Food Chemistry 256 (2018) 53–61

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Application of an advanced and wide scope non-target screening workflow with LC-ESI-QTOF-MS and chemometrics for the classification of the Greek olive oil varieties

Natasa P. Kalogiouri, Reza Aalizadeh, Nikolaos S. Thomaidis*

Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

ARTICLE INFO

Keywords:
Extra virgin olive oils
Varietal classification
Greece
HRMS
Non-target screening
Chemometrics

ABSTRACT

An optimized and validated LC-ESI-QTOF-MS method with an integrated non-target screening workflow was applied in the investigation of the metabolomic profile of 51 Greek monovarietal extra virgin olive oils (EVOOs) from the varieties: Manaki, Ladoelia, Koroneiki, Amfissis, Chalkidikis and Kolovi. Data processing was carried out with the R language and XCMS package. A local database consisting of 1608 compounds naturally occurring in different organs of *Olea Europa L.* was compiled in order to accelerate the identification workflow. The preliminary examination of the distribution of EVOOs toward their cultivars was achieved by Principal Component Analysis (PCA). Ant Colony Optimization-Random Forest (ACO-RF) was developed to prioritize over 250 features and to establish a classification tree. Apigenin, vanillic acid, luteolin 7-methyl ether and oleocanthal were suggested as the markers responsible for the classification of Greek EVOOs' cultivars.

1. Introduction

The importance of phenolic compounds is related to their antioxidant activity and to their contribution to health benefits associated with extra virgin olive oil (EVOO) consumption (Ghanbari, Anwar, Alkharfy, Gilani, & Saari, 2012). EVOO composition determines its intrinsic quality and could be influenced by several factors, including agronomical and technological factors, such as olive cultivar (Tura et al., 2007), the climate (Baccouri et al., 2008), the degree of maturation (Cerretani et al., 2005), crop season (Alkan, Tokatli, & Ozen, 2011) and the production process (Alkan et al., 2011). However, geographical area is greatly responsible for the specific characteristics of EVOOs (Petrakis, Agiomyrgianaki, Christophoridou, Spyros, & Dais, 2008). Olive cultivars, the geographical region along with environmental factors have been reported as the main parameters affecting the chemical profile of EVOOs dominantly (Bajoub et al., 2016; Ballus et al., 2015).

The olive tree (*Olea Europaea L.*) has diverged naturally to many cultivars and is cultivated mainly in the Mediterranean region; Spain, Italy, Greece, Tunisia, Turkey, Morocco and Algeria (Bakhouche et al., 2013). The cultivar defines the quality of the drupe and the olive oil (Kosma et al., 2016). Greece is among the leading olive producing countries of the world, ranked third after Spain and Italy. The number of Greek cultivars is greater than 40 and more than 90% of the territory is cultivated with 20 cultivars. Olive oil produced in Greece has excellent quality and this is because of the local climatic and soil conditions. According to the International Olive Oil Council (http://www.internationaloliveoil.org), 70% of Greek production is categorized as EVOO while almost the 35% is exported. Thus, it is imperative for Greece to characterize and authenticate EVOOs based on cultivar and geographical origin in an effort to establish a brand name in the international market.

The European Union has adopted a series of regulations providing guidelines to maintain the Protected Designation of Origin (PDO) and Protected Geographical Indication (PGI). These include characterization of olive oils based on cultivar and geographical origin to reassure that the quality of the product is closely linked to its territorial and botanical origin and consequently to increase its market value (Council Regulation (EC) No. 510/2006). This regulation states that there is an economic basis for the identification of markers that distinguish PDO EVOOs. Thus, there is an understanding need to enforce the above regulation and develop analytical methods for the authentication of EVOOs and to reassure that the product is closely linked to its territorial origin.

During the past decade, there have been intensive studies for the determination of the cultivar of EVOOs on the basis of different olive oil constituents, including fatty acids, triacylglycerols, sterols, volatiles and phenolic compounds with different analytical methodologies, such as

* Corresponding author.
E-mail address: ntho@chem.uoa.gr (N.S. Thomaidis).

https://doi.org/10.1016/j.foodchem.2018.02.101
Received 26 July 2017; Received in revised form 16 February 2018; Accepted 19 February 2018
Available online 21 February 2018

Liquid Chromatography-Mass Spectrometry (LC-MS) (Bakhouche et al., 2013) LC coupled to Time of Flight (TOFMS) (Ballus et al., 2015), Gas Chromatography (GC) coupled to Flame Ionization Detector (FID) (Karabagias et al., 2013) and MS (Longobardi et al., 2012; Pouliarekou et al., 2011), Nuclear Magnetic Resonance (NMR) (Petrakis et al., 2008) and HPLC coupled to UV (Allalout et al., 2009). Most of these methods combined chemometrics, as complimentary tool. Recently, Bajoub et al. (2017) made an interesting effort applying k-Nearest Neighborhood (k-NN), Partial Least Square-Discriminant Analysis (PLS-DA) and Soft Independent Modeling of Class Analogies (SIMCA). It was concluded that the application of LC coupled to chemometrics, for data treatment, can define the EVOOs varieties with acceptable accuracy. This study, however, could not address the important markers and their contribution behind the classification models. In this field, the literature survey indicates gaps in information, which should be filled in the near future.

LC coupled to High Resolution Mass Spectrometry (HRMS) followed by non-target screening strategies and chemometrics could fulfill this gap. LC-HRMS has been widely applied to analyze complex mixtures with wide polarities owing to its high separation efficiency and sensitivity in the identification of compounds at low concentration levels. Recently, this method was successfully applied in two authenticity studies of EVOOs concerning the organoleptic profile (Kalogiouri, Alygizakis, Aalizadeh, & Thomaidis, 2016) and the production type (Kalogiouri, Aalizadeh, & Thomaidis, 2017), suggesting markers. One important step in non-target HRMS methods is the prioritization of MS features. In our previous works, two different prioritization tools were introduced to prioritize the MS features and extract the markers; Variable Importance in Projection (VIP) for the suggestion of markers and discrimination between defective olive oils and EVOOs (Kalogiouri et al., 2016) and Ant Colony Optimization-Random Forest (ACO-RF) for the classification of conventional and organic EVOOs. A simple and robust decision tree was established and could define production type and guarantee EVOOs authenticity. The decision tree based approach coupled to High Resolution Mass Spectrometry (HRMS) could open new horizons in authenticity studies in the foodomics field.

The objective of this study was to apply a novel, optimized and validated reversed-phase ultra-high-performance liquid chromatography coupled to electrospray ionization-quadrupole-time-of-flight mass spectrometric (RP-UHPLC-ESI-QTOF-MS) method, using an integrated non-target screening workflow for the investigation of the whole metabolome of EVOOs via non-target screening and the identification of markers in extra virgin monovarietal Greek olive oils. For this purpose, 51 monovarietal Greek EVOOs labelled as Amfissis, Manaki, Kolovi, Koroneiki, Ladoelia and Chalkidikis produced during the harvesting period 2015-2016 were acquired from different regions in Greece. In data processing, peak picking was carried out using the R language and V-WSP algorithm was used as an unsupervised variable reduction method to decrease the number of the features. A newly-introduced algorithm, Ant Colony Optimization (ACO) was applied as a feature selection technique and revealed efficiently only the meaningful masses (*m/z*) that contribute to the classification. Non-target identification workflow was applied, and in order to accelerate the identification task, a local database consisting of 1608 compounds commonly occurring in *Olea Europa L.* organs and olive oil was compiled, including, molecular formulas, monoisotopic mass information, molecular ions in negative and positive ionization, chemical identifiers together with references was compiled from FooDB (FoodB, The Food Components Database; http://foodb.ca/.) to accelerate the identification of the unknown masses. Finally, RF was employed to classify the EVOOs according to their varieties.

2. Experimental Section

2.1. Chemicals and standards

All standards and reagents were of high-purity grade (> 95%). MeOH of LC-MS grade and sodium hydroxide (> 99%) were purchased from Merck (Darmstadt, Germany). Ammonium acetate (≥99.0%) for HPLC and formic acid (LC-MS Ultra) were purchased from Fluka (Buchs, Switzerland). Isopropanol was acquired from Fisher Scientific (Geel, Belgium). Distilled water was provided by a Milli-Q purification apparatus (Millipore Direct-Q UV, Bedford, MA, USA). Syringaldehyde 98% was acquired from Sigma-Aldrich (Stenheim, Germany) and used as an internal standard. Oleuropein 98%, vanillic acid 97% and pinoresinol 95% were purchased from Sigma-Aldrich (Stenheim, Germany) and luteolin 98% was purchased from Santa Cruz Biotecnologies. Apigenin (4,5,7 trihydroxyflavone) 97% and tyrosol (2-(4-hydroxyphenyl)ethanol) 98% were purchased from Alfa Aesar (Karlsruche, Germany). Stock standard solutions of individual compounds (1000 mg L^{-1}) were solubilized in methanol and stored at −20 °C in dark brown glass. All intermediate standard solutions containing the analytes were prepared by dilution of the stock solutions in methanol.

2.2. Sampling

51 monovarietal EVOOs belonging to five different cultivars were collected from local producers from various regions in Greece. Taking into consideration that the harvest period and the production processing affects the phenolic profile of the EVOOs, all EVOOs under study were collected between December and January 2015–2016. All the samples acquired were cultivated with conventional type of farming, and they were processed with three phase centrifugation technique. Grinding mills were used for grinding in all cases, and the malaxation time was between 45 and 60 min. In total, 11 samples of the variety Kolovi from Lesvos Island, 9 samples of Chontrolia Chalikidikis, 5 samples of Amfissis, 17 of Koroneiki (8 samples were acquired from Crete and 9 of Peloponnese) 4 samples of Ladoelia and 5 samples of Manaki. Table S1 in the Supplementary Material summarizes the geographical origin of the samples. All samples were collected and stored in dark glass bottles, protected from light and humidity. Nitrogen was inserted as an inert gas to better preserve olive oils and increase the resistance to autoxidation.

2.3. Sample extraction

Sample preparation was carried out using liquid liquid Micro extraction (LLME) as it has been previously described by our group (Kalogiouri et al., 2017), using MeOH:H_2O (80:20, *v/v*) as the extraction solvent. Finally, 5 μL of the extract was injected into the chromatographic system. Procedural blanks were prepared and processed in the chromatographic system to detect any potential contamination.

2.4. Quality control

Quality control (QC) samples were prepared to confirm that the analytical system has been stabilized before the batch of samples and to evaluate its performance. The quality control sample was prepared by mixing EVOO aliquots and was spiked with a standard solution mix (2 mg L^{-1}) that comprised vanillic acid, oleuropein, luteolin and pinoresinol, so that the final concentration of the QC sample was 1 mg L^{-1}. It was injected at the beginning of the analysis (five times for conditioning), and afterwards, it was injected at regular intervals (every ten sample injections). The %RSDs for the peak areas of the standard compounds were less than 5% (n = 10). The retention time shift was in the range 0.09–0.23% RSD (n = 10) and mass error was less than 0.28 ppm (n = 10), confirming the good performance of the analytical system. The quality control results are summarized in Supplementary Material, Table S2.

2.5. Instrumental analysis

The conditions of the Reversed Phase (RP) chromatographic analysis using a UHPLC system with an HPG-3400 pump (Dionex UltiMate 3000 RSLC, Thermo Fisher Scientific, Germany) using an Acclaim RSLC C18 column (2.1 × 100 mm, 2.2 μm) purchased from Thermo Fisher Scientific (Driesch, Germany) with a pre-column of ACQUITY UPLC BEH C18 (1.7 μm, VanGuard Pre-Column, Waters (Ireland)), as well as the operation of the Q-TOF mass spectrometer (Maxis Impact, Bruker Daltonics, Bremen, Germany) in negative electrospray ionization mode, have been already described in a previous work of our research group (Kalogiouri et al., 2017).

2.6. Method validation

The validation procedure of the RP-UHPLC-ESI-QTOF-MS method has already been described in a previous work of our group (Kalogiouri et al., 2017). Standard addition curve was constructed for the quantification of vanillic acid. The standard compound was spiked in real EVOO samples at concentrations between 0.02 and 10 mg kg^{-1} (10 calibration levels with 3 replicates at each level). The calibration curve was constructed with the use of the peak area of the spiked analyte subtracted by the peak area of a neat sample and divided by the peak area of the internal standard (syringaldehyde 1.30 mg kg^{-1}). Limits of detection (LODs) and limits of quantification (LOQs) were calculated at the lowest concentration range of the analytes (0.02–1 mg kg^{-1}), according to the following equations:

$$LOD = \frac{3.3 \times S_a}{b} \quad (1)$$

$$LOD = \frac{10 \times S_a}{b} \quad (2)$$

where: S_a is the standard error of the intercept (a) and b is the slope of the calibration curve. For vanillic acid, LOD was calculated 0.031 mg kg^{-1} and LOQ 0.095 mg kg^{-1}.

2.7. Non-target screening protocol

All 51 samples analyzed were converted to mzXML files using ProteoWizard. Further on, these files transferred to R environment to perform peak picking. Among the peak picking workflows (Smith, Want, O'Maille, Abagyan, & Siuzdak, 2006; Tengstrand, Lindberg, & Åberg, 2014), XCMS has been widely used in LC-HRMS data processing, owing to high efficiency of *centWave* algorithm, which is proved to have high performance due to its robust and sensitive detection of potential region-of-interesting mass traces (ROIs). Moreover, noise and baseline correction can be estimated locally for each ROIs offering high F-score (combined measure of recall and precision, calculated from the ground truth features). XCMS has three main internal parameters of *ppm* (which is the tolerated mass deviation), minimum and maximum chromatographic peak width, and *snthresh* ratio which defines the chromatographic signal-to-noise threshold. Preferably, prefilter (the threshold for an m/z to be considered as a peak if it appears in k consecutive scan at J intensity threshold (k,I)) can be applied to discard false peaks in detected ROIs. A general peak picking workflow also needs an additional step of retention time correction and alignment (here we used the non-linear retention time alignment wrapping algorithm by loess) as well as peaks group across samples. Filling any missing peaks across samples and the also annotation of extract m/z features are highly needed to prevent adducts/isotopic peaks to cofound with their molecular ions. Here we optimized ppm (23.3), minimum (17.5) and maximum (40) peak width using IPO package in R environment (Libiseller et al., 2015). Signal-to-noise threshold was also set at default value of 10, and prefilter was adjusted at 3-1000. The response surface of these parameters can be found in the Supplementary Material, Fig. S1. Annotations of selected peaks were also done using CAMERA package. A matrix of 51 samples and 287 features (m/z) was generated based on the optimized XCMS object and proceeded to identification and development of classification model. For further identification of these peaks, a new prioritization tool so called Ant Colony Optimization (ACO) was used to limit the searching space of m/z from 287 to least 4 ones.

Afterwards, the non-target screening workflow was applied. This workflow involves the identification of the selected peaks according to mass accuracy (less than 5 ppm) and isotopic pattern of the precursor ion (less than 100 mSigma), their fragmentation pattern, and the retention time of the chromatographic peak. Extracted ion chromatograms (EICs) were obtained and MS/MS spectra were examined and interpreted. “SmartFormula Manually” tool was applied in Data Analysis 4.1 (Bruker Daltonics, Bremen, Germany) to assign plausible molecular formula(s) to the mass of interest and suggest elemental compositions of the precursor and fragment ions. Then, the prepared local database consisting only the metabolites and natural products that commonly occur in olives or olive oil was uploaded in Metfrag (Wolf, Schmidt, Muller-Hannemann, & Neumann, 2010). Then, the exact mass and molecular formula were inserted and the mass error for searching the chemical database was set to 5 ppm. Moreover, the MS/MS fragments with relative intensity were added to elucidate the best candidate (s). Further on, the chromatographic retention time of the tentative candidates was predicted using an in silico approach that is based on quantitative structure retention relationships (QSRR) (Aalizadeh, Thomaidis, Bletsou, & Gago-Ferrero, 2016).

The level of confidence achieved in the identification of the detected compounds was established according to Schymanski et al. (2014) to ease the communication of identification confidence. Initially, a mass (m/z) of interest corresponding to an unknown compound starts at Level 5 (exact mass of interest). If it is possible to unambiguously assign a molecular formula to this m/z, then it will be upgraded to level 4 (unequivocal molecular formula). If there is sufficient MS (exact mass, isotope or adducts) and experimental information (eg. t_R), non-target components can gain in confidence through level 3 (Tentative Candidate). This level indicates that evidence exists for one or more possible structure(s), but insufficient information is available to eliminate other possible structural candidates (isomers etc.). Nonetheless, if there is a spectral library match for one single structure or if diagnostic evidence is present to exclude all other possible structures from consideration, the compound can reach level 2 (probable structure). Level 2 includes two sublevels; level 2a eg. evidence by matching MS/MS information with literature or spectral library and level 2b denotes diagnostic evidence, such as agreement between predicted and experimental t_R. Finally, if the structure can be confirmed via appropriate measurement of a reference standard with MS, MS/MS fragments and t_R matching, level of identification is 1.

2.8. Database preparation

A database consisting of 1608 compounds commonly occurring in different organs and by-products of the olive tree, *Olea Europa L.*, such as drupes, stems, leaves, olive oil etc. was compiled from FooDB. This database including chemical identifiers, predicted retention time and MS information can be found in the Excel Supplementary Material. This list was chemically curated (removing the duplicates, metals, salts, solvents and ambiguous bonding between atoms), following eight main steps (Aalizadeh, von der Ohe & Thomaidis, 2017): (1) the initially retrieved chemical identifiers (CAS number or SMILES) were unified into InChI; (2) 2D structures of the InChI were created and the dative bonds (*e.g.* nitro group) were standardized using Open Babel (http://openbabel.org/docs/current/) (O'Boyle et al., 2011); (3) salts, metals and solvents were removed from the chemical structure; (4) the octect number was fixed and hydrogens were added; (5) 2D structures were created using Open Babel and 3D structures were obtained out of

various tautomer forms (the tautomer with the lowest energy was retained to get one structure out of different forms of a duplicate entries) using Balloon (Vainio & Johnson, 2007); (6) A SDF file with optimized 3D structures for all entries was created; (7) optimized InChI chemical identifier were derived from the SDF file; (8) duplicates were identified and removed by comparing their optimized InChI from the SDF file. This list can be directly used in MetFrag (Wolf et al., 2010) to elucidate the structure more appropriately and relevantly than other databases, by limiting the searching space to compounds that occurring in olive/olive oil. The compiled database includes the monoisotopic mass, $[M + H]^+$ and $[M-H]^-$, predicted retention time (t_R), molecular formula and chemical identifiers together with reference.

2.9. Data processing

Using only annotation results created by CAMERA package and excluding molecular ion adducts may not be as effective as removing them based on their intercorrelation profile. In other words, retaining adducts in the final list of *m/z* features and keeping track of their respective molecular ion is better if the adducts give reasonably high intensity. To further prevent the highly cofounded features prior performing the classification model, V-WSP algorithm was used as an unsupervised variable reduction method (Ballabio et al., 2014). This method allows the selection of a representative set of variables based on linear correlation (here we set the correlation threshold to 0.8), so that multicollinearity and redundant information in the data can be reduced. Using V-WSP, the features were reduced from 287 to 250.

2.10. Ant Colony Optimization (ACO)

The use of a feature selection technique with a fitness function (in this case, it is the misclassification in 51 olive oil samples) can efficiently reveal only the meaningful *m/z* that contribute to the classification. Further prioritization of *m/z* was done by ACO. ACO is a swarm intelligence algorithm that is based on the behavior of the ants searching for the food resources using pheromone deposition (Dorigo, Birattari, & Stützle, 2006; Dorigo & Blum, 2005). This enables ants to be adoptable to any environmental changes, and find a new shortest path to the resources (Dorigo & Blum, 2005). ACO is preferably a good method to handle features selection related problems because ants can derive the best combination of subsets that has the minimum fitness objective (here is the miss-classification error). In a typical ACO based features selection case, the algorithm begins with the generation of certain number of ants (here we set this at 100 ants) placed randomly on the graph, which represents the possible combinations of every *m/z*. Thus, each node (in a graph) relates to a *m/z*, and each edge shows the traversal of an ant from one *m/z* to another. The number of artificial pheromone [0,1] for an edge is associated with the popularity of the particular traversal by previous ants. Therefore, ants could make probabilistic decisions to stay at which node and select which edge, based on the artificial pheromone and related traversal degree. This will continue until the minimum degree for the mis-classification error has been reached, otherwise all process will be iterated again (Dorigo et al., 2006; Dorigo & Blum, 2005). The maximum number of iteration was set to 100 and the desired number of features was set up to 7 features. Evaporation Rate (ER) was also set to 0.05 (this value is kept constant during performing ACO and generally is low value (0.01–0.05)) (Dorigo & Blum, 2005). ER causes uniformly decrease in all the pheromone values. From a practical point of view, pheromone evaporation is required to prevent a rapid convergence of the algorithm towards a sub-optimal space. ACO algorithm was written and performed in MATLAB.

2.11. Random Forest (RF)

RF was used for the classification of EVOOs based on their cultivar. RF was introduced by Breiman (2001) and applied for both regression and classification problems. More information concerning RF can be found in the Supplementary Material, Section S1.

Classification model based on ACO-RF was achieved using miss-classification error in leave-one-out cross validation as fitness. The predictive power of the proposed classification model was evaluated independently using a set of 11 external samples that were not part of the initial training set and confusion matrix was calculated to derive error rate, class specificity and sensitivity (Ballabio & Consonni, 2013). The division into training and test set was achieved by Kennard-Stone algorithm (Kennard & Stone, 1969). Kennard-Stone algorithm starts by selecting the pair of points (i samples and *m/z* features) that are the furthest apart. The selected samples were assigned to the training sets and removed from the list of samples. Then, the next pair of samples, which are furthest apart, are assigned to the test set. In a third step, the procedure assigns each remaining sample alternatively to the training and test sets based on the distance to the previously selected sample. The distance function used is Euclidean distance. Moreover, Receiver Operating Characteristics (ROC) was calculated to control the accuracy and error rate of proposed model. ROC curves were derived for each class by plotting the sensitivity versus 1-specificity in six cultivars. A reliable classification model would yield a point in the upper left corner of the ROC area, representing maximum sensitivity and specificity, while a random one causes points to be along the diagonal line from the left bottom to the top right corner (Ballabio & Consonni, 2013).

2.12. Principal Component Analysis (PCA)

The usefulness of feature selection was addressed using unsupervised classification method like Principal Component Analysis (PCA). PCA was applied before and after performing feature selection to investigate whether the covariance explained by PCs increases or not. All the data processing, pretreatment and classification were performed by a homemade program so called ChemoTrAMS, in MATLAB environment.

3. Results and discussion

3.1. Non-target screening identification

Using 40 EVOOs and 250 features in the training set along with leave-one-out cross validation analysis as fitness function to identify potential *m/z* features, ACO selected the four most relevant *m/z*s that could explain the distribution of samples based on their varieties. These selected *m/z* features could also create a final classification model with miss-classification error of zero for each class. These features were *m/z*: 167.0345/t_R = 2.43, *m/z*: 299.0561/t_R = 8.11, *m/z*: 269.0456/t_R = 8.04 and *m/z*: 303.1237/t_R = 6.50. In an attempt to identify these masses, an inclusion list was created and QTOF system operated in Auto MS/MS mode to obtain the MS/MS spectra of the unknown analytes. Following the non-target screening workflow, EICs were generated in Data Analysis and the most plausible molecular formulas were determined showing high mass accuracy (less than 2.97 ppm) and acceptable isotopic fit values (less than 15.9 mSigma). The determined molecular formulas were elucidated to certain chemical structures with mass accuracy of ± 0.001 ppm.

Specifically, for the mass detected at *m/z*: 167.0345, the molecular formula $C_8H_8O_4$ was assigned to it using “SmartFormula Manually”, according to the criteria of mass accuracy (2.4 ppm) and isotopic fit (5.9 mSigma). In a further step, the prepared local database search, as introduced in **2.8**, was loaded in Metfrag (Wolf et al., 2010). The MS/MS spectra were examined and verified resulting in 4 candidate compounds. Only 1 tentative compound was scored with 1.0 in Metfrag (Wolf et al., 2010) with all 7 fragments explained, vanillic acid. Predicted t_R with QSRR (3.97 min) was close to the experimental. The corresponding standard was purchased and the presence of vanillic acid in the samples was verified. Vanillic acid is an antioxidant with

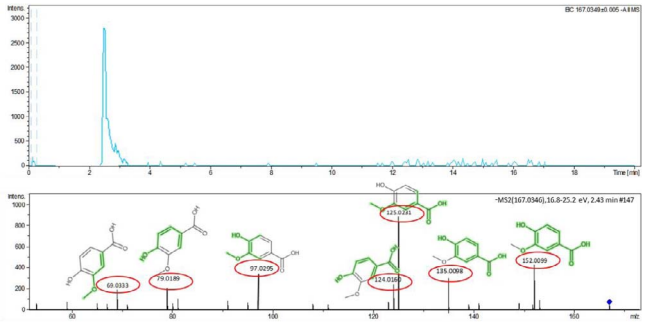

Fig. 1. EIC and MS/MS spectra with 7 explained fragments of vanillic acid.

antioxibacterial, antimicrobial and antifungal activity (Obied et al., 2005). The EIC and MS/MS spectrum of vanillic acid (identification level: 1) is presented in Fig. 1.

For the mass with *m/z*: 269.0456, the molecular formula $C_{15}H_{10}O_5$ was assigned to it using "SmartFormula Manually", with mass accuracy: 2.97 and isotopic fit: 15.9 mSigma). The local database search resulted in three candidate compounds for that molecular formula. Performing in silico fragmentation with Metfrag (Wolf et al., 2010) using the molecular formula and measured MS/MS revealed 1 tentative candidate with high score (1.0) and all fragments explained, apigenin. The predicted t_R for apigenin with QSRR (6.99 min) was close to the experimental. Finally, the identity of apigenin in the samples was confirmed with a standard. The EIC and MS/MS spectrum of apigenin (identification level: 1) is shown in Fig. 2.

A peak corresponding to *m/z*: 299.0561 was detected. After applying mass accuracy and isotopic fit filters (mass accuracy: 1.08 and isotopic fit: 9.3 Sigma), the molecular formula $C_{15}H_{12}O_6$ was assigned to it. The local database provides 4 possible compounds for this molecular formula. These 4 substances were able to explain all the fragments found in the MS/MS spectrum. In this case, QSRR provided the tentative candidate, luteolin 7-methyl ether, the predicted t_R with QSRR (7.01 min) was close to the experimental. Fig. 3 illustrates the EIC and MS/MS spectrum of luteolin 7-methyl ether (identification level: 2b).

For the mass with *m/z*: 303.1237, the molecular formula $C_{17}H_{20}O_5$ was assigned to it using "SmartFormula Manually", with mass accuracy: 2.64 and isotopic fit: 6.8 mSigma). The local database search resulted in only one candidate compound for that molecular formula, oleocanthal. Performing in silico fragmentation with Metfrag (Wolf et al., 2010) using the molecular formula and measured MS/MS all the fragments were explained. The predicted t_R for oleocanthal with QSRR (7.17 min) was close to the experimental. The peak at *m/z*: 165.0556 corresponding to $C_9H_9O_3$ has been reported by Dierkes et al. (2012). In addition, the peak at *m/z*: 183.0663 corresponding to $C_9H_{11}O_4$ has been reported by Dierkes et al. (2012) and Bajoub et al. (2016). Oleocanthal shares unique perceptual and anti-inflammatory characteristics with Ibuprofen (Beauchamp et al., 2005). The EIC and MS/MS spectrum of oleocanthal (identification level: 2a) is shown in Fig. 4.

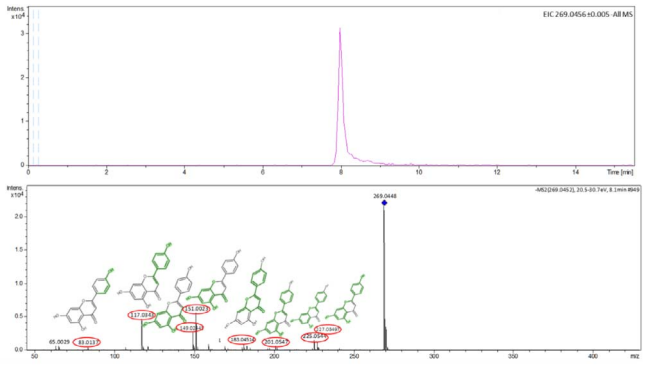

Fig. 2. EIC and MS/MS spectrum with 8 explained fragments of apigenin.

All the compounds identified via non-target screening belong to the applicability domain of the model and the predicted retention time results are highly reliable. More information about the QSRR model and ita applicability domain can be found in the Supplementary Material, Section S2.

3.2. Quantification and semi-quantification results

Standard addition calibration curves were constructed for the quantification and semi-quantification of the results. All standard addition calibration curves were constructed with the use of the peak area of the spiked analyte subtracted by the peak area of a neat sample and divided by the peak area of the internal standard (syringaldehyde 1.30 mg kg^{-1}). Vanillic acid and apigenin were quantified based on the standard addition curves of their commercial standards. These two standards were spiked in real EVOO samples at concentrations between 0.02 and 10 mg kg^{-1} (10 calibration levels with 3 replicates at each level) and the equations of the curves were: $y = [(-0.08 \pm 0.07) + (0.73 \pm 0.02)\times]$ and $y = [(0.37 \pm 1.21) + (12.05 \pm 0.26)\times]$. For the semi-quantification of luteolin 7-methyl ether, luteolin was spiked in real EVOO samples at concentrations between 0.1 and 20 mg kg^{-1} (10 calibration levels with 3 replicates at each level). The standard addition curve of luteolin was: $y = [(0.69 \pm 0.57) + (4.28 \pm 0.29)\times]$. Oleocanthal was found to have structural similarity with tyrosol, as it has already been reported by Kalogiouri et al. (2017). For the semi-quantification of oleocanthal, standard addition calibration curve of tyrosol was constructed over the range 1–100 mg kg^{-1} and the equation was: $[y = (-2.17 \pm 0.03) + (4.41 \pm 0.07)\times]$. The analytical curves presented an adequate fit when submitted to the lack-of-fit test ($F_{calculated}$ was less than $F_{tabulated}$ in all cases) and r^2 above 0.99, proving that they can be used for the quantification of the phenolic compounds. The quantification and semi-quantification results of the identified markers are presented in mg kg^{-1} (mean values ± SD (n = 3)) in the Supplementary Material, Table S3.

3.3. Principal Component Analysis (PCA)

All in all, two PCs explained 59% of variance and showed appropriate grouping of samples belonging to Manaki, Amfissis and Chalkidikis EVOOs variety. These results are shown in Fig. 5(a). This plot is generated by XCMS online and is based on the intensity of the MS selected by "centWave" algorithm with the same parameters used in peak picking step. It is clearly can be seen that PCA is not capable of separating and grouping the samples based on their varieties using all MS features. Surprisingly, a significant increase in variance (80.8%) is observed after the selection of four features, followed by their identification and quantification. According to Fig. 5(b), EVOOs that belonged to the Amfissis variety were distributed as a separate class, and the EVOOs of the varieties Manaki and Ladoelia were grouped as separate classes, as well. However, the EVOOs belonging to the varieties Chalkidikis, Koroneiki and Kolovi could not be differentiated and were grouped together. This proves the requirement of feature selection tool to avoid adding false positive MS features inside the loading variables. Having identified all the selected *m/z* by ACO, their quantification results were used to build the decision tree using RF. ACO-RF as a validated classification approach generated a graph with a threshold for each identified compound. The validation was done using ROC curve showing the accuracy, specificity and selectivity for each variety along with the error associated with leave-one-out and k-fold cross validation results (Ballabio & Consonni, 2013). K-fold cross validation is a good parameter to judge validity and over-fitting of a classification model as instead of 1 sample per analysis, it excludes several samples out and tries to calculate the error associated with classification model. The number of k was set to 10 and cross validation was performed. The

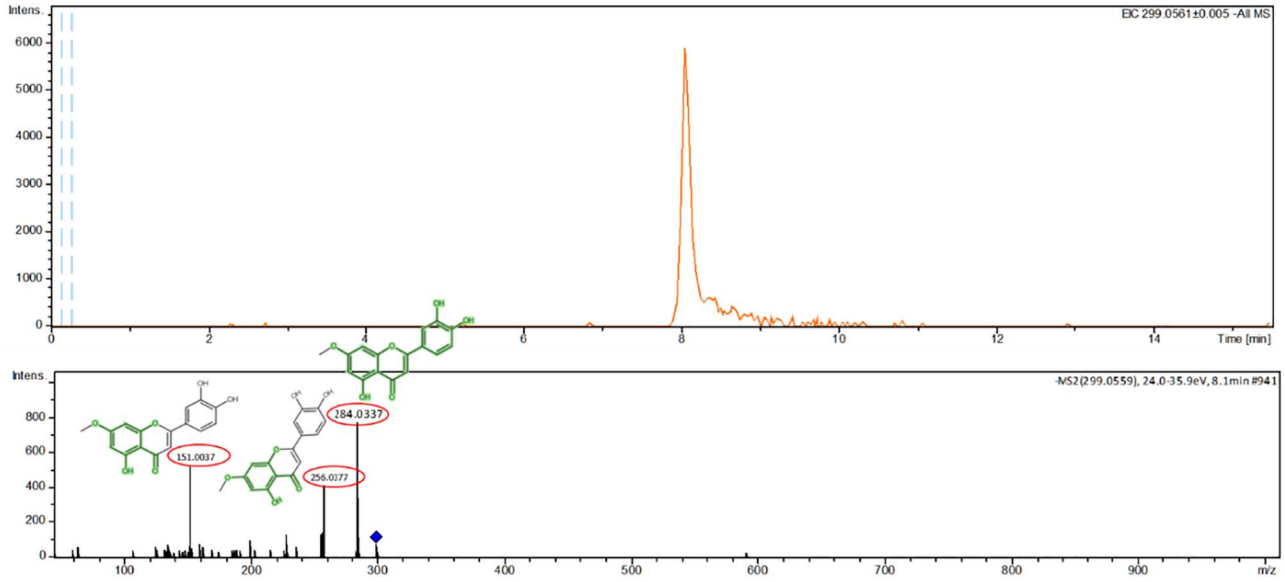

Fig. 3. EIC and MS/MS spectrum with 3 explained fragments of luteolin 7-methyl ether.

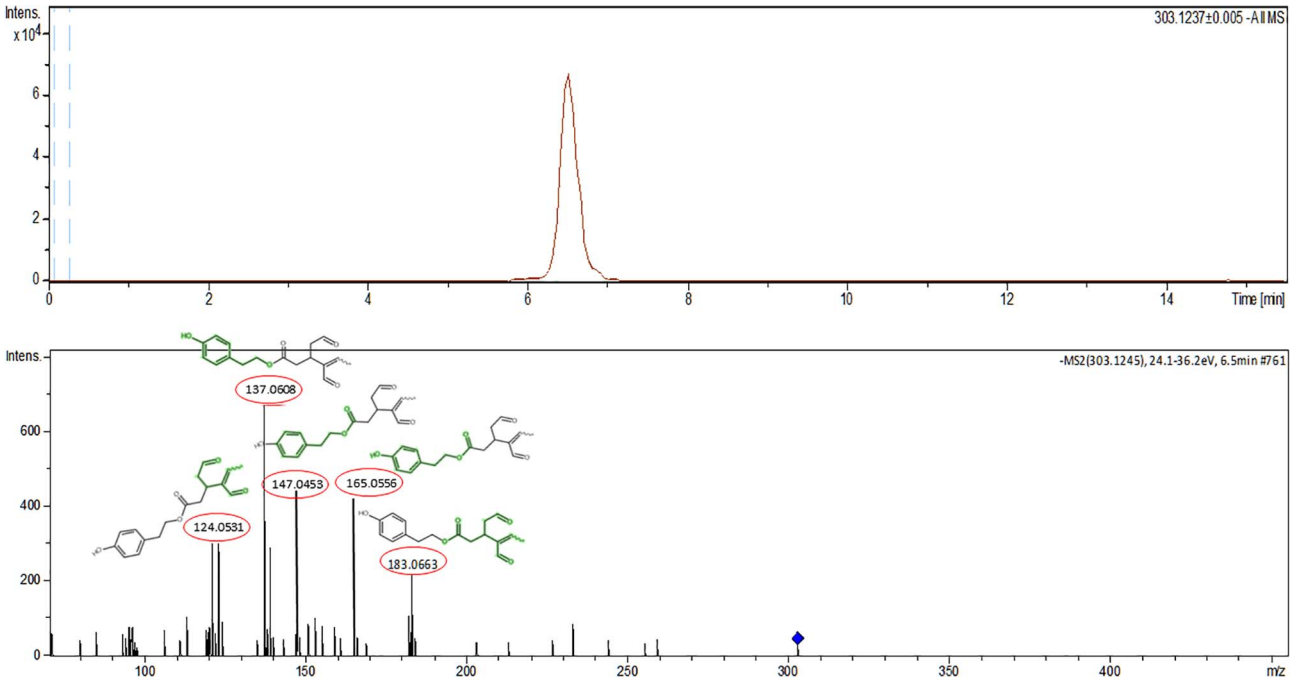

Fig. 4. EIC and MS/MS spectrum with 3 explained fragments of oleocanthal.

errors after leave-one-out and k-fold cross validation were calculated and found to be very low at 0.075 and 0.175, respectively. ROC curves also described area under curve (AUC), accuracy, specificity and selectivity at 1.00, which together with cross validation show that the classification model is not over-fitted and can be applied to a suspect external sample. The decision tree developed for the classification of EVOOs according to their cultivar is presented in Fig. 6.

According to the decision tree established by ACO-RF, oleocanthal and apigenin play dominant roles. Oleocanthal is important for the discrimination of EVOOs labeled as Manaki or Chalkidikis and Ladoelia or Koroneiki after justifying if they have high or low content of the flavonoid apigenin. Therefore, if the concentration of apigenin is higher than 2.16 mg kg^{-1}, it belongs to the cultivars of Ladoelia, Koroneiki or Amfissis; and if its concentration of apigenin is less than 2.16 mg kg^{-1}, then, it belongs to Manaki, Chalkidikis or Kolovi. Interestingly, vanillic acid was found at lowest concentration (below 1.56 mg kg^{-1}) in EVOOs of Kolovi. On the other hand, when the concentration of apigenin is above 2.16 mg kg^{-1} and the concentration of luteolin 7-methyl ether is above 13.20 mg kg^{-1}, the EVOOs belong to the variety of Amfissis. This is also observed from PCA (Fig. 5(b)) where the loading plot showed high content of luteolin 7-methyl ether and apigenin, causing Amfissis EVOOs to group together. None of the EVOOs belonging to the other varieties showed similarly high content of luteolin 7-methyl ether. It is also observed that EVOOs with higher content of oleocanthal and apigenin, but lower content of luteolin 7-methyl ether belong to Ladoelia variety, otherwise (if the concentration of oleocanthal is less than 86.20 mg kg^{-1}) they would be classified as Koroneiki. In addition. The PCA loading plot (Fig. 5(b)) showed that Ladoelia EVOOs grouped together, presenting high concentration valued for oleocanthal. EVOOs with higher concentrations of vanillic acid (more than 1.56 mg kg^{-1}),

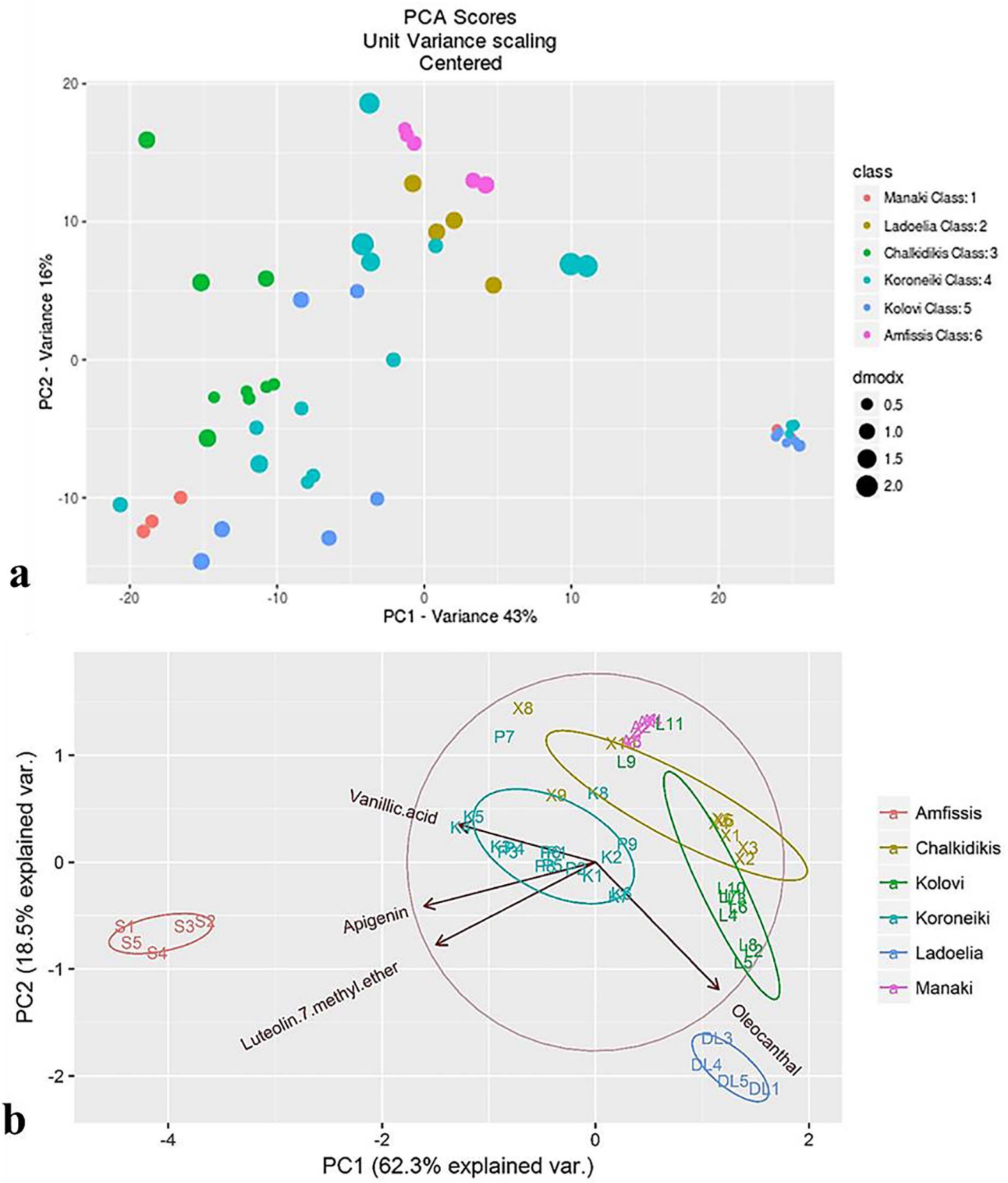

Fig. 5. PCA with color shows the varietal (a) before (b) after MS features prioritization;

but lower concentration of apigenin (less than 2.16 mg kg^{-1}) belong to either Manaki or Chalkidikis. To discriminate between Manaki and Chalkidikis, the decision tree used again oleocanthal and indicated that EVOOs of the Manaki cultivar have lower oleocanthal content containing (less than 31.50 mg kg^{-1}).

Therefore, the decision tree could simply and easily apply discrimination rule to understand how the EVOO varieties correspond to the chemical profile, while PCA as a commonly used chemometrics tool failed to distribute all the EVOOs based on their varieties.

4. Conclusions

This study contributes to the field of food authenticity and guarantees the classification of Greek PDO EVOOs with the application of a non-target screening RP-UHPLC-ESI-QTOFMS method combined with ACO-RF. The proposed method was successfully applied in 51 EVOOs of

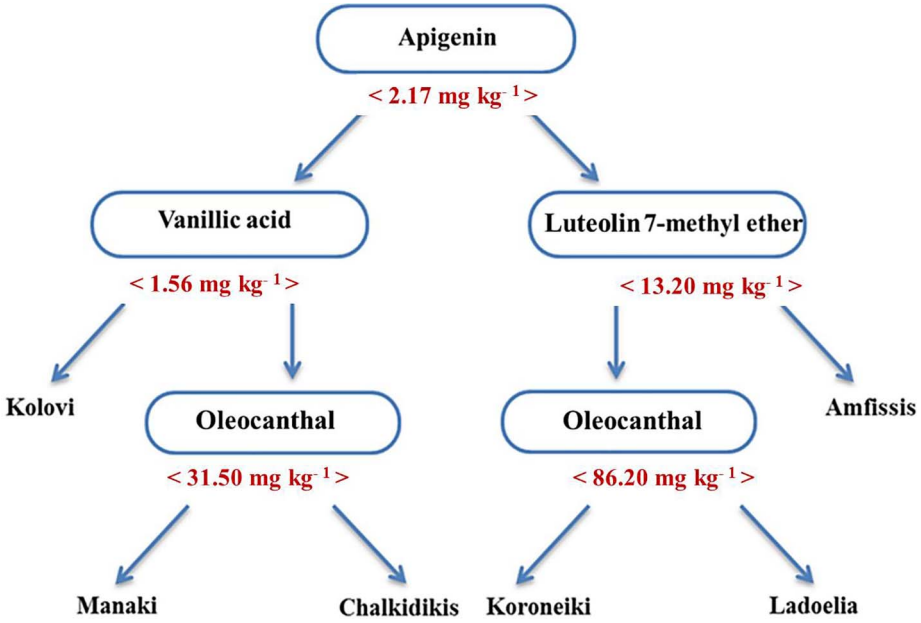

Fig. 6. Classification of EVOOs according to ACO-RF decision tree.

the Greek cultivars: Amfissis, Chalkidikis, Kolovi, Koroneiki, Ladoelia and Manaki that were produced during the harvesting year 2015–2016.

A peak list consisting of 280 features was generated using the XCMS package, and was processed with chemometrics. PCA failed to distribute the samples based on their cultivars according to the initial non-target list, showing that further *m/z* prioritization is needed to prevent the incorporation of false positive features, which affect negatively the distribution of the samples. After the implication of ACO and the selection of the 4 most important features, PCA exhibited higher variance and better sample distribution.

Non-target identification workflow was applied in order to identify these 4 markers. In order to accelerate the identification task, a local database consisting of 1608 compounds commonly occurring in olive matrices was compiled, and 4 markers, apigenin, vanillic acid, luteolin 7-methyl ether and oleocanthal were identified.

Finally, RF established a robust classification that could successfully classify Greek EVOOs, harvested in 2015–2016, into 6 Greek cultivars, setting a concentration threshold for each selected marker. This tree was based on the selection of the 4 markers that were identified as apigenin, vanillic acid, luteolin 7-methyl ether and oleocanthal. Based on ACO-RF, it was concluded that the concentration of oleocanthal changes dramatically across Greek olive oil cultivars and has distinguished quantification threshold between Ladoelia (containing more than 86.20 mg kg^{-1}) and Manaki (containing less than 31.50 mg kg^{-1}). Interestingly, apigenin was found to play a crucial role in the prediction of the cultivars. This method sets several concentration thresholds (based on the quantification results) over the markers identified, making the authentication task simple.

Appendix

Chemometric tools

XCMS: An R package to perform peak picking in data analyzed by High Resolution Mass Spectrometry (HRMS).

***centWave* algorithm**: Highly sensitive feature detection algorithm for high resolution LC/MS that is based on detecting regions of interest (ROI) in the *m/z* domain.

CAMERA: An R package for componentization and deconvolution of mass spectrum.

Ant Colony Optimization (ACO): A nature inspired algorithm to select important features among pool of variables (here, applied over peaks-list to prioritize *m/z*s based on their contributions over each variety).

MetFrag: An in silico fragmentation technique to assign fragments to mass spectra and to subsequently rank the plausible candidates.

V-WSP algorithm: An unsupervised variable reduction method to detect cofounded and redundant *m/z* in peaks-list and excludes them prior performing the classification model.

Random Forest (RF): RF is a multi-class supervised classification technique that is based on decision tree and here was used for classification of EVOOs based on their cultivar.

Kennard-Stone algorithm: This is a robust technique to create representative subset (test set) which is needed for blind evaluation of any classification models built.

Receiver Operating Characteristics (ROC) curve: This is a method to control the accuracy and error rate of any newly proposed classification model.

Principal Component Analysis (PCA): An unsupervised chemometric technique for exploratory data analysis. It projects the data into a reduced hyperspace, defined by orthogonal principal components.

Quantitative Structure-Retention relationship (QSRR): This is a technique to explore relationship between chemical structures and their liquid chromatographic retention time. Here, it was used during assignment of identification level for plausible candidates for a given *m/z*.

ChemoTrAMS: An in-house executable program to perform several chemometric methods (peaks-list pretreatment, *m/z* prioritization, unsupervised and supervised classification/regression methods and quality control of the models) on HRMS results.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.foodchem.2018.02.101.

References

Aalizadeh, R., Thomaidis, N. S., Bletsou, A. A., & Gago-Ferrero, P. (2016). Quantitative structure-retention relationship models to support nontarget high-resolution mass spectrometric screening of emerging contaminants in environmental samples. *Journal of Chemical Information and Modeling, 56*(7), 1384–1398.

Aalizadeh, R., von der Ohe, P. C., & Thomaidis, N. S. (2017). Prediction of acute toxicity of emerging contaminants on the water flea *Daphnia magna* by Ant Colony

Optimization-Support Vector Machine QSTR models. *Environmental Science: Processes & Impacts, 19*(3), 438–448.
Alkan, D., Tokatli, F., & Ozen, B. (2011). Phenolic characterization and geographical classification of commercial extra virgin olive oils produced in Turkey. *Journal of the American Oil Chemists' Society, 89*(2), 261–268.
Allalout, A., Krichène, D., Methenni, K., Taamalli, A., Oueslati, I., Daoud, D., & Zarrouk, M. (2009). Characterization of virgin olive oil from Super Intensive Spanish and Greek varieties grown in northern Tunisia. *Scientia Horticulturae, 120*(1), 77–83.
Baccouri, O., Guerfel, M., Baccouri, B., Cerretani, L., Bendini, A., Lercker, G., ... Daoud Ben Miled, D. (2008). Chemical composition and oxidative stability of Tunisian monovarietal virgin olive oils with regard to fruit ripening. *Food Chemistry, 109*(4), 743–754.
Bajoub, A., Medina-Rodriguez, S., Gomez-Romero, M., Ajal el, A., Bagur-Gonzalez, M. G., Fernandez-Gutierrez, A., & Carrasco-Pancorbo, A. (2017). Assessing the varietal origin of extra-virgin olive oil using liquid chromatography fingerprints of phenolic compound, data fusion and chemometrics. *Food Chemistry, 215*, 245–255.
Bajoub, A., Pacchiarotta, T., Hurtado-Fernandez, E., Olmo-Garcia, L., Garcia-Villalba, R., Fernandez-Gutierrez, A., ... Carrasco-Pancorbo, A. (2016). Comparing two metabolic profiling approaches (liquid chromatography and gas chromatography coupled to mass spectrometry) for extra-virgin olive oil phenolic compounds analysis: A botanical classification perspective. *Journal of Chromatography A, 1428*, 267–279.
Bakhouche, A., Lozano-Sánchez, J., Beltrán-Debón, R., Joven, J., Segura-Carretero, A., & Fernández-Gutiérrez, A. (2013). Phenolic characterization and geographical classification of commercial Arbequina extra-virgin olive oils produced in southern Catalonia. *Food Research International, 50*(1), 401–408.
Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods, 5*(16), 3790–3798.
Ballabio, D., Consonni, V., Mauri, A., Claeys-Bruno, M., Sergent, M., & Todeschini, R. (2014). A novel variable reduction method adapted from space-filling designs. *Chemometrics and Intelligent Laboratory Systems, 136*, 147–154.
Ballus, C. A., Quirantes-Pine, R., Bakhouche, A., da Silva, L. F., de Oliveira, A. F., Coutinho, E. F., ... Godoy, H. T. (2015). Profile of phenolic compounds of Brazilian virgin olive oils by rapid resolution liquid chromatography coupled to electrospray ionisation time-of-flight mass spectrometry (RRLC-ESI-TOF-MS). *Food Chemistry, 170*, 366–377.
Beauchamp, G. K., Keast, R. S. J., Morel, D., Lin, J., Pika, J., Han, Q., ... Breslin, P. A. S. (2005). Phytochemistry: ibuprofen-like activity in extra-virgin olive oil. *Nature, 437*, 45–46.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Cerretani, L., Bendini, A., Caro, A. D., Piga, A., Vacca, V., Caboni, M. F., & Gallina Toschi, T. (2005). Preliminary characterisation of virgin olive oils obtained from different cultivars in Sardinia. *European Food Research and Technology, 222*(3–4), 354–361.
Council Regulation (EC) no. 510, 2006 (2006). on the protection of geographical indications and designations of origin for agricultural products and foodstuffs. *Official Journal of the European Union, L93*, 12–25.
Dierkes, G., Krieger, S., Duck, R., Bongartz, A., Schmitz, O. J., & Hayen, H. (2012). High-performance liquid chromatography-mass spectrometry profiling of phenolic compounds for evaluation of olive oil bitterness and pungency. *Journal of Agricultural and Food Chemistry, 60*(31), 7597–7606.
Dorigo, M., Birattari, M., & Stützle, T. (2006). Ant colony optimization artificial ants as a computational intelligence technique. *IEEE Computational Intelligence Magazine, 1*(4), 28–39.
Dorigo, M., & Blum, C. (2005). Ant colony optimization theory: a survey. *Theoretical Computer Science, 344*(2–3), 243–278.
FooDB, The Food Components Database. http://foodb.ca/. Accessed 13.02.207.
Ghanbari, R., Anwar, F., Alkharfy, K. M., Gilani, A.-H., & Saari, N. (2012). Valuable nutrients and functional bioactives in different parts of olive (Olea europaea L.)—a review. *International Journal of Molecular Sciences, 13*(3), 3291.
Kalogiouri, N. P., Aalizadeh, R., & Thomaidis, N. S. (2017). Investigating the organic and conventional production type of olive oil with target and suspect screening by LC-QTOF-MS, a novel and semi-quantification method using chemical similarity and advanced chemometrics. *Analytical and Bioanalytical Chemistry, 409*, 5413–5426.
Kalogiouri, N. P., Alygizakis, N. A., Aalizadeh, R., & Thomaidis, N. S. (2016). Olive Oil authenticity studies by target and nontarget LC-QTOF combined with advanced chemometric techniques. *Analytical and Bioanalytical Chemistry, 408*(28), 7955–7970.
Karabagias, I., Michos, C., Badeka, A., Kontakos, S., Stratis, I., & Kontominas, M. G. (2013). Classification of Western Greek virgin olive oils according to geographical origin based on chromatographic, spectroscopic, conventional and chemometric analyses. *Food Research International, 54*(2), 1950–1958.
Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics, 11*(1), 137–148.
Kosma, I., Vavoura, M., Kontakos, S., Karabagias, I., Kontominas, M., Apostolos, K., & Badeka, A. (2016). Characterization and classification of extra virgin olive oil from five less well-known greek olive cultivars. *Journal of the American Oil Chemists' Society, 93*(6), 837–848.
Libiseller, G., Dvorzak, M., Kleb, U., Gander, E., Eisenberg, T., Madeo, F., ... Magnes, C. (2015). IPO: A tool for automated optimization of XCMS parameters. *BMC Bioinformatics, 16*(1), 118.
Longobardi, F., Ventrella, A., Casiello, G., Sacco, D., Tasioula-Margari, M., Kiritsakis, A. K., & Kontominas, M. G. (2012). Characterisation of the geographical origin of Western Greek virgin olive oils based on instrumental and multivariate statistical analysis. *Food Chemistry, 133*(1), 169–175.
O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics, 3*(1), 33.
Obied, H. K., Allen, M. S., Bedgood, D. R., Prenzler, P. D., Robards, K., & Stockmann, R. (2005). Bioactivity and analysis of biophenols recovered from olive mill waste. *Journal of Agricultural and Food Chemistry, 53*(4), 823–837.
Petrakis, P., Agiomyrgianaki, A., Christophoridou, S., Spyros, A., & Dais, P. (2008). Geographical characterization of greek virgin olive oils (Cv. Koroneiki) using 1H and 31P NMR fingerprinting with canonical discriminant analysis and classification binary trees. *Journal of Agricultural and Food Chemistry, 56*, 3200–3207.
Pouliarekou, E., Badeka, A., Tasioula-Margari, M., Kontakos, S., Longobardi, F., & Kontominas, M. G. (2011). Characterization and classification of Western Greek olive oils according to cultivar and geographical origin based on volatile compounds. *Journal of Chromatography A, 1218*(42), 7534–7542.
Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science & Technology, 48*(4), 2097–2098.
Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry, 78*(3), 779–787.
Tengstrand, E., Lindberg, J., & Åberg, K. M. (2014). TracMass 2—A modular suite of tools for processing chromatography-full scan mass spectrometry data. *Analytical Chemistry, 86*(7), 3435–3442.
Tura, D., Gigliotti, C., Pedò, S., Failla, O., Bassi, D., & Serraiocco, A. (2007). Influence of cultivar and site of cultivation on levels of lipophilic and hydrophilic antioxidants in virgin olive oils (Olea Europea L.) and correlations with oxidative stability. *Scientia Horticulturae, 112*(1), 108–119.
Vainio, M. J., & Johnson, M. S. (2007). Generating conformer ensembles using a multi-objective genetic algorithm. *Journal of Chemical Information and Modeling, 47*(6), 2462–2474.
Wolf, S., Schmidt, S., Muller-Hannemann, M., & Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics, 11*, 148.