Available online at www.banglajol.info

Bangladesh J. Sci. Ind. Res. **52(2)**, 73-80, 2017

BCSIR

BANGLADESH JOURNAL OF SCIENTIFIC AND INDUSTRIAL RESEARCH

E-mail: bjsir07@gmail.com

Chemometrics assisted method for classification of mango juice by FTIR spectroscopic data

M. N. Uddin[1]*, A. K. Majumder[2], S. Ahamed[1], B. K. Saha[3] and B. Mumtaz[3]

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

[2]*Jahangirnagar University, Savar, Dhaka-1342, Bangladesh*

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

Abstract

Commercial mango juices are adulterated with heavy use of simple sugars in Bangladesh which poses a serious threat to public health. The present study is aimed to develop chemometrics assisted method for classification of commercial mango juices as adulterated or not with excessive use of glucose, fructose and sucrose with FTIR spectral data. Two statistical techniques, Artificial Neural Network (ANN) and Partial Least Squares-Discriminant Analysis (PLS-DA) have been assessed for their efficiencies in classification in this regard. Before calibration, spectral data were preprocessed with de-noising techniques, Savitzky-Golay (S-G) filtering. Concentration of simple sugars were classified as within or over certain limits. Here spectral values of 64 synthetic mixture solutions are used as training data to develop models and 15 spectral data of real mango juice are used as test data. PLS-DA shows better classification performance over lowercase ANN. From the findings, we develop a method for classification of mango juices adulterated with heavy use of simple sugars (glucose, fructose and sucrose). Therefore, it is a simple and cheap method to classify mango juices as adulterated or safe for consumers, manufacturers and quality regulating authorities.

Keywords : Mango juice; Simple sugars; Chemometrics; Classification; ANN; PLS-DA

Introduction

Excessive sweetening agents are used in processed fruit juice to deceive consumers of real taste of the food. Too much intake of sweeteners are causing threat to our health. Packaged or bottled "fruit juices" under different popular brand names sold in the market contain no real fruit extract but artificial flavour, poisonous colours, high level of chemical preservatives and sweetening agents. The Bangladesh Standard and Testing Institute (BSTI) carried out the test on randomly selected samples at the request of the Consumers' Association of Bangladesh (CAB) ('Fruit Juices', 2005, June 27). It is evident from different research findings that there is a relationship between intake of sugar in different form and increase of frequency of major chronic metabolic diseases, i.e., diabetes, cardiovascular disease, high triglycerides, and hypertension (Lustig *et al*., 2012; Tappy, 2012; Basu *et al.,* 2013; Te Morenga *et al*., 2013). There is an association between high sugar consumption with risk of cardiovascular mortality (Yang *et al*., 2014). Besides, mango juices are being exported after fulfilling local demand. So, simple, cost effective methods are necessary to assess the qualification of the products for the sake of consumers, producers as well as traders.

Classification is one of the important applications of chemometric analysis of chemical data through mathematical and statistical modeling (Brerton, 2007). Many applications require that samples be assigned to predefined categories, or "classes". This may involve determining whether a sample is good or bad, or predicting an unknown sample as belonging to one of several distinct groups. A classification model is used to predict a sample class by comparing the sample to a previously analyzed experience set, in which categories are already known.

Application of artificial neural networks (ANN) is a technique for data and knowledge processing, characterized by its analogy with a biological neuron (Otto, 1999). ANN can deal with nonlinear relationships between variables. It can be used both for quantification of a component in a product and for classification. Use of ANN in food authentication studies has been gaining popularity among food scientists, food processing industries and food quality regulating authorities. ANN is used in food science for classification of olive oil for their geographical region and types (Bucci et al., 2002), source of ingredients and manufacturing process of wine (Li-Xian *et al*., 1997; Kavuri *et al*., 2011).

*Corresponding author e-mail: m2nashir@yahoo.com

In spectroscopic instruments, absorbance of light by certain food against wave length or wave number is the main consideration for spectral data. In each spectrum there are huge number of absorbance value for each wave point. Each wave point or data point is considered as spectroscopic variable. These variables are huge in number and are mutually correlated. In this situation we cannot use Ordinary Least Square (OLS) method as problem of singularity exists. So, we use Principal Component Analysis (PCA), Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). Partial Least Square-Discriminant Analysis (PLS-DA) is an extension of PLSR used for classification. For food authentication studies PLS-DA is getting popularity. It has been used for classification of Vinegrs, coffees, Yogurt (Guerrero *et al.,* 2009; Ribeiro *et al.,* 2010; Cruz *et al*., 2013). We exercise classification efficiencies of mixture solutions and real mango juices as adulterated or not with excess use of sugars in them by PLS-DA.

Fourier transform infrared spectroscopy (FTIR) is an analytical technique, which measures the infrared intensity versus wave number of light. The resulting spectrum is characteristic of the organic molecules, which absorb infrared energy at specific frequencies so that the basic structure of compounds can be determined by the spectral locations of their infrared (IR) absorptions. Numerous reports have shown that the FTIR has been applied to evaluate the freshness of virgin olive oils in combination with multivariate analysis to determine fatty acid profile of virgin olive oil (Maggio *et al*., 2009), to analyze the free fatty acid content of Atlantic salmon skin lipid (Aryee *et al*., 2009) and to test the authentication of fruit juice (Vardin *et al*., 2008; Jha and Gunasekaran, 2010). However, classification of mango juice as adulterated or safe with over use of simple sugars with FTIR spectroscopic data and chemometric techniques is not reported yet.

Therefore, the objective of the study is to develop a easy, fast and cheap method to classify commercial mango juices as adulterated or not with over use of glucose, fructose and sucrose by using FTIR spectroscopic data and chemometric techniques like ANN and PLS-DA. This is an alternative of existing methods which would reduce the use of chemical standards, rigorous sample preparation hassles and would not generate chemical waste.

Materials and methods

Preparation of mixture solutions and collection of commercial mango juice

Standard mixture solutions of three sugars available in mango juice, i.e., glucose, fructose and sucrose, were prepared. Eight different concentrations of glucose (0.5, 1.0, 1.5, 2.0, 2.5, 3, 5, 10 percent), fructose (0.5, 1.0, 1.5, 2.0, 2.5, 3, 5, 10 percent) and sucrose (7.0, 7.5, 8.0, 8.5, 9.0, 9.5, 10.0, 15.0 percent) were used to prepare synthetic mixture solutions. Here "Orthogonal Experimental Design" was used to statistically maximize the information in the outputs. Thus, in total 64 mixtures were prepared with different concentrations of glucose, fructose and sucrose. According to the combination of concentrations from experimental design, the sugars were dissolved into de-ionized water to make solutions. Next, we collected 15 commercially available mango juices of different locally manufacturing companies. Then concentrations of glucose, fructose and sucrose in commercial mango juices were measured at laboratory by standard AOAC method (Horwitz, 2005). Both mixture solutions and commercial juices were used in scientific instrument, Fourier Infrared (FTIR) spectrophotometer, to get spectral data from the instrument. Finally, known concentrations of simple sugars and spectral data of synthetic mixture solutions were used to develop and validate a method, and spectra of real mango juices were used to test the method for prediction of glucose, fructose and sucrose in mango juices and classifying them.

FTIR measurements

In Fourier Transform Infrared (FTIR) spectroscopy, IR radiation is passed through a sample. Some of the infrared radiation is absorbed by the sample and some of it is passed through (transmitted). The resulting spectrum represents the molecular absorption and transmission, creating a molecular fingerprint of the sample. Like a fingerprint no two unique molecular structures produce the same infrared spectrum. This makes infrared spectroscopy useful for several types of analysis.

FTIR spectrometer (Shimadzu, Model: IRAffinity1) connected to software of IRSolution Operating system (Version 1.40) was used to obtain FTIR spectra of samples. The samples were placed in contact with Attenuated Total Reflectance (ATR) element at controlled ambient temperature. Finally, the mixture solutions and real mango juices were run in FTIR to get their respective spectra. FTIR spectra were collected in frequency 4000-650 cm^{-1} by co-adding 30 scans and at resolution of 4 cm^{-1}. All spectra were rationed against a background of air spectrum. Before every scan, a new reference air background spectrum was taken. There spectra were recorded as absorbance values at each data point in triplicate. The ATR plate was carefully cleaned *in situ* by wiping it with acetone, and dried with soft tissue before filling in with next sample.

Preprocessing of spectral data

The spectral data acquired from instrument contain spectra background information and noises which are interfered desired relevant quality attributes information. Interfering spectral parameters, such as light scattering, path length variations and random noise resulted from variable physical sample properties or instrumental effects need to be eliminated or reduced in order to obtain reliable, accurate and stable calibration models. Thus, it is very necessary to pre-process spectral data prior to modeling (Rinnan *et al*., 2009).

From spectra of FTIR we can see that there is no spectral peak and almost no variance of absorbance below the wave number 3700 cm^{-1} contain least information for prediction. So, wave number range 3700-648 cm^{-1} has been selected for further analysis. Here, spectral data were de-noised with Savitzky–Golay filtering (Palma *et al*., 2002; Nicolai *et al*., 2006) is used to de-noised the spectral data.

Classes of mango juice

Concentrations of glucose, fructose and sucrose in commercial mango juices were measured at laboratory by standard AOAC method. Average concentration of glucose, fructose and sucrose are 1.6%, 1.7% and 8.4% respectively. So, for the sake of classification we divide the simple sugar concentrations into two groups (Group 1 and Group 2) on the basis of approximation of average concentration of the simple sugars in commercial mango juice.

Table I. Assigned classes of mango juices according to the concentration of simple sugars

Simple sugars	Groups	Range	Number of chemical samples in the data (%)
Glucose	Group 1	< 2%	44 (56%)
	Group 2	≥ 2%	35 (44%)
Fructose	Group 1	< 2%	42 (53%)
	Group 2	≥ 2%	37 (47%)
Sucrose	Group 1	< 8%	28 (35%)
	Group 2	≥ 8%	51 (65%)

Chemometric techniques for classification

It is very often necessary to classify foods on the basis of their raw materials, species of ingredients, brand, geographical origin, category of product etc. to identify their specialty or traceability. This is one the process to identify the authenticity of food. Several classification techniques are used in food authentication studies by food scientists and food quality controlling authorities. Here, ANN and PLS-DA have been used to develop classification model to classify commercial mango juice as adulterated or safe on the basis of simple sugars they contain. Here the inputs are spectral data of mango juices and classification categories (high and low level of concentrations of sugars) are outputs.

Artificial Neural Network (ANN)

An artificial neural network (ANN) is a data processing system based on the structure of the biological neural simulation by learning from the data generated experimentally or using validated models (Bhotmange and Shastri, 2011). A network consists of a sequence of layers with connections between successive layers. Data to the network is presented at input layer and the response of the network to the given data is produced in the output layer (Fig.1) . There may be several layers between these two principal layers, which are called hidden layers. Finally, neural network approach build a predictive model for quantification or classification in a complex system.

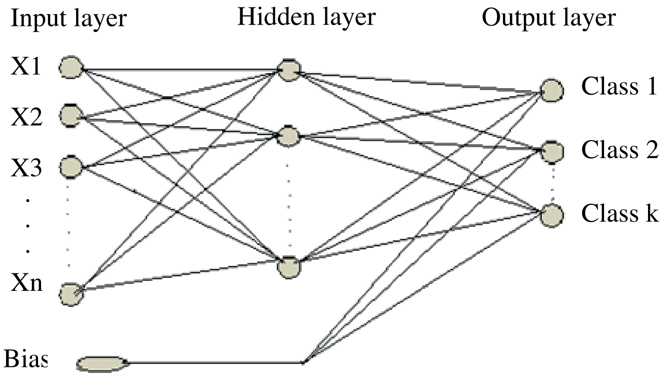

Fig. 1. Structure of an artificial neural network

An ANN consists of three basic steps: Signal propagation, Training the network and Testing the network are shown below:

1. Signal propagation: The input layer comprises *n* neurons that code for the *n* pieces of input signal (X_1,…,Xn) of the network (independent variables). The number of neurons of the hidden layer is chosen empirically by the user. Finally, the output layer comprises *k* neurons for the *k* classes (dependent variables). In the input layer, the state of each neuron is determined by the input variable; the other neurons (hidden layer and output layer) evaluate the state of the signal from the previous layer as:

$$a_j = \sum_{i=1}^{1} X_i W_{ji}$$

where a_j is the net input of neuron j; X_i is the output value of neuron i of the previous layer; W_{ji} is the weight factor of the connection between neuron i and neuron j. The activity of neurons is usually determined via a sigmoid function:

$$f(a_j) = \frac{1}{1\text{-}exp^{-aj}}$$

2. Training the network: The back-propagation technique is akin to supervised learning as the network is trained with the expected reply/replies. Each iteration modifies the connection weights in order to minimize the error of the reply (expected value-estimated value). Adjustment of the weights, layer by layer, is calculated from the output layer back to the input layer. This correction is made by:

$$\Delta W_{ji} = \eta \delta_j f(a_j)$$

where ΔW_{ji} is the adjustment of weight between neuron j and neuron i from the previous layer; $f(a_i)$ is the output of neuron i, η is the learning rate, and δ_j depends on the layer. For the output layer, δ_j is:

$$\delta_j = (\mathrm{Y_j} - \hat{\mathrm{Y}}_\mathrm{j}) f\square(a_\mathrm{j})$$

where Y_j is the expected value ('observed value') and $\hat{\mathrm{Y}}_\mathrm{j}$ is the current output value ('estimated value') of neuron j. For the hidden layer, δ_j is:

$$\delta_j = f\square(a_\mathrm{j}) \sum_{\mathrm{k=1}}^{\mathrm{k}} \delta_k W_{kj}$$

where k is the number of neurons in the next layer. The training, performed on a representative data set, runs until the sum squared of errors (SSE) is minimized:

$$\mathrm{SSE} = \frac{1}{2} \sum_{\mathrm{p=1}}^{\mathrm{p}} \sum_{\mathrm{j=1}}^{\mathrm{N}} (Y_{pj} - \hat{Y}_{\mathrm{pj}})^2$$

where: Y_{pj} is the expected output value, $\hat{Y}_{pj}$ is the estimated value by the network, $j=1..N$ is the number of records and $p=1..K$ is the number of neurons in the output layer.

3. Testing the network: After training, the performance of the network has to be tested. As in discriminant analysis, a first indication is given by the percentage of correct classifications of the training set records. Nevertheless, the performance of the network with a test set (set of similar data unused during training) is more relevant. In the test step, the input data are fed into the network and the desired values are compared to the network's output values. The agreement or disagreement of the results thus give an indication of the performance of the trained network.

In practical, Levenberg-Marquardt back-propagation neural network is used in the present study. Number of hidden neurons is 10 and the sigmoid activation function is used in each training.

Partial Least Square-Discriminant Analysis (PLS-DA)

PLS-DA is a variant used when the dependent variable is categorical. It is a partial least squares regression (PLSR) of a set of binary variables describing the categories of a categorical variable on a set of predictor variables. It is a compromise between the usual discriminant analysis and a discriminant analysis on the significant principal components of the predictor variables.

In PLSR, instead of directly fitting a model between X and Y, the PLS decomposes X and Y into low-dimensional space (so called laten variable space) first:

$$\mathrm{X = T{*}P' + E_0}, \text{ and}$$
$$\mathrm{Y = U{*}Q' + F_0}$$

where P and Q are orthogonal matrices, i.e. P'*P=I, Q'*Q=I, T and U has the same number of columns, a, which is much less than the number of columns of X. Then, a least squares regression is performed between T and U:

$$\mathrm{U = T{*}B + F_1}$$

At the end, the overall regression model is

$$\mathrm{Y = X{*}(P{*}B{*}Q') + F}$$

i.e. the overall regression coefficient is P*B*Q'.

The reason to perform PLS instead of total LS regression is that the data sets X and Y may contain random noises, which should be excluded from regression. Decomposing X and Y into laten space can ensure the regression is performed based on most reliable variation.

Classification efficiency of PLS-DA assessed in the study to classify mango juice as adulterated or not with over use of simple sugars on the basis of FTIR spectroscopic data in terms of error rate cv (coefficient of variation).

Computational software

SPSS (version 22.0) have been used for developing the Orthogonal Experimental Design. All the programs used to generate the results were written in MATLAB (version 8.1.0.604, The Math Works Inc., Natick, USA) for model development, validation and test with FTIR spectroscopic data.

Results and discussion

ANN for classification

In the present study, ANN was used for classification of simple sugars (Glucose, Fructose and Sucrose). Network performance is expressed by two statistics; they are Root Mean Squared Error (RMSE) and Prediction Error Percentage for misclassification (% E). RMSE is the square root of average squared difference between outputs and targets. Lower values are better. Zero means no error. Percent error indicates the fraction of samples which are misclassified (% E). A value of 0 means no misclassifications, 100 indicates maximum misclassifications.

Table II shows misclassification rate which is bit higher for validation and test data sets. So, from these findings, ANN cannot be suggested for classification of mango juices as adulterated or not with over use of simple sugars.

PLS-DA for classification

For optimal component selection by PLS-DA, we drew line graphs of error rate cv (coefficient of variation) against number of latent variables and presented in figure 2.

Table II. Classification of simple sugars by ANN

Simple sugars	Samples (Size)	RMSE	Percent Error of Misclassification (%E)
Glucose	Training (52)	0.0577	0.000
	Validation (12)	0.4297	16.667
	Test (15)	0.3112	16.667
Fructose	Training (52)	0.0498	0.000
	Validation (12)	0.3018	8.333
	Test (15)	0.2502	8.333
Sucrose	Training (52)	0.2020	1.818
	Validation (12)	0.3616	16.667
	Test (15)	0.3286	8.333

Table III. Classification of glucose, fructose and sucrose by PLS-DA

Simple sugars	Dataset	No. of components	Explained variance (%)	Error rate	Non error rate	Error rate CV	Non error rate CV
	Training (64)	8	100	0.02	0.98	0.29	0.71
Glucose	Test (15)	8	99	0	1	0.67	0.33
	Training (64)						
Fructose	Test (15)	8	99	0	1	0.17	0.83
	Training (64)	8	100	0	1	0.12	0.88
Sucrose	Test (15)	8	99	0	1	0	1

From figures 1, error rate cv is minimum at latent variable number 8 for glucose, fructose and sucrose. So, first 8 latent variables were used in the PLS-DA for classification of mango juice as adulterated or not with over or within range percentage of glucose, fructose and sucrose in them (Table III).

Therefore, PLS-DA can be used for classification of in mango juices for over or within range using glucose, fructose and sucrose in them.

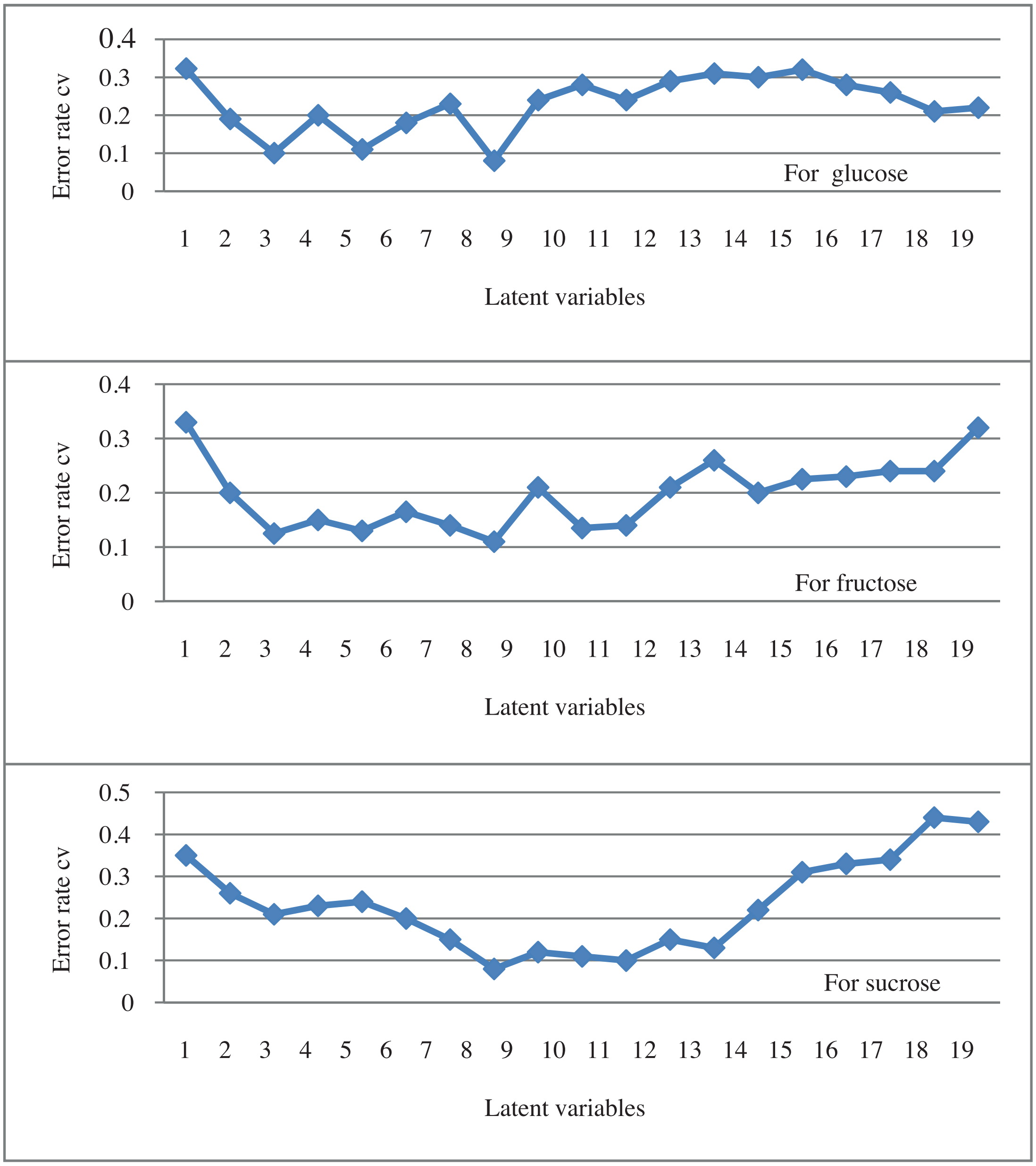

Fig. 2. Optimal number of latent variables for PLS-DA for glucose, fructose and sucrose

Error rate (misclassification error rate) is zero for model building with training and test data except classifying glucose and fructose with training data, and the error rate in these two cases are 2 percent only.

Conclusion

It is evident from results of the study that misclassification error rate is very high (upto about 17 percent) in classification of sugar solutions and real mango juices by ANN for training, validation and test data sets. On the other hand, PLS-DA for simple sugar concentration data can be suggested for classifying mango juices as adulterated with excessive use of these elements due to low misclassification error rate (less than 2 percent). From the analysis results of instrumental data, we can select PLS-DA for classification of commercial mango juice as adulterated or not with heavy use of simple sugars on the basis of FTIR spectral data.

Finally, a chemometric method has been developed with PLS-DA and FTIR spectral data for classification of mango juice which is cost effective, time saving and do not generate chemical waste as in this method no chemical standard is used. Newly proposed chemometric methods could save a huge amount of quality testing cost for mango juice producing companies, quality regulating authorities and food testing laboratories.

Acknowledgement

The research work was funded by the Bangabandhu Fellowship on Science & ICT under Ministry of Science & Technology, Bangladesh.

References

Aryee ANA, Van de Voort FR and Simpson B K (2009), FTIR determination of free fatty acids in fish oils intended for biodiesel production, *Process Biochemistry* **44**: 401–405.

Basu S, Yoffe P, Hills N and Lustig RH (2013), The relationship of sugar to population-level diabetes prevalence: An econometric analysis of repeated cross-sectional data, *PLOS ONE* **8**(2): e57873.

Bhotmange M and Shastri P (2011), Application of Artificial Neural Networks to Food and Fermentation Technology, Artificial Neural Networks- Industrial and Control Engineering Applications, Ed. Prof. Kenji Suzuki, ISBN: 978-953-307-220-3, pp 201-222.

Brerton RG (2007). Applied Chemoetrics for Scientists, John Wiley & Sons Ltd. England, pp-145.

Bucci R; Magri AD, Magri AL, Marini D and Marini F (2002), Chemical Authentication of Extra Virgin Olive Oil Varieties by Supervised Chemometric Procedures, *J. Agric.Food Chem*. **55**: 413-418.

Cruz AG, Cadena RS, Alvaro MBVB, Sant'Ana AS, Oliveira CAF and Faria JAF (2013), Assessing the use of different chemometric techniques to discriminate low-fat and full-fat yogurts, LWT-*Food Science and Technology* **50**: 210-214.

Fruit Juice missing the Fruits (2005, June 27), *The Daily Star*.

Guerrero ED, Mejias RC, Marin RN, Lovillo MP and Barroso CG (2010), A new FT-IR method combined with multivariate analysis for the classification of vinegars from different raw materials and production processes, *J Sci Food Agri*. **90**: 712-718.

Horwitz W (2005), Association of Official Agricultural Chemists (AOAC), *Official Method of Analysis*, 18th Ed., Chapter 44, AOAC International, Maryland, USA, pp 9-11.

Jha SN and Gunasekaran S (2010), Authentication of sweetness of mango juice using Fourier transform infrared-attenuated total reflection spectroscopy, *Journal of Food Engineering* **101**: 337-342.

Kavuri NC and Kundu M (2011), ART1 Network: Application in Wine Classification, *International Journal of Chemical Engineering and Applications* **2**(3): 189-195.

Li-Xian S; Danzer K and Thiel G (1997), Classification of wine samples by means of artificial neural networks and discrimination analytical methods, *Fresenius J Aal Chem*. **359**: 143-149.

Lustig RH, Schmidt LA and Brindis CD (2012), Public health: The toxic truth about sugar, *Nature* **482**: 27-29.

Maggio RM, Kaufman TS, Carlo MD, Cerretani L, Bendini A, Cichelli A and Compagnone D (2009), Monitoring of fatty acid composition in virgin olive oil by Fourier transformed infrared spectroscopy coupled with partial least squares, *Food Chemistry* **114**: 1549–1554.

Nicolai BM, Theron KI and Lammertyn J (2006), Kernel PLS regression on wavelet transformed NIR spectra for prediction of sugar content of apple, *Chemometrics and Intelligent Laboratory Systems* **85**: 243-252.

Otto M (1999), Chemometrics: Statistics and computer application in analytical chemistry, Wiley-VCH, Hoboken, NJ, pp 314.

Palma M and Barroso CG (2002), Application of FT-IR spectroscopy to the characterization and classification of wines, brandies and other distilled drinks, *Talanta* **58**: 265-271.

Ribeiro JS, Salva TJ and Ferreira MMC (2010), Chemometric studies for quality control of processed Brazilian coffees using drifts, *Journal of Food Quality* **33**: 212-227.

Rinnan A, Van Den Berg F and Englsen SB (2009), Review of the most common pre-processing techniques for near-infrared spectra, *Trends in Analytical Chemistry* **28** (10): 1201-1222.

Tappy L (2012), Q&A: Toxic effects of sugar: Should we be afraid of fructose? *BMC Biology* **10**(1): 42.

Te Morenga L, Mallard S and Mann J (2013), Dietary sugars and body weight: Systematic review and meta-analyses of randomised controlled trials and cohort studies, *BMJ* **346**:e 7492.

Vardin H, Tay A, Ozen B and Mauer L (2008), Authentication of pomegranate juice concentrate using FTIR Spectroscopy and Chemometrics, *Food Chemistry* **108**:742-748.

Yang Q, Zhang Z, Gregg EW, Flanders W, Merritt R, and Hu FB (2014), Added sugar intake and cardiovascular diseases mortality among US adults, *JAMA Internal Medicine* **174**(4): 516-524.

Received: 3 October 2016; Revised: 14 November 2016; Accepted: 28 December 2016.