International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 3 No. 2 June 2013, pp. 493-503

http://www.issr-journals.org/ijias/

Classification of olives from Moroccan regions by using direct FT-IR analysis: Application of support vector machines (SVM)

Wafa Terouzi[1], Stefan Platikanov[2], Anna de Juan Capdevila[3], and Abdelkhalek Oussama[1]

[1]Laboratory of applied and environmental Spectrochemistry,
Faculty of Science and Technology of Beni Mellal,
University of Sultan Moulay slimane,
21000- Beni Mellal, Morocco

[2]Department of Environmental Chemistry,
IDAEA-CSIC, Jordi Girona 18-26,
08034 Barcelona, Catalonia, Spain

[3]Department of Analytical Chemistry,
Faculty of Chemistry, University of Barcelona,
Martí i Franquès 1-11, 08028 Barcelona, Catalonia, Spain

ABSTRACT: The aim of this work was to characterize and classify three close regions of olives by direct analysis on the olive without any preliminary treatment. This study was focused on the olive samples picked in the three zones: named Bazaza, oled ayad and oled hamdan, in the Moroccan region of Beni Mellal. All samples were also analysed by FT-IR spectroscopy, the spectral data were subjected to a preliminary derivative transform based on the gap segment algorithm to reduce the noise and extract a largest number of analytical information from spectra. A multivariate statistical procedure based on cluster analysis (CA) coupled to support vector machines (SVM), was elaborated, providing an effective classification method. On the basis of a hierarchical agglomerative CA and principal component analysis (PCA), three distinctive clusters were recognized. The SVM procedure was then applied to classify samples from the same regions. The model resulted able to separate the three classes and classify new objects into the appropriate defined classes with a percentage prediction of 93%. The results showed that FTIR spectroscopy coupled with chemiometric methods are an interesting technique for classifying olive samples according to their geographical origins.

KEYWORDS: Olives, FT-IR spectroscopy, Chemometrics, geographical origin, Support Vector Machines.

1 INTRODUCTION

Nowadays, one of the major issues regarding food products is to develop objective tools to determine the origin of raw materials as well as finished products in order to ensure their traceability [1].

Traceability is particularly relevant to assessing origin and content of the olive oil but also to protect and prevent frauds at different stages in oil production. In this case, the quality of virgin olive oil is principally a function of four parameters: varietal and geographic origin, olive fruit quality and extraction process. The oil content in olive fruit was shown to have a high variability between cultivars [2].

A denomination of a cultivar is recognized using different morphological descriptors (tree, leaf, fruit, and stone). However, on the basis of these criteria, only very dedicated specialists are able to perform a strict differentiation between all

Corresponding Author: Wafa Terouzi (wafaterouzi@yahoo.fr)

the cultivated varieties. The origin and the authenticity of virgin olive oils have been the object of many studies in the past few years [3], [4].

The application of spectroscopy which includes IR and Raman techniques combined with chemometric methods is a relatively new approach in food characterization studies [5]. Fourier transform infrared spectroscopy (FTIR) has been successfully used to quantify a number of olive oil parameters [6]. This technique is fast, simple to perform and does not require sample pre-treatment.

Fourier transform infrared (FTIR) spectroscopy, a widely used and well-established tool for structure elucidation and quality control in various industries application, has gradually entered into the identification and classification of natural products like herbal medicines [7], microorganisms [8] and foods [9]. The obvious advantage of FT IR is not only in the effective specificity but also its rapid and nondestructive nature.

FTIR has been proposed to authenticate extra virgin olive oils or to detect adulteration of virgin olive oil [10], [11]. Some attempts in using FTIR to distinguish olive oils from different geographical origin and different genetic varieties have been proposed [6], [12], [13], [14]. This approach has demonstrated to be very useful in many applications, due to the ability in achieving the spectral resolution of the FTIR signals [15]. But, to our knowledge, little has been done to determine the origin of raw materials with morphological characterization [16], to identify olive varieties [17], [18], [19].

The aim of this study was to show the advantages of combined MIR spectroscopy associated with chemometric treatment for a direct and rapid test method used firstly to provide geographical origin recognition of olive fruits and secondly to demonstrate the capability of this new technique to distinguish from regions very close.

2 MATERIALS AND METHODS

2.1 SAMPLING

The study area is confined to province of Beni-Mellal in central Morocco, expanded on a surface of nearly 7100 Km². Investigation was focused on the olive samples (of Picholine Marocaine variety) picked in the zones named Ouled Hamdan, ouled Ayad and Bazzaza. Altitude of these regions is 600 m, temperature ranges from 3.5 °C to 48 °C and precipitation rate is 300 to 750 mm.

Sampling is an important step. Indeed, the reliability and robustness of the method adopted repose largely on the choice and number of samples. We created two different collections of samples. The first includes 24 olives were manually harvested from three farms belong to the three areas.The second consists of 24 olives were randomly collected from an extraction unit of olive oil.

48 olives chosen from the harvested olives based on their size, maturity and the absence of surface defects. Similar size olives were chosen to minimize the effect of size on spectral measurements.

A series of 14 samples (from second collection) was used as an external validation set. This last series was used to establish the robustness of the SVM model. whilst the remaining 34 sample were selected to build up the calibration model

The olive samples were kept in cold storage (7°C) during the nights between the days of measurements. Spectroscopic measurements were taken from the olives after they had been brought into equilibrium with the room temperature of 25°C.

2.2 SPECTROSCOPIC MEASUREMENTS

Spectra were recorded from 4000cm-1to 600cm-1, with 4cm−1 resolution and 98 scans on a “Vector 22 Bruker” spectrometer, equipped with a DGTS detector, an Globar source and a KBr/Germanium beam splitter. Olive samples were directly deposited between two well-polished KBr plates, without preparation on an Attenuated Total Reflectance cell provided with a diamond crystal. The background spectrum was recorded on air for of each sample.

Spectra were scanned in the absorbance mode from 4000 to 600 cm^{-1} and the data were handled with OPUS logiciel. The software (Opus 4.0 MSD) fitted to the infrared spectrometer Fourier transform used in this study allows the automatic acquisition of the spectra without any form of computer manipulation may impair the quality of results. The Fourier transform is automatically calculated by the software prior to the acquisition of spectra.

Between spectra, the ATR plate was cleaned in situ by scrubbing with ethanol solution, enabling to dry the ATR.

For ATR-FTIR measurements, it was necessary to keep a controlled pressure, to ensure good contact between the sample and the diamond surface.

2.3 CHEMOMETRIC METHODS

2.3.1 CLUSTER ANALYSIS

Cluster analysis is a non-supervised technique, represents a series of multivariate methods which provide means for classifying a given population into groups (clusters), based on similarity or closeness measures. The objective principle of the distance is adopted for this aim. The agglomerative hierarchical clustering is nowadays one of the most cited methods in literature [20], providing intuitive similarity relationships between each sample and the entire data set.

In hierarchical clustering, each cluster is subdivided into smaller clusters, forming a tree-shaped data structure or dendrogram. Agglomerative hierarchical clustering starts with the single-gene clusters and successively joins the closet clusters until all genes have been joined into the supercluster: The sample grouping is illustrated by a dendrogram that permits a global vision of the similarity among the objects. In fact, there is a whole family of clustering methods, differing only in the way intercluster distance is defined [21].

In this work, as hierarchic agglomerative cluster algorithm, the complete linkage (largest distance between any two members) algorithm was adopted to process the similarity and the distance elaboration was performed using correletion distance. This distance is based on the Pearson correlation coefficient that is calculated from the sample values and their standard deviations. The correlation coefficient r takes values from –1 (large, negative correlation) to +1 (large, positive correlation). Effectively, the Pearson distance **dp** is computed as

$$\mathbf{dp = 1- r} \qquad (1)$$

And lies between 0 (when correlation coefficient is +1, i.e. the two samples are most similar) and 2 (when correlation coefficient is -1). Note that the data are centered by subtracting the mean, and scaled by dividing by the standard deviation [22], [23].

2.3.2 PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal component analysis (PCA) is a non-supervised statistical tool commonly used for classification of data. The main aim of PCA is to reduce a large number of variables to a much smaller number of principal components (PCs) that capture the vast majority of variance in the data. This reduces the dimensionality of the data considerably, enabling effective visualization, regression and classification of multivariate data [24], [25].

2.3.3 SUPPORT VECTOR MACHINES (SVM)

SVM is a supervised learning technique for classification and regression that uses linear or non-linear kernel-functions to project the data into a high-dimensional feature space. Correlation is then performed in this hyperspace based on the structural risk minimization principle; *i.e.*, aiming to increase the generalization ability of a model [26], [27].

Two SVM classification types are available in The Unscrambler logiciel which are based on different means of minimizing the error function of the classification:

- c-SVC: also known as Classification SVM Type 1.
- nu-SVC: also known as Classification SVM Type 2.

In the c-SVM classification, a capacity factor, C, can be defined. The value of C should be chosen based on knowledge of the noise in the data being modeled. Its value can be optimized through cross-validation procedures. When using nu-SVM classification, the “nu” value must be defined (default value = 0.5). Nu serves as the upper bound of the fraction of errors and is the lower bound for the fraction of support vectors. Increasing “nu” will allow more errors, while increasing the margin of class separation [22].

The choice of SVM as classification method is justified by the results obtained in References [28], [29], [30], [31], where the great performance of SVM becomes evident, and several authors have shown, that support vector machines provide a fast and effective means for classification.

2.3.4 SOFTWARE

The chemometric applications were performed by using the Unscrambler software version 10.2 from CAMO (Computer Aided Modeling, Trondheim, Norway).

3 RESULTS AND DISCUSSION

Fourier transform infrared (FTIR) spectra of 48 olive samples were recorded and divided in two sets: a calibration set of 34 samples and a prediction set of 14 samples. A mean spectrum was calculated for each region of calibration set. The resultant spectra are shown in Fig.1.

Fig.1 shows the mean FTIR spectra of the studied olives. The differences among them were clearly small and occurred only in limited regions of the spectra. The obtained spectra are dominated by typical bands of holocellulosic materials in the 900-1200 cm^{-1} region [32].The significant bands of water are clearly visible in the olive spectra at 3400 cm^{-1}. The band of aromatic ring stretch of lignin should appear at 1604 cm^{-1}. However, this region was obscured by the strong water deformation band centered at 1638 cm^{-1}. The typical infrared pattern of lignocellulosic materials is observed in the region 900-1200 cm^{-1}. The two bands at 2924 cm^{-1} and 2848 cm^{-1} are characteristic of olive oil, while the range 2400 - 2300 cm^{-1} is due to CO2.

The use of single peaks or narrow wavelength ranges to obtain information useful to distinguish the olives seemed very hard. These data were so conveniently handled by multivariate statistical techniques. With the aim to obtain more information from the FTIR spectral data, the spectra were firstly subjected to mathematical elaboration. In particular, derivative transformations were applied [33], [34].The best improvement in data variance was reached when the derivative function through the Gap segment algorithm was used. Best results were obtained by fixing the following parameters: 2nd order, gap size 7 and segment size 5, with mean-centered data.

3.1 CLUSTER ANALYSIS

The calibration data set obtained from derivative transformation of the FTIR data was employed to perform CA, applying complete linkage clustering. Results were reported in the form of dendrogram, shown in **Fig.2**. On the basis of the connecting distances three distinctive clusters were defined.CA proved highly selective in grouping the olive samples. In fact, while belonging to the same variety (Picholine Marocaine), the method was able in aggregating the samples from different regions.

It is noteworthy that, for the analytical responses, the water content in these samples could be considered significant, as demonstrated by the relevant change in the broad band between 4000 and 3000 cm^{-1}, or due to the variation of the oil content and fatty acid relative rates of samples of each class. In the other hand, since the three regions are characterized by the same climatic conditions, we can say that the distinction between the classes is the result of a change in the composition of the soil.

3.2 PCA MODELING

PCA model was built by the multivariate decomposition of the FTIR data in the ranges 4400–2400 and 2300–600 cm^{-1}. When the model was validated by full cross-validation procedure [35], PCA showed a clear separation of the three classes. The explained variance (%) obtained from the full cross-validation of the PCA model and three PCs were selected for a complete description of the variance in the spectral data set: 98%. In the PCA model, the first two components achieved an explained variance of 94%, which is enough to cluster the samples in the three classes, as can be seen in the score plot of Fig. 3.

Thus, we see on plot loadings that the first PC has 82% of the variance (Fig.3 (a)), while the second PC only 12% (Fig.3 (b)). These two loadings plot highlight the variables responsible for the distinction between the three classes, it is the spectral range characteristic of oleic acid composition and characteristic band of water.

3.3 CLASSIFICATION MODEL: SVM MODELING

The SVM model was built by considering, as X variables, the spectra in the range 4000-2400 and 2300-600cm^{-1}; and the classification model was validated by Cross validation with segment = 30.

We constructed SVM model with a nu-SVC classification, different kernels have been tested on these data, and the results showed that the best choice is the linear kernel, to determine the hyperplane that give best separates the classes. The optimal parameter for "nu" which lies in the range 0-1, is then selected as the value that give the maximum correct classification rate, nu = 0.5. Consequently, a larger number of calibration samples are retained as support vectors, it is 28, where 10 of Bazzaza samples, 9 of Oled ayad samples and 9 of Oled hamdan samples.

Application of SVM with Cross validation on a set of thirty-four samples allowed a classification with accuracy of 100% in training and 97% in validation, which can be considered satisfactory.

The main result of the SVM is the confusion matrix, which indicates how many samples were classified is each class, and the prediction matrix, which indicates the classification determined for each sample in the training set.

The confusion matrix is a matrix used for visualization for classification results from supervised method, support vector machine classification. It carries information about the predicted and actual classifications of samples, with each row showing the instances in a predicted class, and each column representing the instances in an actual class. Look at the confusion matrix (Table 1). All the samples are well classified.

This result is confirmed by classification plot (Fig.4), in fact, based on three characteristic wavelengths of water and oleic acid, detected as significant variables responsible for the distinction between the three classes in loadings plot of PCA, all the classes resulted perfectly separated from the other ones.

3.4 CLASSIFICATION OF NEW SAMPLES

In this step, the model was subdued to validation procedure by classifying the new objects in to the classes previously established.

The SVM model was applied to a group of unknown samples from different olives of three regions (14 external olive samples), the results are listed in Table 2.

Table 2 shows the classification results with the comparison between the predicted results of each class and the theoretical reference classes. The rate of correct classification was 93% within the test set. In particular13 samples were safely assigned in the three classes, while one sample O8 resulted classified in another class.

4 TABLES AND FIGURES

4.1 TABLES

Table 1. Confusion matrix of calibration set, carried out by SVM

Confusion matrix	BAZ	OAY	OHD
Predicted	**1**	**2**	**3**
BAZ	10	0	0
OAY	0	10	0
OHD	0	0	14

Table 2. Classification of olive samples of the prediction set by using SVM model

Samples	Predicted	Reference class
O1	BAZ	BAZ
O2	BAZ	BAZ
O3	BAZ	BAZ
O4	BAZ	BAZ
O5	OAY	OAY
O6	OAY	OAY
O7	OAY	OAY
O8	OH	OAY
O9	OH	OH
O10	OH	OH
O11	OH	OH
O12	OH	OH
O13	OH	OH
O14	OH	OH

4.2 FIGURES

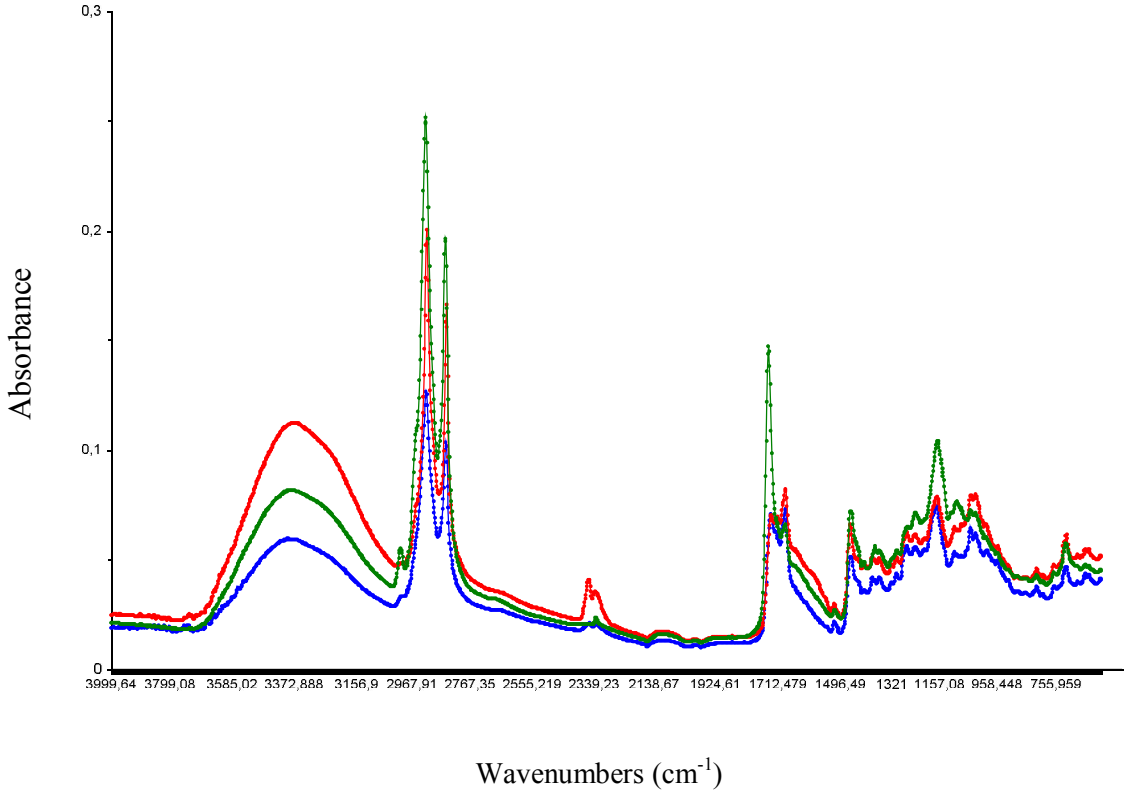


Fig. 1. The mean spectra calculated for each class: Oled hamdan (Oh), Oled ayad (Oay) and Bazaza (Baz)

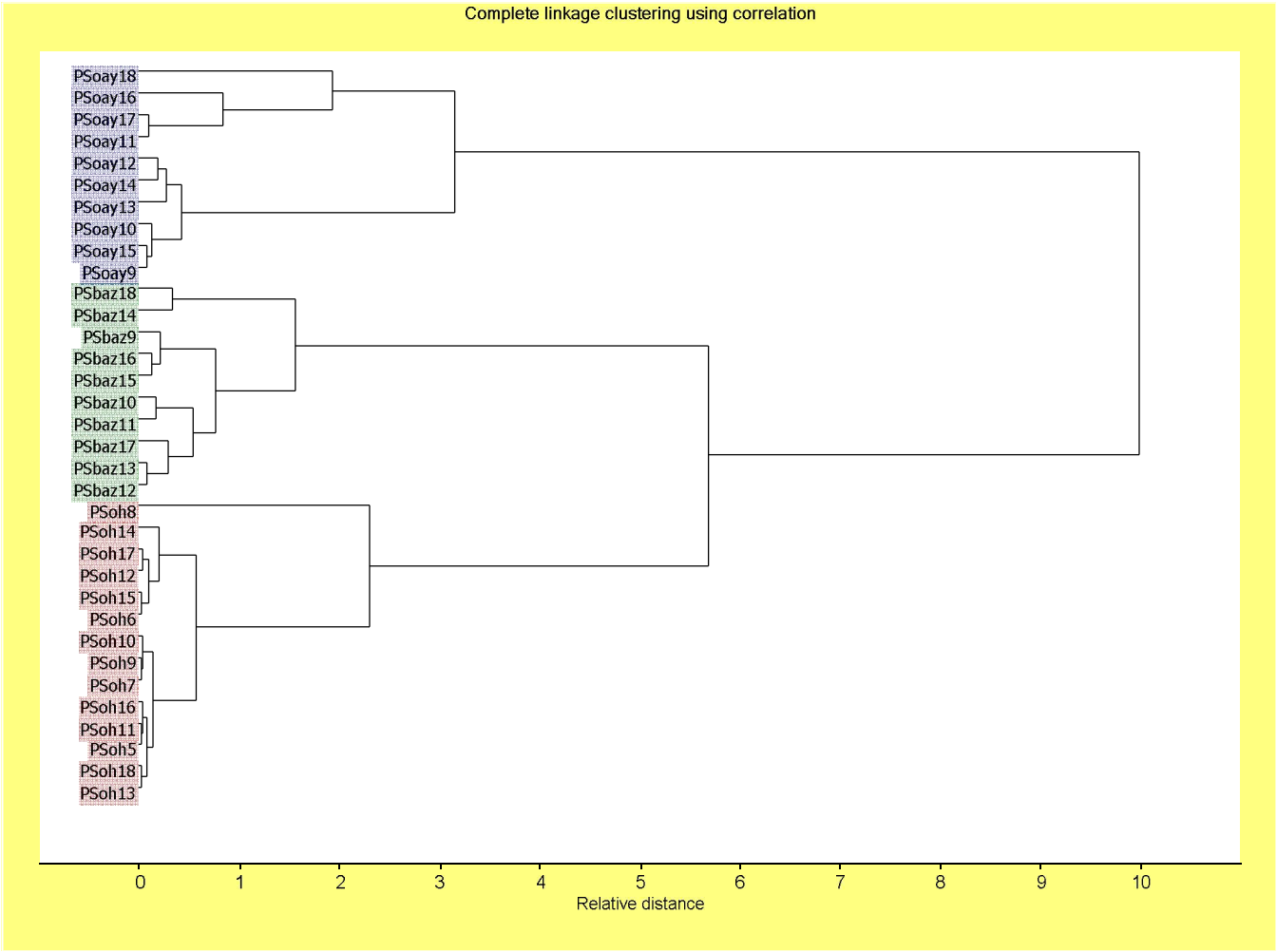


Fig. 2. Dendrogram by CA analysis on the calibration set

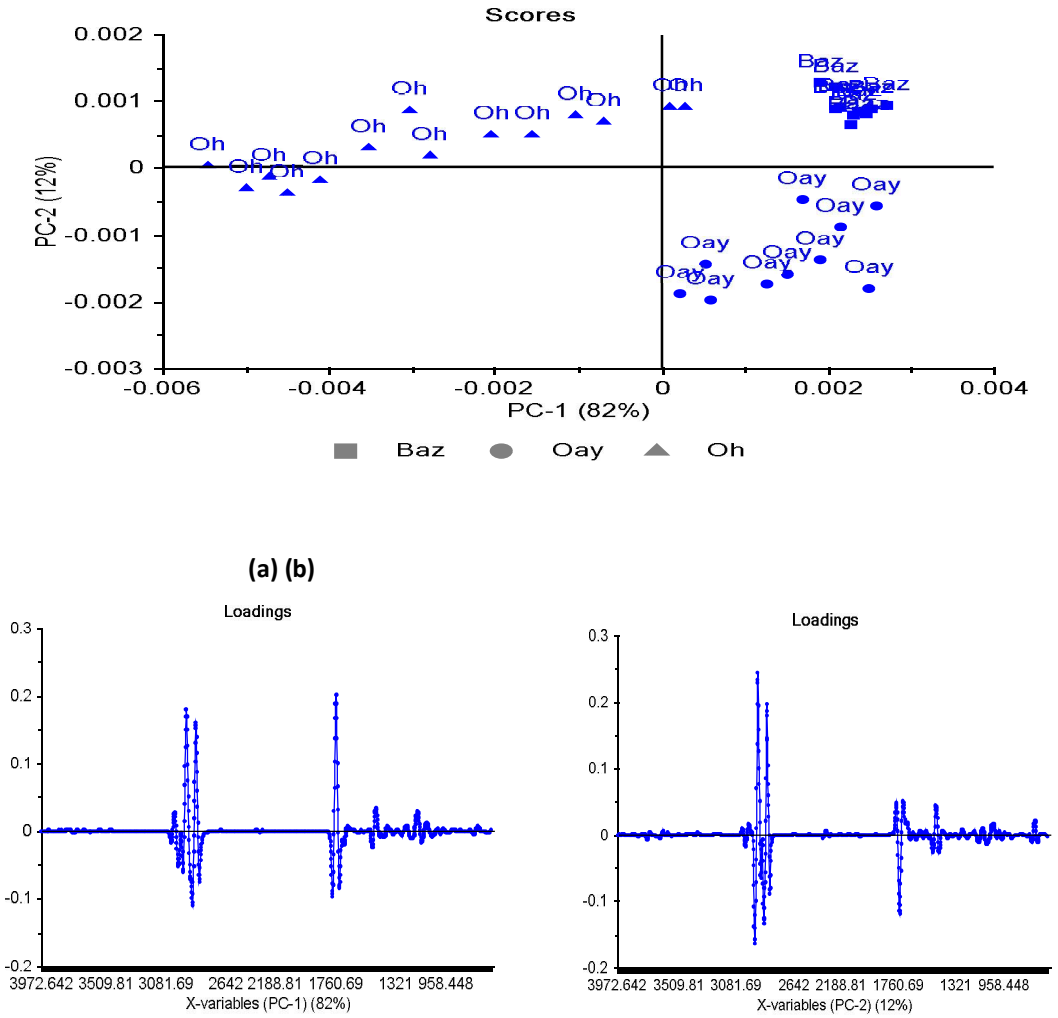
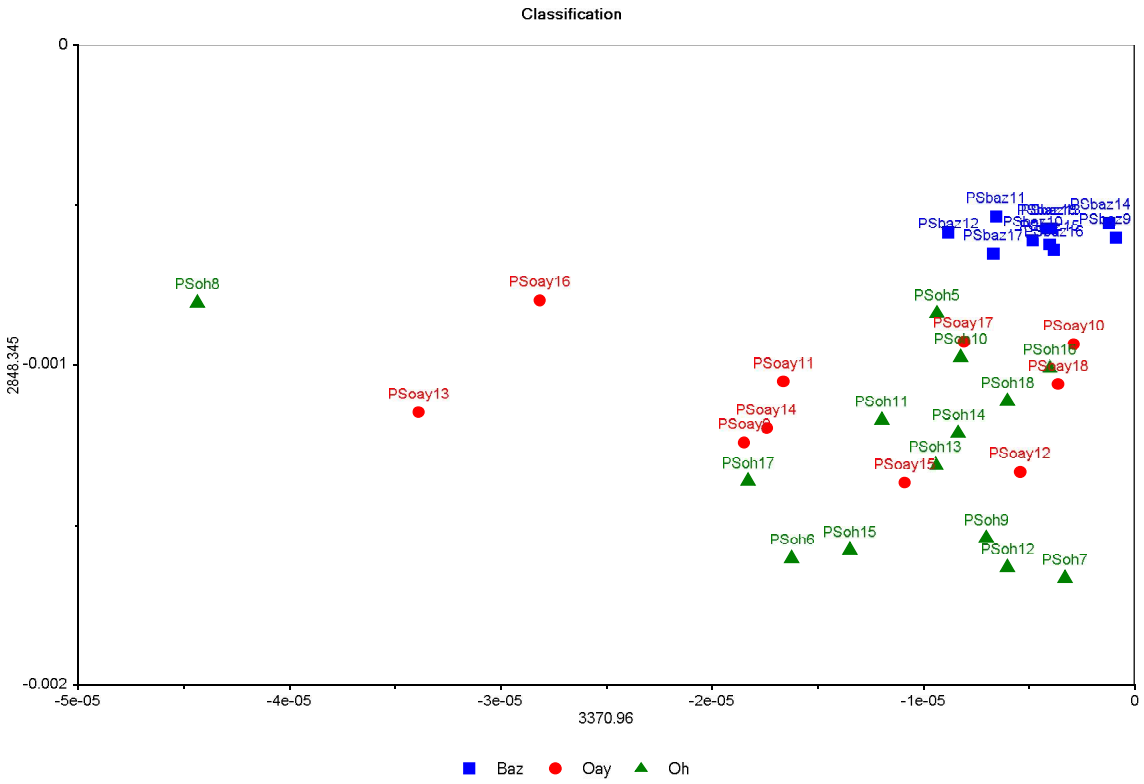


Fig. 3. PC1 / PC2 Score plot by PCA analysis on the calibration set: Oled hamdan (Oh), Oled ayad (Oay) and Bazaza (Baz); (a) first principal component; (b) second principal component

(a)



(b)

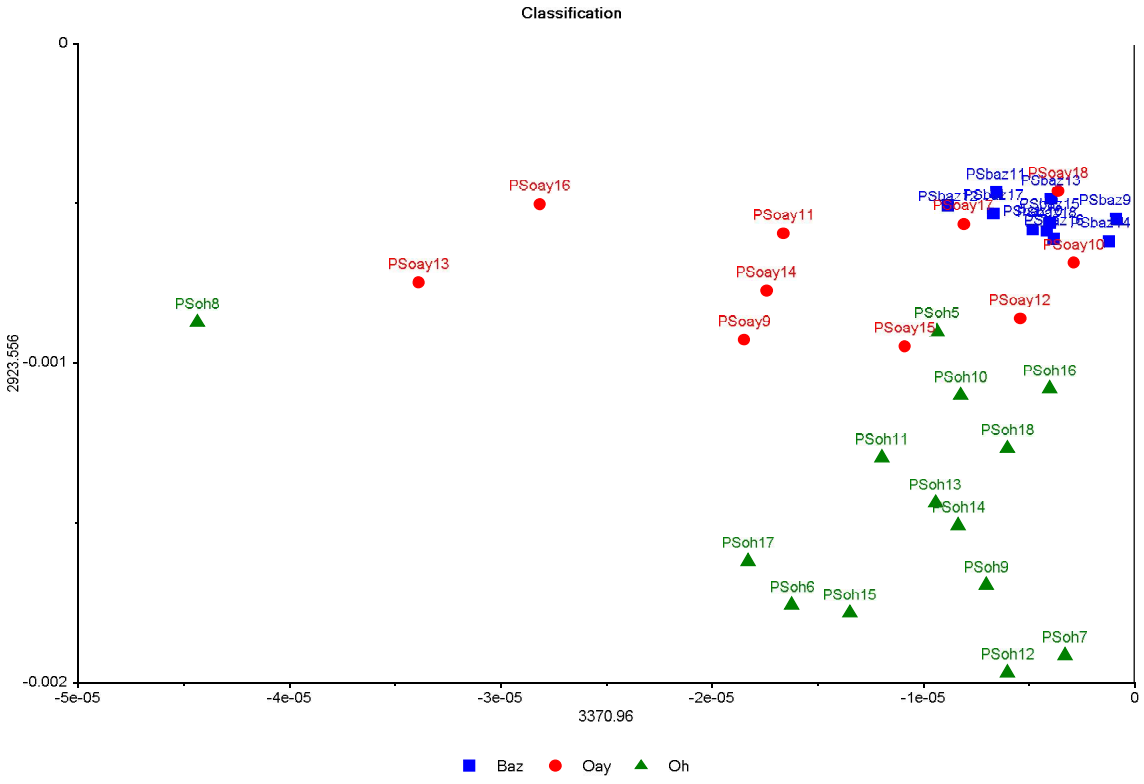


Fig. 4. 2D score plot of classification results by SVM on calibration. (a) with 2 wavelenths 3370 / 2848 cm^{-1}; (b) with 2 wavelenths 3370 / 2923 cm^{-1}

5 CONCLUSION

This study shows that olives of three regions close can be discriminated by differences in their FTIR spectra. Synchronous IR spectroscopy combined with multi-dimensional chemometric techniques is successfully applied to the classification of olives according to their geographical origin. The method presented in this study can be used in olive oil production facilities for the rapid quality control of raw material based on olives as spectra are acquired from samples 'as-received' without any pretreatment.

Then, the spectroscopic methods, combined with chemometric strategies, could represent a reliable, cheap and fast classification tool, able to draw a complete fingerprint of a food product, describing its traceability.

REFERENCES

[1] Baeten, V., Fernandez Pierna, J.A., Dardenne, P., Meurens, M., García González,D.L. and Aparicio Ruiz, R. (2005). Detection of the presence of hazelnut oil in olive oil by FT-raman and FT-MIR spectroscopy. J. Agric. Food Chem. 53 (16), 6201-6206.

[2] Uceda, M., Hermoso, M., Garcia-Ortiz, A., Jimenez, A., et al. Intraspecific variation of oil content and the characteristics of oils in olive cultivars. Acta Hortic. ISHS, 1999, 474.

[3] Bertran, E., Blanco, M., Iturriaga, H., Maspoch, S., et al. Near infrared spectometry and pattern recognition as screening methods for the authentification of virgin olive oils of very close geographical origins. J. Near Infrared Spec., 2000, 8, 45–52.

[4] Galtier, O., Dupuy, N., Le Dre´au, Y., Ollivier, D., et al. Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra. Anal. Chim. Acta, 2007, 595, 136–144.

[5] Yang, H., Irudayaraj, J., Paradkar, M.M. Food Chem. 2005, 93, 25–32.

[6] Tapp, H. S., Defernez, M., Kemsley, E. K. FTIR Spectroscopy and multivariate analysis can distinguish the geographical origin of extra virgin olive oils. J. Agric. Food Chem., 2003, 51, 6110–6115.

[7] Liu, H. X. and Sun, S.Q. Modern Instrum., 2005, 11, 6.

[8] Maquelin, K., Kirschner, C., Choo-Smith, L.P. and Microbiol, J. Methods, 2002, 51, 255.

[9] Reid, L.M., O'Donnell, C.P. and Downey, G. Trends Food Sci. Technol., 2006, 17, 344.

[10] Lerma-García, M.J., Ramis-Ramos, G., Herrero-Martínez, J.M. and Simó-Alfonso, E.F. Food Chem. 2010, 118, 78-83.

[11] Ozen, B.F. and Mauer, L.J. J. Agric. Food Chem., 2002, 50, 3898-3901.

[12] Sinelli, N., Cerretani, L., Di Egidio, V., Bendini, A. and Casiraghi, E. Food Res. Int. , 2010, 43, 369-375.

[13] De Luca, M., Terouzi, W., Ioele, G., Kzaiber, F., Oussama, A., Oliverio, F., Tauler, R., Ragno, G. Food Chem., 2011. doi:10.1016/j.foodchem.2010.07.010.

[14] Gurdeniz, G., Tokatli, F. and Ozen, B. Eur. J. Lipid Sci. Technol., 2007, 109, 1194-1202.

[15] Zagonel, G.F., Peralta-Zamora, P. and Ramos, L.P. Talanta, 2004, 63, 1021-1025.

[16] Pinheiro, P. B. M. and da Silva, J. Chemometric classification of olives from three Portuguese cultivars of Olea europaea L. Anal. Chim. Acta, , 2005, 544, 229–235.

[17] Terouzi, W., DeLuca, M., Bolli, A., Oussama, A., Patumi, M., Ioele , G., Ragno, G., (2011). A discriminant method for classification of Moroccan olive varieties by using direct FT-IR analysis of the mesocarp section, Vib. Spectrosc., doi:10.1016/j.vibspec.2011.01.004

[18] DeLuca, M., Terouzi, W., Kzaiber, F., Ioele , G., Oussama, A. and Ragno, G. Classification of moroccan olive cultivars by linear discriminant analysis applied to ATR–FTIR spectra of endocarps, International Journal of Food Science and Technology, 2012. doi:10.1111/j.1365-2621.2012.02972.x

[19] Dupuy, N., Galtier, O., Le Dréau, Y., Pinatel, C., Kister, J. and Artaud, J. Chemometric analysis of combined NIR and MIR spectra to characterize French olives. Eur. J. Lipid Sci. Technol. 2010, 112, 463–475.

[20] Reynolds, A. P., Richards, G., de la Iglesia, B., and Rayward-Smith, V. J. (2006). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modelling and Algorithms, 5, 475–504.

[21] Patrik D'haeseleer. How does gene expression clustering work?. Journal of Nature biotechnology, volume 23, number 12, December 2005, pp 1499-1501.

[22] Esbensen, K. H. Multivariate data analysis – In practice. An introduction to multivariate data analysis and experimental design. Oslo: CAMO Process, 2002.

[23] Michiel de Hoon, Seiya Imoto, Satoru Miyano. The C Clustering Library for cDNA microarray data. The University of Tokyo, institute of midecal science, human genome center, 5 July 2008, p 3.

[24] Wold, S., Esbensen, K. and Geladi, P. Chem. Intell. Lab. Syst., 1987, 2, 37.

[25] Brereton, R.G. Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies, Elsevier Science, Netherlands, 1992.
[26] Vapnik, V.The Nature of Statistical Learning Theory. Second edition. New York: Springer, 1999.
[27] Drucker, H., Burges, CJ., Kaufman, L., Smola, A. and Vapnik, V. Support Vector Regression Machines. Adv Neural Inf Process Syst, 1997, 9:155-161.
[28] Belousov AI, Verzakov SA, von Frese J. Support Vector Machines: A Versatile and Powerful Approach to Data Analysis, Poster at the Gordon Conf. On Statistics and Chemical Engineering, Williamstown, MA, 2001.
[Online] Available: http://www. Amstat.org./sections/spes/GRC2001.htm (30 July 2004)
[29] Belousov AI, Verzakov SA, von Frese J. Applicational aspects of support vector machines. J. Chemometrics 2002; 16: 482–489.
[30] Ding, C.Q. and Dubchak, I. Multi-class protein fold recognition using support vector machines and neural network. Bioinformatics, vol 17, n° 4 (2001), 349-358.
[31] Broun, M.P.S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T.S., Manuel Ares, Jr. and Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS, vol. 97, n° 1, 262-267, Jannuary 4, 2000.
[32] Iñón, F. A., Garrigues, J. M., Garrigues, S., Molina, A., and de la Guardia, M. Selection of calibration set samples in determination of olive oil acidity by partial least squares-attenuated total reflectance-Fourier transform infrared spectroscopy. Analytica Chimica Acta, 2003, 489, 59–75.
[33] Maggio, R. M., Kaufman, T. S., Del Carlo, M., Cerretanil, L., Bendini, A., Cichelli, A., et al. Monitoring of fatty acid composition in virgin olive oil by Fourier transformed infrared spectroscopy coupled with partial least squares. Food Chemistry, 2009, 114, 1549–1554.
[34] Pandey, K. K.A. Study of Chemical Structure of Soft and Hardwood and Wood Polymers by FTIR Spectroscopy. J. Appl. Polym. Sci., 1999, 71, 1969–1975.
[35] Sathle, L. and Wold, S. Partial least squares analysis with crossvalidation for the two-class problem: a Monte Carlo study. Journal of Chemometrics, 1987, 1, 185–196.