RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

Detection and quantification of food colorant adulteration in saffron sample using chemometric analysis of FT-IR spectra

Sadegh Karimi,[a]*, Javad Feizy[b] ,Fatemeh Mehrrjo[a], Maryam Farrokhnia[c]

[a] *Department of Chemistry ,College of Science, Persian Gulf University, Bushehr, Iran*

[b] *Research Institute of Food Science and Technology, P.O. Box 91735-147, Mashhad, Iran*

[c] *The Persian Gulf Marine Biotechnology Research Center, Bushehr University of Medical Sciences, Bushehr, Iran*

ABSTRACT

The aim of present study is to investigate the combination of Fourier transform infrared (FT-IR) spectroscopy with pattern recognition to recognize the standard saffron from those which have been adulterated with various types of food colorants. Transmittance FT-IR spectra have been obtained for standard saffron and six mixed samples with food colorants including Tartrazine, Sunset yellow, Azorubine, Quinoline-yellow, Allura red and Sudan II. Genetic algorithm –linear discriminant analysis (GA-LDA) based on clustering of variable concept has been applied to transmittance FT-IR spectra for classification of standard saffron from fraud ones. Analysis of the selected clusters of variables indicates that three regions band corresponding to 1800-1830 , 2600-2900 and 3700-3850 cm^{-1} are responsible for differentiation of standard samples from their fraud ones. Regression analysis has been introduced in order to have information related to amount of food colorant. Combination of FT-IR and clustering of variable concept resulted in the best performances for calibration and external test set with 100% sensitivity and specificity.

Keywords: Saffron; Quality control; Adulteration; Food colorant; Pattern recognition; clustering of variable concept

Corresponding author, E-mail address: sakarimi@pgu.ac.ir ; karimi.sadegh@ gmail.com

1. Introduction

Food quality control (authenticity) is one of the increasingly important and sometimes vital subjects for consumers, regulatory agencies, and food industry. One of the main characteristics of authentication is to find a way for finding those economically motivated adulteration in food products which usually are more readily available and less expensive substitute. However their identifications are very difficult by routine analytical methodologies[1]. Meanwhile fraud detection by routine analytical methodologies is usually time-consuming.

Saffron has long been used as a coloring and flavoring agent in food. It is also known for a wide range of health benefits [2, 3]. It consists of dried stigmas of the cultivated species *Crocus sativus L*. On the other hand, saffron is one the most expensive species in food industry. In addition, this product is just produced in a few countries such as Iran and Spain. These two factors cause that the saffron can be good candidate for adulteration conducted for economic gain and has been subjected to various types of adulteration over the centuries [4, 5]. A good review of the different types of saffron adulteration has been collected by Consonni and their coworkers [6]. As they mentioned, different spectroscopic (UV-Vis) and several chromatographic methods have been used for saffron adulterations, however, each method has its limitation.

Food colorant like Azorubine, Quinoline, Sunset yellow, Sudan II, Allura red and Tartrazine are another area for authentication of saffron samples. Regardless of the experimental practice and design, the detection of food colorant frauds in saffron is a challenging task since changes in physical and color properties are not always easily identifiable. FT-IR spectroscopy is a simple analytical technique largely applied for its rapidity and reproducibility in food fraud detection[7]. Another characteristic of FT-IR spectroscopy is the potential for high-throughput

analysis with minimal sample pretreatment[8, 9]. Transmittance FT-IR spectroscopy based fingerprinting may identify the differences that often exist between authentic samples from normal products ones. As an example, Fourier-transform mid-infrared (FT-MIR) spectroscopy has been recently used to investigate how the typical Fourier-transform mid-infrared (FT-MIR) spectrum of saffron changes as a result of storage under different conditions [10]. However the FT-IR spectra of samples have owned complexity and limitations for analysis. For example a FT-IR spectrum, usually consists of hundreds or even thousands of measurements or channels, containing information for a classification. Almost these measurements contain the redundant or irrelevant information. Therefore, powerful methods should be used to extract the fingerprint of the analytes or properties from the total signal.

Due to high similarities between transmittance FT-IR spectra of samples, it is impossible to discriminate them using visual inspection of their spectra. This problem encouraged us to apply powerful multivariate pattern recognition for analysis of such data set[7]. On the other hand transmittance FT-IR spectra have high throughput out for each samples, this subject led to create small sample size problem (the ratio of variable to sample is high) for pattern recognition analysis method. In this condition, the classification methods such as LDA have a tendency to show over-fitting result [11]. This subject can be solved using clustering of variable concept for transmittance FT-IR spectra before LDA analysis. Finally, most of the wavenumbers are irrelevant to class distinction and should be discarded or removed. The current study presents an approach to discriminate adulterations in saffron by means of transmittance FT-IR and pattern recognition method. The efficiency of new pattern recognition algorithm [11] which has been worked by clustering of variable is demonstrated in the present study. GA-LDA based on self-

organizing map (SOM) approach is based on combination of data dimension reduction and variable selection algorithms. Besides, the obtained results have been compared with partial least square – discriminant analysis (PLS-DA) as a conventional pattern recognition method. Different studies related to chemical composition and geographic origin has been done in saffron study[12-14]. To the best of our knowledge, there are no published reports providing information about discrimination and quantification of standard saffron samples from food colorant adulteration using FT-IR spectroscopy.

2. Materials and methods

2.1. Preparation of fraud saffron samples

Stigma (without styles attached) of pure Iranian saffron have been collected from flowers of Torbat Heydareyeh farms (harvest 2014) in the Khorasan Razavi province. All of the standard saffron samples and food colorant, were finely ground in a mortar. In order to create the adulterated (synthetic data set) saffron samples, artificial spurious mixtures containing 0.5% up to 30.0% (w/w) of food colorant adulterant were prepared. Overall, 20 mixtures with different amount of food colorant were used for each adulterant and thus seven classes were defined, including the standard and authentic saffron samples. The chemical structures and their FT-IR spectra of six food colorant can be found in Figures (S1) and (S2) from supplementary section. As we can see from Figure (S2) the FT-IR spectral data of food colorants are very similar and fraud kind detection is not so straightforward. On the other hand, the spectral look similar but they have shown some differences (the height and area of the peaks are different) when compared in details, because different quantity of food colorant is added to each sample. In other words the differences exist in the intensity of spectra not in the shape.

Saffron can be considered as a complex compound, so its FT-IR (transmittance or absorbance) spectrum shows extensive overlap of various compounds. The FT-IR spectra of representative saffron sample from the investigated ones are shown in Figure 2a. As it has been reported in the literature [14-18] the broad peak which has been centered around 3400 cm^{-1} is due to hydroxyl (–OH) groups. The spectral region related to 3000–2830 cm^{-1} presents two peaks (2929 and 2851 cm^{-1}) which correspond to C–H stretching [19, 20]. Moreover the spectral region 1800–1500 cm^{-1} is the characteristic groups region. The carbonyl (–C=O) group (esters, ketones, aldehydes), the non-removed water and the aromatic ring absorb in this region [19, 20]. The region 1500–800 cm^{-1} is the 'fingerprint region'. The peaks in this region are associated with the skeletal vibrations of the components and have been attributed to $–CH_2–$, $CH_3–$, –OH, C–C, C–O, C–O–C groups [17, 18]. Particularly, the 1200–800 cm^{-1} spectral region has been correlated with the presence of sugars and polysaccharides [21].

2.2. Transmittance spectral measurements

Fourier transform infrared (FT-IR) spectra have been recorded on a Bruker Vector22 spectrometer, operating in the region 4000-400 cm^{-1} in the transmittance mode. A total of 16 scans with 4 cm^{-1} resolution were acquired for each spectrum. For FT-IR transmittance measurements, all samples were mixed with KBr (suitable ratio (w/w)) and homogenized. This mixture for each synthetic sample was then compressed under a pressure of ca. 200 MPa for 1 min to form a thin KBr disc. Also the spectrum of a clean KBr disc (without saffron) was used for background subtraction. It is worthy to mention that the time required for the preparation of KBr pellet for each sample is approximately 5 minutes and totally 10 minutes with scanning the wavenumber. So that in comparison with other methods for example HPLC, ELISA ... this one is

simple, economical and very low time consuming approach. The spectrometer was located in an air-conditioned room (25 ^{0}C). The spectra were stored using the OPUS software supplied from the same manufacturer.

2.3. Multivariate data analysis

Principal Components Analysis (PCA) and GA-LDA based on SOM have been performed with auto scaling as preprocessing algorithm. The basic idea behind the PCA is to visualize the data in the low dimensional space. For this purpose, PCA transform the data from a high dimensional space onto lower dimensional ones, without losing much information. The principal components are constructed in such a way that the first explains most of the data variance; the second is orthogonal to the first and describes most of the variance not explained by the first PC, and so on. Finally, samples are distributed in this low space (two and three) based on their similarities.

2.4 Linear discriminant analysis

Among traditional classifiers algorithms, linear discriminant analysis is probably the most known method [22]. The method can be considered as the probabilistic parametric classification technique which separates the objects into classes by maximizing the between-class variance and minimizing the within-class variance. Furthermore LDA makes the assumption that the classes have identical covariance matrices and fits a multivariate normal density to each group with a pooled estimate of the covariance. Because a pooled covariance matrix is calculated, the number of objects must be significantly greater than the number of variables. In other word, when the class object sizes are small compared to the dimension of the measurement space (the number of variables), the inversion of covariance matrices became difficult [23]. Also in the case of highly

correlated variables the, i.e. in presence of multicollinearty, discriminant analysis led to over fitting results.

2.5. Partial least square –discriminant analysis

Partial least square –discriminant analysis (PLS-DA) can be considered an extension form of LDA algorithm which uses the latent variables for predicting one (or several) binary responses(s) **y** from a set of variables in **D** [23]. Similar to PLS regression, PLS-DA performs a dimension reduction; however the extracted scores are used to discriminate the calibration and prediction samples. Thus, PLS-DA needs the class-variable of the objects and extracted scores not only retain the maximal variances of the original variables but also are correlated with the class-variable.

2.6. Kohonen self-organizing map (SOM)

A Kohonen self- organizing map (SOM) is a two dimensional array of neurons, which each neuron containing a weight vector that has the same dimension as the experimental variable data set. A SOM is trained to reflect as much as possible the relationship between individual pieces of data. There are able to map multidimensional information into a surface (the 2D array). Similarly to principal component analysis, SOM reduces multidimensional information to two dimensions with maintaining the topology of information. However, in contrast to PCA, SOM has advantages to use the nonlinear relationship between variables in data matrix. Figure 1 shows the structural design of a Kohonen network. Each column in the grid map represents a neuron and each box in such a column represents a weight (a number). In this case, the objects are the samples and the variables are wavenumber. Before the starting training, the weights take the random values. It should be noted that the learning is a competitive and iterative process. This step includes the adjustment of the weight during the training phase. The procedure of SOM can

be summarized as follow. (1) A variable from training set is introduced to the network (2) The neuron which its weight vector is the most similar (determined using the Euclidean distance) to the input variable is called the wining neuron or the best matching unit (BMU). (3) The weights of winning neuron modifies by network to become much more similar to the input variables. (4) With the same aim, neighborhood neurons are also corrected. However the amounts of these corrections depend on their distance from the winning neuron. (5) All these steps repeat iteratively to reach a predefined number of cycles (epoch) finally and then the process stops. Finally, when the entire wavenumber are entered in the Kohonen network and the process is completed, similar input (in our case similar spectral information) vectors are clustered based on their similarities [11].

2.7. Description of GA-LDA based on SOM

Almost chemical data which has been obtained from laboratory have many variables in comparison with samples. For analysis (multivariate calibration and classification) of such data sets, we should careful about over fitting problem. Suppose we have a data matrix (D) with *m* rows (the samples) and *n* columns (the wavenumbers). The proposed algorithm can be illustrated using the subsequent steps:

1- In the first part, the whole wavenumbers has been divided in *q* cluster using Kohonen self-organizing map (SOM). Clustering of wavenumbers into different sub-matrix (D_i) has been performed according to their similarities in information.

$$D = \left[[D_1][D_2] \dots [D_q]\right] \qquad (1)$$

2- Afterward, in order to obtain the principal components and loadings of each sub-matrix, PCA can be applied in each sub-matrix (D_i) separately.

$$D_i = T_i P_i^T \qquad i = 1{:}q \qquad (2)$$

The matrices T_i and P_i are the principal components and loadings of the each sub-matrix (D_i) respectively. The superscript "T" indicates the matrix transpose notation.

3- Substitution the equation (2) into equation (1) gives the reduced data set (D_r) :

$$D_r = \left[[T_1 P_1^T][T_2 P_2^T] \dots [T_q P_q^T] \right] \qquad (3)$$

Obviously the column of this reduced data set, D_r, consists of all the obtained PCs from different clusters. So that, the dimensions of D_r is ($m \times r$), where m is the number of samples and r is the total number of principal components obtained from previous step. Equation (3) indicates that one can be able to separate the PCs and loadings of different clusters. By this approach, three main purposes have been obtained. The first one is that the most information of original data matrix has been maintained. The second one, which is the most important for LDA analysis, is that the dimension of data has been reduced. Lastly, the information in the PCs of original data set has been divided into different, useful and redundant, parts.

4- Finally the LDA classifier has been applied, on the reduced data set (D_r) and the classification score for training sample (x_i) is defined as:

$$\text{classification score}\,(x_i) = (x_i - \mu_k) \sum_{pooled}^{-1} (x_i - \mu_k)^{-1} \qquad (4)$$

The $\sum_{pooled}^{-1}$,is the inverse of class covariance matrix , $\mu_k$is the mean vector of class k. It should be noted again that, for ill-condition situation, the number of wavenumbers is higher than the number of objects, the estimations of the class covariance matrix become highly uncertain, which is not true in our case.

5- The reduced data (D_{ru}), for prediction step [11], can be constructed as:

$$t_{ru} = [t_{1u}\, t_{2u} \ldots t_{qu}] = D_u V^{+} \qquad (5)$$

The superscript '+' represents the matrix pseudo-inversion.

Two important subjects must be considered in the mentioned algorithm. The first one is the type of clustering algorithms and the second ones is the cluster size (q). Recently we have shown that non-linear clustering such as Kohonen SOM has superiority respect to other clustering algorithms for regression modeling [24]. The cluster size (q) should be optimized by trial and errors such that all classification models have been performed on the any network size and the obtained results have been compared for their prediction abilities. The performance evaluation of the each cluster size from LDA classification models has been used based on Not-Error Rate (NER) values, evaluated both on cross-validation groups and external test samples. The validation of the presented classification models is based on leave many out (LMO) cross-validation (1/5 being excluded during each run).

As it is noted previously, the PCs of different cluster and corresponding loadings contain the useful and redundant one for classification. Our effort is to get rid of second ones which reduces the calibration and prediction efficiency of our models. On the other hands the useful PCs which can be improve the classification model should be extract. This can be done by applying the PC selection algorithm such as genetic algorithm (GA) on the reduced data set (D_r). For any network size of SOM, the classification models have been constructed based on selected PCs and statistical parameter have been used to compare the network sizes. Variable selection algorithm (GA) used in this paper is described by Leardi et.al, [25] in PLS regression, except that in the proposed algorithm, GAs are coupled directly with LDA to improve the power of the

classification algorithm. The selection of variables is performed by repeating GAs, *t* times and then including the variables on the basis of the frequencies of selection.

2.8. Clustering of variable –PLS regression

Recently we have introduced a new strategy for variable selection in PLS regression using clustering of variable concept [26]. This algorithm which we called it clustering of variable –partial least square, CLoVA-PLS, consists of two steps. Like the first step of GA-LDA based SOM (Eq.1), in this algorithm whole spectral region has been divided into some clusters based on their similarities using Kohonen self -organizing map. As we mentioned in section (2.7) the number of clusters can be varied from 1 to the number of variables. For example if one set the number of cluster size (q) to 1 , all the variables contribute in model building and can be considered as the conventional PLS . In practice, the number of clusters size can be optimized by gradually increasing the cluster size (q) and followed the statistical parameter to find a model with the satisfied result. In other word, for each cluster size, in order to find the most useful cluster of variable, all of the produced sub-matrix (clusters) has been investigated using PLS regression separately. The statistical parameters, usually RMSCV and RMSEP, of constructed model from each cluster, are used to judgment for selecting the informative one(s). It is worthy to mention that, calibration samples are responsible for training and select the useful variables, while test samples have never used during the optimization stage and there subsequently predicted by means of the models optimized in the training samples. For more details about this algorithm, the interested readers can refer to our previous publications [26].

Data analysis has been performed in a MATLAB environment (MathWorks, Inc., Natick, MA, USA, version 7.2). GA-LDA is based on GA-PLS of Leardi which is modified for classification problem. The LDA classification and Kohonen self-organizing map toolboxes

provided by Ballabio were downloaded for free from the website of Milano Chemometrics and QSAR research group (http://michem.disat.unimib.it/chm/download/kohoneninfo.htm). PLS calibrations were based on the PLS Toolbox version 4 from Eigenvector Research

3. Result and discussion

The transmittance FT-IR spectra of all studied saffron samples (standard and fraud ones) have been collected in a data matrix D of the dimension of ($n_s \times n_w$), where n_s and n_w are the number of sample and wavenumber respectively. Thus, each row of D (d_i) is the transmittance FT-IR spectrum of a specified sample. Since we have obtained 20 spectra for each kind of saffron adulteration size n_s is considered 140. The ability of transmittance FT-IR spectroscopy in combination with multivariate pattern recognition for discrimination between standard and fraud saffron samples have been investigated. Data matrix (D) has been divided into the calibration and prediction sets by the DUPLEX algorithm[27]. Summary DUPLEX algorithm is start as follows: first the two points which are furthest away from each other are selected for the calibration set; from the remaining points, the two objects which are furthest away from each other are included in the prediction set; then the remaining point which is furthest away from the two previously selected for the calibration set is included in the calibration set. The procedure is repeated for the test set which is furthest from the existing points in that set. In conclusion, points representing both training and test sets were distributed uniformly within the whole space which is constructed via the entire dataset. Based on DUPLEX strategy, one can assure that the composition of the training set and the test set is representative, at the same time the imbalance of the two datasets is avoided.

In our case 98 samples have been included in the training and the remaining 42 samples have been selected as test. The transmittance FT-IR spectra of the standard saffron samples and

fraud ones are represented in Figure 2. As it is evidence from these figures they are very similar spectra to each other; so that visual inspection of the spectra is impossible. The major bands in the typical FT-IR spectra of saffron samples can be found in Table S1 from supporting information. Consider that saffron is a complex mixture of different chemical compounds so that it is difficult to assign all of the bands directly to specific constituents. On the other hand, since our goal was to identify wavenumber region(s) with high effect on discrimination of standard saffron from fraud ones, we investigated all spectral region(s) in the infrared data using LDA based on clustering of variable concept.

3.1 PCA overview of transmittance FT-IR data

To get an overview of the saffron data set, PCA has been applied to the extracted the meaningful PC's. The results of application of PCA on the transmittance spectral data matrix of whole samples set are given in Table S2 from supplementary section. Different strategies exist for adequate PC selection in the literature [28]. Therefore, in this table, the Eigen value, percent of variances in the data matrix is explained by each PC and the cumulative percent of variances (CPV) are reported. The first five principal components could explain, 98.45% of variance in data set. In the other words, 42 saffron (standard and fraud) samples can be visualized in four principal components instead of 1868 dimension wavenumbers. A plot of the first two principal component scores for auto scaled data, which are corresponding to 94.92% of the original variance, is shown in Figure 3. Due to high similarity between the transmittance FT-IR spectra of the saffron samples (standard and different fraud ones) there is no evidence of discrimination between seven classes along the first two principal components. Because, most of the wavenumber are unrelated to class of our samples, more extraction of PCs is not useful.

Although PCA is the powerful and versatile method, it just uses transmittance data matrix and consequently gives some overview of complex multivariate data

3.2. GA-LDA based on SOM: combining data dimension reduction and classification

The transmittance FT-IR spectra of saffron samples data matrix is composed of 1868 variables (wavenumber). Definitely, not all parts of the presented wavenumber have useful information about the class information of samples. In the first step of GA-LDA based on SOM, Kohonen SOM is applied to cluster wavenumber based on their similarity. Different clustering algorithms can be used in this step, but we have shown that Kohonen SOM has superiority respect to other clustering algorithm[24].

One of the important parameters of Kohonen SOM which should be optimized is the number of Kohonen sizes (nodes). Each n-node Kohonen SOM model leads to ($n \times n$) cluster of variables. Therefore, the number of clusters (q) produced by each Kohonen map model is equal to n^2. The wavenumber which is located in each cluster are considered as one cluster of variables has similar information. In order to characterize the obtained result of Kohonen map, each cluster can be Nomenclature as $S_{i,j}$, where i and j are row and columns of clusters, respectively. Seven Kohonen SOM networks from the node sizes of 2×2 to 8×8 have been checked. Figure (S3) from supporting information shows the distribution of wavenumbers in the (4×4) Kohonen SOM network. In spite of other interval based pattern recognition methods [7] which divide the variables equally, in clustering of variable strategy each cluster includes different numbers of variable. This is clearly identified in Figure (S3). As it is shown from this figure clusters $S_{2,1}$, $S_{1,2}$,$S_{1,3}$ and $S_{4,3}$ contain a high number of wavenumbers and some of them, such as $S_{3,2}$,$S_{4,2}$ and $S_{3,4}$ **,** have low number of wavenumbers. This is due to fact that the variables have been cluster based on

their similarities (similar information). In the next stage, the meaningful PCs and corresponding loadings of each cluster are extracted by applying the PCA on each cluster separately. Extracted PC of whole clusters builds a new data matrix (D_r, step 3 of GA-LDA based SOM theory) which their columns are significant principal components retained from produced clusters. In other words the columns of original data matrix (wavenumbers) have been replaced with principal components which are extracted from different clusters. This procedure has been done for all Kohonen network sizes. This new data set (D_r) along with LDA algorithm, have been used to constructions the classification algorithms using linear relation by genetic algorithm PC selection. Table 1 lists the statistical classification parameters of the models obtained from different number of clusters through Kohonen SOM method. This table includes the number of total PCs which are extracted from the clusters (N_{EPC}) and the number of selected PCs in the final LDA model (N_{SPC}) using genetic algorithm. The statistical parameters (NER_{cal}, NER_{val} and NER_{pre}) obtained from different GA-LDA based SOM models for saffron discrimination has been shown in the last three columns of Table 1. Now the question is that which of them is the best model? We had a complete discussion in our previous publication related to best model selection [11]. In summary both calibration and prediction results should be considered for optimum model selections. Based on the obtained result in Table 1, it is evident that the number of extracted PCs is increased (from 30 to 86) when the number of clusters or Kohonen nodes is increased (from 4 to 64). However, the number of selected PCs in the LDA analysis remains relatively constant (3 to 4) and are independent of the number of clusters. The Not Error Rate (NER) of calibration, validation and prediction statistics shown in Table I reveal that GA-LDA model obtained from Kohonen nodes q=4, (16 cluster) is the optimum one for both calibration and prediction classification ability. This 16-cluster GA-LDA model which uses three PCs out of

64 extracted PCs, has very high degree of correctly assigned sample (NER) 1.000, 1.000 and 1.000 for calibration, cross-validation and prediction, respectively. The same conclusion can be achieved by looking at Table II. According to the results presented in this table, GA-LDA of network size 4 has higher values of sensitivity, sensitivity describes the model ability to correctly recognize objects belonging to g^{th} class, and specificity for both cross-validation and test set samples than other network sizes.

The clusters and their selected PCs used in the GA-LDA based on SOM modeling of Saffron data are presented in Table III. Figure S3 reveals that the first segment ($S_{1,2}$) is located at the top-left, and $S_{4,1}$ is located at the down left corner of the distribution plane of the wavenumbers. Interestingly, selected PCs cluster $S_{4,1}$ of variables do not have the highest variable and they are chosen based on their correlations with the class information. The selected PCs are representative of the wavenumbers that appeared in these clusters. These wavenumbers have spectral information that is more correlated with class information. To know which subset of wavenumbers are more useful for classification of the saffron samples from adulteration ones, the corresponding loadings of the selected PCs have been searched for variables (wavenumbers) of the highest loading values and those detected are shown in the last column of Table III. It should be noted that GA-LDA based on SOM does not build classification model based on the selected wavenumbers and uses all wavenumbers of the selected clusters for model building. However, it has the ability to identify the most important ones.

Selected wavenumbers region have been shown in Figure 4. As we can see from this figure the wavenumber 1800-1830, 2600-2900 and 3700-3850 cm^{-1} corresponding to C=O stretching of aldehyde and ketone, stretching H-acidic and –OH phenolic can be proposed for food colorant adulteration detection. Finally, the discriminant function plot (DF1) of Kohonen

network size q = 4 is given in Figure 5. As it is evident, a clear separation between samples from the LDA plot of this cluster size is observed. That is, the selected wavenumbers in Figure 4 has high efficiency related to detection of adulteration in saffron samples. Moreover the statistical parameter of PLS-DA has been reported in the last row of Table II. As we can see the PLS-DA analysis (Figure (S4)) also led to promising result but three main purposes have been obtained using proposed algorithm. The first one is that the most information of original data matrix has been maintained. The second one, which is the most important, is that the dimension of data has been reduced. Lastly, the information in the PCs of original data set has been divided into different parts. Moreover with this strategy the small sample size problem of LDA classifier can be solved.

3.3 Regression analysis based on clustering of variable concept

Finally, in order to have the information related to amount of food colorants in saffron samples, regression analysis has been introduced. Since the absorbance spectra has the linear relation with concentration (Beer's law), the absorbance spectral data of infrared has been used for regression propose. Figure (S5) shows the absorbance spectral data of saffron samples which has been contaminate with different amount of food colorants. Here, seven SOM network sizes (2-8) have been examined. The maximum latent variables were set to 10, and the optimum number of latent variables was obtained by five-fold cross validation. In accordance with the results of Table (IV), cluster $S_{4,4}$ from Kohonen network size q=4 has lower error especially for prediction step and can been considered as the most informative ones. This cluster possesses root mean square errors of 0.084, 0.112 and 0.087 for calibration, cross-validation and prediction, respectively. In other word, the variables of this cluster are more informative than the full spectral data for regression

model. Finally we found that the obtained prediction error of this cluster decreases (13.8%) in comparison with conventional PLS regression.

4. Conclusion

In the present study, the application of clustering of variable concept combined with transmittance FT-IR spectra has been used in the quality control of standard saffron samples from food colorant adulteration. The effect of six typical and well known food colorant (Tartrazine, Sunset yellow, Azorubine, Quinoline-yellow, Allura red and Sudan II) have been investigated. Powerful pattern recognition as a useful alternative way, instead of more complex analytical tools for the detection of adulteration can been proposed. Moreover the analysis of such "high correlated" dataset has been introduced for quality control diagnosis and food chemistry using proposed method. The obtained results demonstrate that it is possible to split the information in transmittance FT-IR spectra into informative and redundant ones and used the first ones.

References

1. Petrakis, E. A.; Cagliani, L. R.; Polissiou, M. G.; Consonni, R., Evaluation of saffron (Crocus sativus L.) adulteration with plant adulterants by 1 H NMR metabolite fingerprinting. *Food chemistry* **2015,** 173, 890-896.

2. Melnyk, J. P.; Wang, S.; Marcone, M. F., Chemical and biological properties of the world's most expensive spice: Saffron. *Food Research International* **2010,** 43, (8), 1981-1989.

3. Winterhalter, P.; Straubinger, M., Saffron renewed interest in an ancient spice. *Food Reviews International* **2000,** 16, (1), 39-59.

4. Hagh-Nazari, S.; Keifi, N. In *Saffron and various fraud manners in its production and trades*, II International Symposium on Saffron Biology and Technology 739, 2006; 2006; pp 411-416.

5. Torelli, A.; Marieschi, M.; Bruni, R., Authentication of saffron (Crocus sativus L.) in different processed, retail products by means of SCAR markers. *Food Control* **2014,** 36, (1), 126-131.

6. Eleftherios A. Petrakis, L. R. C., Moschos G. Polissiou, Roberto Consonni, Evaluation of saffron (Crocus sativus L.) adulteration with plant adulterants by 1H NMR metabolite fingerprinting. *Food Chemistry* **2014**.

7. Javidnia, K.; Parish, M.; Karimi, S.; Hemmateenejad, B., Discrimination of edible oils and fats by combination of multivariate pattern recognition and FT-IR spectroscopy: A comparative study between different modeling methods. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2013,** 104, 175-181.

8. Sherazi, S. T. H.; Kandhro, A.; Mahesar, S. A.; Bhanger, M. I.; Talpur, M. Y.; Arain, S., Application of transmission FT-IR spectroscopy for the trans fat determination in the industrially processed edible oils. *Food chemistry* **2009,** 114, (1), 323-327.

9. Yang, H.; Irudayaraj, J.; Paradkar, M. M., Discriminant analysis of edible oils and fats by FTIR, FT-NIR and FT-Raman spectroscopy. *Food chemistry* **2005,** 93, (1), 25-32.

10. Ordoudi, S. A.; de los Mozos Pascual, M.; Tsimidou, M. Z., On the quality control of traded saffron by means of transmission Fourier-transform mid-infrared (FT-MIR) spectroscopy and chemometrics. *Food chemistry* **2014,** 150, 414-421.

11. Karimi, S.; Farrokhnia, M., Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique. *Chemometrics and Intelligent Laboratory Systems* **2014,** 139, 6-14.

12. Anastasaki, E.; Kanakis, C.; Pappas, C.; Maggi, L.; Del Campo, C. P.; Carmona, M.; Alonso, G. L.; Polissiou, M. G., Differentiation of saffron from four countries by mid-infrared spectroscopy and multivariate analysis. *European Food Research and Technology* **2010,** 230, (4), 571-577.

13. Cagliani, L. R.; Culeddu, N.; Chessa, M.; Consonni, R., NMR investigations for a quality assessment of Italian PDO saffron (Crocus sativus L.). *Food Control* **2015,** 50, 342-348.

14. Zalacain, A.; Ordoudi, S. A.; Doaz-Plaza, E. M.; Carmona, M.; Blazquez, I.; Tsimidou, M. Z.; Alonso, G. L., Near-infrared spectroscopy in saffron quality control: determination of chemical composition and geographical origin. *Journal of agricultural and food chemistry* **2005,** 53, (24), 9337-9341.

15 Coates, J. Interpretation of infrared spectra, a practical approach. In R. A. Meyers (Ed.), *Encyclopedia of analytical chemistry* ,**2000**, (pp. 10815–10837). John Wiley & Sons Ltd.

16 Kanou, M., Nakanishi, K., Hashimoto, A., & Kameoka, T. Influences of monosaccharides and its glycosidic linkage on infrared spectral characteristics of disaccharides in aqueous solutions. *Applied Spectroscopy* **2005**, 59, 885–892.

17 Nikonenko, N. A., Buslov, D. K., Sushko, N. I., & Zhbankov, R. G. Spectroscopic manifestation of stretching vibrations of glycosidic linkage in polysaccharides. *Journal of Molecular Structure* **2005**, 752, 20–24.

18 Sun, D.-W. *Infrared spectroscopy for food quality analysis and control* (1st ed.). **2009** New York: Elsevier (chap. 4).

19. Nakanishi K, Solomon PA *Infrared absorption spectroscopy*. **1977** Holden-Day, San Francisco

20. Socrates G *Infrared characteristic group frequencies*. **1997** Wiley, Chichester

21. Pappas CS, Tarantilis PA, Harizanis PC, Polissiou MG. New method for pollen identification by FT-IR spectroscopy. *Applied Spectroscopy* **2003**, 57, 23–27

22. McLachlan, G, Discriminant analysis and statistical pattern recognition, Wiley.com, 2004.

23. Ballabio D, Skov T, Leardi, R, Bro, R, Classification of GC–MS measurements of wines by combining data dimension reduction and variable selection techniques, *Journal of Chemometrics*. **2008**, 22, 457–463.

24. Hemmateenejad, B.; Karimi, S.; Mobaraki, N., Clustering of variables in regression analysis: a comparative study between different algorithms. *Journal of Chemometrics* **2013,** 27, (10), 306-317.

25. Leardi R, Lupiz Gonzalez, A. Genetic algorithms applied to feature selection in PLS regression: how and when to use them, *Chemometrics and Intelligent Laboratory Systems*. **1998**, 41, 195–207.

26. Farrokhnia, M, Karimi ,S. Variable selection in multivariate calibration based on clustering of variable concept. *Analytical Chimica Acta*. 10.1016/j.aca.2015.11.002

27. Snee, R. D., Validation of regression models: methods and examples. *Technometrics* **1977,** 19, (4), 415-428.

28. Smilde, R. B. A. K., Principal component analysis. *Analytical Methods* **2014,** 6, 2812-2831.

Figure legend:

Figure 1. Architecture of a Kohonen self-organizing map or Kohonen network

Figure 2. Transmittance FT-IR spectra of the saffron samples used in this study: (a) Standard saffron (b) Azorubine (c) Quinoline yellow (d) Allura red (e) Sudan (II) (f) Sunset yellow (g) Tartrazine (h) Extended multiplicative scatter correction preprocessed saffron data set.

Figure 3. Distribution pattern of the saffron samples in the PCA factor spaces of their transmittance FTIR spectra for extended multiplicative scatter correction.

Figure 4. Selection the important variables using GA-LDA based dimension reduction for discrimination of saffron samples

Figure 5. Classification using GA-LDA based on dimension reduction technique for adulteration in saffron data set. The circles and asterisks have been used to show the calibration and prediction samples respectively.

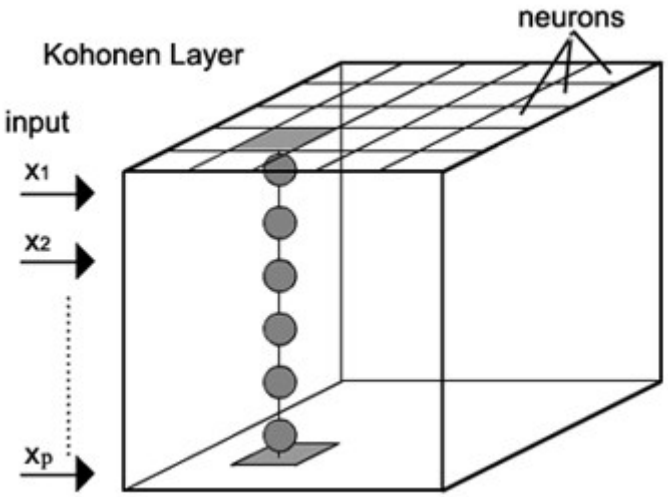
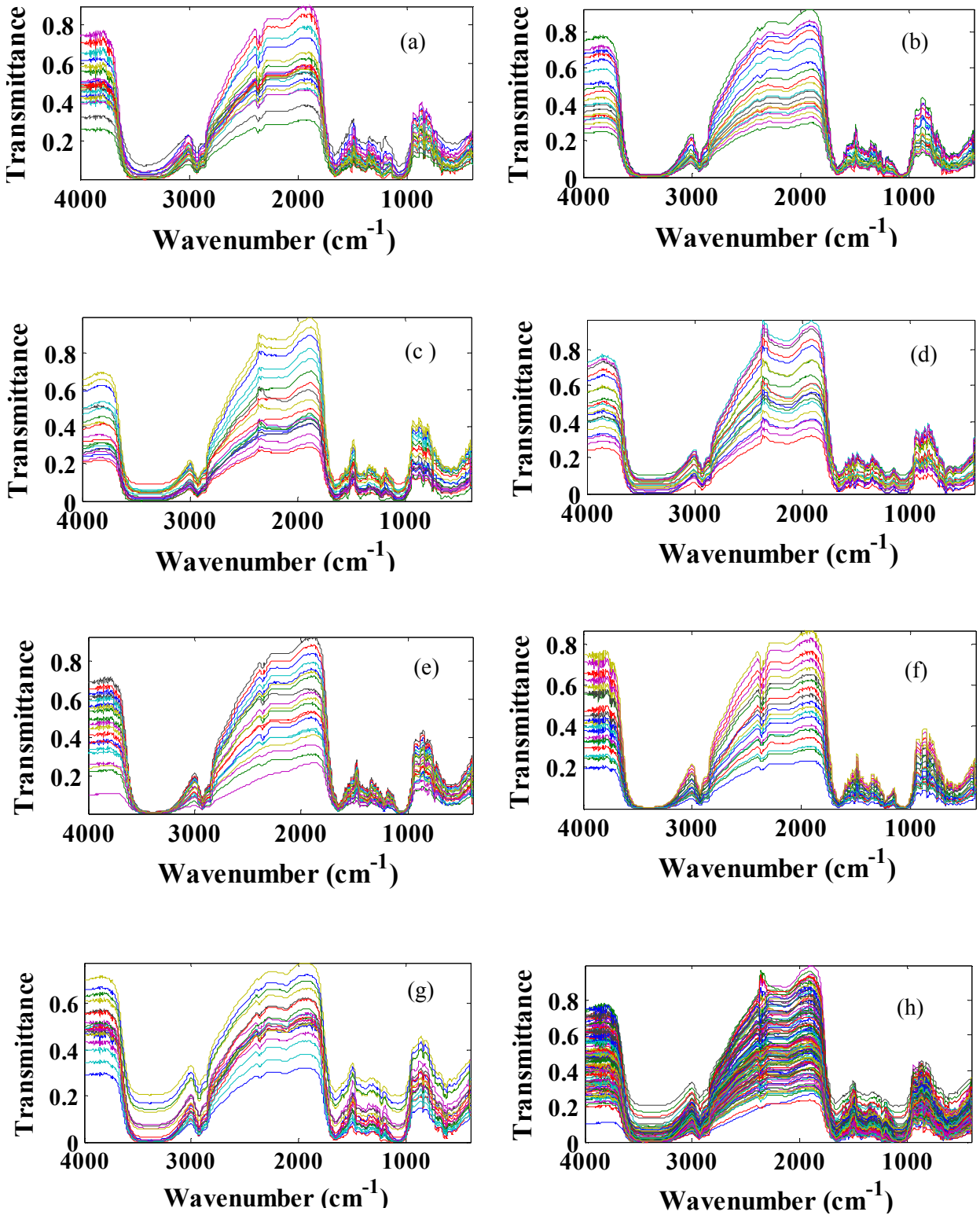


Figure 1



548 **Figure (2)**

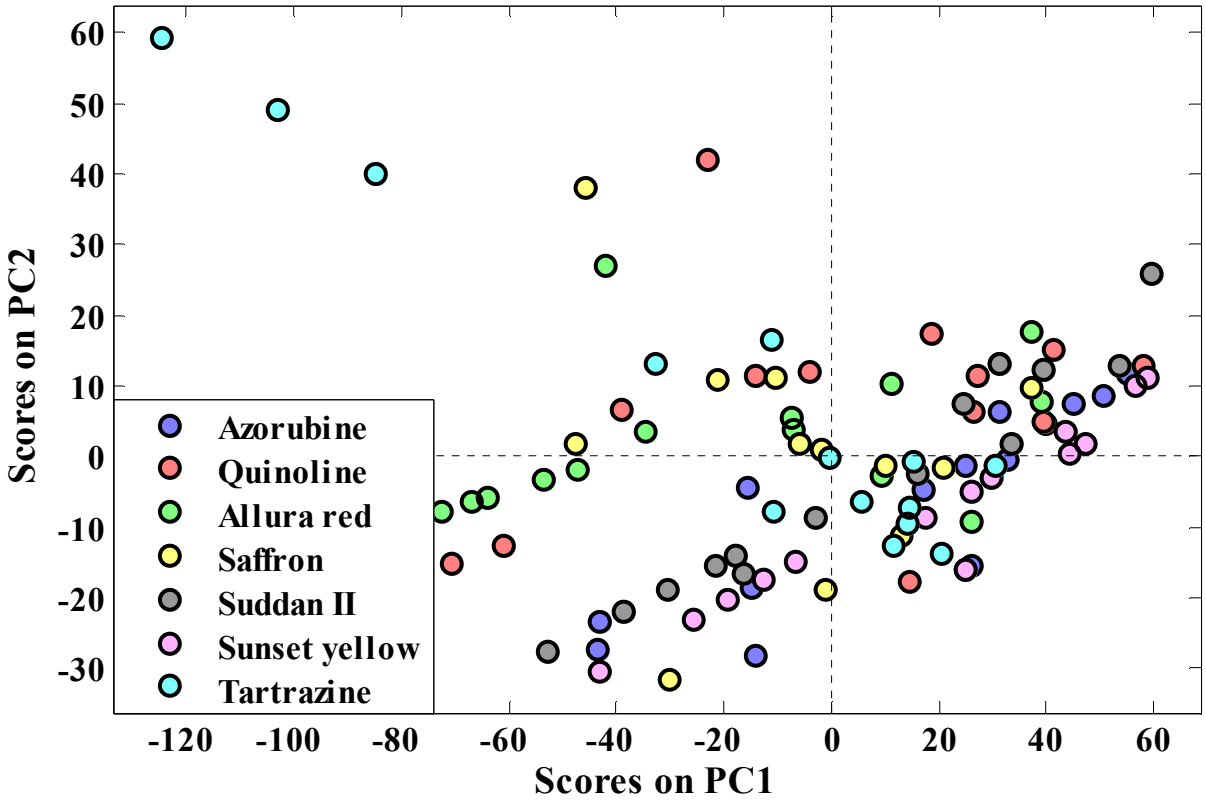
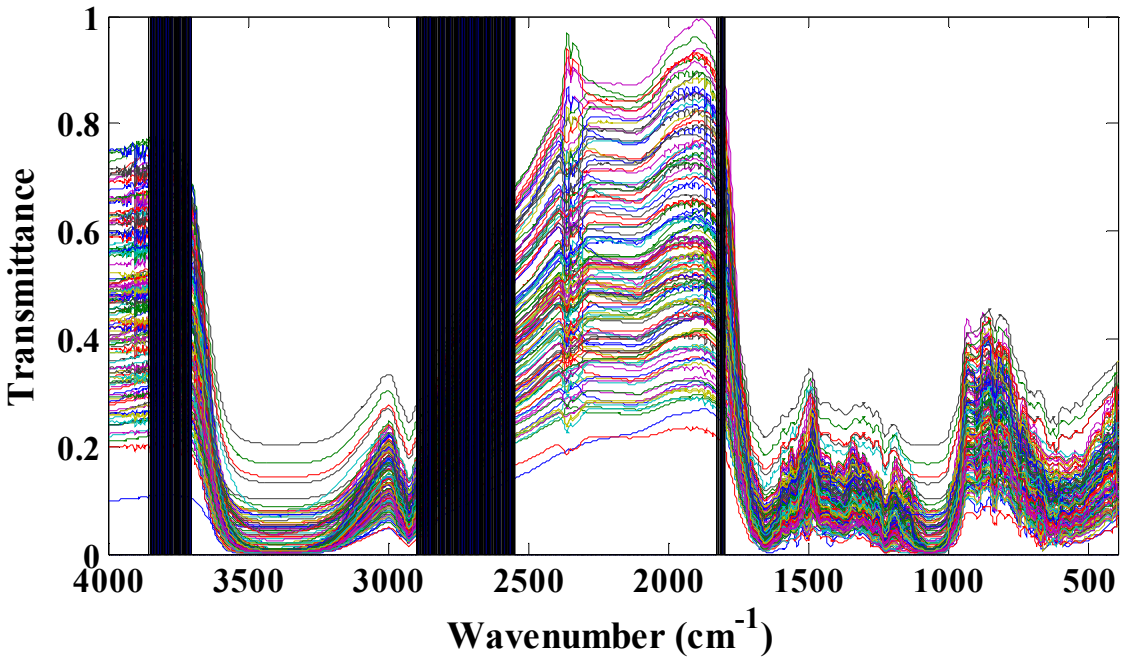


Figure 3

559



560

561

562 **Figure 4**

563

564

565

566

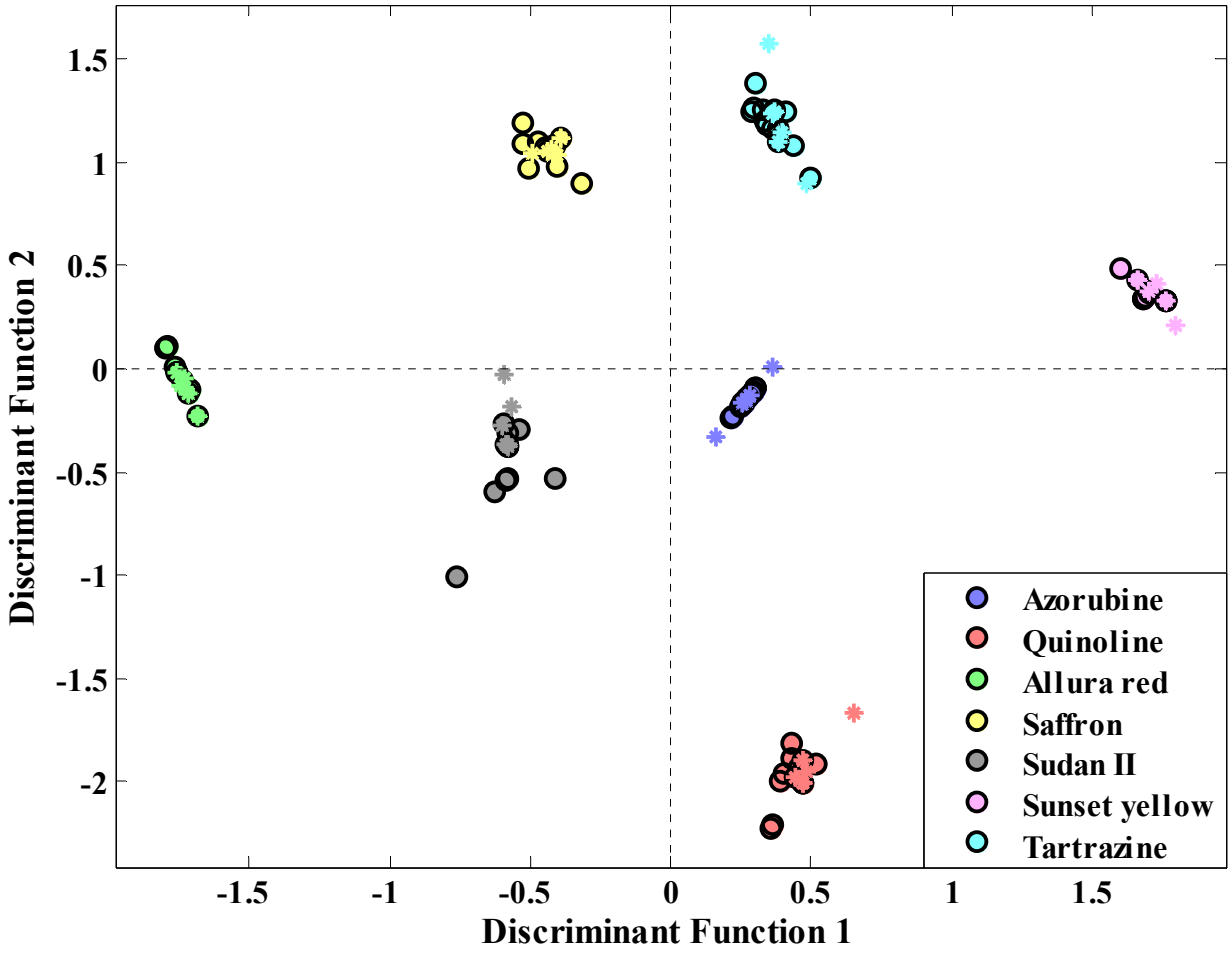


Figure 5

Table 1. Statistical parameters of the GA-LDA based dimension reduction models obtained from different clusters (nodes of the Kohonen network): Transmittance FT-IR of saffron data set

Number of segments (Kohonen nodes)	N_{EPC} [a]	N_{SPC} [b]	NER_{cal} [c]	NE_{val} [d]	NER_{pre} [e]
4 (2×2)	30	2	0.94	0.84	0.80
9 (3×3)	41	3	1.00	0.86	0.85
16 (4×4)	**64**	**3**	**1.00**	**1.00**	**1.00**
25 (5×5)	70	3	1.00	0.93	0.92
36 (6×6)	75	3	1.00	0.87	0.85
49 (7×7)	80	4	0.98	0.84	0.80
64 (8×8)	86	4	0.97	0.82	0.75
PLS-DA model	---	4	1.00	0.93	0.91

[a] Number of the extracted PCs from all clusters.

[b] Number of selected PC.

[c] Not error rate based on leave many out cross validation for calibration set

[d] Not error rate based on leave many out cross validation for validation set

[e] Not error rate based on leave many out cross validation for prediction set

Table II. Sensitivity (S_n)[a] and specificity (S_p)[b] achieved by different cluster size for proposed algorithm.

	(2×2)		(3×3)		(4×4)		(5×5)		(6×6)		(7×7)		(8×8)	
	CV[c]	Test	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test	CV	Test
Specificity	0.85	0.75	0.87	0.79	**1.0**	**1.0**	0.91	0.85	0.88	0.80	0.85	0.75	0.82	0.70
Sensitivity	0.86	0.85	0.90	0.88	**1.0**	**1.0**	0.95	0.93	0.93	0.88	0.86	0.85	0.80	0.75

[a]Class sensitivity (S_n) describes the model ability to correctly recognize samples belonging to the g^{th} class, i.e. if all the samples belonging to *g* are correctly assigned, S_n is equal to 1.

[b]Class specificity (S_p) describes the model ability to reject samples of all the other classes from class g^{th}, i.e. if samples not belonging to *g* are never assigned to *g*, S_p is equal to 1.

[C] Cross validation

Table III. The Analysis of the clusters used in the 16-cluster GA-LDA based SOM model of saffron data set considering the number of wavelengths in the clusters (N_W), number of extracted and selected PCs (N_{EPC} and N_{SPC}, respectively) and the selected wavenumbers of the highest loading value (SW)

Cluster	N_W	N_{EPC}	N_{SPC}	Selected PC[a]	SW (cm^{-1})
$S_{1,2}$	161	5	1	PC_1	1800-1830
$S_{4,1}$	160	5	2	PC_1-PC_2	3700-3850-2600-2900

[a] The subscript numbers denotes the order of PCs with respect to the variance explained of their corresponding sub-matrix

631

632

633 **Table IV**. Statistical parameter of the optimum cluster of network size (q=4) in CLoVA – PLS
634 regression and conventional PLS for food colorant adulteration
635
636

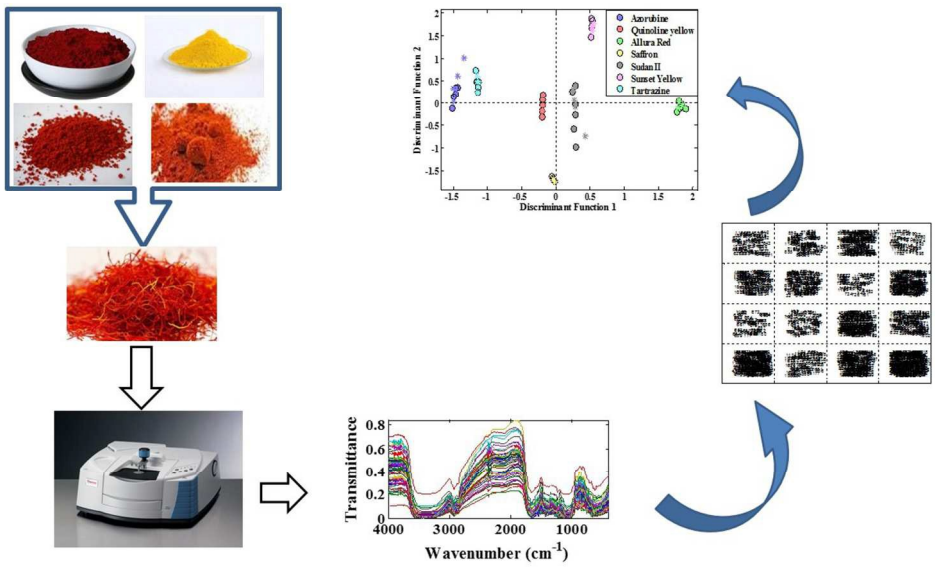
Regression model	N_w[a]	R^2_C	RMSC	RMSCV	R^2_P	RMSEP
CLoVA-PLS ($s_{4,4}$ of network size q=4)	141	0.928	0.084	0.112	0.926	0.087
PLS	1868	0.875	0.111	0.135	0.844	0.101

637 [a] Number of wavenumber

638

639

640



338x190mm (96 x 96 DPI)