

Accepted Manuscript

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

Giacomo Squeo, Silvia Grassi, Vito Michele Paradiso, Cristina Alamprese,
Francesco Caponio



PII: S0956-7135(19)30126-4
DOI: 10.1016/j.foodcont.2019.03.027
Reference: JFCO 6575
To appear in: *Food Control*
Received Date: 17 January 2019
Accepted Date: 20 March 2019

Please cite this article as: Giacomo Squeo, Silvia Grassi, Vito Michele Paradiso, Cristina Alamprese, Francesco Caponio, FT-IR extra virgin olive oil classification based on ethyl ester content, *Food Control* (2019), doi: 10.1016/j.foodcont.2019.03.027

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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

Giacomo Squeo^{1*}, Silvia Grassi^{2*}, Vito Michele Paradiso¹, Cristina Alamprese², 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.

17 **ABSTRACT**

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

38

39 **KEYWORDS:** FT-IR; LDA; SIMCA; Fatty acid ethyl esters; Extra virgin olive oil;
40 Authentication.

41 1. Introduction

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

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

64 According to the official method (Commission Regulation (EU) No 61/2011), FAAE
65 determination requires their separation from triacylglycerols and other oil constituents by

66 chromatography on a hydrated silica gel column using Sudan 1 (1-phenylazo-2-naphthol) as
67 indicator for the elution. Then, the FAAE fraction is collected, dried and re-suspended in n-
68 heptane or iso-octane. Alkyl esters are finally separated by capillary gas-chromatography.
69 Quantification is achieved by the addition of a proper internal standard. Overall, around 5
70 hours are needed to complete the analysis, without considering the preparation steps such as
71 silica conditioning. Besides, a large amount of organic solvents is used for the determination,
72 mainly n-hexane. Indeed, approximately 250 mL are required for one determination.
73 Considering that the analysis should be performed at least in duplicate in order to obtain
74 reliable results, around 1 L of solvent is required for the analysis of only two oil samples.

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

82 Spectroscopic techniques are non-destructive, green, fast and easy to use. Among them, mid-
83 infrared (MIR) spectroscopy is one of the most used, having an illustrious history in lipid
84 chemistry, and it has experienced growing interest and applications thanks to the introduction
85 of the Fourier transform instruments (FT-IR) (Dobson, 2001). The MIR range goes from
86 around 2.5 to 25 μm or, as most commonly reported, from 4000 to 400 cm^{-1} . Absorption of a
87 MIR photon typically excites one of the fundamental vibrations, associated with a change of
88 the dipole moment of an oscillating molecule (Sikorska, Khmelinskii, & Sikorski, 2014).
89 Despite the complexity of spectra collected along the food systems, the association of MIR
90 spectroscopy with chemometrics allows the extraction of the significant and valuable

91 information (Gómez-Caravaca, Maggio, & Cerretani, 2016). Indeed, when spectra are
92 recorded from real food samples, they contain information about different components of the
93 sample matrix together with their interactions, and multivariate methods are successfully used
94 in interpreting the spectra signals for analytical purposes (Bro, 2003; Kjeldahl & Bro, 2010,
95 Sikorska, Khmelinskii, & Sikorski, 2014). Several chemometric approaches might be used,
96 falling in two main classes: qualitative and quantitative methods. As regard to alkyl esters, in
97 a previous study by Valli et al. (2013), Partial Least Square (PLS) regression models were
98 tentatively developed for the quantification of FAAE based on VOO FT-IR spectra. However,
99 they were aimed at the quantification of the sum of ethyl and methyl esters as well as their
100 ratio, parameters that are no longer considered for the EVOO classification.

101 After the introduction of the Commission Delegated Regulation (EU) 2016/2095, few authors
102 have taken interest in FAAE determination by green methods. Indeed, near infrared (NIR) and
103 Vis-NIR spectroscopy has been used to develop regression models for measuring total FAAE
104 content, as well as FAEE and FAME content separately (Cayuela, 2017; Garrido-Varo,
105 Sánchez, De la Haba, Torres, & Pérez-Marín, 2017). However, even though chemometric
106 approaches can overcome the overlapping NIR signals resulting from first and second
107 overtones and combinations of the fundamental vibrations, more accurate assignments of
108 absorption bands can be reached by MIR spectroscopy. This is particularly relevant when
109 assessing differences among molecules having similar bonds that scatter in a complex matrix
110 such as oil. As far as we know, despite the importance of FAEE, no other attempts have been
111 carried out by IR spectroscopy to develop a rapid procedure for their analysis. **Starting from
112 these considerations, the aim of this work was the application of IR spectroscopy to the
113 development of classification models (based on Linear Discriminant Analysis and Soft
114 Independent Modelling of Class Analogy) able to discriminate between EVOO and non-
115 EVOO based on FAEE content.** Though FAEE is a continuous variable, a classification

116 approach was chosen instead of quantification since, by a practical point of view, the
117 proposed method should address a discrimination issue regarding the authentication of
118 EVOO, where authentication is intended as the compliance of a food with its label description
119 (Danezis, Tsagkaris, Camin, Brusica, & Georgiou, 2016). A similar approach based on
120 discriminant classification techniques has already been applied in the literature in order to
121 develop fast sorting tests for olive oils, based on the content of α -tocopherol or squalene
122 (Cayuela & Garcia, 2018; Cayuela & Garcia, 2017). Supervised classification techniques use
123 the information about the known class membership of training samples in order to create
124 classification rules able to assign new unknown samples to one of the defined classes, based
125 on their fingerprint measurement (Berrueta, Alonso-Salces, & Héberger, 2007). Thus, these
126 chemometric techniques perfectly fit in authentication issues where the goal is to verify if a
127 sample belongs to a predefined class, such in the case of EVOO and non-EVOO differentiated
128 by the FAEE content.

129

130 **2. Materials and methods**

131 *2.1. Sampling*

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

138

139 *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 $^{\circ}\text{C}$) and a DB-5HT (15 m \times 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 $^{\circ}\text{C}$ for 1 min; from 80 to 140 $^{\circ}\text{C}$ at 20 $^{\circ}\text{C min}^{-1}$; from
155 140 to 335 $^{\circ}\text{C}$ at 5 $^{\circ}\text{C min}^{-1}$; 335 $^{\circ}\text{C}$ 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 $^{\circ}\text{C}$),
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:

$$180 \quad FW = \frac{(\bar{x}_{v1} - \bar{x}_{v2})^2}{\sum_{i=1}^{I_1} \frac{(x_{iv1} - \bar{x}_{v1})^2}{I_1} + \sum_{i=1}^{I_2} \frac{(x_{iv2} - \bar{x}_{v2})^2}{I_2}}$$

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

184 The discrimination of EVOO (class 1; FAEE ≤ 35 mg kg⁻¹) from non-EVOO (class 2; FAEE
185 > 35 mg kg⁻¹) samples was performed by applying two classification techniques on the
186 reduced spectral datasets, i.e. Linear Discriminant Analysis (LDA), a supervised pattern
187 recognition technique based on discriminant canonicals, and Soft Independent Modelling of
188 Class Analogy (SIMCA), a supervised class-modelling technique. Classification models were

189 validated both in cross-validation by 5 cancellation groups (CV) and in prediction using three
190 different external test sets, randomly created, each containing about 30% samples of the
191 whole dataset.

192 Since LDA maximises the separation among classes by the construction of an optimal *a-*
193 *posteriori* classification rule that assigns every object to one unique class, the LDA models'
194 performance was evaluated by correct classification percentage, i.e. by the percentage of
195 samples correctly assigned by the *a-posteriori* rule to the *a-priori* defined class. In the case of
196 SIMCA, sensitivity (true positive rate) and specificity (false positive rate) were also
197 calculated to evaluate the effectiveness of the classification models. Indeed, in a class-
198 modelling technique, such as SIMCA, a sample can be assigned: (1) correctly and exclusively
199 to the actual class; (2) exclusively to one class which differs from the actual class; (3) to no
200 classes; or (4) to more than one class. Thus, while evaluating a SIMCA model, it comes the
201 need to consider the latter three cases, sources of classification errors, through the calculation
202 of the above-mentioned figures of merit (sensitivity and specificity).

203 Data elaboration was performed by The Unscrambler X software (v. 10.2, Camo Software
204 AS, Oslo, Norway) and the V-Parvus package (Forina, Lanteri, Armanino, Casolino, Casale,
205 & Oliveri, 2008).

206

207 **3. Results and discussion**

208 *3.1. Fatty acid ethyl ester content of olive oil samples*

209 Table 1 reports the descriptive statistics of the olive oil samples, divided by classes: EVOO
210 (class 1), non-EVOO (class 2). The number of objects was quite different between the two
211 classes since no previous information about the amount of FAEE in the collected samples
212 were provided by producers. In any case, this situation is representative of a real scenario,
213 because during the authentication of EVOO a low number of non-EVOO should be expected.

214 Class 2 (non-EVOO), although made up of a lower number of samples, was characterised by
215 a higher variability, covering a wide range of FAEE content, from 37.7 to 298 mg kg⁻¹. On the
216 contrary, the variability of class 1 (EVOO samples) was quite low, as most samples had a
217 FAEE content in the range 1-10 mg kg⁻¹.

218

219 *3.2. FT-IR spectra and data exploration*

220 Fig. 1 depicts the average FT-IR spectra for the two olive oil classes (i.e. EVOO and non-
221 EVOO). Overall, spectra were similar to those reported for edible oils and olive oils (Dobson,
222 2001; Sikorska et al., 2014). Although the spectra were almost overlapping, EVOO (class 1)
223 presented slightly higher band intensities roughly along the whole spectral range. More
224 evident differences in intensity were observed for particular bands, namely in the regions
225 from around 3030 to 2800 cm⁻¹, corresponding to different stretching vibrations, around 1238
226 and 1160 cm⁻¹, corresponding to -C-O, -CH₂- stretching and bending, and around 1118, 1097
227 and 1030 cm⁻¹, corresponding to -C-O stretching (Sikorska et al., 2014). No considerable
228 differences were observable for those bands reported as typical of the ester linkage (2730,
229 2677 cm⁻¹), with the exception of the peak around 1746 cm⁻¹ (highlighted in Fig. 1), which
230 might be potentially influenced by the amount of alkyl esters. The little differences observed
231 between the spectra of the two oil classes confirm the necessity of chemometric tools to
232 extract useful information.

233 As a first step, an exploratory analysis was performed by PCA on both the raw and the
234 different pre-treated data. Fig. 2a reports the score plot resulting from the raw data analysis.
235 Few samples belonging to EVOO class resulted isolated from the sample cloud. However, no
236 sample subsets were identified; EVOO and non-EVOO classes appeared confused. Sample
237 distribution in the space of the first two principal components (PC) was very similar for any
238 of the pre-treated dataset explored (data not shown).

239 The PC1 loadings plot (Fig. 2b) showed that the most informative variables affecting the
240 sample distribution, i.e. the ones farer from zero, corresponded to the most evident differences
241 in the absorbance observed in the raw spectra. PC2 loadings trend (Fig. 2b) remarked a strong
242 negative influence in sample separation of the band from 3770 to 3070 cm^{-1} , likely due to -
243 OH group of oxidation products (Sikorska et al., 2014), and a strong positive influence of
244 variables around 3030 - 2800 cm^{-1} and 1743 cm^{-1} . Those regions were highlighted also in the
245 VOO spectra recorded by Valli et al. (2013) and pre-treated by multiplicative scatter
246 correction and mean centering.

247

248 3.3. Classification models

249 As the whole spectral range did not look promising in EVOO and non-EVOO discrimination,
250 a variable selection strategy was investigated. The SELECT algorithm allowed to identify the
251 30 most informative variables, among the original 5478, to be used in the following
252 classification procedures. Variable selection and classification model development were
253 performed on both raw and pre-treated data (SMOOTH, SMOOTH-SNV, and SMOOTH-
254 SNV-d1).

255 The 30 variables selected from the raw dataset are reported in Fig. 3a. Half of the selected
256 variables belongs to the high informative fingerprint region, i.e. between 1500 and 650 cm^{-1} .
257 The signals from 750 to 679 cm^{-1} can be linked to the CH_2 rocking vibrations occurring at 723
258 cm^{-1} (Guillen & Cabo, 1997; Lerma-García, Simó-Alfonso, Ramis-Ramos, & Herrero-
259 Martínez, 2011). The variables selected between 1000 and 900 cm^{-1} could be linked to C-H
260 out of plane deformation occurring at 968 cm^{-1} due to isolated *trans* double bonds presence
261 (Guillen & Cabo, 1997). The 1119 cm^{-1} variable could originate from signal due to the C-O
262 stretching vibration (Guillen & Cabo, 1997) as well as the variables selected in the range
263 1300-1270 cm^{-1} , where also the C-H bending vibration is expected to occur. The fourth range

264 of selected variables includes wavenumbers 1474-1472 cm^{-1} that are related to CH_2 and CH_3
265 scissoring vibrations (Guillen & Cabo, 1997). The other selected variables lay outside the IR
266 fingerprint region. Between 1626 cm^{-1} and 1580 cm^{-1} the selected signals can be related to C-
267 C vibration in aromatic structures. The absorption of carbonyl aldehydes and ketones are in
268 the selected range 1730-1645 cm^{-1} (Van de Voort, Sedman, & Russin, 2001). The C=O
269 stretching vibration characterises the variables selected between 1775 and 1713 cm^{-1} ; indeed,
270 these are related to an intense peak that is normally observed at 1746 cm^{-1} and known to be
271 related to carbonyl group vibration in ester linkage between fatty acid and glycerol (Sikorska
272 et al., 2014). Other four variables were selected in the range 4000-3600 cm^{-1} ; normally, the
273 vibration of C=O stretching belonging to triacylglycerol ester bonds occurs also in this range
274 (Sikorska et al., 2014).

275 In Fig. 3b the 30 variables selected for the SMOOTH-SNV-d1 dataset are reported. Apart
276 from the variables selected in the regions already discussed above, relevant variables appear
277 between 3500 and 2600 cm^{-1} . In the region 3000-2800 cm^{-1} , C-H stretching vibrations are
278 present, such as the stretching of CH_3 in the range 2962-2873 cm^{-1} (Vlachos, Skopelitis,
279 Psaroudaki, Konstantinidou, Chatzilazarou, & Tegou, 2006) and two intense bands at 2924
280 and 2853 cm^{-1} linked to CH_2 stretching of the fatty acid chains in triacylglycerols (Guillèn &
281 Cabo, 1997). Also C=O Fermi resonance of ester groups occurs at 2677 and 2730 cm^{-1}
282 (Guillèn & Cabo, 1997; Lerma-García et al., 2011). The last variables selected from the
283 transformed dataset belong to the 4000-3500 cm^{-1} range, where the relevant bands are related
284 to C=O overtone and stretching vibrations (Vlachos et al., 2006).

285 Concerning LDA, good classification performances were obtained using both raw and
286 SMOOTH-SNV-d1 pre-treated data, while less promising results were achieved with
287 SMOOTH or SMOOTH-SNV data (Table 2). In details, for both raw and SMOOTH-SNV-d1
288 datasets the correct classification rate in calibration was 100% no matter the external test set

289 used, whereas the SMOOTH-SNV model resulted affected by the data excluded from the
290 calibration sets, ranging the correct classification from 98.9 to 99.1% depending on the
291 considered test set. Smoothed data also provided very good classification rate in calibration
292 even though less accurate than those of raw and SMOOTH-SNV-d1 datasets. With internal
293 validation (cross-validation), the correct classification rates remained high for the SMOOTH-
294 SNV-d1 dataset (97.3-100%), while decreased for raw data (96.4-97.3%) as well as for
295 SMOOTH-SNV data (90.1-94.6%). Data smoothing gave the best results just after those of
296 SMOOTH-SNV-d1 dataset (95.8-99.1%). The prediction step confirmed the reliability of the
297 models developed with raw and SMOOTH-SNV-d1 datasets; indeed the correct classification
298 rates resulted between 95.8 and 100%, whereas the SMOOTH and SMOOTH-SNV data
299 reached as high as 97.9 and 93.7% of correct classification, respectively.

300 Considering that the dearth or soft mathematical pre-treatment should be preferred (Grassi,
301 Casiraghi, & Alamprese, 2018), LDA results based on the selected variables of raw spectra
302 looked the more promising among the developed models. In addition, the cross-validation
303 results of raw data, being more stable when changing the external test set, revealed a more
304 robust model. These results look much more promising than those obtained by Cayuela
305 (2017) using NIR (100-2300 nm) spectra collected on a huge set of virgin (extra, virgin and
306 lampante) olive oils. The author's main purpose was to perform PLS regression models to
307 quantify FAAE, FAEE, and FAME. Furthermore, an oil classification based on each
308 parameter was performed, but no information about the classification algorithm was reported.
309 Moreover, the performance evaluation was based on the correspondence between the
310 regression predicted values and the defined classes of the samples from reference
311 measurements. In any case, correct classification rates of the models based on FAAE (two
312 classes with 75 mg kg⁻¹ threshold), FAEE (two classes with 35 mg kg⁻¹ threshold), and FAME
313 (two classes with 35 mg kg⁻¹ threshold) were 78.65-94.8%, 70.0-88.7%, and 95.2%,

314 respectively. Thus, it is difficult to deeply compare results by Cayuela (2017) with the LDA
315 models obtained in this work.

316 Class-modelling techniques are often more appropriate in food authentication than
317 discriminant methods, because they aim at answering to the essential question in food
318 authenticity problems: if a product is sold with a specific label, it is important to verify if it is
319 really consistent with the claimed characteristics (Oliveri, Di Egidio, Woodcock, & Downey,
320 2011). Moreover, this approach does not force a sample to belong to one specific class,
321 resulting in a more severe but realistic scenario in real-life applications when a large number
322 of non-target classes is possibly present. Thus, SIMCA models for olive oil discrimination
323 based on FAEE classes were developed and the related results are provided in Table 3.
324 Differently from what observed with LDA, the best classification performances were obtained
325 on pre-treated data instead of raw ones. In details, models obtained with SMOOTH-SNV-d1
326 treated spectra gave higher percentages of classification ability than those developed with
327 SMOOTH or SMOOTH-SNV data, ranging from around 91 to slightly less than 96%.
328 Looking at the prediction performances of the models, the correct classification ability was,
329 on average, higher when smoothing was applied to spectral data although the absolute best
330 performance was observed applying first derivative in combination with smoothing and SNV
331 (up to 81.25%). Sensitivity of all the models was promising, showing a percentage of true
332 positive assignments ranging from 75 to about 92%, being again, on average, SMOOTH
333 model the best. Indeed, when comparing the Cooman's plot (Fig. 4) of the best performing
334 SIMCA models, it can be noticed that a similar and low number of samples is distributed
335 outside the 95% model thresholds for SMOOTH and SMOOTH-SNV-d1 datasets (Fig. 4a and
336 4c). It means that those samples were not accepted by any of the class models (false
337 negatives). However, the most important issue concerned the low specificity affecting models.
338 The average specificity was around 50% (Table 3) considering the different pre-treatments,

339 meaning that around the half of the test samples belonging to one of the class was not rejected
340 from the assignment to the other class. Referring to the Cooman's plots (Fig. 4), these
341 samples are located in the small square on the bottom resulting from the intersection of 95%
342 threshold lines of class 1 and class 2 models. Those samples were accepted by both the model
343 of EVOO and non-EVOO samples. In particular, the lower specificity values of class 2 (non-
344 EVOO), which were on average between 40 and 50% for the three considered external sets
345 (data not shown), implicates a high probability of assignment of non-EVOO to the EVOO
346 class. Thus, it is clear that the application of SIMCA class-modelling could be hazardous for
347 EVOO labelling based on FAEE content, and the discriminant strategy (LDA) should be
348 preferred. Actually, when only two well-defined classes are considered, i.e. $FAEE \leq 35 \text{ mg}$
349 kg^{-1} (EVOO class) and $FAEE > 35 \text{ mg kg}^{-1}$ (non-EVOO class), a discriminant-based strategy
350 as LDA is a reliable option. Indeed, a discriminant classification method divides the sample
351 domain into a number of sub-areas corresponding to the spaces of the considered classes. A
352 sample is always assigned to one of the considered classes, even if it is located very far from
353 all the classes. However, when only two well-defined classes are considered, such in this
354 study, the classification-discriminant method can give reliable results thanks to the low
355 complexity of the dataset (Di Egidio, Oliveri, Woodcock, & Downey, 2011).

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

361

362 4. Conclusions

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

376

377 **Acknowledgements**

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

381

382 **References**

383 Alcalá, S., Ocaña, M. T., Cárdenas, J. R., Miquel, M. Á., Vilar, J., Espínola, F., & Moya, M.
384 (2017). Alkyl esters content and other quality parameters in oil mill: A response surface
385 methodology study. *European Journal of Lipid Science and Technology*, 119, 1600026.
386 <https://doi.org/10.1002/ejlt.201600026>.

- 387 Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition
388 in food analysis. *Journal of Chromatography A*, *1158*, 196-214.
389 <https://doi.org/10.1016/j.chroma.2007.05.024>.
- 390 Bro, R. (2003). Multivariate calibration: What is in chemometrics for the analytical
391 chemistry? *Analytica Chimica Acta*, *500*, 185-194. [https://doi.org/10.1016/S0003-](https://doi.org/10.1016/S0003-2670(03)00681-0)
392 [2670\(03\)00681-0](https://doi.org/10.1016/S0003-2670(03)00681-0).
- 393 Caponio, F., Squeo, G., Curci, M., Silletti, R., Paradiso, V. M., Summo, C., Crecchio, C., &
394 Pasqualone, A. (2018). Calcium carbonate effect on alkyl esters and enzymatic activities
395 during olive processing. *Italian Journal of Food Science*, *30*, 381-392.
396 <https://doi.org/10.14674/IJFS-1087>.
- 397 Cayuela, J. A. (2017). Rapid NIR determination of alkyl esters in virgin olive oil. *Grasas y*
398 *Aceites*, *68*, e195. <http://dx.doi.org/10.3989/gya.1275162>.
- 399 Cayuela, J. A., & García, J. F. (2017). Sorting olive oil based on alpha-tocopherol and total
400 tocopherol content using near-infra-red spectroscopy (NIRS) analysis. *Journal of Food*
401 *Engineering*, *202*, 79-88. <https://doi.org/10.1016/j.jfoodeng.2017.01.015>.
- 402 Cayuela, J. A., & García, J. F. (2018). Nondestructive measurement of squalene in olive oil
403 by near infrared spectroscopy. *LWT - Food Science and Technology*, *88*, 103-108.
404 <https://doi.org/10.1016/j.lwt.2017.09.047>.
- 405 Commission Delegated Regulation (EU) 2016/2095 of 26 September 2016 amending
406 Regulation (EEC) No 2568/91 on the characteristics of olive oil and olive-residue oil and
407 on the relevant methods of analysis. *Official Journal of the European Union*, L 326/1-6.
- 408 Commission Regulation (EU) No 61/2011, of 24 January 2011 amending Regulation (EEC)
409 No 2568/91 on the characteristics of olive oil and olive-residue oil and on the relevant
410 methods of analysis. *Official Journal of the European Union*, L 23/1-14.

- 411 Council Regulation (EC) No 1513/2001, of 23 July 2001 amending Regulations No
412 136/66/EEC and (EC) No 1638/98 as regards the extension of the period of validity of the
413 aid scheme and the quality strategy for olive oil. *Official Journal of the European*
414 *Communities*, L 201/4-7.
- 415 Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusica, V., & Georgiou, C. A. (2016). Food
416 authentication: Techniques, trends & emerging approaches. *TrAC Trends in Analytical*
417 *Chemistry*, 85, 123-132. <https://doi.org/10.1016/j.trac.2016.02.026>.
- 418 Di Egidio, V., Oliveri, P., Woodcock, T., & Downey, G. (2011). Confirmation of brand
419 identity in foods by near infrared transfectance spectroscopy using classification and
420 class-modelling chemometric techniques - The example of a Belgian beer. *Food Research*
421 *International*, 44, 544-549. <https://doi.org/10.1016/j.foodres.2010.11.021>.
- 422 Dobson, G. (2001). Spectroscopy and spectrometry of lipids - Part 1. *European Journal of*
423 *Lipid Science and Technology*, 103, 815-840. [https://doi.org/10.1002/1438-](https://doi.org/10.1002/1438-9312(200112)103:12<815::AID-EJLT815>3.0.CO;2-X)
424 [9312\(200112\)103:12<815::AID-EJLT815>3.0.CO;2-X](https://doi.org/10.1002/1438-9312(200112)103:12<815::AID-EJLT815>3.0.CO;2-X)
- 425 Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of*
426 *Eugenics*, 7(2), 179-188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>.
- 427 Forina, M., Lanteri, S., Armanino, C., Casolino, C., Casale, M., & Oliveri, P. (2008). V-
428 PARVUS 2008. University of Genova. <http://www.parvus.unige.it>.
- 429 Forina, M., Lanteri, S., Casale, M., & Cerrato Oliveros, M. C. (2007). Stepwise
430 orthogonalization of predictors in classification and regression techniques: an “old”
431 technique revisited. *Chemometrics and Intelligent Laboratory Systems*, 87, 252-261.
432 <https://doi.org/10.1016/j.chemolab.2007.03.003>.
- 433 Garrido-Varo, A., Sánchez, M. T., De la Haba, M. J., Torres, I., & Pérez-Marín, D. (2017).
434 Fast, low-cost and non-destructive physico-chemical analysis of virgin olive oils using

- 435 near-infrared reflectance spectroscopy. *Sensors*, 17, 26-42.
436 <https://doi.org/10.3390/s17112642>.
- 437 Gómez-Caravaca, A. M., Maggio, R. M., & Cerretani, L. (2016). Chemometric applications
438 to assess quality and critical parameters of virgin and extra-virgin olive oil. A review.
439 *Analytica Chimica Acta*, 913, 1-21. <https://doi.org/10.1016/j.aca.2016.01.025>.
- 440 Grassi, S., Casiraghi, E., & Alamprese, C. (2018). Handheld NIR device: A non-targeted
441 approach to assess authenticity of fish fillets and patties. *Food Chemistry*, 243, 382-388.
442 <https://doi.org/10.1016/j.foodchem.2017.09.145>.
- 443 Guillen, M. D., & Cabo, N. (1997). Characterization of edible oils and lard by Fourier
444 transform infrared spectroscopy. Relationships between composition and frequency of
445 concrete bands in the fingerprint region. *Journal of the American Oil Chemists' Society*, 74,
446 1281-1286. <https://doi.org/10.1007/s11746-997-0058-4>.
- 447 Kjeldahl, K., & Bro, R. (2010). Some common misunderstandings in chemometrics. *Journal*
448 *of Chemometrics*, 24, 558-564. <https://doi.org/10.1002/cem.1346>.
- 449 Kowalski, B. R., & Bender, C. F. (1976). An orthogonal feature selection method. *Pattern*
450 *Recognition*, 8, 1-4. [https://doi.org/10.1016/0031-3203\(76\)90023-6](https://doi.org/10.1016/0031-3203(76)90023-6).
- 451 Lerma-García, M.J., Simó-Alfonso, E.F., Ramis-Ramos, G. & Herrero-Martínez, J.M. (2007).
452 Determination of tocopherols in vegetable oils by CEC using methacrylate ester-based
453 monolithic columns. *Electrophoresis*, 22, 4128-4135.
454 <https://doi.org/10.1002/elps.200700195>.
- 455 Oliveri, P., Di Egidio, V., Woodcock, T., & Downey, G. (2011). Application of class-
456 modelling techniques to near infrared data for food authentication purposes. *Food*
457 *Chemistry*, 125, 1450-1456. <https://doi.org/10.1016/j.foodchem.2010.10.047>.
- 458 Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16
459 December 2008 on classification, labelling and packaging of substances and mixtures,

- 460 amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending
461 Regulation (EC) No 1907/2006. *Official Journal of the European Union*, L. 353/1-1355.
- 462 Sikorska, E., Khmelinskii, I., & Sikorski, M. (2014). Vibrational and electronic spectroscopy
463 and chemometrics in analysis of edible oils. In R. Cruz, I. Khmelinskii, & M. Vieira (Eds.),
464 *Methods in food analysis* (pp. 201-234). Boca Raton: CRC Press.
- 465 Squeo, G., Silletti, R., Summo, C., Paradiso, V. M., Pasqualone, A., & Caponio, F. (2017).
466 Fatty acids methyl and ethyl esters behaviour during olives processing by means of
467 technological coadjuvants. *Italian Journal of Food Science*, 29, 370-376.
468 <https://doi.org/10.14674/1120-1770/ijfs.v747>.
- 469 United Nations (2016). Sustainable Development Goals Report 2016. UN.
470 [http://ggim.un.org/documents/The%20Sustainable%20Development%20Goals%20Report](http://ggim.un.org/documents/The%20Sustainable%20Development%20Goals%20Report%202016.pdf)
471 [%202016.pdf](http://ggim.un.org/documents/The%20Sustainable%20Development%20Goals%20Report%202016.pdf). Accessed 1 December 2018.
- 472 Valli, E., Bendini, A., Maggio, R. M., Cerretani, L., Toschi, T. G., Casiraghi, E., & Lercker,
473 G. (2013). Detection of low-quality extra virgin olive oils by fatty acid alkyl esters
474 evaluation: a preliminary and fast mid-infrared spectroscopy discrimination by a
475 chemometric approach. *International Journal of Food Science & Technology*, 48, 548-555.
476 <https://doi.org/10.1111/j.1365-2621.2012.03220.x>.
- 477 Van de Voort, F.R., Sedman, J. & Russin, T. (2001). Lipid analysis by vibrational
478 spectroscopy. *European Journal of Lipid Science and Technology*, 103, 815-826.
479 [https://doi.org/10.1002/1438-9312\(200112\)103:12<815::AID-EJLT1111815>3.0.CO;2-P](https://doi.org/10.1002/1438-9312(200112)103:12<815::AID-EJLT1111815>3.0.CO;2-P).
- 480 Vlachos, N., Skopelitis, Y., Psaroudaki, M., Konstantinidou, V., Chatzilazarou, A. & Tegou,
481 E. (2006). Applications of Fourier transform-infrared spectroscopy to edible oils. *Analitica*
482 *Chimica Acta*, 573-574, 459-465. <https://doi.org/10.1016/j.aca.2006.05.034>.

483

484 **Figure captions**

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

488
489 **Fig. 2.** Principal Component Analysis results of raw FT-IR spectra collected on extra virgin
490 (EVOO) and non-extra virgin (nonEVOO) olive oil samples: a) PC1 vs PC2 score plot and b)
491 loading plot for PC1 and PC2.

492
493 **Fig. 3.** FT-IR variables (cm^{-1}) selected by SELECT algorithm before to perform classification
494 of extra virgin and non-extra virgin olive oil samples: a) 30 variables selected from the raw
495 dataset, b) 30 variables selected from the data pre-treated with smoothing, standard normal
496 variate and first derivative.

497
498 **Fig. 4.** Soft Independent Modelling of Class Analogy results for extra virgin (EVOO) and
499 non-extra virgin (nonEVOO) olive oil samples. Cooman's plots obtained with the 30 FT-IR
500 selected features and the external test set no. 3: a) smoothed data (SMOOTH; b) data pre-
501 treated with smoothing and standard normal variate (SMOOTH-SNV); c) data pre-treated
502 with smoothing, standard normal variate and first derivative (SMOOTH-SNV-d1).

Figure 1.

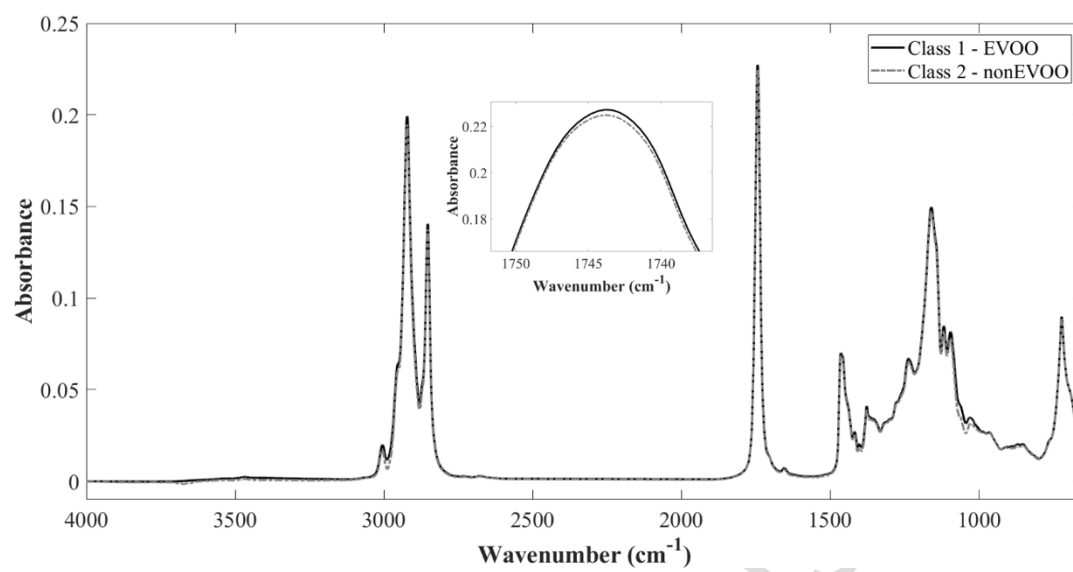


Figure 2.

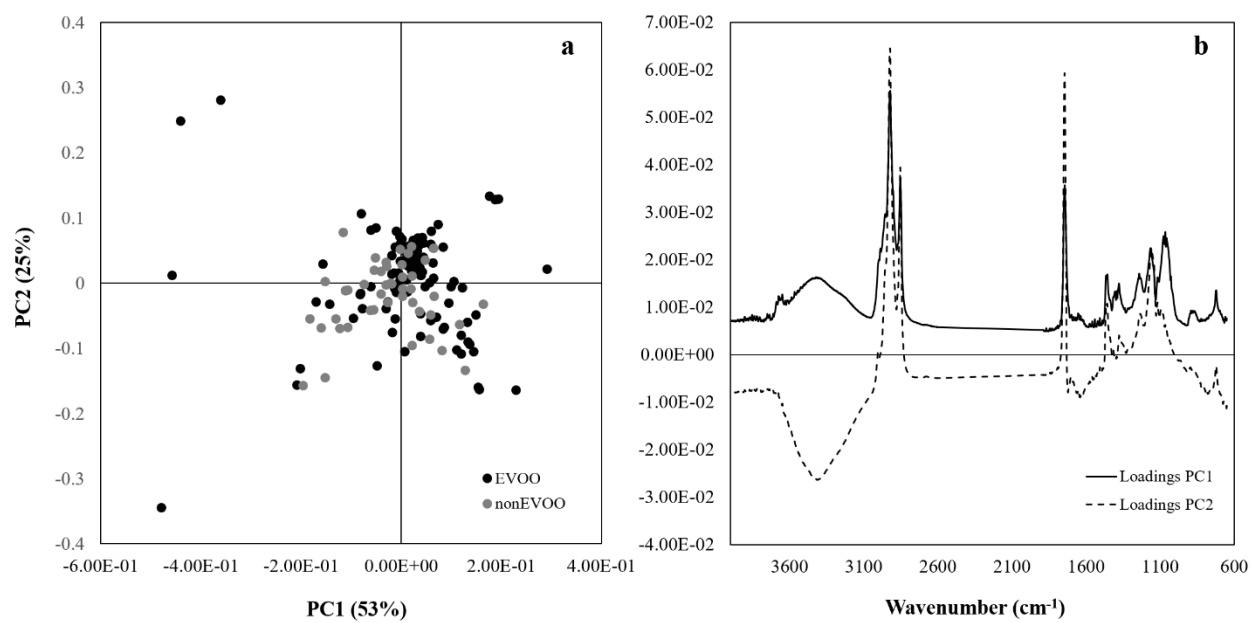


Figure 3.

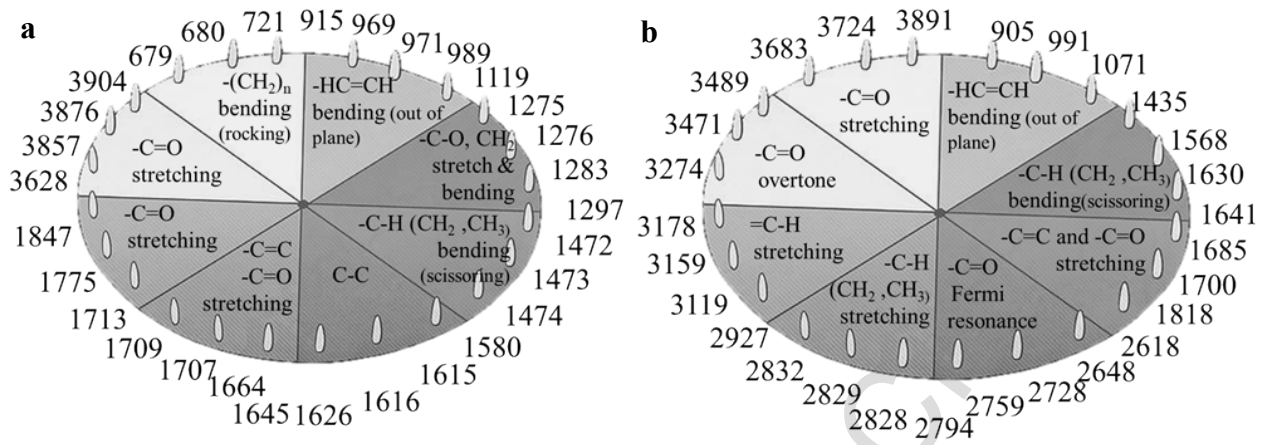
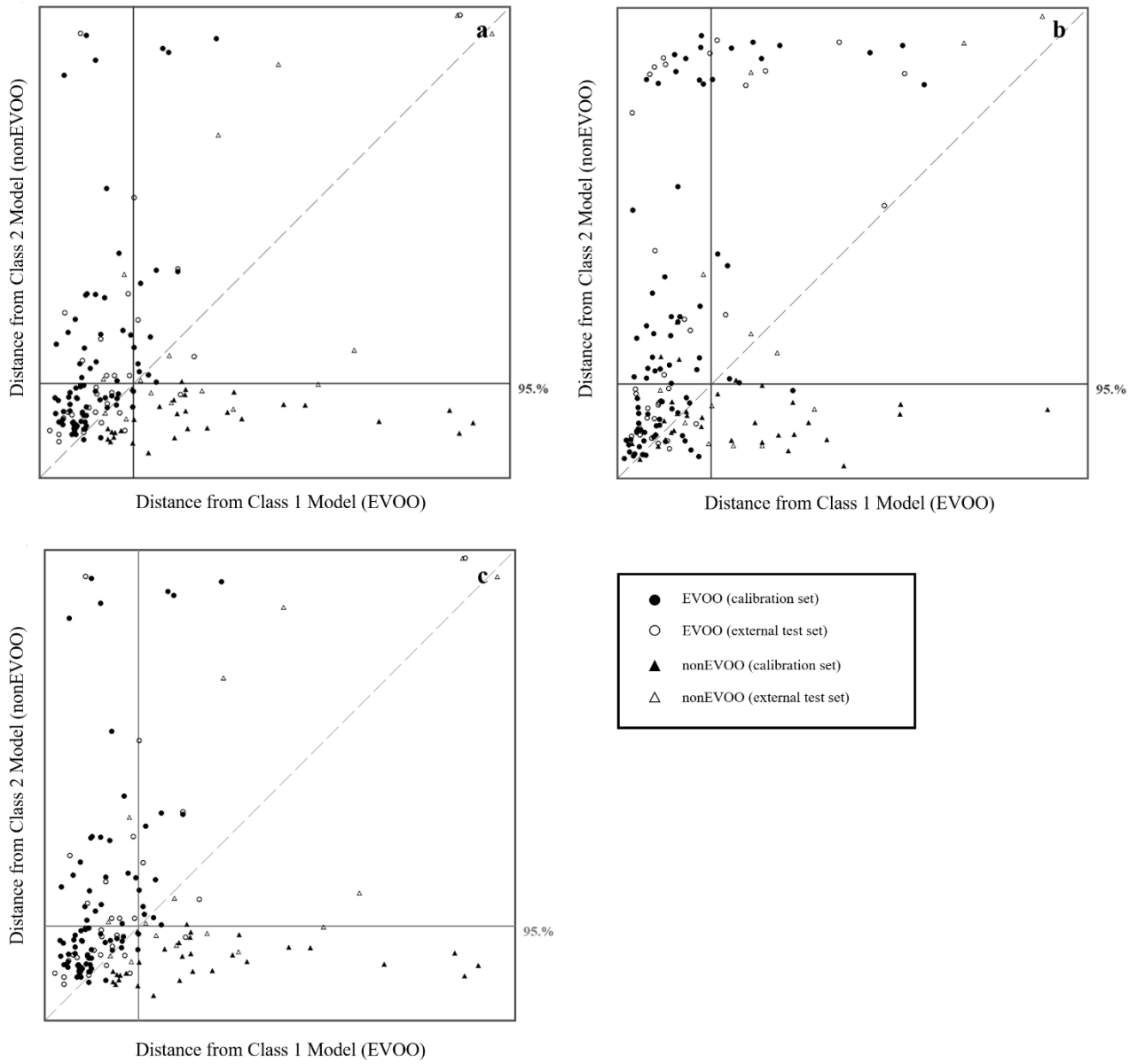


Figure 4.



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).

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.