linear Discrimination Analysis with 9 selected nutritional and physic-chemical properties of mango.

Conclusion

As nutritional values and tastes of mango vary greatly according to their varieties, classifying varieties of mango and its products is a great concern of consumers and producers. Mango could be classified on the basis of nutritional and physic-chemical parameters to their original varieties. Cost, time and efforts are involved with getting measured values of each parameter. So, selection of appropriate physicochemical parameters is one of the most important tasks before classification of mango varieties. The study used Factor Analysis to reduce the dimension of the data. This method efficiently reduced 18 physiochemical parameters to its half. Then Artificial Neural Network, a nonlinear prediction method was used to classify the varieties of mango with 9 selected parameters. ANN could classify mangoes with 3.7 percent misclassification error both in training and test data. Next, another method, Linear Discriminant Analysis, was used for classification of mango, and this method classifies each specimen mango to their correct varieties. Therefore, 96.3 percent of varieties are correctly classified by ANN whereas classification rate by LDA is 100 percent in this regard. Finally, a new analytical method is being proposed where mango pulp could be classified to their respective varieties simply by using measurement values of 9 parameters (total energy, crude fibre, total fat, TSS, total protein, Ash, total carbohydrate and moisture content) and LDA. This method could contribute to improve the nutritional quality of processed commercial mango products like juice, jam, jelly, pickles etc through selection of appropriate varieties of mango.

References

Anderson D and McNeill G (1992), Artificial Neural Network Technology. Data and Analysis Center for Software.

Ara R, Motalab M, Uddin MN, Fakhruddin ANM and Saha BK (2014), Nutritional evaluation of different mango varieties available in Bangladesh*, International Food Research Journal* **21**(6): 2169-2174.

Bro R (2013), Chemometrics in Food Chemistry, Elsevier B.V., p. 172.

Hair JF, Anderson RE, Tatham RL and Black WC (1998), Multivariate Data Analysis (5th Ed). Prentice- Hall, Inc., p. 90.

Jilani MS, Bibi F, Waseem K and Khan MA (2010), Evaluation of physico-chemical characteristics of mango (*Mangifera indica* L.) varieties grown in D. I. Khan, *J. Agric. Res.* **48**(2): 201-207.

Kobra K, Hossain MA, Talukder MAH and Bhuyan MAJ (2012), Performance of twelve mango cultivars grown in different agro-ecological zones of Bangladesh, *Bangladesh J. Agric. Res.* **37**(4): 691-710.

Kramer R (1998), Chemometric Techniques for Quantitative Analysis, Marcel Dekker, Inc.

Lattin J, Carroll D and Green P (2003), Analyzing multivariate data, New York: Duxbury.

Nehdi M, Jebbar YD and Khan A (2011), Neural network model for cellular concrete, *ACI Mat. J.* **98**(5): 402-409.

Sanchez Lopez FJ, Gil Garcia MD, Martinez Vidal JL, Aguilera PA and Garrido Frenich A (2004), Assessment of metal contamination in Donana National Park Spain using crayfish (Procamburas clarkii), *Environmental Monitoring and Assessment* **93**(1-3):17-29.

Shafqat A, Haq CA and Hussain S (1992), Physico-chemical studies of varieties of mango grown at Shujabad, *Pakistan J. Agric. Sci.* **13**(4): 350-355.

Shyam N, Jha, Jaiswal P, Narsaiah K, Kumar R, Sharma R, Gupta M, Bhardwaj R, Ashish and K Singh. (2013), Authentication of Mango Varieties Using Near-Infrared Spectroscopy, *Agric. Res.* **2**(3):229-235.

Warren SS (1994), Neural Networks and Statistical Models, Preceedings of the 19th Annual SAS Users Group International conference.

Won JO, Won IL, Tae TK and Won JL (1999), Application of neural networks for proportioning of concrete mixes, *ACI Mat. J.* **96**(1): 61-67.

Wunderlin DA, Diaz MDP, Ame MV, Pesce SF, Hued AC and Bistoni MD (2011), Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia River Basin (Cordoba Argentina), *Water Research* **35:** 2881-2894.

Received: 04 April 2016; Revised: 17 May 2016; Accepted: 28 June 2016.