



# Chemometric discrimination of Turkish olive oils by variety and region using PCA and comparison of classification viability of SIMCA and PLS-DA

Onur Özdikicierler<sup>1</sup>

Received: 27 May 2020 / Revised: 9 September 2020 / Accepted: 13 September 2020 / Published online: 21 September 2020  
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

## Abstract

Virgin olive oil samples of eight varieties from four regions (North Aegean, South Aegean, Mediterranean, and South-eastern) of Turkey were discriminated using fatty acid and sterol composition. Principle component analysis represented a separation of South Aegean olive oils from the rest of the sample groups, that mainly depend on Stigmasterol,  $\beta$ -sitosterol,  $\Delta 5,24$ -Stigmastadienol,  $\Delta 7$ -Avenasterol, C17:0, and C17:1 variables. Except few overlaps, North Aegean samples were also discriminated with Mediterranean and Southeastern samples. The varietal separation was not interpretable by itself but since all samples from South Aegean region were Memecik variety, regional separation has become clearer. Soft independent modeling of class analogy shows good separation between North and South Aegean samples with only a few exceptions. The number of misestimated samples was high at Mediterranean and Southeastern models on Coomans' plots because of high variance within each group. Partial least squares discrimination analysis was more successful than Soft independent modeling of class analogy. The prediction capabilities of South Aegean and North Aegean models were better than others. Root mean squared error of prediction and goodness of prediction were 0.092 and 0.961 for South Aegean, 0.182, and 0.853 for North Aegean, respectively. Unlikely to soft independent modeling of class analogy, Southeastern and Mediterranean samples were not rejected but remained as "uncertain" on partial least squares discrimination analysis with the help of its algorithm.

**Keywords** Olive oil · Sterol composition · Fatty acid composition · PCA · SIMCA · PLS-DA

## Introduction

Olive oil has been an emblematic product for the producing countries and has gained an increasing worldwide reputation due to its sensorial and nutritional properties last few years. Because the growth of olive fruit is highly climate-dependent, most of the olive and olive oil production in the world is carried out by countries that have a coast to the Mediterranean Sea. Over the past few decades, public awareness of olive oil authenticity has increased because of its vulnerability to economically motivated adulteration and/or mislabeling which has great importance not only for olive oil producers, retailers, and consumers but also for lawmakers

and regulators. To meet the demands and the quality expectations of consumers, producers focus on highlighting the authentic properties of the product such as variety and geographical origin of olive oil [1]. As the value of the olive oils with specific features increased, olive oil become a target for fraudsters to gain unfair profit [2].

Verification of the geographic origin of an olive oil is a way to protect consumers and producers from economical and health frauds. Registration procedure of geographic indication of olive oil requires a good determination of targeted geographical area, the specialty of the olive oil such as determination of variety and production methodology, of which environmental factors directly affect the chemical composition of olive oil [3]. Today, a vast majority of the geographic indicated olive oils registered with the eAmbrosia system of the European Commission are monovarietal or composition of two varieties with one dominant in volume [4]. The price of these products is generally higher than extra virgin olive oils with no specific characteristics, hence fraudulent

✉ Onur Özdikicierler  
onur.ozdikicierler@ege.edu.tr

<sup>1</sup> Food Engineering Department, Faculty of Engineering, Ege University, Bornova, İzmir, Turkey

activities regarding a false claim of the geographical origin of the product on the label is spreading and has not been fully avoided yet [3, 5]. Beyond financial losses, such frauds cause a threat to human health, such as allergic reactions, due to incorrectly or non-declared substances and processes [6].

In general, the possible environmental factors that cause chemical and sensorial diversity among olive oils may be summarized as olive variety, the region where the olive tree grows, climatic conditions, cultivation, harvest, and processing techniques. Even olive oils produced from the same variety but in different regions generally differ from each other in their chemical composition [7–11]. Although several chromatographic techniques have been used to reveal the so-called effects of these environmental factors, interpretation, and integration of data samples, when dealing with such a complex matrix like olive oil, are not straightforward tasks. A precise evaluation requires a favorable tool to estimate the geographical area and variety of olive oil not only during the submission of new geographic indications but also for inspection procedures. In this context, chemometrics can be utilized to reveal correlations between the contents of chemical components and numerous variables such as agronomic, technological, and environmental factors that may affect the chemical composition of olive oil [12].

In many recent studies, discriminative chemometric methods were used for evaluation and interpretation of the possible classifications among different olive oils [1, 3, 13–19], detection of adulteration or method development for measurement of a specific constituent of olive oil [20]. The chemometric techniques used in these studies were generally; principle component analysis (PCA), soft independent modeling of class analogy (SIMCA), partial least squares discriminant analysis (PLS-DA), linear discrimination analysis (LDA), ward method, hierarchical cluster analysis (HCA). In particular, unsupervised chemometric techniques such as PCA, is used as an explanatory approach to oversee the discriminative predisposition of the data. In SIMCA, a supervised classification can be maintained by developing PCA models for each class using a training sample set and as new observations are projected into each PC model, residual distances are calculated. A sample is assigned to the model class if its residual distance is lower than the statistical limit. PLS-DA uses models developed by classical partial least squares regression where the response variable is a dummy variable level to express the class membership of the samples [21, 22].

Due to the increasing regional interest of the Turkish olive oils, the classification and characterization using above mentioned chemometric techniques has been discussed in several studies. Fatty acid composition and spectral data of olive oils were used for discrimination according to variety and harvest year [19], trace element contents for variety and

region in another study [23], triglyceride composition for region and variety [24] and fatty acid composition for variety [25]. However, the regional and varietal classification of Turkish olive oils covering all olive producing regions in Turkey using sterol and fatty acid compositions combined were not studied in a previous study.

Our study aims to extend the efforts on the discrimination and classification of olive oils according to their variety and region. The possible groupings within Turkish virgin olive oil (VOO) samples were evaluated using PCA as an explanatory approach and modeled with respect to their region prior to classifying using SIMCA. As a second approach, PLS-DA was used as another supervised classification method. The classification results of SIMCA and PLS-DA were comparatively discussed.

## Materials and methods

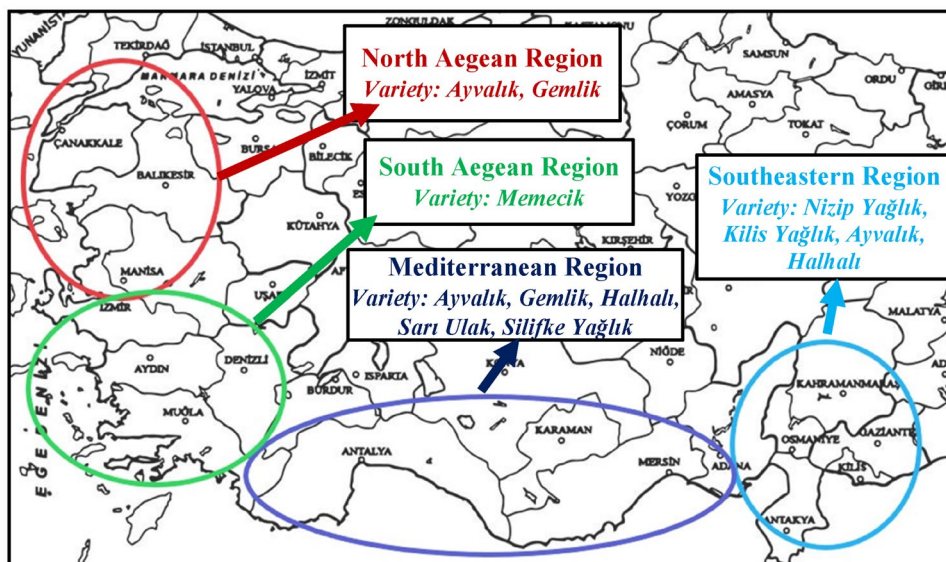
VOO samples were kindly donated by the Tariş Olive and Olive Oil Agriculture Sales Cooperatives Union refinery in Çiğli (İzmir, Turkey) and samples were stored at  $-20^{\circ}\text{C}$  temperature until analysis. The study was started with a total 97 VOO samples produced at different regions in Turkey in 2013 (63 samples) and 2014 (34 samples) harvest seasons. Major olive oil producing areas in Turkey were divided into 4 sub-regions as North Aegean (30 samples), South Aegean (32 samples), Mediterranean (20 samples), and Southeastern (15 samples). Regions were selected according to varying characteristics in geographical and climatic conditions in Turkey (Fig. 1). Selection of smaller regions in our study was avoided due to the information reported by Kritiotti et al. and Karabagias et al., which clarifies it would be an unrealistic aim using chromatographic data for discrimination of VOO samples from smaller areas [26, 27].

All virgin olive oil samples were monovarietal obtained from Ayvalık (44 samples), Gemlik (8 samples), Halhalı (4 samples), Kilis Yağlık (3 samples), Memecik (33 samples), Nizip Yağlık (2 samples), Sarı Ulak (2 samples) and Silifke Yağlık (one sample) varieties. Since VOO's were commercial samples, they possibly had some variations resulting from uncontrolled weather conditions and different cultivation practices. The free fatty acidity (FFA) analysis revealed that FFA levels of 17 out of 97 VOO samples were between 0.8 and 2%. The FFA content of the rest of the samples were below 0.8%. The total size of the data matrix was  $27 \times 97$  (variables  $\times$  samples) in the beginning of the study.

## Chemical analyses

To determine the fatty acid composition, the methyl-esters were prepared by vigorous shaking of a oil solution in *n*-heptane (0.5 ml in 7 ml) with 1 ml of 2 N methanolic

**Fig. 1** Regions of VOO samples together with associated varieties for each region



potash and analyzed with an Agilent (7820) gas chromatography equipped with a flame ionization detector. A Supelco Silica capillary column (60 m length  $\times$  0.25–0.32 mm i.d.) coated with cynopropylsilicone phase (0.1–0.3  $\mu$ m thickness) was used for analysis. Supelco 37 Component FAME Mix (CRM47885) was used for the identification of fatty acids. Hydrogen was used as carrier gas with a flow rate through the column at 1 ml/min. The temperature of the oven, injector, and detector was set at 165  $^{\circ}$ C, 250  $^{\circ}$ C and 270  $^{\circ}$ C, respectively. An injection volume of 1  $\mu$ l was used (COI/T20/Doc.17).

Sterol composition (%) was determined by the same gas chromatography equipment with fatty acid composition with a capillary column (20–30 m length  $\times$  0.25–0.32 mm i.d.) coated with SE-52 (0.30  $\mu$ m thickness). A solution of 0.2%  $\alpha$ -cholestanol (m/V) in chloroform was used as an internal standard for the identification of sterols. Working conditions were as follows: carrier gas, hydrogen; flow through the column, 1.3 ml/min; injector temperature, 280  $^{\circ}$ C; detector temperature, 290  $^{\circ}$ C; oven temperature, 260  $^{\circ}$ C; injection volume 1  $\mu$ l (COI/T20/Doc 10).

### Statistical analysis and chemometric methods

All statistical analyses were conducted in Unscrambler X v10.4 (Camo Analytics, Oslo, Norway). As well as the fatty acid composition and sterol composition, apparent  $\beta$ -sitosterol (sum of the  $\beta$ -sitosterol, sitostanol,  $\Delta$ 5,24-stigmastadienol, clerosterol, and  $\Delta$ 5-avenasterol) and total sterol content were also imported into a data matrix. Region, and variety were defined as categorical variables. Prior to PCA, mean centering was applied to data as preprocessing. Data were mean-centered and randomized, cross-validation method was used for internal validation. All values were

weighted as reciprocal of the standard deviation to eliminate the size difference effect on the results. NIPALS was selected as the PCA algorithm. This procedure was applied to the whole dataset of VOO's for unbiased explanatory discrimination with respect to region and variety as well as regional sub-groups of data individually to obtain PCA models prior to SIMCA. The first two principal component (PC) were visualized on two-dimensional biplots to determine the discrimination on the score plots. Hotelling's  $T^2$  ellipses with a 95% confidence limit used to reveal potential outliers.

Since regional classification in SIMCA and PLS-DA of the VOO samples was one of the aims of the study, the dataset was randomly divided into two groups as calibration and classification sets, after exclusion of detected outliers with extreme leverage according to PCA. Calibration samples were used to develop models and classification samples were used as testing samples. Consequently, 15 samples from North Aegean, 16 samples from South Aegean, 10 samples from the Mediterranean, and 7 samples from Southeastern regions were selected randomly as calibration samples. The rest were grouped as classification samples. A standard procedure for PLS-DA algorithm was used for decisions and arbitrary thresholds [1]. Application details of SIMCA and PLS-DA were explained in detail below.

### Results and discussion

Sharing the exact fatty acid composition and sterol composition of the samples in the text was avoided in the frame of confidentiality. Alternatively, summary tables with descriptive statistics of the general fatty acid and sterol composition with respect to region and variety of VOO's were presented.

This descriptive documentation of the general fatty acid and sterol profiles according to the region and variety of VOO samples is valuable to understand the chemical characteristics of the samples (Tables 1, 2).

Mean values with standard deviations of the sterol composition, apparent  $\beta$ -sitosterol, total sterol content, and fatty acid composition were presented for four different regions (Table 1) and eight different VOO varieties (Table 2) separately. The mean value of  $\beta$ -sitosterol, the abundant sterol of VOO's, fluctuated around 79.78% and 87.96%. Total sterol content showed a significant difference not only between regions but also between VOO varieties. The highest mean sterol contents within each group were determined as 2083 mg/kg at North Aegean VOO's and 2086 mg/kg for Ayvalik varieties. The mean value of oleic acid (C18:1), major fatty acid for olive oils,

fluctuated between 68.25% and 73.09%. It was impossible to make good discrimination among samples only considering the means of individual fatty acids. Five major fatty acids were C18:1, C16:0, C18:2, C18:0, and C16:1 for VOO's. The first PCA was performed to identify uninformative variables and possible outliers that have high leverage (results not presented). For fatty acid composition, C14:0, C22:0 and C22:1, and for sterol composition 24-methylenecholesterol, campastanol and clerosterol does not vary enough to provide an adequate signal for PCA model, which were excluded from PCA dataset later. In addition, one VOO sample of Memecik variety from the Southeastern region harvested in 2014 was determined as an outlier and dropped from the dataset due to high score leverage. Eventually, after the exclusion of so-called variables and one observation, another PCA, of which

**Table 1** Average values of sterol composition, apparent  $\beta$ -sitosterol, total sterol content and fatty acid composition by means of regions

	Mediterranean olive oil samples	North Aegean olive oil samples	South Aegean olive oil samples	Southeastern olive oil samples
Sterol composition (%)				
Cholesterol	0.11 ± 0.04	0.07 ± 0.03	0.13 ± 0.05	0.15 ± 0.05
24- Methylenecholesterol	0.07 ± 0.04	0.07 ± 0.02	0.08 ± 0.02	0.08 ± 0.03
Campesterol	3.08 ± 0.59	2.92 ± 0.18	3.09 ± 0.2	3.39 ± 0.41
Campastanol	0.06 ± 0.02	0.06 ± 0.08	0.06 ± 0.03	0.05 ± 0.02
Stigmasterol	0.62 ± 0.2	0.35 ± 0.15	1.38 ± 0.36	0.85 ± 0.16
Clerosterol	1.07 ± 0.09	0.96 ± 0.17	1.13 ± 0.14	1.1 ± 0.1
$\beta$ -sitosterol	83.54 ± 2.79	81.33 ± 1.75	87.95 ± 1.41	85.41 ± 1.71
Sitositanol	0.55 ± 0.31	0.47 ± 0.18	0.42 ± 0.12	0.78 ± 0.32
$\Delta$ 5-Avenasterol	8.37 ± 3.07	10.49 ± 1.55	4.57 ± 1.45	6.27 ± 1.92
$\Delta$ 5,24-Stigmastadienol	0.96 ± 0.31	1.58 ± 0.24	0.43 ± 0.06	0.7 ± 0.35
$\Delta$ 7-Stigmastenol	0.53 ± 0.16	0.47 ± 0.07	0.36 ± 0.06	0.46 ± 0.09
$\Delta$ 7-Avenasterol	1.01 ± 0.2	1.2 ± 0.2	0.4 ± 0.11	0.76 ± 0.15
Apparent $\beta$ -Sitosterol (%)	94.49 ± 0.5	94.82 ± 0.27	94.51 ± 0.37	94.26 ± 0.36
Total sterol (mg/kg)	1853 ± 297	2083 ± 257	1196 ± 168	1693 ± 473
Fatty acid composition (%) <sup>a</sup>				
C14:0	0.02 ± 0	0.02 ± 0	0.02 ± 0	0.02 ± 0.01
C16:0	14.17 ± 0.57	13.47 ± 0.5	12.57 ± 0.64	14.98 ± 0.76
C16:1	1.1 ± 0.19	0.86 ± 0.06	0.87 ± 0.1	1.22 ± 0.19
C17:0	0.16 ± 0.03	0.15 ± 0.01	0.05 ± 0.02	0.15 ± 0.02
C17:1	0.26 ± 0.04	0.25 ± 0.01	0.08 ± 0.03	0.23 ± 0.04
C18:0	2.95 ± 0.38	2.6 ± 0.12	2.63 ± 0.17	3.14 ± 0.64
C18:1	69.67 ± 1.85	70.32 ± 0.75	73.09 ± 1.53	68.56 ± 2.1
C18:2	10.08 ± 1.76	10.82 ± 0.6	9.09 ± 0.97	9.99 ± 2.5
C18:3	0.66 ± 0.06	0.56 ± 0.05	0.68 ± 0.1	0.75 ± 0.11
C20:0	0.44 ± 0.05	0.44 ± 0.04	0.45 ± 0.08	0.48 ± 0.07
C20:1	0.28 ± 0.04	0.3 ± 0.02	0.31 ± 0.02	0.28 ± 0.04
C22:0	0.11 ± 0.03	0.12 ± 0.03	0.11 ± 0.03	0.13 ± 0.04
C22:1	0.08 ± 0.07	0.08 ± 0.02	0.07 ± 0.02	0.05 ± 0.02

<sup>a</sup>C16:0 palmitic acid, C16:1 palmitoleic acid, C17:0 margaric acid, C17:1 margoleic acid, C18:0 stearic acid, C18:1 oleic acid, C18:1 vaccenic acid, C18:2 linoleic acid, C18:3 linolenic acid, C20:0 arachidic acid, C20:1 gondoic acid, C22:0 behenic acid, C24:0 lignoceric acid

**Table 2** Average values of sterol composition, apparent  $\beta$ -sitosterol, total sterol content and fatty acid composition by means of VOO variety

	Ayvalık	Memecik	Gemlik	Halhalı	Kilis yağlık	Nizip yağlık	Sarı ulak	Silifke yağlık
Sterol composition (%)								
Cholesterol	0.08 ± 0.03	0.13 ± 0.05	0.11 ± 0.05	0.19 ± 0.06	0.16 ± 0.02	0.15 ± 0.04	0.15 ± 0.02	0.16 ± 0.01
24-Methylenecholesterol	0.07 ± 0.03	0.07 ± 0.02	0.08 ± 0.04	0.09 ± 0.03	0.08 ± 0.02	0.09 ± 0.08	0.09 ± 0.06	0.06 ± 0.01
Campesterol	3.09 ± 0.31	3.1 ± 0.21	3.11 ± 0.56	3.14 ± 0.53	3.11 ± 1.05	2.14 ± 0.25	3.19 ± 0.75	2.83 ± 0.01
Campastanol	0.06 ± 0.06	0.06 ± 0.03	0.05 ± 0.02	0.08 ± 0.02	0.07 ± 0.01	0.07 ± 0.06	0.05 ± 0.01	0.08 ± 0.00
Stigmasterol	0.43 ± 0.2	1.36 ± 0.38	0.69 ± 0.21	0.83 ± 0.07	0.75 ± 0.16	1.05 ± 0.06	1.07 ± 0.18	0.56 ± 0.01
Clerosterol	1 ± 0.16	1.14 ± 0.14	1.09 ± 0.13	1.06 ± 0.08	1.03 ± 0.04	1.08 ± 0.06	1.13 ± 0.11	1.07 ± 0.02
$\beta$ -sitosterol	82.53 ± 2.38	87.96 ± 1.39	84.03 ± 1.64	85.46 ± 2.5	81.64 ± 4.82	79.78 ± 4.41	84.64 ± 0.51	83.21 ± 0.21
Sitositanol	0.47 ± 0.18	0.43 ± 0.12	0.53 ± 0.32	0.9 ± 0.32	1.11 ± 0.35	0.52 ± 0.22	1.03 ± 0.5	0.81 ± 0.01
$\Delta$ 5-Avenasterol	9.23 ± 2.39	4.53 ± 1.44	7.93 ± 1.59	6.59 ± 2.53	10.29 ± 5.1	13.02 ± 5.59	6.6 ± 0.3	8.4 ± 0.01
$\Delta$ 5,24-Stigmastadienol	1.44 ± 0.33	0.43 ± 0.06	0.79 ± 0.27	0.5 ± 0.11	0.54 ± 0.21	0.66 ± 0.1	0.63 ± 0.07	0.9 ± 0.01
$\Delta$ 7-Stigmastenol	0.48 ± 0.07	0.37 ± 0.07	0.51 ± 0.19	0.45 ± 0.16	0.38 ± 0.07	0.48 ± 0.33	0.49 ± 0	0.74 ± 0.01
$\Delta$ 7-Avenasterol	1.11 ± 0.24	0.41 ± 0.14	0.99 ± 0.26	0.72 ± 0.15	0.8 ± 0.3	0.98 ± 0.4	0.94 ± 0.04	1.19 ± 0.01
Apparent $\beta$ -Sitosterol (%)	94.66 ± 0.36	94.49 ± 0.37	94.36 ± 0.45	94.51 ± 0.32	94.62 ± 0.88	95.05 ± 0.91	94.02 ± 0.66	94.39 ± 0.56
Total Sterol (mg/kg)	2086 ± 263	1226 ± 236	1660 ± 198	1439 ± 199	1397 ± 191	1535 ± 280	1291 ± 179	2080 ± 121
Fatty acid composition (%)								
C14:0	0.02 ± 0.01	0.02 ± 0	0.02 ± 0	0.02 ± 0.01	0.02 ± 0.01	0.01 ± 0	0.02 ± 0.01	0.02 ± 0.00
C16:0	13.86 ± 0.83	12.59 ± 0.64	14.34 ± 0.56	14.43 ± 0.7	14.78 ± 0.56	14.17 ± 0.26	15.4 ± 0.09	12.95 ± 0.04
C16:1	0.96 ± 0.18	0.87 ± 0.1	1.09 ± 0.2	1.01 ± 0.16	1.16 ± 0.16	1.48 ± 0.08	1.31 ± 0.23	0.97 ± 0.01
C17:0	0.15 ± 0.01	0.05 ± 0.02	0.18 ± 0.05	0.17 ± 0.01	0.16 ± 0.01	0.15 ± 0.01	0.17 ± 0.03	0.18 ± 0.01
C17:1	0.25 ± 0.01	0.08 ± 0.04	0.28 ± 0.06	0.21 ± 0.02	0.2 ± 0.02	0.27 ± 0.01	0.32 ± 0.01	0.29 ± 0.01
C18:0	2.63 ± 0.13	2.64 ± 0.18	2.96 ± 0.42	3.69 ± 0.45	3.93 ± 0.5	3.11 ± 0.05	2.82 ± 0.06	3.28 ± 0.02
C18:1	69.96 ± 0.98	72.79 ± 2.27	69.18 ± 2.63	69.31 ± 2.13	70.19 ± 1.87	71.87 ± 0.69	68.25 ± 1.77	68.89 ± 0.21
C18:2	10.63 ± 0.73	9.34 ± 1.72	10.36 ± 2.17	9.49 ± 1.88	7.98 ± 0.79	7.32 ± 0.2	10.05 ± 2.17	11.74 ± 0.11
C18:3	0.6 ± 0.08	0.69 ± 0.12	0.69 ± 0.07	0.7 ± 0.09	0.66 ± 0.05	0.68 ± 0.01	0.73 ± 0.04	0.76 ± 0.01
C20:0	0.44 ± 0.04	0.45 ± 0.08	0.44 ± 0.05	0.5 ± 0.08	0.51 ± 0.16	0.42 ± 0.02	0.48 ± 0.1	0.44 ± 0.01
C20:1	0.3 ± 0.02	0.31 ± 0.02	0.29 ± 0.03	0.29 ± 0.07	0.23 ± 0.05	0.24 ± 0.03	0.29 ± 0.04	0.28 ± 0.01
C22:0	0.12 ± 0.03	0.11 ± 0.03	0.13 ± 0.02	0.13 ± 0.04	0.12 ± 0.03	0.1 ± 0.02	0.13 ± 0.01	0.1 ± 0.01
C22:1	0.07 ± 0.02	0.07 ± 0.02	0.05 ± 0.03	0.06 ± 0.03	0.06 ± 0.01	0.21 ± 0.18	0.06 ± 0	0.1 ± 0.01

score plots (Fig. 2a, b) and loading plot (Fig. 2c) were presented, was carried out with the same settings to the whole remaining data matrix.

The first three PC's explained 38.8%, 17.9%, and 15.4% the total sample variance, respectively. Consequently 72.1% as the cumulative variance explained. When the sample groupings were labeled for different regions, it was determined that South Aegean VOO's were clearly discriminated from the rest of the samples mainly on the axis of PC1. Southeastern, Mediterranean, and North Aegean VOO samples were generally had negative loadings on the PC1 but showed discrimination from each other on the PC2 axis with slight overlaps. PC3 which contributes the lowest variance was ineffective for visualizing discrimination among sample groups by regions. No sample was located outside of the Hotelling  $T^2$  ellipses on the score plot between PC1 and PC2. Discrimination between Mediterranean and Southeastern regions was compelling since these were close to each other geographically.

In relation to that, Kritiotti et al. reported that fatty acid composition alone did not provide adequate information when coupled with hierarchical cluster analysis to classify VOO's obtained from southern region of Cyprus since the total area was small. Researcher indicated that, the climatic diversity was not large enough to make a significant impact on composition of fatty acid composition [26]. Similarly, Karabagias et al. sampled VOO's from 4 different Greek islands on Ionian Sea and used chromatographic data to develop canonical discriminant functions to discriminate VOO's according to their geographical origins. Only VOO's from Korfu island was discriminated from Lefkada, Kefalonya and Zakynthos islands' VOO's, since these three islands were located closer to each other while Korfu was located more distant from these three islands [27]. These findings support the thesis put forward by Kritati et al. that it is difficult to achieve regional discrimination for small sampling areas using the chromatographic data alone with chemometric techniques.



**Table 3** Absolute values of variable correlation loadings

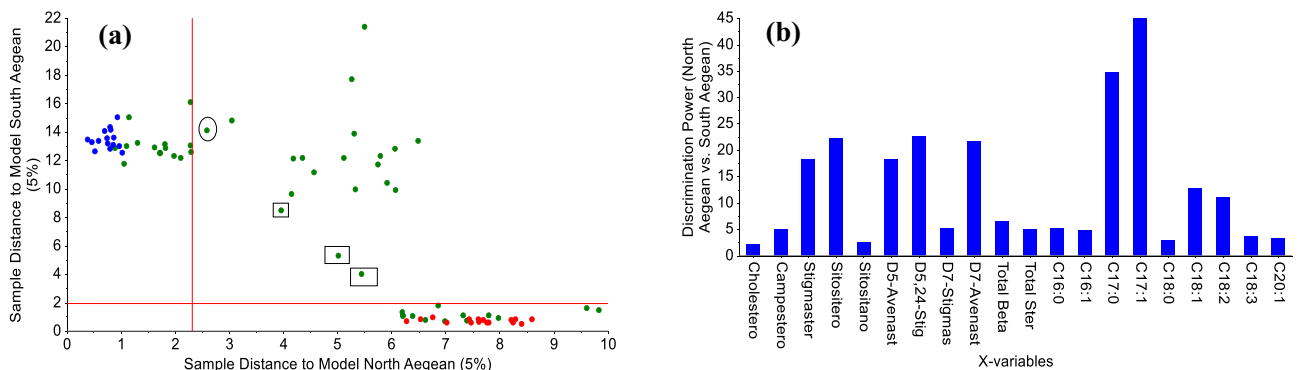
Variables	PC1	PC2	PC3 <sup>a</sup>	Variables	PC1	PC2	PC3 <sup>a</sup>
Cholesterol	0.396	0.336	0.520	C16:0	0.533	0.686	0.060
Campasterol	0.209	0.424	0.563	C16:1	0.223	0.660	0.296
Stigmasterol	0.862	0.188	0.030	C17:0	0.912	0.261	0.064
$\beta$ -sitositerol	0.830	0.318	0.321	C17:1	0.931	0.223	0.006
Sitositanol	0.180	0.422	0.590	C18:0	0.118	0.496	0.550
$\Delta$ 5-Avenasterol	0.773	0.398	0.381	C18:1	0.686	0.573	0.311
$\Delta$ 5,24-Stigmastadienol	0.801	0.391	0.190	C18:2	0.523	0.109	0.649
$\Delta$ 7-Stigmastenol	0.584	0.300	0.207	C18:3	0.335	0.690	0.185
$\Delta$ 7-Avenasterol	0.937	0.121	0.031	C20:0	0.259	0.266	0.687
Apparent $\beta$ -Sitosterol	0.225	0.635	0.473				
Total Sterol	0.790	0.039	0.344				

<sup>a</sup>Not plotted in the text

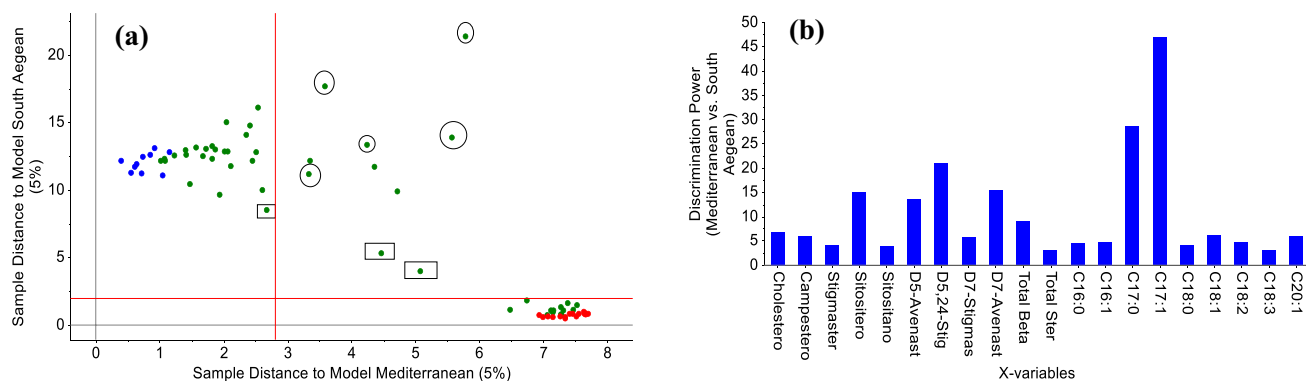
of VOO in the samples made it more difficult to achieve an interpretable classification result.

SIMCA was one of the approaches for VOO classification in regions. Prior to SIMCA, the sample set was divided into calibration and classification datasets as described in methods. PCA was applied to all VOO sample groups for each region separately. Data were mean-centered, randomized, cross-validated, and values were weighted as reciprocal of standard deviation, and NIPALS was selected as PCA algorithm. After saving all PCA models, SIMCA was applied to the classification dataset. To evaluate the results of SIMCA classification, the Coomans' plot, is a useful tool for determination of sample groupings on two axes that represents the distance of every validation sample to the individual PCA model. The critical levels were indicated as straight red lines displaying the boundaries for significant model similarity. When a sample has a position to a range greater than the critical distance, it is determined at outside of the class model, therefore rejected as an outlier for the specific group [19]. If a sample rejected by its correct sample group model, mentioned as "false negative" in this article.

Likewise, if a sample gets accepted by a model of another sample group incorrectly it was mentioned as "false positive". SIMCA did not give a good classification result for our VOO dataset especially for the Mediterranean, North Aegean and Southeastern samples that were classified with many false positives and negatives. Hence, Coomans' plots of these not given. Two different Coomans' plots representing the separation of South Aegean VOO's from North Aegean and Mediterranean VOO's was shown in Figs. 3 and 4. The red and blue dots indicate calibration samples for each PCA model. Green dots were classification samples and samples in circles are false negatives for the classification model given on the *x*-axis, while samples in rectangles are false negatives for the classification model given on *y*-axis for both Coomans' plot. South Aegean VOO's were separated from North Aegean and Mediterranean VOO's with few false negatives according to the SIMCA. Main responsible variables were C17:0, C17:1, Stigmasterol,  $\beta$ -sitositerol,  $\Delta$ 5-Avenasterol,  $\Delta$ 7-Avenasterol and  $\Delta$ 5,24-Stigmastadienol for discrimination between North Aegean and South Aegean VOO's. The big variance within calibration samples of



**Fig. 3** **a** Coomans' plot between North Aegean and South Aegean VOO models, red lines indicate 5% significance level for each model. **b** Sample discrimination power of *x* variables between North Aegean and South Aegean VOO's



**Fig. 4** **a** Coomans' plot between Mediterranean and South Aegean VOO models, red lines indicate 5% critical level for each model. **b** Sample discrimination power of  $x$  variables between the Mediter-

nean and South Aegean VOO's. *RMSEP* root mean square error of prediction,  $Q^2$  goodness of prediction

the same region causes an increase in the critical level for acceptance. The variance was so high in Southeastern and Mediterranean VOO models that even calibration samples fall into the wrong region in Coomans' plots of SIMCA (plot not given). The same situation was observed between North Aegean and Mediterranean models. On the Coomans' plot of the Mediterranean versus North Aegean models, all Mediterranean VOO's were rejected by the North Aegean model. However, 13 out of 15 North Aegean VOO calibration samples were false positive for Mediterranean model. According to the Coomans' plot of the Mediterranean versus Southeastern models, 2 out of 7 Southeastern calibration samples were false positive for the Mediterranean model, while all Mediterranean calibration set was accepted by Southeastern model. Therefore, there were numerous false positives for Southeastern and Mediterranean models as seen in the classification results of SIMCA (Table 4). The results showed that the weakest model in SIMCA was Southeastern model according to the classification table and Coomans'

plots. On the other hand, Mediterranean model was found to be weakest in PLS-DA when the *RMSEP* and  $Q^2$  values were compared. The performance of the SIMCA classification according to VOO regions was poorer in our study than that of samples were classified according to variety using fatty acid composition and mid-IR data which was studied and reported previously [19]. In another study using FTIR spectral data to classify olive oils according to their variety and geographical origin, it was reported that LDA classified more than 98% of the olive oil samples according to variety and with 96% of them according to origin [30]. These previous findings showed that spectroscopic data may provide more information on VOO's for classification purposes using chemometric techniques. The results of SIMCA was given together with PLS-DA results to represent a clear comparison between methods in Table 4.

PLS regression was applied to all calibration VOO samples of different regions separately prior to discrimination analysis. To build the models with a sufficient number of representative

**Table 4** VOO classification results of SIMCA and PLS-DA

Mediterranean ( $n = 10$ )		North Aegean ( $n = 15$ )		South Aegean ( $n = 16$ )		Southeastern ( $n = 7$ )	
<b>SIMCA</b>							
Acc: 4	False positives	Acc: 12	False positives	Acc: 14	False positives	Acc: 7	False positives
Unc: n/a	Med: –	Unc: n/a	Med: 1	Unc: n/a	Med: –	Unc: n/a	Med: 8
Rej: 6	NAE: 15	Rej: 3	NAE: –	Rej: 2	NAE: –	Rej: –	NAE: 16
	SAE: –		SAE: –		SAE: –		SAE: 6
	South: 2		South: –		South: –		South: –
<b>PLS-DA</b>							
Acc: –	False positives	Acc: 9	False positives	Acc: 14	False positives	Acc: –	False positives
Unc: 7	Med: –	Unc: 6	Med: 2	Unc: 2	Med: –	Unc: 7	Med: 1
Rej: 3	NAE: –	Rej: –	NAE: –	Rej: –	NAE: –	Rej: –	NAE: –
	SAE: –		SAE: –		SAE: –		SAE: –
	South: –		South: –		South: –		South: –

*Acc* accepted, *Unc* uncertain, *Rej* rejected, *n/a* not applicable, *Med*, *NAE*, *SAE* and *South* are calibration samples of Mediterranean, North Aegean, South Aegean, and Southeastern regions, respectively



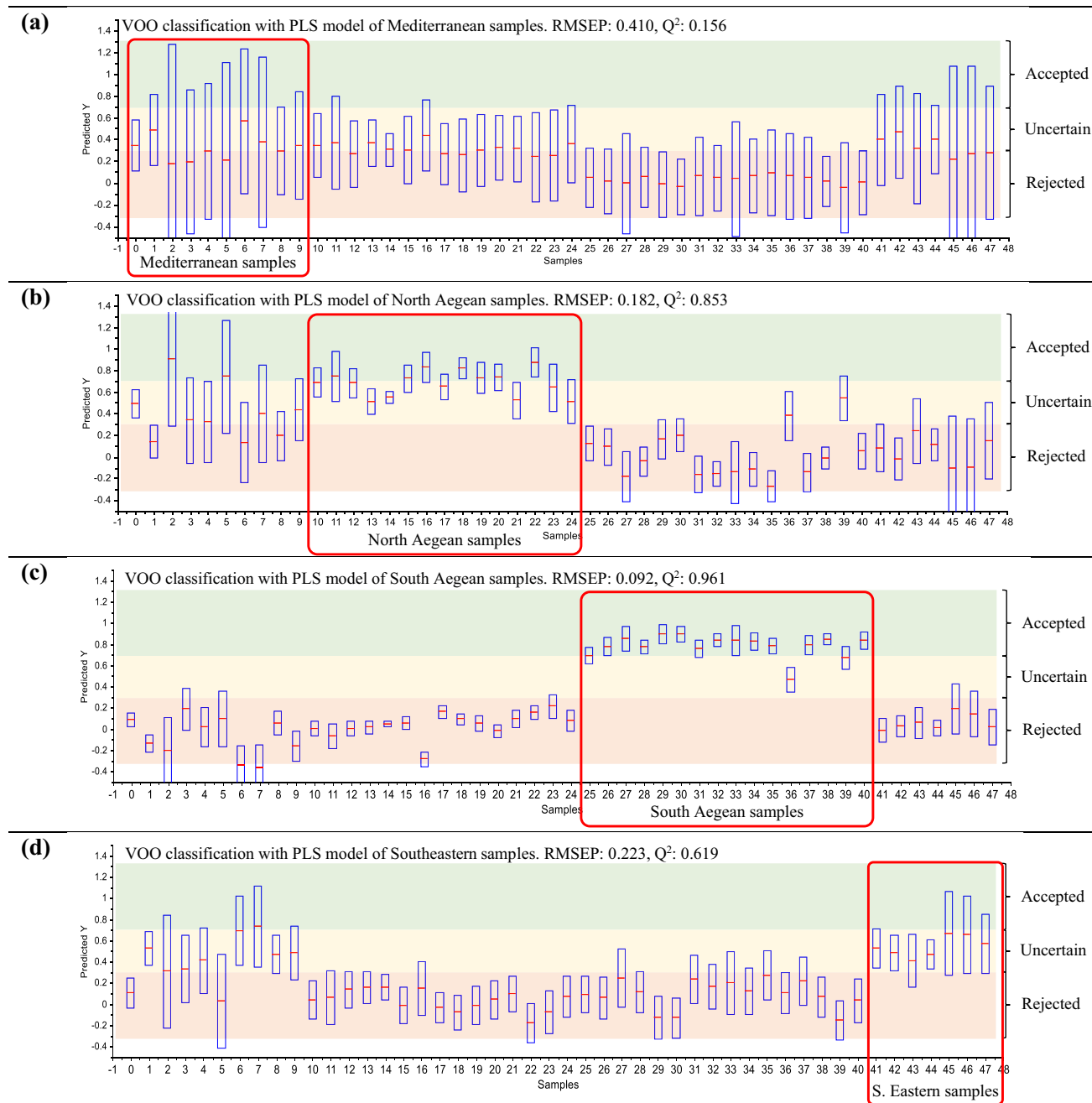
samples from each variety, approximately half of the samples were randomly selected and used as a calibration set to train the PLS models likewise SIMCA. Every PLS regression model was built to predict particular variety against all calibration sets with a binary coding indicating if a sample belongs to modeled variety to give a value of 1, else to give a value of 0. However, depending on the degree of residuals and regression coefficients, the calibration resulted in a  $Y$ -score between 0 and 1 in general. Therefore, a classical PLS-DA arbitrary threshold with uncertainty zone was used as a limit to accept or reject classification samples properly. The intervals suggested in a previous study were used with conservative thresholds and samples were determined as accepted to the modeled variety if their predicted  $Y$ -score was between 0.7 and 1.3. Likewise, the sample rejected by the modeled variety if they positioned within the band of  $-0.3$  and  $0.3$  [1]. Samples, of which predicted  $Y$ -score falls outside of these acceptance and rejection zones, could not be clearly classified to modeled variety and classified as “uncertain samples”. It was useful to interpret the PLS-DA  $Y$ -scores of classification samples together with their standard deviations because as the deviation gets smaller PLS model gets more robust which is significant a successful classification. RMSEP is an indicator of the reliability and predictive ability of the model while  $Q^2$  indicated the predicted relevance of the model where low levels of RMSEP and high values of  $Q^2$  (close to 1) mean the prediction-classification capability of the model is good [31]. In our study, all PLS models were developed with three components since models had the highest  $Q^2$  values at first three components. At first glance, South Aegean and North Aegean models were more robust than others by comparing the RMSEP and  $Q^2$  statistics which were 0.092 and 0.961 for the South Aegean model, 0.182 and 0.853 for North Aegean model, respectively. Mediterranean model was weaker than the Southeastern model, unlikely to the outcomes of SIMCA, according to their RMSEP and  $Q^2$  statistics which were 0.410, 0.156 for Mediterranean model and 0.223, 0.619 for Southeastern model, respectively. As expected, the best classification results with PLS-DA were achieved for South Aegean samples, where 14 samples out of 16 were identified, because of the low deviation within the group. Because of the availability of the discrimination rule that we set to consider the uncertainty, PLS-DA did not completely reject remained 2 South Aegean samples that was not accepted, while SIMCA directly rejected those because of its algorithm. Similarly good geographical discrimination of south Aegean Memecik olive oils from the rest of the Aegean olive oils by applying discriminant analysis only on fatty acid composition was reported in a previous study [25]. Alkan et al. indicated that discrimination of South Aegean and North Aegean VOO's was possible when PLS-DA was applied on the phenolic profiles of the samples, especially this differentiation was deepened since the varieties of these two regions were also different [32].

PLS-DA classified 9 out of 15 North Aegean VOO's with 2 false negatives from Mediterranean samples. PLS-DA did not offer satisfactory results for Mediterranean and Southeastern VOO models, as no sample from these regions was accepted by the related models. High deviation in the  $Y$ -scores was observed for Mediterranean VOO's while the Southeastern VOO samples were only able to be dragged into uncertainty zone by its own model without any acceptance. This may be associated with two factors. First, North Aegean samples were mainly consisting of Ayvalık variety while South Aegean samples were completely from Memecik variety which may become a significant subsidiary factor to increase the robustness of South Aegean and North Aegean models. Second, if there is an unbalance in the number of the sample of each group and the number of modeled region's samples are fewer than the sum of samples from the other regions, this causes lower  $Y$ -scores from the modeled cultivar than expected level of 1, which fits Southeastern and Mediterranean VOO results in Fig. 5 [33]. Latter may overcome through setting balanced calibration sets by selecting the same number of samples from the modeled region and other regions for each PLS model [1]. However, while such customization is proposed on the structure of sample sets for every model in various cases to increase the accuracy of the expected classification, this application is not feasible if chemometric approaches are offered as a determinative tool not only for scientific purposes but also for legal cases, which is a prudential target for such studies.

The classification approach was highly determinant for the results when SIMCA and PLS-DA results compared. SIMCA relies on PCA models and the success of the classification is directly dependent on discrimination of the sample groups on PCA of whole sample dataset. Separate regression models for each calibration set were developed with PLS to give a binary level of response. Although SIMCA was more successful on acceptance counts, especially for Mediterranean, North Aegean, and Southeastern samples, these acceptances come with many false positives when compared with PLS-DA. In PLS-DA, the acceptance numbers were comparably lower but because of the arbitrary threshold that was set for uncertainty, unaccepted samples not directly rejected and a second determination may be applied to these uncertain samples to verify their region more accurately. In addition, when the low number of false positives and lower rejected VOO counts considered, PLS-DA was more successful than SIMCA for the classification of VOO's.

## Conclusion

In this study, Turkish VOO's of eight different varieties from four major producing regions were discriminated and classified using chemometric techniques, namely PCA, SIMCA,



**Fig. 5** Y scores with standard deviations of VOO samples for each region model

and PLS-DA, to reveal possible discrimination using chromatographic data. Initial PCA results were useful for the identification of possible outliers and variable selection for further determination. Secondary PCA results also showed a fair grouping of South Aegean samples away from the rest of the VOO samples, where Stigmasterol,  $\beta$ -sitosterol,  $\Delta 5,24$ -Stigmastadienol,  $\Delta 7$ -Avenasterol, C17:0 and C17:1 found to be mainly responsible for this separation. North Aegean samples were also grouped and separated except few overlaps with Mediterranean and Southeastern VOO samples.

Although, there was a visual separation between Mediterranean and Southeastern VOO samples, the distribution within each group was noteworthy. Groupings or discrimination when samples were labeled according to their varieties was not clear as like the regions. SIMCA was applied after developing PCA models for each region separately. While there was good separation between North Aegean and South Aegean models with only a few false negatives, many false positives were accepted by the Mediterranean model on Coomans' plot because of high disturbances. Even

calibration samples could not be separated within the significance level when Southeastern models were involved in Coomans' plots. PLS-DA classification was offered comparably successful results than SIMCA. However, for Southeastern and Mediterranean models, their own samples were not rejected but remained as "uncertain" because of the evaluation method, which gives a second chance to false-negative samples. There was no surprise when the results of all chemometrics considered, generally, the degree of discrimination followed a similar pattern through the study. Although today chromatographic data offer main information for the characterization of VOO's, the extent of the computational power of chemometrics may handle bigger datasets consists of more variables. Thus, new spectroscopic techniques may offer more informative data after a proper preprocessing and selection method to develop robust models with higher classification capability.

**Acknowledgements** Thanks to the Tariş Olive and Olive Oil Agriculture Sales Cooperatives Union refinery in Çiğli, İzmir, for providing samples and the kind contribution in chemical analyses.

**Funding** Not applicable.

## Compliance with ethical standards

**Conflict of interest** There is no conflict of interest.

**Availability of data and material** Raw data not presented.

## References

- Maléchaux A, Le Dréau Y, Vanloot P et al (2019) Discrimination of extra virgin olive oils from five French cultivars: en route to a control chart approach. *Food Control* 106:106691. <https://doi.org/10.1016/j.foodcont.2019.06.017>
- Tsimidou MZ, Ordoudi SA, Nenadis N, Mourtzinos I (2015) *Food Fraud*, 1st edn. Elsevier Ltd., USA
- Cecchi L, Migliorini M, Giambanelli E et al (2020) Authentication of the geographical origin of virgin olive oils from the main worldwide producing countries: a new combination of HS-SPME-GC-MS analysis of volatile compounds and chemometrics applied to 1217 samples. *Food Control* 112:107156. <https://doi.org/10.1016/j.foodcont.2020.107156>
- European Commission (2020) eAmbrosia—the EU geographical indications register. <https://ec.europa.eu/info/food-farming-fisheries/food-safety-and-quality/certification/quality-labels/geographical-indications-register/>. Accessed 8 Apr 2020
- Esteki M, Simal-Gandara J, Shahsavari Z et al (2018) A review on the application of chromatographic methods, coupled to chemometrics, for food authentication. *Food Control* 93:165–182. <https://doi.org/10.1016/j.foodcont.2018.06.015>
- Gerbig S, Neese S, Penner A et al (2017) Real-time food authentication using a miniature mass spectrometer. *Anal Chem* 89:10717–10725. <https://doi.org/10.1021/acs.analchem.7b01689>
- Mele MA, Islam MZ, Kang HM, Giuffrè AM (2018) Pre-and post-harvest factors and their impact on oil composition and quality of olive fruit. *Emirates J Food Agric* 30:592–603. <https://doi.org/10.9755/ejfa.2018.v30.i7.1742>
- Kalua CM, Allen MS, Bedgood DR et al (2007) Olive oil volatile compounds, flavour development and quality: a critical review. *Food Chem* 100:273–286. <https://doi.org/10.1016/j.foodchem.2005.09.059>
- Abdallah M, Vergara-Barberán M, Lerma-García MJ et al (2016) Classification of Tunisian extra virgin olive oils according to their genetic variety and maturity index using fatty acid profiles established by direct infusion mass spectrometry. *Eur J Lipid Sci Technol* 118:735–743. <https://doi.org/10.1002/ejlt.201500230>
- Bozdogan Konuskan D, Mungan B (2016) Effects of variety, maturation and growing region on chemical properties, fatty acid and sterol compositions of virgin olive oils. *J Am Oil Chem Soc* 93:1499–1508. <https://doi.org/10.1007/s11746-016-2904-8>
- Li X, Flynn JD, Wang SC (2019) The effects of variety, growing region, and drought stress on fatty acid and sterol compositions of California olive oil. *JAOCs J Am Oil Chem Soc* 96:215–230. <https://doi.org/10.1002/aocs.12192>
- Gómez-Caravaca AM, Maggio RM, Cerretani L (2016) Chemometric applications to assess quality and critical parameters of virgin and extra-virgin olive oil. A review. *Anal Chim Acta* 913:1–21. <https://doi.org/10.1016/j.aca.2016.01.025>
- Ün İ, Ok S (2018) Analysis of olive oil for authentication and shelf life determination. *J Food Sci Technol* 55:2476–2487. <https://doi.org/10.1007/s13197-018-3165-3>
- Ghisoni S, Lucini L, Angilletta F et al (2019) Discrimination of extra-virgin-olive oils from different cultivars and geographical origins by untargeted metabolomics. *Food Res Int* 121:746–753. <https://doi.org/10.1016/j.foodres.2018.12.052>
- El Riachy M, Hamade A, Ayoub R et al (2019) Oil content, fatty acid and phenolic profiles of some olive varieties growing in Lebanon. *Front Nutr* 6:94. <https://doi.org/10.3389/fnut.2019.00094>
- Jiménez-Carvelo AM, Cruz CM, Olivieri AC et al (2019) Classification of olive oils according to their cultivars based on second-order data using LC-DAD. *Talanta* 195:69–76. <https://doi.org/10.1016/j.talanta.2018.11.033>
- Crizel RL, Hoffmann JF, Zandoná GP et al (2020) Characterization of extra virgin olive oil from Southern Brazil. *Eur J Lipid Sci Technol* 1900347:11–14. <https://doi.org/10.1002/ejlt.201900347>
- Martakos I, Kostakis M, Dasenaki M et al (2020) Simultaneous determination of pigments, tocopherols, and squalene in Greek olive oils: a study of the influence of cultivation and oil-production parameters. *Foods* 9:31. <https://doi.org/10.3390/foods9010031>
- Gurdeniz G, Ozen B, Tokatli F (2008) Classification of Turkish olive oils with respect to cultivar, geographic origin and harvest year, using fatty acid profile and mid-IR spectroscopy. *Eur Food Res Technol* 227:1275–1281. <https://doi.org/10.1007/s00217-008-0845-7>
- Deiana P, Santona M, Dettori S et al (2019) Multivariate approach to assess the chemical composition of Italian virgin olive oils as a function of variety and harvest period. *Food Chem* 300:125243. <https://doi.org/10.1016/j.foodchem.2019.125243>
- Medina S, Perestrelo R, Silva P et al (2019) Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends Food Sci Technol* 85:163–176. <https://doi.org/10.1016/j.tifs.2019.01.017>
- Lee LC, Liang CY, Jemain AA (2018) Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst* 143:3526–3539. <https://doi.org/10.1039/c8an00599k>
- Gumus ZP, Celenk VU, Tekin S et al (2017) Determination of trace elements and stable carbon isotope ratios in virgin olive oils from Western Turkey to authenticate geographical origin with a

- chemometric approach. *Eur Food Res Technol* 243:1719–1727. <https://doi.org/10.1007/s00217-017-2876-4>
24. Gökçebağ M, Dıraman H, Özdemir D (2013) Classification of Turkish monocultivar (Ayvalık and Memecik cv.) virgin olive oils from north and south zones of Aegean region based on their triacylglycerol profiles. *JAOCS J Am Oil Chem Soc* 90:1661–1671. <https://doi.org/10.1007/s11746-013-2308-y>
  25. Dıraman H, Saygı H, Hışıl Y (2011) Geographical classification of Turkish virgin olive oils from the Aegean Region for two harvest years based on their fatty acid profiles. *JAOCS J Am Oil Chem Soc* 88:1905–1915. <https://doi.org/10.1007/s11746-011-1862-4>
  26. Kritiöti A, Menexes G, Drouza C (2018) Chemometric characterization of virgin olive oils of the two major Cypriot cultivars based on their fatty acid composition. *Food Res Int* 103:426–437. <https://doi.org/10.1016/j.foodres.2017.10.064>
  27. Karabagias I, Michos C, Badeka A et al (2013) Classification of Western Greek virgin olive oils according to geographical origin based on chromatographic, spectroscopic, conventional and chemometric analyses. *Food Res Int* 54:1950–1958. <https://doi.org/10.1016/j.foodres.2013.09.023>
  28. Ben TS, Manai H, Methenni K et al (2008) Sterolic composition of Chétoui virgin olive oil: influence of geographical origin. *Food Chem* 110:368–374. <https://doi.org/10.1016/j.foodchem.2008.02.012>
  29. Guissous M, Le Dréau Y, Boukhroune H et al (2018) Chemometric characterization of eight monovarietal Algerian Virgin Olive Oils. *JAOCS J Am Oil Chem Soc* 95:267–281. <https://doi.org/10.1002/aocs.12030>
  30. Tapp HS, Defernez M, Kemsley EK (2003) FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils. *J Agric Food Chem* 51:6110–6115. <https://doi.org/10.1021/jf030232s>
  31. Hair et.al (2019) *Multivariate data analysis*. 8th edn. Annabel Ainscow
  32. Alkan D, Tokatli F, Ozen B (2012) Phenolic characterization and geographical classification of commercial extra virgin olive oils produced in Turkey. *JAOCS J Am Oil Chem Soc* 89:261–268. <https://doi.org/10.1007/s11746-011-1917-6>
  33. Borràs E, Ferré J, Boqué R et al (2016) Olive oil sensory defects classification with data fusion of instrumental techniques and multivariate analysis (PLS-DA). *Food Chem* 203:314–322. <https://doi.org/10.1016/j.foodchem.2016.02.038>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.