Corresponding Author: Dr. Giacomo Squeo, Ph.D.

Corresponding Author's Institution: University of Bari

First Author: Giacomo Squeo, Ph.D.

Order of Authors: Giacomo Squeo, Ph.D.; Silvia Grassi; Vito M Paradiso; Cristina Alamprese; Francesco Caponio

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.

Authors:

Giacomo Squeo
Department of Soil, Plant and Food Sciences, Food Science and Technology Unit, University of Bari Aldo Moro, Italy.

Silvia Grassi
Department of Food, Environmental and Nutritional Sciences (DeFENS), Università degli Studi di Milano, Italy.

Vito Michele Paradiso
Department of Soil, Plant and Food Sciences, Food Science and Technology Unit, University of Bari Aldo Moro, Italy.

Cristina Alamprese
Department of Food, Environmental and Nutritional Sciences (DeFENS), Università degli Studi di Milano, Italy.

Francesco Caponio
Department of Soil, Plant and Food Sciences, Food Science and Technology Unit, University of Bari Aldo Moro, Italy.

Dear G. Campbell-Platt,
Editor-in-Chief
Food Control

We are pleased to submit an original research article entitled "**FT-IR extra virgin olive oil classification based on ethyl ester content**" for consideration for publication in Food Control.

We believe that this manuscript is appropriate for publication by Food Control since our findings give an original contribute to the topic of food quality assurance with a particular regard to the case of Extra Virgin Olive Oil (EVOO) one of the most appreciate food of the Mediterranean diet. Due to its great value, EVOO is often subjected to commercial fraud aimed at deal non-EVOO for EVOO. To assure that only best quality oils became EVOO several analytical parameters must be checked, some of them by tedious and not sustainable analysis as it is the case of fatty acid ethyl esters. Our aim has been the development of a rapid classification method based on FT-IR spectroscopy. It is our opinion that the development of such models is relevant not only for contributing to the production of scientific knowledge about that topic, but also matches the worldwide needs of more sustainable approaches. Although some papers have been published on related topic, as far as we know, they were related to not up-to-date parameters and did not use the same chemometric approach.

This manuscript has not been published and is not under consideration for publication elsewhere.

The authors certify that there is no conflict of interest with any financial/research/academic organization, with regards to the content/research work discussed in the manuscript.

Thank you for your consideration.

Corresponding Authors:

Giacomo Squeo, giacomo.squeo@uniba.it, Department of Soil, Plant and Food Sciences, Food Science and Technology Unit, University of Bari Aldo Moro, 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.

Manuscript Number: FOODCONT-D-19-00134

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

To the Editorial Board of Food Control,

I am submitting the revised version of Ms. Ref. No.: FOODCONT-D-19-00134 entitled "FT-IR extra virgin olive oil classification based on ethyl ester content" by G. Squeo, S. Grassi, V.M. Paradiso, C. Alamprese and F. Caponio.

Following the Reviewers' comments the manuscript has been improved and all the changes introduced have been highlighted in red.

Sincerely,

Giacomo Squeo

Reviewers' comments:

Reviewer #1: In my opinion the subject of the manuscript has the scientific interest and the originality necessary for its publication.
The authors wish to thank the Reviewer for her/his positive evaluation and for the usefulness suggestions.

1.   However, unfortunately the authors have considered a wrong legal limit for FAEE. The FAEE limit for seasons after 2015 is 30 mg kg-1, according to the current regulation EU No 1348/2013 of 16 December. This must be considered and corrected through the entire document. Consequently, it seems logical to say the tests should be remade.
To the best of our knowledge and as reported in lines 60-63 of the manuscript, the current regulation about FAEE limit is the Commission Delegated Regulation (EU) 2016/2095, which were released after the Reg. EU No 1348/2013 and fixes the ethyl esters limit at 35 mg kg^{-1} for the EVOO class. In the light of this, the authors think that the experimental approach of the manuscript is valid and no corrections are needed.

2.   One important idea I think is, if the question is the commercial class to which the product matches according to its content in alkyl esters... this, in my opinion, does not match the concept 'authentication'. This concept refers to guarantee the nature of the product (olive), its geographical origin, varietal, etc. Therefore, in this context the word authentication, does not fairly match and disorients.

The authors refer to the word authentication as the compliance with the food label description, as reported by a number of reviews covering food authentication issues. For instance, Danezis et al. (2016) stated that "Food authentication is the process that verifies that a food is in compliance with its label description. This may include, among others, the origin (species, geographical or genetic), production method (conventional, organic, traditional procedures, free range), or processing technologies (irradiation, freezing, microwave heating). The declaration of specific quality attributes in high-value products is of particular interest since these products are often target of fraudulent labelling". In order to clarify this point, the reported definition has been added in lines 118-119.

(Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusic, V., & Georgiou, C. A. (2016). Food authentication: Techniques, trends & emerging approaches. TrAC Trends in Analytical Chemistry, 85, 123-132.)

3.   The end of the Introduction lacks the presentation of the objectives.

The objectives of the work are reported at lines 111-115 and have been specified better.

4.   A detailed description of the analysis method of FAEE is required.

Thank you for the suggestion. The description of the analytical method for FAEE determination has been added (lines 141-156).

5.   The Conclusions should wrap more on the results, before to declare opinions. SIMCA, LDA...?

As suggested by the referee, some more details about the obtained results have been added in Conclusions (lines 364-368).

6.   Comments in the attached manuscript pdf notes:

a. The FAEE limit for seasons after 2015 is 30 mg kg^{-1}. This must be considered and corrected trough the entire document. European Commission Implementing Regulation. (2013). EU No 1348/2013 of 16 December 2013 amending Regulation No 2568/91/EEC on the characteristics of olive oil and olive-residue oil and on the relevant methods of analysis. Official Journal of the European Community, L338, 1e37. Attachment 1 (p 36)

Please, look at the answer to the point 1.

b. The details shown in L77-80 are little significant (everyone who has a minimum lab. experience using hexane, knows the other mentioned solvents are worse to health).

As suggested by the Reviewer, the details about the risk statements have been removed and the sentence arranged accordingly (lines 75-76).

c. Approachs for olive oil classification as referred in L118-119 have been reported. The studies following deals similar approach in olive oils regarding other quality parameters. It could be logical reference them:

Cayuela, J.A., García, J.F. 2017 - Sorting olive oil based on alpha-tocopherol and total tocopherol content using Near-Infra-Red Spectroscopy (NIRS) analysis. J. Food Engineering. 202, 79-88.

Cayuela, J.A., García, J.F. 2018. Nondestructive measurement of squalene in olive oil by near infrared spectroscopy. LWT - Food Science and Technology 88, 103-108.

As suggested by the Reviewer, the references have been added (lines 119-122).

d. If the question is the commercial class to which the product matches according to its content in alkyl esters... This, in my opinion, does not match the concept 'authentication'. This concept refers to guarantee the nature of the product (olive), its geographical origin, varietal, etc. Therefore, in this context the word authentication, disorients.

Please, look at the answer to the point 2.

e. The meaning's nuance of 'authentication' does not fairly match with the manuscript's issue.

Please, look at the answer to the point 2.

f. The end of the Introduction lacks the presentation of the objectives.

Please, look at the answer to the point 3.

g. Collected...where? Some more details are required. Were the samples obtained in industries or was their origin commercial? Please, clarify their respective origin and detail.

As requested, further details have been added (lines 133-137).

h. A detailed description of the analysis method is required. (Fatty acid ethyl esters determination)

Please, look at the answer to the point 4.

i. L158-159: Please reference Fisher classification weight.

As suggested by the Reviewer, Fisher classification weight has been defined (lines 179-182) and a reference has been added (line 182).

(Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179-188.)

l. The Conclusions should wrap more on the results, before to declare opinions. SIMCA, LDA...?

Please, look at the answer to the point 5.

Reviewer #3: *The paper presents interesting and innovative result on classification of extra virgin*

olive oil according to the ethyl ester content using FTIR and chemometrics. Paper is generally well written, however several issues should be clarify before publication.

The Authors wish to thank the Reviewer for her/his positive evaluation and for the useful suggestions.

1. Lines 151-159. Why only these, limited pre-processing procedures were tested? Did authors test other methods for example, MSC, second derivative, and others and their combination.

The authors have tested different pre-processing strategies and they have decided to show the most relevant ones in term of robustness of the results. Concerning the methods mentioned by the Reviewer, it should be considered that standard normal variate (SNV) and multiplicative scatter correction (MSC) often give very similar results and are widely regarded as exchangeable. Only the first derivative was tested as we do believe that softer pre-processing strategies should be preferred when they give good results.

*2. Line 157. More detailed description should be provide for SELECT algorithm*

Thank you for the suggestion. Additional information have been added in the text (lines 178-183).

3. Line 209-222. It would be interesting to see results and discussion of PCA analysis for variables selected by SELECT algorithm.

As suggested by the Reviewer, we prepared two Figures for the results of PCA made with the selected variables of the raw and the smooth-SNV-d1 datasets. However, we think that these results do not improve the manuscript, being the score plots very similar to the one obtained with all the variables and reported in Fig. 2. On the contrary, we thank the Reviewer for the suggestion at point 4: the comments about the selected variables do improve the result discussion.

In any case, the Figures about the requested PCA are reported here below. If the Reviewer still think that they are useful, we will be glad to add them in the manuscript.

Principal Component Analysis results of variables selected from raw FT-IR spectra of 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.
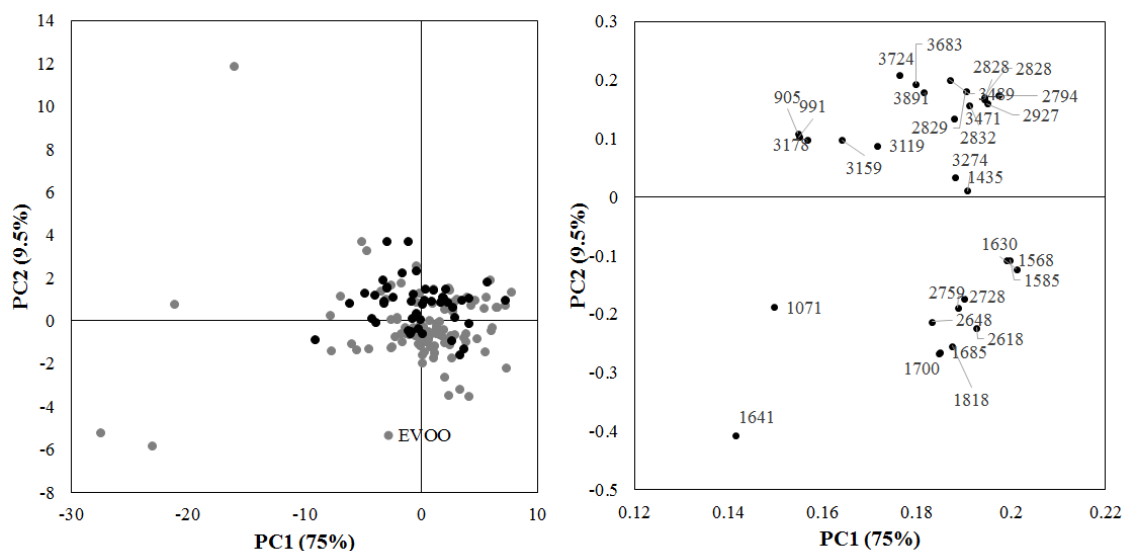
Principal Component Analysis results of variables selected from FT-IR spectra of extra virgin (EVOO) and non-extra virgin (nonEVOO) olive oil samples, after pre-treatment with smoothing, standard normal variate and first derivative: a) PC1 vs PC2 score plot and b) loading plot for PC1 and PC2.
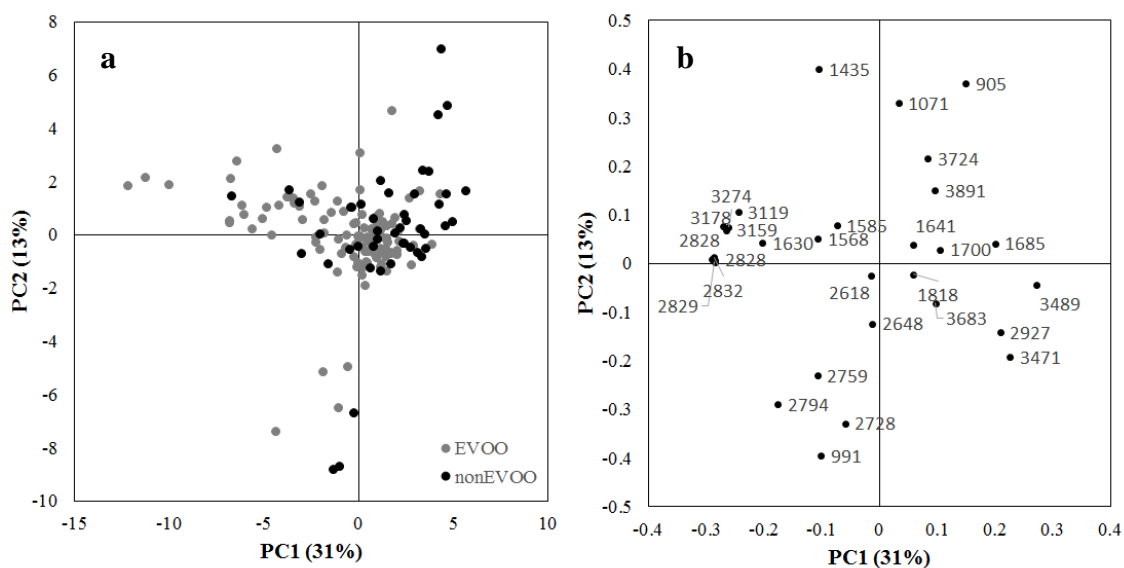
4. Line 225-227. There is lack of important information, which 30 variables were selected for classification analysis. These variables should be described and interpreted in the terms of chemical composition of oils, especially in regard to the signals from ethyl ester.

Thank you for the valuable suggestion. Details about the selected variables have been added in the text (lines 254-283). Moreover, a new Figure (now Fig. 3) has been added to better show the selected variables for the raw and the smooth-SNV-d1 datasets, i.e. the ones giving better results in both LDA and SIMCA approaches.

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

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

2

3   Giacomo Squeo[1*], Silvia Grassi[2*], Vito Michele Paradiso[1], Cristina Alamprese[2], Francesco

4   Caponio[1]

5

6 [1]Department of Soil, Plant and Food Sciences, Food Science and Technology Unit, Università

7   degli Studi di Bari, Via Amendola, 165/A, 70126 Bari, Italy

8   [2]Department of Food, Environmental and Nutritional Sciences (DeFENS), Università degli

9   Studi di Milano, Via Celoria 2, 20133 Milano, Italy

10

11 ***CORRESPONDING AUTHORS:**

12 Giacomo Squeo, giacomo.squeo@uniba.it, Department of Soil, Plant and Food Sciences,

13 Food Science and Technology Unit, Università degli Studi di Bari, Via Amendola, 165/A,

14 70126 Bari, Italy.

15 Silvia Grassi, silvia.grassi@unimi.it, Department of Food, Environmental and Nutritional

16 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 ≤ 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

3

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, Khmelinskii, & 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

5

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

140 The analysis of FAEE was carried out according to the official method (Commission

141 Regulation (EU) No 61/2011). Briefly, for each determination, 15 g of pre-conditioned silica

142 gel was suspended in n-hexane and introduced in a glass column for liquid chromatography.

143 Samples were prepared by adding to 500 mg of oil 250 μL of internal standard (methyl

144 heptadecanoate 0.02% w/v in iso-octane) and 50 μL of Sudan I solution (1% w/v). In order to

145 remove impurities, 30 mL of n-hexane were percolated through the column before loading the

146 sample. About 230 mL of n-hexane/ethyl ether mixture (99:1) was percolated through the

147 column, with an elution flow of about 15 drops every 10 s, till the Sudan dye reached the

148 bottom of the column. After the elution, solvents were evaporated and the remaining fraction

149 containing the methyl and ethyl esters was diluted with 2 mL of iso-octane. The iso-octane

150 solution (1 μL) was then injected directly on-column in the GC-FID system composed by an

151 Agilent gas chromatograph (7890B, Agilent Technologies, Santa Clara, CA, USA) equipped

152 with a FID detector (set at 350 °C) and a DB-5HT (15 m × 0.32 mm, 0.1 μm film thickness)

153 nonpolar capillary column (Agilent Technologies, Santa Clara, CA, USA). The temperature

154 gradient was programmed as follows: 80 °C for 1 min; from 80 to 140 °C at 20 °C min^{-1}; from

155 140 to 335 °C at 5 °C min^{-1}; 335 °C for 20 min. Helium was the carrier gas at a flow rate of 2

156 mL min^{-1}.

157

158 *2.3. FT-IR spectra acquisition*

159 FT-IR spectra were collected by means of an ATR module on a Nicolet iS50 spectrometer

160 (Thermo Fisher Scientific Inc., Waltham, MA, USA) under the following conditions: 4000-

161 600 cm^{-1} spectral range, 4 cm^{-1} resolution, 32 scans for both sample and background. The

162 equipment was controlled by OMNIC software (Thermo Fisher Scientific Inc., Waltham, MA,

163 USA). Six (6) spectra per each sample were collected at room temperature (around 25 °C),

164 recording a new background every 3 spectra. Before each measurement, the ATR crystal was

165 cleaned with alcohol by a cotton wipe. Spectral acquisition was concomitant with the alkyl

166 esters determination.

167

168 *2.4. Data analysis*

169 Dataset descriptive statistics were calculated by means of Minitab 17 software (Minitab Inc.,

170 State College, PA, USA). The replicates of FT-IR spectra collected from each sample were

171 averaged before data elaboration. Spectral range was reduced in order to eliminate the noisiest

172 and the least informative regions (600-650 cm^{-1}; 1890-2600 cm^{-1}). The resulting dataset (159

173 samples x 5478 spectral variables) was pre-treated with smoothing (moving average, 15

174 wavenumbers window size; SMOOTH), eventually followed by standard normal variate

175 (SNV) or SNV coupled with first derivative (second polynomial order, 15 wavenumbers

176 window size; d1), and explored by Principal Component Analysis (PCA). Then, a thirty-

177 variable selection was performed by SELECT algorithm (Forina, Lanteri, Casale, & Cerrato

178 Oliveros, 2007; Kowalski, & Bender, 1976) that searches for the variable with the largest

179 Fisher classification weight (FW), defined for the variable *v* and the two categories 1 and 2 as:

$$FW = \frac{(\bar{x}_{v1} - \bar{x}_{v2})^2}{\sum_{i=1}^{I_1} \frac{(x_{iv1} - \bar{x}_{v1})^2}{I_1} + \sum_{i=1}^{I_2} \frac{(x_{iv2} - \bar{x}_{v2})^2}{I_2}}$$

180 where $\bar{x}_{vc}$ is the mean of variable *v* in category *c*; I_c is the number of objects in category c

181 (Fisher, 1936). The variable is selected and decorrelated from the remaining predictors and

182 then the process continues to iterate until all the given variables are selected.

183 The discrimination of EVOO (class 1; FAEE ≤ 35 mg kg^{-1}) from non-EVOO (class 2; FAEE

184 > 35 mg kg^{-1}) samples was performed by applying two classification techniques on the

185 reduced spectral datasets, i.e. Linear Discriminant Analysis (LDA), a supervised pattern

186 recognition technique based on discriminant canonicals, and Soft Independent Modelling of

187 Class Analogy (SIMCA), a supervised class-modelling technique. Classification models were

188 validated both in cross-validation by 5 cancellation groups (CV) and in prediction using three

189 different external test sets, randomly created, each containing about 30% samples of the

190 whole dataset.

191 Since LDA maximises the separation among classes by the construction of an optimal *a-*

192 *posteriori* classification rule that assigns every object to one unique class, the LDA models'

193 performance was evaluated by correct classification percentage, i.e. by the percentage of

194 samples correctly assigned by the *a-posteriori* rule to the *a-priori* defined class. In the case of

195 SIMCA, sensitivity (true positive rate) and specificity (false positive rate) were also

196 calculated to evaluate the effectiveness of the classification models. Indeed, in a class-

197 modelling technique, such as SIMCA, a sample can be assigned: (1) correctly and exclusively

198 to the actual class; (2) exclusively to one class which differs from the actual class; (3) to no

199 classes; or (4) to more than one class. Thus, while evaluating a SIMCA model, it comes the

200 need to consider the latter three cases, sources of classification errors, through the calculation

201 of the above-mentioned figures of merit (sensitivity and specificity).

202 Data elaboration was performed by The Unscrambler X software (v. 10.2, Camo Software

203 AS, Oslo, Norway) and the V-Parvus package (Forina, Lanteri, Armanino, Casolino, Casale,

204 & Oliveri, 2008).

205

206 **3. Results and discussion**

207 *3.1. Fatty acid ethyl ester content of olive oil samples*

208 Table 1 reports the descriptive statistics of the olive oil samples, divided by classes: EVOO

209 (class 1), non-EVOO (class 2). The number of objects was quite different between the two

210 classes since no previous information about the amount of FAEE in the collected samples

211 were provided by producers. In any case, this situation is representative of a real scenario,

212 because during the authentication of EVOO a low number of non-EVOO should be expected.

9

213 Class 2 (non-EVOO), although made up of a lower number of samples, was characterised by

214 a higher variability, covering a wide range of FAEE content, from 37.7 to 298 mg kg^{-1}. On the

215 contrary, the variability of class 1 (EVOO samples) was quite low, as most samples had a

216 FAEE content in the range 1-10 mg kg^{-1}.

217

218 *3.2. FT-IR spectra and data exploration*

219 Fig. 1 depicts the average FT-IR spectra for the two olive oil classes (i.e. EVOO and non-

220 EVOO). Overall, spectra were similar to those reported for edible oils and olive oils (Dobson,

221 2001; Sikorska et al., 2014). Although the spectra were almost overlapping, EVOO (class 1)

222 presented slightly higher band intensities roughly along the whole spectral range. More

223 evident differences in intensity were observed for particular bands, namely in the regions

224 from around 3030 to 2800 cm^{-1}, corresponding to different stretching vibrations, around 1238

225 and 1160 cm^{-1}, corresponding to -C-O, -CH$_2$- stretching and bending, and around 1118, 1097

226 and 1030 cm^{-1}, corresponding to -C-O stretching (Sikorska et al., 2014). No considerable

227 differences were observable for those bands reported as typical of the ester linkage (2730,

228 2677 cm^{-1}), with the exception of the peak around 1746 cm^{-1} (highlighted in Fig. 1), which

229 might be potentially influenced by the amount of alkyl esters. The little differences observed

230 between the spectra of the two oil classes confirm the necessity of chemometric tools to

231 extract useful information.

232 As a first step, an exploratory analysis was performed by PCA on both the raw and the

233 different pre-treated data. Fig. 2a reports the score plot resulting from the raw data analysis.

234 Few samples belonging to EVOO class resulted isolated from the sample cloud. However, no

235 sample subsets were identified; EVOO and non-EVOO classes appeared confused. Sample

236 distribution in the space of the first two principal components (PC) was very similar for any

237 of the pre-treated dataset explored (data not shown).

238 The PC1 loadings plot (Fig. 2b) showed that the most informative variables affecting the

239 sample distribution, i.e. the ones farer from zero, corresponded to the most evident differences

240 in the absorbance observed in the raw spectra. PC2 loadings trend (Fig. 2b) remarked a strong

241 negative influence in sample separation of the band from 3770 to 3070 cm^{-1}, likely due to -

242 OH group of oxidation products (Sikorska et al., 2014), and a strong positive influence of

243 variables around 3030 - 2800 cm^{-1} and 1743 cm^{-1}. Those regions were highlighted also in the

244 VOO spectra recorded by Valli et al. (2013) and pre-treated by multiplicative scatter

245 correction and mean centering.

246

247 *3.3. Classification models*

248 As the whole spectral range did not look promising in EVOO and non-EVOO discrimination,

249 a variable selection strategy was investigated. The SELECT algorithm allowed to identify the

250 30 most informative variables, among the original 5478, to be used in the following

251 classification procedures. Variable selection and classification model development were

252 performed on both raw and pre-treated data (SMOOTH, SMOOTH-SNV, and SMOOTH-

253 SNV-d1).

254 The 30 variables selected from the raw dataset are reported in Fig. 3a. Half of the selected

255 variables belongs to the high informative fingerprint region, i.e. between 1500 and 650 cm^{-1}.

256 The signals from 750 to 679 cm^{-1} can be linked to the CH$_2$ rocking vibrations occurring at 723

257 cm^{-1} (Guillen & Cabo, 1997; Lerma-García, Simó-Alfonso, Ramis-Ramos, & Herrero-

258 Martínez, 2011). The variables selected between 1000 and 900 cm^{-1} could be linked to C-H

259 out of plane deformation occurring at 968 cm^{-1} due to isolated *trans* double bonds presence

260 (Guillen & Cabo, 1997). The 1119 cm^{-1} variable could originate from signal due to the C-O

261 stretching vibration (Guillen & Cabo, 1997) as well as the variables selected in the range

262 1300-1270 cm^{-1}, where also the C-H bending vibration is expected to occur. The fourth range

263 of selected variables includes wavenumbers 1474-1472 cm^{-1} that are related to CH$_2$ and CH$_3$

264 scissoring vibrations (Guillen & Cabo, 1997). The other selected variables lay outside the IR

265 fingerprint region. Between 1626 cm^{-1} and 1580 cm^{-1} the selected signals can be related to C-

266 C vibration in aromatic structures. The absorption of carbonyl aldehydes and ketones are in

267 the selected range 1730-1645 cm^{-1} (Van de Voort, Sedman, & Russin, 2001). The C=O

268 stretching vibration characterises the variables selected between 1775 and 1713 cm^{-1}; indeed,

269 these are related to an intense peak that is normally observed at 1746 cm^{-1} and known to be

270 related to carbonyl group vibration in ester linkage between fatty acid and glycerol (Sikorska

271 et al., 2014). Other four variables were selected in the range 4000-3600 cm^{-1}; normally, the

272 vibration of C=O stretching belonging to triacylglycerol ester bonds occurs also in this range

273 (Sikorska et al., 2014).

274 In Fig. 3b the 30 variables selected for the SMOOTH-SNV-d1 dataset are reported. Apart

275 from the variables selected in the regions already discussed above, relevant variables appear

276 between 3500 and 2600 cm^{-1}. In the region 3000-2800 cm^{-1}, C-H stretching vibrations are

277 present, such as the stretching of CH$_3$ in the range 2962-2873 cm^{-1} (Vlachos, Skopelitis,

278 Psaroudaki, Konstantinidou, Chatzilazarou, & Tegou, 2006) and two intense bands at 2924

279 and 2853 cm^{-1} linked to CH$_2$ stretching of the fatty acid chains in triacylglycerols (Guillèn &

280 Cabo, 1997). Also C=O Fermi resonance of ester groups occurs at 2677 and 2730 cm^{-1}

281 (Guillèn & Cabo, 1997; Lerma-García et al., 2011). The last variables selected from the

282 transformed dataset belong to the 4000-3500 cm^{-1} range, where the relevant bands are related

283 to C=O overtone and stretching vibrations (Vlachos et al., 2006).

284 Concerning LDA, good classification performances were obtained using both raw and

285 SMOOTH-SNV-d1 pre-treated data, while less promising results were achieved with

286 SMOOTH or SMOOTH-SNV data (Table 2). In details, for both raw and SMOOTH-SNV-d1

287 datasets the correct classification rate in calibration was 100% no matter the external test set

288   used, whereas the SMOOTH-SNV model resulted affected by the data excluded from the

289   calibration sets, ranging the correct classification from 98.9 to 99.1% depending on the

290   considered test set. Smoothed data also provided very good classification rate in calibration

291   even though less accurate than those of raw and SMOOTH-SNV-d1 datasets. With internal

292   validation (cross-validation), the correct classification rates remained high for the SMOOTH-

293   SNV-d1 dataset (97.3-100%), while decreased for raw data (96.4-97.3%) as well as for

294   SMOOTH-SNV data (90.1-94.6%). Data smoothing gave the best results just after those of

295   SMOOTH-SNV-d1 dataset (95.8-99.1%). The prediction step confirmed the reliability of the

296   models developed with raw and SMOOTH-SNV-d1 datasets; indeed the correct classification

297   rates resulted between 95.8 and 100%, whereas the SMOOTH and SMOOTH-SNV data

298   reached as high as 97.9 and 93.7% of correct classification, respectively.

299   Considering that the dearth or soft mathematical pre-treatment should be preferred (Grassi,

300   Casiraghi, & Alamprese, 2018), LDA results based on the selected variables of raw spectra

301   looked the more promising among the developed models. In addition, the cross-validation

302   results of raw data, being more stable when changing the external test set, revealed a more

303   robust model. These results look much more promising than those obtained by Cayuela

304   (2017) using NIR (100-2300 nm) spectra collected on a huge set of virgin (extra, virgin and

305   lampante) olive oils. The author's main purpose was to perform PLS regression models to

306   quantify FAAE, FAEE, and FAME. Furthermore, an oil classification based on each

307   parameter was performed, but no information about the classification algorithm was reported.

308   Moreover, the performance evaluation was based on the correspondence between the

309   regression predicted values and the defined classes of the samples from reference

310   measurements. In any case, correct classification rates of the models based on FAAE (two

311   classes with 75 mg kg^{-1} threshold), FAEE (two classes with 35 mg kg^{-1} threshold), and FAME

312   (two classes with 35 mg kg^{-1} threshold) were 78.65-94.8%, 70.0-88.7%, and 95.2%,

13

respectively. Thus, it is difficult to deeply compare results by Cayuela (2017) with the LDA

models obtained in this work.

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,

14

338   meaning that around the half of the test samples belonging to one of the class was not rejected

339   from the assignment to the other class. Referring to the Cooman's plots (Fig. 4), these

340   samples are located in the small square on the bottom resulting from the intersection of 95%

341   threshold lines of class 1 and class 2 models. Those samples were accepted by both the model

342   of EVOO and non-EVOO samples. In particular, the lower specificity values of class 2 (non-

343   EVOO), which were on average between 40 and 50% for the three considered external sets

344   (data not shown), implicates a high probability of assignment of non-EVOO to the EVOO

345   class. Thus, it is clear that the application of SIMCA class-modelling could be hazardous for

346   EVOO labelling based on FAEE content, and the discriminant strategy (LDA) should be

347   preferred. Actually, when only two well-defined classes are considered, i.e. FAEE ≤ 35 mg

348   kg^{-1} (EVOO class) and FAEE > 35 mg kg^{-1} (non-EVOO class), a discriminant-based strategy

349   as LDA is a reliable option. Indeed, a discriminant classification method divides the sample

350   domain into a number of sub-areas corresponding to the spaces of the considered classes. A

351   sample is always assigned to one of the considered classes, even if it is located very far from

352   all the classes. However, when only two well-defined classes are considered, such in this

353   study, the classification-discriminant method can give reliable results thanks to the low

354   complexity of the dataset (Di Egidio, Oliveri, Woodcock, & Downey, 2011).

355   Future perspectives can consider the possibility to enlarge the VOO dataset by collecting

356   samples from different Italian regions and covering different cultivars and storage strategies.

357   Moreover, the non-EVOO class could be extended accordingly. From an instrumental point of

358   view, a multichannel IR-instrument based on 30 or fewer wavenumbers could be developed

359   from commercial prototypes and its reliability compared with the presented models.

360

361   **4. Conclusions**

362 The potential of FT-IR spectroscopy coupled with chemometrics as a tool for a rapid and fast

363 discrimination of extra virgin olive oils and virgin olive oils based on fatty acid ethyl ester

364 content has been assessed. The legal limit for FAEE content enables to highlight two well-

365 defined classes of products, thus a discriminant classification approach can be considered as

366 the most suitable in this context, to be preferred over the SIMCA class-modelling algorithm.

367 LDA models based on selected features of the raw spectra datasets gave good and robust

368 results, with percentages of correct classification in prediction higher than 95%. Being the

369 discriminant classification results so promising, the application of FT-IR spectroscopy should

370 be encouraged in the olive oil sector. Moreover, since the variable selection strategy was

371 effective, the development of simplified and cheap instruments can boost their use also in

372 small enterprises. Such instruments can be calibrated in order to give also results about oil

373 composition, thus providing producers and control bodies with many important information

374 about olive oil samples.

375

380

381 **References**

382 Alcala, S., Ocaña, M. T., Cárdenas, J. R., Miquel, M. Á., Vilar, J., Espínola, F., & Moya, M.

383   (2017). Alkyl esters content and other quality parameters in oil mill: A response surface

384   methodology study. *European Journal of Lipid Science and Technology, 119,* 1600026.

385   https://doi.org/10.1002/ejlt.201600026.

16

386 Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition

387 in food analysis. *Journal of Chromatography A, 1158,* 196-214.

388 https://doi.org/10.1016/j.chroma.2007.05.024.

389 Bro, R. (2003). Multivariate calibration: What is in chemometrics for the analytical

390 chemistry? *Analytica Chimica Acta, 500,* 185-194. https://doi.org/10.1016/S0003-

391 2670(03)00681-0.

392 Caponio, F., Squeo, G., Curci, M., Silletti, R., Paradiso, V. M., Summo, C., Crecchio, C., &

393 Pasqualone, A. (2018). Calcium carbonate effect on alkyl esters and enzymatic activities

394 during olive processing. *Italian Journal of Food Science, 30,* 381-392.

395 https://doi.org/10.14674/IJFS-1087.

396 Cayuela, J. A. (2017). Rapid NIR determination of alkyl esters in virgin olive oil. *Grasas y*

397 *Aceites, 68,* e195. http://dx.doi.org/10.3989/gya.1275162.

398 Cayuela, J. A., & García, J. F. (2017). Sorting olive oil based on alpha-tocopherol and total

399 tocopherol content using near-infra-red spectroscopy (NIRS) analysis. *Journal of Food*

400 *Engineering, 202,* 79-88. https://doi.org/10.1016/j.jfoodeng.2017.01.015.

401 Cayuela, J. A., & García, J. F. (2018). Nondestructive measurement of squalene in olive oil

402 by near infrared spectroscopy. *LWT - Food Science and Technology, 88,* 103-108.

403 https://doi.org/10.1016/j.lwt.2017.09.047.

404 Commission Delegated Regulation (EU) 2016/2095 of 26 September 2016 amending

405 Regulation (EEC) No 2568/91 on the characteristics of olive oil and olive-residue oil and

406 on the relevant methods of analysis. *Official Journal of the European Union,* L 326/1-6.

407 Commission Regulation (EU) No 61/2011, of 24 January 2011 amending Regulation (EEC)

408 No 2568/91 on the characteristics of olive oil and olive-residue oil and on the relevant

409 methods of analysis. *Official Journal of the European Union,* L 23/1-14.

Council Regulation (EC) No 1513/2001, of 23 July 2001 amending Regulations No 136/66/EEC and (EC) No 1638/98 as regards the extension of the period of validity of the aid scheme and the quality strategy for olive oil. *Official Journal of the European Communities,* L 201/4-7.

Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusic, V., & Georgiou, C. A. (2016). Food authentication: Techniques, trends & emerging approaches. *TrAC Trends in Analytical Chemistry, 85*, 123-132. https://doi.org/10.1016/j.trac.2016.02.026.

Di Egidio, V., Oliveri, P., Woodcock, T., & Downey, G. (2011). Confirmation of brand identity in foods by near infrared transflectance spectroscopy using classification and class-modelling chemometric techniques - The example of a Belgian beer. *Food Research International, 44,* 544-549. https://doi.org/10.1016/j.foodres.2010.11.021.

Dobson, G. (2001). Spectroscopy and spectrometry of lipids - Part 1. *European Journal of Lipid Science and Technology, 103,* 815-840. https://doi.org/10.1002/1438-9312(200112)103:12<815::AID-EJLT815>3.0.CO;2-X

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7(2),* 179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Forina, M., Lanteri, S., Armanino, C., Casolino, C., Casale, M., & Oliveri, P. (2008). V-PARVUS 2008. University of Genova. http://www.parvus.unige.it.

Forina, M., Lanteri, S., Casale, M., & Cerrato Oliveros, M. C. (2007). Stepwise orthogonalization of predictors in classification and regression techniques: an "old" technique revisited. *Chemometrics and Intelligent Laboratory Systems, 87,* 252-261. https://doi.org/10.1016/j.chemolab.2007.03.003.

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

434   near-infrared   reflectance   spectroscopy.   *Sensors,   17,*   26-42.

435   https://doi.org/10.3390/s17112642.

436   Gómez-Caravaca, A. M., Maggio, R. M., & Cerretani, L. (2016). Chemometric applications

437   to assess quality and critical parameters of virgin and extra-virgin olive oil. A review.

438   *Analytica Chimica Acta, 913,* 1-21. https://doi.org/10.1016/j.aca.2016.01.025.

439   Grassi, S., Casiraghi, E., & Alamprese, C. (2018). Handheld NIR device: A non-targeted

440   approach to assess authenticity of fish fillets and patties. *Food Chemistry, 243,* 382-388.

441   https://doi.org/10.1016/j.foodchem.2017.09.145.

442   Guillen, M. D., & Cabo, N. (1997). Characterization of edible oils and lard by Fourier

443   transform infrared spectroscopy. Relationships between composition and frequency of

444   concrete bands in the fingerprint region. *Journal of the American Oil Chemists' Society, 74,*

445   1281-1286. https://doi.org/10.1007/s11746-997-0058-4.

446   Kjeldahl, K., & Bro, R. (2010). Some common misunderstandings in chemometrics. *Journal*

447   *of Chemometrics, 24,* 558-564. https://doi.org/10.1002/cem.1346.

448   Kowalski, B. R., & Bender, C. F. (1976). An orthogonal feature selection method. *Pattern*

449   *Recognition, 8,* 1-4. https://doi.org/10.1016/0031-3203(76)90023-6.

450   Lerma-García, M.J., Simó-Alfonso, E.F., Ramis-Ramos, G. & Herrero-Martínez, J.M. (2007).

451   Determination of tocopherols in vegetable oils by CEC using methacrylate ester-based

452   monolithic   columns.   *Electrophoresis*,   *22*,   4128–4135.

453   https://doi.org/10.1002/elps.200700195.

454   Oliveri, P., Di Egidio, V., Woodcock, T., & Downey, G. (2011). Application of class-

455   modelling techniques to near infrared data for food authentication purposes. *Food*

456   *Chemistry, 125,* 1450-1456. https://doi.org/10.1016/j.foodchem.2010.10.047.

457   Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16

458   December 2008 on classification, labelling and packaging of substances and mixtures,

amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending
Regulation (EC) No 1907/2006. *Official Journal of the European Union,* L. 353/1-1355.

Sikorska, E., Khmelinskii, I., & Sikorski, M. (2014). Vibrational and electronic spectroscopy
and chemometrics in analysis of edible oils. In R. Cruz, I. Khmelinskii, & M. Vieira (Eds.),
Methods in food analysis (pp. 201-234). Boca Raton: CRC Press.

Squeo, G., Silletti, R., Summo, C., Paradiso, V. M., Pasqualone, A., & Caponio, F. (2017).
Fatty acids methyl and ethyl esters behaviour during olives processing by means of
technological coadjuvants. *Italian Journal of Food Science, 29,* 370-376.
https://doi.org/10.14674/1120-1770/ijfs.v747.

United Nations (2016). Sustainable Development Goals Report 2016. UN.
http://ggim.un.org/documents/The%20Sustainable%20Development%20Goals%20Report
%202016.pdf. Accessed 1 December 2018.

Valli, E., Bendini, A., Maggio, R. M., Cerretani, L., Toschi, T. G., Casiraghi, E., & Lercker,
G. (2013). Detection of low-quality extra virgin olive oils by fatty acid alkyl esters
evaluation: a preliminary and fast mid-infrared spectroscopy discrimination by a
chemometric approach. *International Journal of Food Science & Technology, 48,* 548-555.
https://doi.org/10.1111/j.1365-2621.2012.03220.x.

Van de Voort, F.R., Sedman, J. & Russin, T. (2001). Lipid analysis by vibrational
spectroscopy. *European Journal of Lipid Science and Technology, 103,* 815–826.
https://doi.org/10.1002/1438-9312(200112)103:12<815::AID-EJLT1111815>3.0.CO;2-P.

Vlachos, N., Skopelitis, Y., Psaroudaki, M., Konstantinidou, V., Chatzilazarou, A. & Tegou,
E. (2006). Applications of Fourier transform-infrared spectroscopy to edible oils. *Analitica
Chimica Acta, 573–574*, 459–465. https://doi.org/10.1016/j.aca.2006.05.034.

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).

Table

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).

Class	N	Mean	SD*	CV*	Min	Median	Max	Range	IQR
1	113	6.68	4.95	74.21	1.81	5.34	31.36	29.54	3.93
2	46	106.90	77.20	72.25	37.70	72.50	298.80	261.10	114.60

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.

Pre-treatment	External test set	Correct classification (%)		
		Calibration	Cross-validation	Prediction
Raw	1	100	97.30	95.83
	2	100	96.40	95.83
	3	100	96.61	100
SMOOTH	1	100	97.30	97.92
	2	99.28	99.10	91.67
	3	99.49	95.76	97.56
SMOOTH-SNV	1	98.92	90.10	93.75
	2	99.10	94.60	87.50
	3	99.15	91.53	90.24
SMOOTH-SNV-d1	1	100	97.30	100
	2	100	100	95.83
	3	100	99.15	100

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.

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

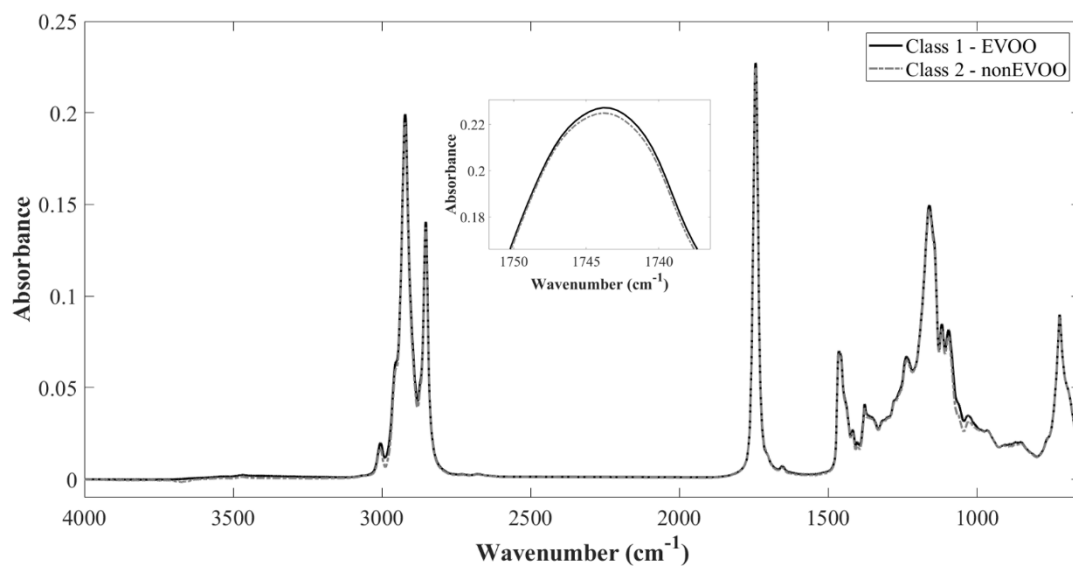SMOOTH, moving average smoothing; SNV, standard normal variate; d1, first derivative.
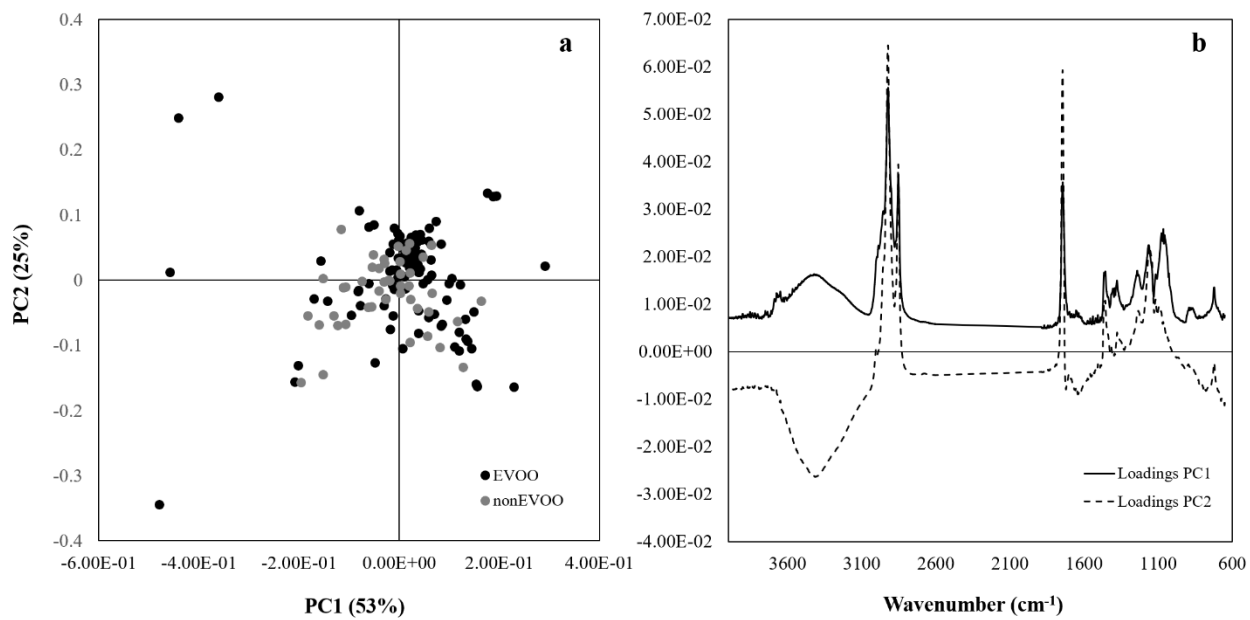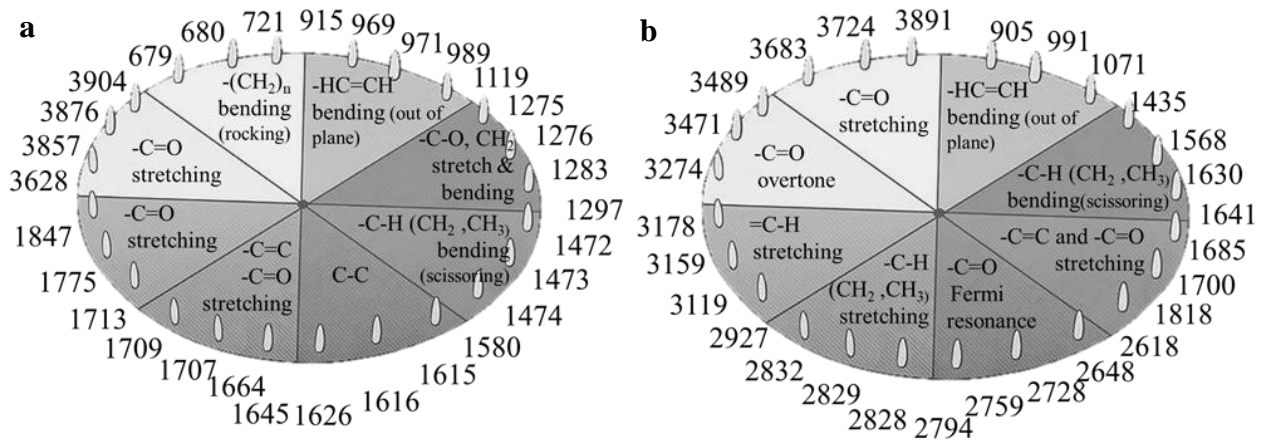
Figure 1.

Figure 2.

Figure 3.

Figure 4.