FT-IR extra virgin olive oil classification based on ethyl ester content

Giacomo Squeo¹*, Silvia Grassi²*, Vito Michele Paradiso¹, Cristina Alamrese², Francesco Caponio¹

¹Department of Soil, Plant and Food Sciences, Food Science and Technology Unit, Università
degli Studi di Bari, Via Amendola, 165/A, 70126 Bari, Italy
²Department of Food, Environmental and Nutritional Sciences (DeFENS), Università degli
Studi di Milano, Via Celoria 2, 20133 Milano, Italy

*CORRESPONDING AUTHORS:
Giacomo Squeo, giacomo.squeo@uniba.it, Department of Soil, Plant and Food Sciences,
Food Science and Technology Unit, Università degli Studi di Bari, Via Amendola, 165/A,
70126 Bari, Italy.
Silvia Grassi, silvia.grassi@unimi.it, Department of Food, Environmental and Nutritional
Sciences (DeFENS), Università degli Studi di Milano, Via Celoria 2, 20133 Milano, Italy.
ABSTRACT

According to the Regulation (EU) 2016/2095, extra virgin olive oils (EVOO) must contain a maximum of 35 mg kg$^{-1}$ of fatty acid ethyl esters (FAEE). The official method for FAEE quantification is time-consuming and requires the use of a large amount of toxic solvents. Thus, the aim of this work was the application of FT-IR spectroscopy for the development of classification models (i.e. Linear Discriminant Analysis, LDA; Soft Independent Modelling of Class Analogy, SIMCA) able to discriminate EVOO from non-EVOO based on FAEE content. To the aim, 113 EVOO and 46 non-EVOO samples were analysed. Since the Principal Component Analysis revealed that the whole FT-IR spectral range (both raw or pre-treated) was not promising in EVOO and non-EVOO distinction, a variable selection strategy was applied (i.e. SELECT algorithm). All the classification models were validated both by cross validation and with three different external test sets. The best and more robust LDA model was obtained with the raw FT-IR selected variables, reaching 96-100% of correct classification in prediction. SIMCA models resulted less reliable. In particular, the low specificity values (40-67%) revealed that there is a high probability of assignment of non-EVOO to the EVOO class. In conclusion, FT-IR spectroscopy coupled with a discriminant classification approach is a useful tool for a rapid and fast discrimination of EVOO and non-EVOO based on FAEE content. Since the variable selection strategy was effective, the development of simplified and cheap instruments can boost the FT-IR spectroscopy application also in small enterprises, giving the opportunity to acquire many important information about olive oils.

KEYWORDS: FT-IR; LDA; SIMCA; Fatty acid ethyl esters; Extra virgin olive oil; Authentication.
1. Introduction

Extra-virgin olive oil (EVOO) is a premium vegetable oil obtained from fresh fruits only by means of physical and mechanical processes (Council Regulation (EC) No 1513/2001). It has a great market value due to its appreciated features. Over the years, several European Regulations have modified the quality and purity characteristics of virgin olive oils (VOO) for their commercial classification and labelling. In 2011, the European Commission introduced a limit to the content of fatty acid alkyl esters (FAAE) in extra virgin olive oils (Commission Regulation (EU) No 61/2011). According to the mentioned Regulation, a virgin olive oil labelled as EVOO must contain a maximum of 75 mg kg\(^{-1}\) for the sum of ethyl esters (FAEE) and fatty acid methyl esters (FAME) or their sum could be between 75 and 150 mg kg\(^{-1}\) in case their ratio (FAEE/FAME) is guaranteed to be \(\leq 1.5\).

FAAE are neutral lipids originating from the esterification of free low-weight alcohols with free fatty acids. Among involved alcohols, the most important are methanol and ethanol, yielding to FAME and FAEE, respectively. Among fatty acids, palmitic and oleic acids are the most common. Methanol and ethanol rise respectively from the progressive degradation of drupe cell walls and from fermentation processes mainly occurring during olive and/or oil storage in improper conditions. Fatty acids, instead, are commonly found in VOO to some extent, depending on the maturation stage of olives but, mainly, on their quality and integrity. FAAE content could be also affected by the extraction process (Alcalá et al., 2017; Caponio et al., 2018; Squeo, Silletti, Summo, Paradiso, Pasqualone, & Caponio, 2017). The Commission Regulation (EU) No 61/2011 has been finally modified by the Commission Delegated Regulation (EU) 2016/2095 focusing only on the FAEE content and setting the maximum value for EVOO at 35 mg kg\(^{-1}\) of oil.

According to the official method (Commission Regulation (EU) No 61/2011), FAAE determination requires their separation from triacylglycerols and other oil constituents by
chromatography on a hydrated silica gel column using Sudan 1 (1-phenylazo-2-naphthol) as indicator for the elution. Then, the FAAE fraction is collected, dried and re-suspended in n-heptane or iso-octane. Alkyl esters are finally separated by capillary gas-chromatography. Quantification is achieved by the addition of a proper internal standard. Overall, around 5 hours are needed to complete the analysis, without considering the preparation steps such as silica conditioning. Besides, a large amount of organic solvents is used for the determination, mainly n-hexane. Indeed, approximately 250 mL are required for one determination. Considering that the analysis should be performed at least in duplicate in order to obtain reliable results, around 1 L of solvent is required for the analysis of only two oil samples.

Hexane is toxic for humans as well as for the environment as extensively reported in hexane safety data sheet, according to the Regulation (EC) No 1272/2008. Thus, the possibility to significantly reduce the use of this solvent, together with the other organic solvents required for the analysis (diethyl ether, isoctane), matches the sustainability goals desired from Institutions all over the world (United Nations, 2016). The possibility of overcoming these issues (time-consuming analysis, health and environment hazards) lies in green approaches like the use of spectroscopic techniques that is one of the most promising.

Spectroscopic techniques are non-destructive, green, fast and easy to use. Among them, mid-infrared (MIR) spectroscopy is one of the most used, having an illustrious history in lipid chemistry, and it has experienced growing interest and applications thanks to the introduction of the Fourier transform instruments (FT-IR) (Dobson, 2001). The MIR range goes from around 2.5 to 25 μm or, as most commonly reported, from 4000 to 400 cm⁻¹. Absorption of a MIR photon typically excites one of the fundamental vibrations, associated with a change of the dipole moment of an oscillating molecule (Sikorska, Khmelinkii, & Sikorski, 2014). Despite the complexity of spectra collected along the food systems, the association of MIR spectroscopy with chemometrics allows the extraction of the significant and valuable
information (Gómez-Caravaca, Maggio, & Cerretani, 2016). Indeed, when spectra are
recorded from real food samples, they contain information about different components of the
sample matrix together with their interactions, and multivariate methods are successfully used
in interpreting the spectra signals for analytical purposes (Bro, 2003; Kjeldahl & Bro, 2010,
Sikorska, Khmelinskii, & Sikorski, 2014). Several chemometric approaches might be used,
falling in two main classes: qualitative and quantitative methods. As regard to alkyl esters, in
a previous study by Valli et al. (2013), Partial Least Square (PLS) regression models were
tentatively developed for the quantification of FAAE based on VOO FT-IR spectra. However,
they were aimed at the quantification of the sum of ethyl and methyl esters as well as their
ratio, parameters that are no longer considered for the EVOO classification.

After the introduction of the Commission Delegated Regulation (EU) 2016/2095, few authors
have taken interest in FAAE determination by green methods. Indeed, near infrared (NIR) and
Vis-NIR spectroscopy has been used to develop regression models for measuring total FAAE
content, as well as FAEE and FAME content separately (Cayuela, 2017; Garrido-Varo,
Sánchez, De la Haba, Torres, & Pérez-Marín, 2017). However, even though chemometric
approaches can overcome the overlapping NIR signals resulting from first and second
overtones and combinations of the fundamental vibrations, more accurate assignments of
absorption bands can be reached by MIR spectroscopy. This is particularly relevant when
assessing differences among molecules having similar bonds that scatter in a complex matrix
such as oil. As far as we know, despite the importance of FAEE, no other attempts have been
carried out by IR spectroscopy to develop a rapid procedure for their analysis. Starting from
these considerations, the aim of this work was the application of IR spectroscopy to the
development of classification models (based on Linear Discriminant Analysis and Soft
Independent Modelling of Class Analogy) able to discriminate between EVOO and non-
EVOO based on FAEE content. Though FAEE is a continuous variable, a classification
approach was chosen instead of quantification since, by a practical point of view, the proposed method should address a discrimination issue regarding the authentication of EVOO, where authentication is intended as the compliance of a food with its label description (Danezis, Tsagkaris, Camin, Brusic, & Georgiou, 2016). A similar approach based on discriminant classification techniques has already been applied in the literature in order to develop fast sorting tests for olive oils, based on the content of α-tocopherol or squalene (Cayuela & Garcia, 2018; Cayuela & Garcia, 2017). Supervised classification techniques use the information about the known class membership of training samples in order to create classification rules able to assign new unknown samples to one of the defined classes, based on their fingerprint measurement (Berrueta, Alonso-Salces, & Héberger, 2007). Thus, these chemometric techniques perfectly fit in authentication issues where the goal is to verify if a sample belongs to a predefined class, such in the case of EVOO and non-EVOO differentiated by the FAEE content.

2. Materials and methods

2.1. Sampling

A set of 159 VOO (113 extra virgin and 46 virgin) from Apulia region (southeast Italy) were collected during 2016/17 and 2017/18 production seasons directly from olive mills located in different provinces (i.e., 100 samples from Bari province; 15 from Brindisi province; 13 from Barletta-Andria-Trani province; 11 from Foggia province; 11 from Lecce province; 9 from Taranto province). All samples were bulk oils, blends of the principal Apulian olive cultivars, all extracted by continuous plants equipped with decanter centrifuge.

2.2. Fatty acid ethyl esters determination
The analysis of FAEE was carried out according to the official method (Commission Regulation (EU) No 61/2011). Briefly, for each determination, 15 g of pre-conditioned silica gel was suspended in n-hexane and introduced in a glass column for liquid chromatography. Samples were prepared by adding to 500 mg of oil 250 μL of internal standard (methyl heptadecanoate 0.02% w/v in iso-octane) and 50 μL of Sudan I solution (1% w/v). In order to remove impurities, 30 mL of n-hexane were percolated through the column before loading the sample. About 230 mL of n-hexane/ethyl ether mixture (99:1) was percolated through the column, with an elution flow of about 15 drops every 10 s, till the Sudan dye reached the bottom of the column. After the elution, solvents were evaporated and the remaining fraction containing the methyl and ethyl esters was diluted with 2 mL of iso-octane. The iso-octane solution (1 μL) was then injected directly on-column in the GC-FID system composed by an Agilent gas chromatograph (7890B, Agilent Technologies, Santa Clara, CA, USA) equipped with a FID detector (set at 350 °C) and a DB-5HT (15 m × 0.32 mm, 0.1 μm film thickness) nonpolar capillary column (Agilent Technologies, Santa Clara, CA, USA). The temperature gradient was programmed as follows: 80 °C for 1 min; from 80 to 140 °C at 20 °C min⁻¹; from 140 to 335 °C at 5 °C min⁻¹; 335 °C for 20 min. Helium was the carrier gas at a flow rate of 2 mL min⁻¹.

2.3. FT-IR spectra acquisition

FT-IR spectra were collected by means of an ATR module on a Nicolet iS50 spectrometer (Thermo Fisher Scientific Inc., Waltham, MA, USA) under the following conditions: 4000-600 cm⁻¹ spectral range, 4 cm⁻¹ resolution, 32 scans for both sample and background. The equipment was controlled by OMNIC software (Thermo Fisher Scientific Inc., Waltham, MA, USA). Six (6) spectra per each sample were collected at room temperature (around 25 °C), recording a new background every 3 spectra. Before each measurement, the ATR crystal was
cleaned with alcohol by a cotton wipe. Spectral acquisition was concomitant with the alkyl
esters determination.

2.4. Data analysis

Dataset descriptive statistics were calculated by means of Minitab 17 software (Minitab Inc.,
State College, PA, USA). The replicates of FT-IR spectra collected from each sample were
averaged before data elaboration. Spectral range was reduced in order to eliminate the noisiest
and the least informative regions (600-650 cm\(^{-1}\); 1890-2600 cm\(^{-1}\)). The resulting dataset (159
samples x 5478 spectral variables) was pre-treated with smoothing (moving average, 15
wavenumbers window size; SMOOTH), eventually followed by standard normal variate
(SNV) or SNV coupled with first derivative (second polynomial order, 15 wavenumbers
window size; d1), and explored by Principal Component Analysis (PCA). Then, a thirty-
variable selection was performed by SELECT algorithm (Forina, Lanteri, Casale, & Cerrato
Oliveros, 2007; Kowalski, & Bender, 1976) that searches for the variable with the largest
Fisher classification weight (FW), defined for the variable \(v\) and the two categories 1 and 2 as:

\[
FW = \frac{(x_{v1} - x_{v2})^2}{\sum_{i=1}^{I_1} \left(\frac{x_{i,v1} - \bar{x}_{v1}}{I_1}\right)^2 + \sum_{i=1}^{I_2} \left(\frac{x_{i,v2} - \bar{x}_{v2}}{I_2}\right)^2}
\]

where \(\bar{x}_{vc}\) is the mean of variable \(v\) in category \(c\); \(I_c\) is the number of objects in category \(c\)
(Fisher, 1936). The variable is selected and decorrelated from the remaining predictors and
then the process continues to iterate until all the given variables are selected.

The discrimination of EVOO (class 1; FAEE \(\leq\) 35 mg kg\(^{-1}\)) from non-EVOO (class 2; FAEE
\(>\) 35 mg kg\(^{-1}\)) samples was performed by applying two classification techniques on the
reduced spectral datasets, i.e. Linear Discriminant Analysis (LDA), a supervised pattern
recognition technique based on discriminant canonicals, and Soft Independent Modelling of
Class Analogy (SIMCA), a supervised class-modelling technique. Classification models were
validated both in cross-validation by 5 cancellation groups (CV) and in prediction using three
different external test sets, randomly created, each containing about 30% samples of the
whole dataset.
Since LDA maximises the separation among classes by the construction of an optimal \textit{a-posteriori} classification rule that assigns every object to one unique class, the LDA models’
performance was evaluated by correct classification percentage, i.e. by the percentage of
samples correctly assigned by the \textit{a-posteriori} rule to the \textit{a-priori} defined class. In the case of
SIMCA, sensitivity (true positive rate) and specificity (false positive rate) were also
calculated to evaluate the effectiveness of the classification models. Indeed, in a class-
modelling technique, such as SIMCA, a sample can be assigned: (1) correctly and exclusively
to the actual class; (2) exclusively to one class which differs from the actual class; (3) to no
classes; or (4) to more than one class. Thus, while evaluating a SIMCA model, it comes the
need to consider the latter three cases, sources of classification errors, through the calculation
of the above-mentioned figures of merit (sensitivity and specificity).
Data elaboration was performed by The Unscrambler X software (v. 10.2, Camo Software
AS, Oslo, Norway) and the V-Parvus package (Forina, Lanteri, Armanino, Casolino, Casale,
& Oliveri, 2008).

3. Results and discussion

3.1. Fatty acid ethyl ester content of olive oil samples
Table 1 reports the descriptive statistics of the olive oil samples, divided by classes: EVOO
(class 1), non-EVOO (class 2). The number of objects was quite different between the two
classes since no previous information about the amount of FAEE in the collected samples
were provided by producers. In any case, this situation is representative of a real scenario,
because during the authentication of EVOO a low number of non-EVOO should be expected.
Class 2 (non-EVOO), although made up of a lower number of samples, was characterised by a higher variability, covering a wide range of FAEE content, from 37.7 to 298 mg kg\(^{-1}\). On the contrary, the variability of class 1 (EVOO samples) was quite low, as most samples had a FAEE content in the range 1-10 mg kg\(^{-1}\).

3.2. FT-IR spectra and data exploration

Fig. 1 depicts the average FT-IR spectra for the two olive oil classes (i.e. EVOO and non-EVOO). Overall, spectra were similar to those reported for edible oils and olive oils (Dobson, 2001; Sikorska et al., 2014). Although the spectra were almost overlapping, EVOO (class 1) presented slightly higher band intensities roughly along the whole spectral range. More evident differences in intensity were observed for particular bands, namely in the regions from around 3030 to 2800 cm\(^{-1}\), corresponding to different stretching vibrations, around 1238 and 1160 cm\(^{-1}\), corresponding to -C-O, -CH\(_2\)- stretching and bending, and around 1118, 1097 and 1030 cm\(^{-1}\), corresponding to -C-O stretching (Sikorska et al., 2014). No considerable differences were observable for those bands reported as typical of the ester linkage (2730, 2677 cm\(^{-1}\)), with the exception of the peak around 1746 cm\(^{-1}\) (highlighted in Fig. 1), which might be potentially influenced by the amount of alkyl esters. The little differences observed between the spectra of the two oil classes confirm the necessity of chemometric tools to extract useful information.

As a first step, an exploratory analysis was performed by PCA on both the raw and the different pre-treated data. Fig. 2a reports the score plot resulting from the raw data analysis. Few samples belonging to EVOO class resulted isolated from the sample cloud. However, no sample subsets were identified; EVOO and non-EVOO classes appeared confused. Sample distribution in the space of the first two principal components (PC) was very similar for any of the pre-treated dataset explored (data not shown).
The PC1 loadings plot (Fig. 2b) showed that the most informative variables affecting the sample distribution, i.e. the ones farer from zero, corresponded to the most evident differences in the absorbance observed in the raw spectra. PC2 loadings trend (Fig. 2b) remarked a strong negative influence in sample separation of the band from 3770 to 3070 cm\(^{-1}\), likely due to -OH group of oxidation products (Sikorska et al., 2014), and a strong positive influence of variables around 3030 - 2800 cm\(^{-1}\) and 1743 cm\(^{-1}\). Those regions were highlighted also in the VOO spectra recorded by Valli et al. (2013) and pre-treated by multiplicative scatter correction and mean centering.

3.3. Classification models

As the whole spectral range did not look promising in EVOO and non-EVOO discrimination, a variable selection strategy was investigated. The SELECT algorithm allowed to identify the 30 most informative variables, among the original 5478, to be used in the following classification procedures. Variable selection and classification model development were performed on both raw and pre-treated data (SMOOTH, SMOOTH-SNV, and SMOOTH-SNV-d1).

The 30 variables selected from the raw dataset are reported in Fig. 3a. Half of the selected variables belongs to the high informative fingerprint region, i.e. between 1500 and 650 cm\(^{-1}\). The signals from 750 to 679 cm\(^{-1}\) can be linked to the CH\(_2\) rocking vibrations occurring at 723 cm\(^{-1}\) (Guillen & Cabo, 1997; Lerma-García, Simó-Alfonso, Ramis-Ramos, & Herrero-Martínez, 2011). The variables selected between 1000 and 900 cm\(^{-1}\) could be linked to C-H out of plane deformation occurring at 968 cm\(^{-1}\) due to isolated trans double bonds presence (Guillen & Cabo, 1997). The 1119 cm\(^{-1}\) variable could originate from signal due to the C-O stretching vibration (Guillen & Cabo, 1997) as well as the variables selected in the range 1300-1270 cm\(^{-1}\), where also the C-H bending vibration is expected to occur. The fourth range
of selected variables includes wavenumbers 1474-1472 cm\(^{-1}\) that are related to CH\(_2\) and CH\(_3\) scissoring vibrations (Guillen & Cabo, 1997). The other selected variables lay outside the IR fingerprint region. Between 1626 cm\(^{-1}\) and 1580 cm\(^{-1}\) the selected signals can be related to C-C vibration in aromatic structures. The absorption of carbonyl aldehydes and ketones are in the selected range 1730-1645 cm\(^{-1}\) (Van de Voort, Sedman, & Russin, 2001). The C=O stretching vibration characterises the variables selected between 1775 and 1713 cm\(^{-1}\); indeed, these are related to an intense peak that is normally observed at 1746 cm\(^{-1}\) and known to be related to carbonyl group vibration in ester linkage between fatty acid and glycerol (Sikorska et al., 2014). Other four variables were selected in the range 4000-3600 cm\(^{-1}\); normally, the vibration of C=O stretching belonging to triacylglycerol ester bonds occurs also in this range (Sikorska et al., 2014).

In Fig. 3b the 30 variables selected for the SMOOTH-SNV-d1 dataset are reported. Apart from the variables selected in the regions already discussed above, relevant variables appear between 3500 and 2600 cm\(^{-1}\). In the region 3000-2800 cm\(^{-1}\), C-H stretching vibrations are present, such as the stretching of CH\(_3\) in the range 2962-2873 cm\(^{-1}\) (Vlachos, Skopelitis, Psaroudaki, Konstantinidou, Chatzilazarou, & Tegou, 2006) and two intense bands at 2924 and 2853 cm\(^{-1}\) linked to CH\(_2\) stretching of the fatty acid chains in triacylglycerols (Guillén & Cabo, 1997). Also C=O Fermi resonance of ester groups occurs at 2677 and 2730 cm\(^{-1}\) (Guillén & Cabo, 1997; Lerma-García et al., 2011). The last variables selected from the transformed dataset belong to the 4000-3500 cm\(^{-1}\) range, where the relevant bands are related to C=O overtone and stretching vibrations (Vlachos et al., 2006).

Concerning LDA, good classification performances were obtained using both raw and SMOOTH-SNV-d1 pre-treated data, while less promising results were achieved with SMOOTH or SMOOTH-SNV data (Table 2). In details, for both raw and SMOOTH-SNV-d1 datasets the correct classification rate in calibration was 100% no matter the external test set.
used, whereas the SMOOTH-SNV model resulted affected by the data excluded from the calibration sets, ranging the correct classification from 98.9 to 99.1% depending on the considered test set. Smoothed data also provided very good classification rate in calibration even though less accurate than those of raw and SMOOTH-SNV-d1 datasets. With internal validation (cross-validation), the correct classification rates remained high for the SMOOTH-SNV-d1 dataset (97.3-100%), while decreased for raw data (96.4-97.3%) as well as for SMOOTH-SNV data (90.1-94.6%). Data smoothing gave the best results just after those of SMOOTH-SNV-d1 dataset (95.8-99.1%). The prediction step confirmed the reliability of the models developed with raw and SMOOTH-SNV-d1 datasets; indeed the correct classification rates resulted between 95.8 and 100%, whereas the SMOOTH and SMOOTH-SNV data reached as high as 97.9 and 93.7% of correct classification, respectively.

Considering that the dearth or soft mathematical pre-treatment should be preferred (Grassi, Casiraghi, & Alamprese, 2018), LDA results based on the selected variables of raw spectra looked the more promising among the developed models. In addition, the cross-validation results of raw data, being more stable when changing the external test set, revealed a more robust model. These results look much more promising than those obtained by Cayuela (2017) using NIR (100-2300 nm) spectra collected on a huge set of virgin (extra, virgin and lampante) olive oils. The author’s main purpose was to perform PLS regression models to quantify FAAA, FAEE, and FAME. Furthermore, an oil classification based on each parameter was performed, but no information about the classification algorithm was reported. Moreover, the performance evaluation was based on the correspondence between the regression predicted values and the defined classes of the samples from reference measurements. In any case, correct classification rates of the models based on FAAA (two classes with 75 mg kg$^{-1}$ threshold), FAEE (two classes with 35 mg kg$^{-1}$ threshold), and FAME (two classes with 35 mg kg$^{-1}$ threshold) were 78.65-94.8%, 70.0-88.7%, and 95.2%,
respectively. Thus, it is difficult to deeply compare results by Cayuela (2017) with the LDA
text.

Class-modelling techniques are often more appropriate in food authentication than
discriminant methods, because they aim at answering to the essential question in food
authenticity problems: if a product is sold with a specific label, it is important to verify if it is
really consistent with the claimed characteristics (Oliveri, Di Egidio, Woodcock, & Downey,
2011). Moreover, this approach does not force a sample to belong to one specific class,
resulting in a more severe but realistic scenario in real-life applications when a large number
of non-target classes is possibly present. Thus, SIMCA models for olive oil discrimination
based on FAEE classes were developed and the related results are provided in Table 3.

Differently from what observed with LDA, the best classification performances were obtained
on pre-treated data instead of raw ones. In details, models obtained with SMOOTH-SNV-d1
treated spectra gave higher percentages of classification ability than those developed with
SMOOTH or SMOOTH-SNV data, ranging from around 91 to slightly less than 96%.

Looking at the prediction performances of the models, the correct classification ability was,
on average, higher when smoothing was applied to spectral data although the absolute best
performance was observed applying first derivative in combination with smoothing and SNV
(up to 81.25%). Sensitivity of all the models was promising, showing a percentage of true
positive assignments ranging from 75 to about 92%, being again, on average, SMOOTH
model the best. Indeed, when comparing the Cooman’s plot (Fig. 4) of the best performing
SIMCA models, it can be noticed that a similar and low number of samples is distributed
outside the 95% model thresholds for SMOOTH and SMOOTH-SNV-d1 datasets (Fig. 4a and
4c). It means that those samples were not accepted by any of the class models (false
negatives). However, the most important issue concerned the low specificity affecting models.
The average specificity was around 50% (Table 3) considering the different pre-treatments,
meaning that around the half of the test samples belonging to one of the class was not rejected from the assignment to the other class. Referring to the Cooman’s plots (Fig. 4), these samples are located in the small square on the bottom resulting from the intersection of 95% threshold lines of class 1 and class 2 models. Those samples were accepted by both the model of EVOO and non-EVOO samples. In particular, the lower specificity values of class 2 (non-EVOO), which were on average between 40 and 50% for the three considered external sets (data not shown), implicates a high probability of assignment of non-EVOO to the EVOO class. Thus, it is clear that the application of SIMCA class-modelling could be hazardous for EVOO labelling based on FAEE content, and the discriminant strategy (LDA) should be preferred. Actually, when only two well-defined classes are considered, i.e. FAEE ≤ 35 mg kg⁻¹ (EVOO class) and FAEE > 35 mg kg⁻¹ (non-EVOO class), a discriminant-based strategy as LDA is a reliable option. Indeed, a discriminant classification method divides the sample domain into a number of sub-areas corresponding to the spaces of the considered classes. A sample is always assigned to one of the considered classes, even if it is located very far from all the classes. However, when only two well-defined classes are considered, such in this study, the classification-discriminant method can give reliable results thanks to the low complexity of the dataset (Di Egidio, Oliveri, Woodcock, & Downey, 2011).

Future perspectives can consider the possibility to enlarge the VOO dataset by collecting samples from different Italian regions and covering different cultivars and storage strategies. Moreover, the non-EVOO class could be extended accordingly. From an instrumental point of view, a multichannel IR-instrument based on 30 or fewer wavenumbers could be developed from commercial prototypes and its reliability compared with the presented models.

4. Conclusions
The potential of FT-IR spectroscopy coupled with chemometrics as a tool for a rapid and fast discrimination of extra virgin olive oils and virgin olive oils based on fatty acid ethyl ester content has been assessed. The legal limit for FAEE content enables to highlight two well-defined classes of products, thus a discriminant classification approach can be considered as the most suitable in this context, to be preferred over the SIMCA class-modelling algorithm. LDA models based on selected features of the raw spectra datasets gave good and robust results, with percentages of correct classification in prediction higher than 95%. Being the discriminant classification results so promising, the application of FT-IR spectroscopy should be encouraged in the olive oil sector. Moreover, since the variable selection strategy was effective, the development of simplified and cheap instruments can boost their use also in small enterprises. Such instruments can be calibrated in order to give also results about oil composition, thus providing producers and control bodies with many important information about olive oil samples.

Acknowledgements

This work was supported by AGER 2 Project, grant no. 2016-0105. The authors declare that there are no conflicts of interest regarding the publication of this paper. All the authors have contributed to the same extent to the realization of this paper.

References


Garrido-Varo, A., Sánchez, M. T., De la Haba, M. J., Torres, I., & Pérez-Marín, D. (2017). Fast, low-cost and non-destructive physico-chemical analysis of virgin olive oils using...


Figure captions

Fig. 1. Average FT-IR spectra for extra virgin olive oils (Class 1-EVOO; black straight line) and non-extra virgin olive oils (Class 2-nonEVOO; grey dashed-line). In detail the peak around 1746 cm\(^{-1}\).

Fig. 2. Principal Component Analysis results of raw FT-IR spectra collected on extra virgin (EVOO) and non-extra virgin (nonEVOO) olive oil samples: a) PC1 vs PC2 score plot and b) loading plot for PC1 and PC2.

Fig. 3. FT-IR variables (cm\(^{-1}\)) selected by SELECT algorithm before to perform classification of extra virgin and non-extra virgin olive oil samples: a) 30 variables selected from the raw dataset, b) 30 variables selected from the data pre-treated with smoothing, standard normal variate and first derivative.

Fig. 4. Soft Independent Modelling of Class Analogy results for extra virgin (EVOO) and non-extra virgin (nonEVOO) olive oil samples. Cooman's plots obtained with the 30 FT-IR selected features and the external test set no. 3: a) smoothed data (SMOOTH; b) data pre-treated with smoothing and standard normal variate (SMOOTH-SNV); c) data pre-treated with smoothing, standard normal variate and first derivative (SMOOTH-SNV-d1).
Figure 1.
Figure 2.
Figure 3.
Figure 4.
Highlights

- FT-IR spectroscopy discrimination of olive oils based on fatty acid ethyl esters
- FT-IR variable selection improved olive oil discrimination
- Linear Discriminant Analysis models reached up to 100% of correct prediction
- Class modelling had high sensitivity but scarce specificity (<50%) in prediction
Table 1. Descriptive statistics about the fatty acid ethyl ester content (mg kg\(^{-1}\)) of the olive oil samples under study divided by classes: extra virgin olive oils (class 1), non-extra virgin olive oils (class 2).

<table>
<thead>
<tr>
<th>Class</th>
<th>N</th>
<th>Mean</th>
<th>SD*</th>
<th>CV*</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>Range</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>113</td>
<td>6.68</td>
<td>4.95</td>
<td>74.21</td>
<td>1.81</td>
<td>5.34</td>
<td>31.36</td>
<td>29.54</td>
<td>3.93</td>
</tr>
<tr>
<td>2</td>
<td>46</td>
<td>106.90</td>
<td>77.20</td>
<td>72.25</td>
<td>37.70</td>
<td>72.50</td>
<td>298.80</td>
<td>261.10</td>
<td>114.60</td>
</tr>
</tbody>
</table>

N, number of samples; SD, standard deviation; CV, percent coefficient of variation; IQR, interquartile range.

*Standard deviation and coefficient of variation are referred to the class distribution.
Table 2. Results of Linear Discriminant Analysis for olive oil discrimination based on fatty acid ethyl ester content: average correct classification percentages of models based on the 30 most informative variables of FT-IR spectra.

<table>
<thead>
<tr>
<th>Pre-treatment</th>
<th>External test set</th>
<th>Calibration</th>
<th>Cross-validation</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>1</td>
<td>100</td>
<td>97.30</td>
<td>95.83</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>100</td>
<td>96.40</td>
<td>95.83</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>100</td>
<td>96.61</td>
<td>100</td>
</tr>
<tr>
<td>SMOOTH</td>
<td>1</td>
<td>100</td>
<td>97.30</td>
<td>97.92</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>99.28</td>
<td>99.10</td>
<td>91.67</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>99.49</td>
<td>95.76</td>
<td>97.56</td>
</tr>
<tr>
<td>SMOOTH-SNV</td>
<td>1</td>
<td>98.92</td>
<td>90.10</td>
<td>93.75</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>99.10</td>
<td>94.60</td>
<td>87.50</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>99.15</td>
<td>91.53</td>
<td>90.24</td>
</tr>
<tr>
<td>SMOOTH-SNV-d1</td>
<td>1</td>
<td>100</td>
<td>97.30</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>95.83</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>100</td>
<td>99.15</td>
<td>100</td>
</tr>
</tbody>
</table>

SMOOTH, moving average smoothing; SNV, standard normal variate; d1, first derivative.
Table 3. Results of Soft Independent Modeling of Class Analogy for olive oil discrimination based on fatty acid ethyl ester content: average classification, prediction ability, and sensitivity and specificity values in prediction of models based on the 30 most informative variables of FT-IR spectra after different mathematical pre-treatments.

<table>
<thead>
<tr>
<th>Pre-treatment</th>
<th>External test set</th>
<th>Calibration</th>
<th>Prediction</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Classification ability (%)</td>
<td>Prediction ability (%)</td>
<td>Sensitivity (%)</td>
<td>Specificity (%)</td>
</tr>
<tr>
<td>Raw</td>
<td>1</td>
<td>83.78</td>
<td>77.08</td>
<td>87.50</td>
<td>39.58</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>88.29</td>
<td>75.00</td>
<td>91.67</td>
<td>45.83</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>87.29</td>
<td>73.17</td>
<td>87.81</td>
<td>41.46</td>
</tr>
<tr>
<td>SMOOTH</td>
<td>1</td>
<td>72.07</td>
<td>79.17</td>
<td>91.67</td>
<td>45.83</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>85.59</td>
<td>75.00</td>
<td>89.58</td>
<td>50.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>88.98</td>
<td>80.49</td>
<td>90.24</td>
<td>53.66</td>
</tr>
<tr>
<td>SMOOTH-SNV</td>
<td>1</td>
<td>83.78</td>
<td>77.08</td>
<td>75.00</td>
<td>54.17</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>88.29</td>
<td>75.00</td>
<td>79.17</td>
<td>66.67</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>89.83</td>
<td>70.73</td>
<td>80.49</td>
<td>48.78</td>
</tr>
<tr>
<td>SMOOTH-SNV-d1</td>
<td>1</td>
<td>95.50</td>
<td>75.00</td>
<td>75.00</td>
<td>52.08</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>90.99</td>
<td>81.25</td>
<td>70.83</td>
<td>62.50</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>95.76</td>
<td>73.17</td>
<td>80.49</td>
<td>43.90</td>
</tr>
</tbody>
</table>

SMOOTH, moving average smoothing; SNV, standard normal variate; d1, first derivative.