EL ARTICULO CORRESPONDIENTE A ESTA VERSION SE ENCUENTRA PUBLICADO EN LA REVISTA FOOD CHEMISTRY CON DOI: https://doi.org/10.1016/j.foodchem.2018.04.019

Combination of complementary data mining methods for geographical characterization of extra virgin olive oils based on mineral composition

Ana Sayago[a,b], Raúl González-Domínguez[a,b,1*], Rafael Beltrán[a,b], Ángeles Fernández-Recamales[a,b]

[a]Department of Chemistry, Faculty of Experimental Sciences, University of Huelva. 21007, Spain. [b]International Campus of Excellence CeiA3, University of Huelva. 21007, Spain. [1]Present address: Biomarkers & Nutrimetabolomics Laboratory, Department of Nutrition, Food Sciences and Gastronomy, Faculty of Pharmacy and Food Sciences, University of Barcelona. 08028, Spain.

***corresponding author:** Raúl González-Domínguez; e-mail address: raul.gonzalez@dqcm.uhu.es; Phone: +34 959219975; Postal address: Department of Chemistry, Faculty of Experimental Sciences, University of Huelva. 21007, Spain.

e-mail addresses: Ana Sayago: ana.sayago@dqcm.uhu.es; Raúl González-Domínguez: raul.gonzalez@dqcm.uhu.es; Rafael Beltrán: beltran@uhu.es; Ángeles Fernández-Recamales: recamale@uhu.es

Abbreviations. EVOO, extra virgin olive oil; ICP-OES, inductively coupled plasma optical emission spectrometry; ICP-MS, inductively coupled plasma mass spectrometry; LDA,

linear discriminant analysis; PCA, principal component analysis; PLS-DA, partial least squares discriminant analysis; RF, random forest; SENS, sensitivity; SPEC, specificity; SVM, support vector machine

Abstract

This work explores the potential of multi-element fingerprinting in combination with advanced data mining strategies to assess the geographical origin of extra virgin olive oil samples. For this purpose, the concentrations of 55 elements were determined in 125 oil samples from multiple Spanish geographic areas. Several unsupervised and supervised multivariate statistical techniques were used to build classification models and investigate the relationship between mineral composition of olive oils and their provenance. Results showed that Spanish extra virgin olive oils exhibit characteristic element profiles, which can be differentiated on the basis of their origin in accordance with three geographical areas: Atlantic coast (Huelva province), Mediterranean coast and inland regions. Furthermore, statistical modeling yielded high sensitivity and specificity, principally when random forest and support vector machines were employed, thus demonstrating the utility of these techniques in food traceability and authenticity research.

Keywords: olive oil; geographical traceability; mineral profile; inductively coupled plasma-mass spectrometry; data mining

1. Introduction

Olive oil is the main source of fat in the Mediterranean diet, which has historically been associated with beneficial effects on health. In this sense, the epidemiology suggests that olive oil might have a role in the prevention of coronary diseases and several cancer types, because of its high levels of monounsaturated fatty acids and polyphenolic compounds (Aguilera, Martin-Cabrejas, & González de Mejia, 2016; Martínez-González & Sanchez-Villegas, 2004). Thereby, the consumption of olive oil has increased in last years, due to its organoleptic and nutritional properties as well as healthy reputation.

Quality, safety and sensorial attributes, such as colour, flavour and taste, are the most important characteristics that determine the commercial value of a food such as olive oil. These features are determined by the chemical composition that, in turn, is affected by pre-harvest and post-harvest factors. Among these pre-harvest factors, olive variety is the main source of variation in composition and sensorial attributes (Aparicio & Harwood, 2013). In addition, quality is also influenced by crop conditions including environmental (climate, soil composition) and agronomic (irrigation, fertilization) factors, sampling time and degree of ripeness (Agiomyrgianaki, Petrakis, & Dais, 2012; Longobardi, Ventrella, Casiello, Sacco, Tasioula-Margari, et al., 2012; Mihailova, Abbado, Kelly, & Pedentchouk, 2015; Romero, Saavedra, Tapia, Sepúlveda, & Aparicio, 2016). On the other hand, manufacturing methods used for olive oil extraction and storage conditions are the most important post-harvest factors impacting quality (Ben-Hassine et al., 2013). All these variables are closely associated with the geographical origin and provoke significant differences in organoleptic characteristics, nutritional composition and nutraceutical value of olive oils. Therefore,

71 the traceability and authenticity of olive oil is an important objective in guaranteeing
72 the quality demanded by consumers, producers and regulatory bodies for this valuable
73 food product.
74 Numerous analytical methodologies have been proposed for the differentiation and
75 classification of olive oils according to the geographical origin, which are usually
76 focused on the determination of different organic compounds (Luykx & van Ruth,
77 2008). For instance, the profile of fatty acids and triglycerides has previously been used
78 to compare Tunisian, Maghrebian and French virgin olive oils (Laroussi-Mezghani et al.,
79 2015), to characterize Apulian virgin olive oils (Longobardi, Ventrella, Casiello, Sacco,
80 Catucci, et al., 2012) or to distinguish different Turkish oils (Arslan, Karabekir, &
81 Schreiner, 2013). Furthermore, the analysis of minor metabolites such as free sterols,
82 aliphatic and terpenic alcohols, phenolic compounds, tocopherols, hydrocarbons and
83 pigments has also successfully been applied for the classification of olive oils according
84 to their geographical origin (Alonso-Salces et al., 2010; Arslan et al., 2013; Longobardi,
85 Ventrella, Casiello, Sacco, Catucci, et al., 2012). Alternatively, authentication can also
86 be carried out by evaluating the mineral composition of olive oils, which is very
87 informative since it is affected by multiple factors. First, the multi-element profile of
88 plant-derived products primarily depends on the biological demand of the plant as well
89 as on the bioavailability and mobility of mineral compounds from soil. However, some
90 elements can also modify their concentrations in response to agronomic practices (e.g.
91 use of fertilizers and pesticides), climatic factors and manufacturing processes. Thus,
92 mineral elements can be considered as good markers for tracing the geographical
93 origin of virgin olive oils. Furthermore, it should be noted that metal content can be
94 determined by using simple and rapid analytical techniques, unlike tedious time-

consuming chromatographic methods needed to analyze organic compounds.

Nowadays, inductively coupled plasma mass spectrometry (ICP-MS) and inductively coupled plasma optical emission spectrometry (ICP-OES) are the most commonly used analytical platforms for obtaining elemental fingerprints. In this context, these techniques have successfully been applied to differentiate olive oils from eight European sites (Camin, Larcher, Nicolini, et al., 2010), four Western Greek islands (Karabagias et al., 2013), and four municipalities from Huelva province, in Southern Spain (Beltrán, Sánchez-Astudillo, Aparicio, & García-González, 2015).

High throughput analytical techniques, such as ICP-MS/OES, generate large data sets that require the use of advanced chemometric tools in order to extract the maximum amount of useful information. Several supervised pattern recognition procedures, such as linear discriminant analysis, soft independent model class analogy or partial least squares discriminant analysis, have frequently been used in food analysis to solve authentication problems (Berrueta, Alonso-Salces, & Héberger, 2007; Roberts & Cozzolino, 2016). Complementarily, new machine learning algorithms such as random forest, artificial neural networks and support vector machines have demonstrated excellent performance during the last decade for the analysis of complex datasets in many research areas, including food science (Batista et al., 2012; Palacios-Morillo, Jurado, Alcázar, & Pablos, 2016).

The aim of this study was to evaluate the performance of several machine-learning algorithms to discriminate extra virgin olive oils (EVOOs) from different origins by investigating the multi-elemental profile. For this purpose, the present work considers 125 samples from multiple locations across the Spanish geography, which were fingerprinted by using ICP-MS and ICP-OES. Then, three complementary supervised

pattern recognition techniques, including partial least squares discriminant analysis (PLS-DA), support vector machine (SVM) and random forest (RF), were applied in order to build statistical models with the aim to discriminate between olive oil groups.

2. Materials and methods

2.1. Extra virgin olive oil samples

Extra virgin olive oil samples from multiple Spanish locations (N=125) were kindly provided by local oil mill stores from each geographical area. Figure 1 shows the sampling zones investigated, together with the number of samples from each, which included three different regions with characteristic climate and geochemistry. The first study area was the province of Huelva, which is characterized by a Mediterranean oceanic climate due to its proximity to the Atlantic Ocean. The second group included olive oil samples from Granada, Málaga, Valencia, Mallorca and Tarragona, all of them coastal areas with a typical Mediterranean climate. Finally, samples from inland provinces (i.e. Córdoba, Jaén, Sevilla, Ciudad Real, Navarra and Lleida) constituted the third study group.

2.2. Sample mineralization

Microwave-assisted acid digestion was performed to mineralize oil samples prior to multi-elemental analysis, following the methodology described by Beltran *et al.* (2015). The digestion was carried out using an Anton Paar microwave oven (Multiwave 3000 SOLV) with programmable power control. For this, 0.5 g of olive oil sample was weighed directly into the digestion vessel and mixed with 5 mL of nitric acid (65%, v/v), 3 mL of hydrogen peroxide (30%, v/v) and 1 mL of hydrochloric acid (37%, v/v) (Sigma-Aldrich, Madrid, Spain). Mineralization was accomplished by ramping temperature until reach 280 ºC and 8000 kPa in 15 min. These conditions were maintained for 20

143 min with minimum ventilation. Afterwards, samples were vented for 15 min and
144 stored at 25 ºC for 12 h. Finally, digested samples were made-up to 25 mL with
145 ultrapure water obtained from a Milli-Q system (Millipore, Bedford, US).

146 **2.3. Multi-element analysis**

147 Forty seven elements (Table S1) were measured by inductively coupled plasma mass
148 spectrometry (ICP-MS) using the Agilent 7700X system (Agilent Technologies, Santa
149 Clara, US), equipped with nickel sampler and skimmer cones. Instrumental conditions
150 were daily optimized by using a tuning aqueous solution containing Li, Co, Y, and Tl at 1
151 µg L^{-1}. The forward power was set at 1.5 kW, and the argon flow rates were fixed at 15
152 L min^{-1} for plasma gas, 0.9 L min^{-1} for auxiliary gas and 1.1 L min^{-1} for carrier gas.
153 Samples were infused at 400 mL min^{-1} using a sampling depth of 9.0 mm, and the
154 nebulization chamber temperature was set at 2 ºC. On the other hand, major elements
155 (i.e. Al, Ca, Fe, K, Mg, Mn, Na, Ti) were determined using a Jobin-Yvon Ultima 2
156 inductively coupled plasma optical emission spectrometer (ICP-OES) equipped with
157 Ultrasonic nebulizer (U6000 AT+, Cetac). The instrument operated using the following
158 conditions: RF frequency, 27 MHz; operating power, 1200 W; plasma argon flow rate,
159 12 L min^{-1}; sample flow rate, 1 mL min^{-1}; and nebulization pressure, 3 bar. Analysis of
160 each sample was done in triplicate and results were expressed as the average of these
161 measurements.

162 **2.4. Analytical quality control.**

163 The methodology was validated according to the UNE 82009-1:1998 normative. The
164 within-laboratory repeatability (within-day precision) and within-laboratory
165 reproducibility (day-to-day precision) were assessed by analyzing an extract six times
166 within the same day and over a period of 1 month in duplicate, respectively.

Calibration curves were built at five concentration levels by diluting a multi-element standard solution from SCP Science (Paris, France) with 10% HNO_3 in ultrapure water. The concentration range was 0.2-60 ng mL^{-1} for all the elements, excepting Ba, Ca, Sr, V and Zr, which were calibrated in a wider range (10-200 ng mL^{-1}). Recoveries were determined in a control sample, prepared by mixing 5 g of each individual oil sample, spiked with a multi-element standard solution.

2.5. Data analysis

The chemometric evaluation of data was performed by means of the application of several pattern recognition techniques, with the aim to explore the relationship between the mineral composition of olive oils and their geographical origins. The investigated data matrix comprised 125 rows (olive oil samples) and 55 columns (minerals determined by ICP-MS/OES). Multivariate data analysis and non-parametric analysis of variance (Kruskal-Wallis test) were performed using the Statistica 8.0 software (StatSoft, Tulsa, UK); while partial least square discriminant analyses (PLS-DA) were accomplished using the SIMCA-P v11.5 software (UMETRICS, Umeå, Sweden). The Kruskal-Wallis test was used to provide a first evaluation of the discriminant efficiency of variables and to find out statistical differences in elemental content between olive oil origins. This non parametric method was employed since most of the variables showed a skewed distribution (checked by normal probability plots) and variances were not homogeneous (checked by Levene's test). Then, results were explored by means of principal components analysis (PCA) in order to get an overview of data and to identify possible outliers and trends towards the grouping of samples. Complementarily, supervised learning methods (linear discriminant analysis, partial least square discriminant analysis, support vector machines and random forest) were

191 also applied for building predictive models, which were subsequently compared for
192 evaluating the recognition and prediction abilities of each one. Linear discriminant
193 analysis (LDA) is a supervised classification tool based on the generation of a number
194 of orthogonal linear discriminant functions equal to the number of categories minus
195 one. The discriminant power of each variable is evaluated by measuring the value of
196 the Wilks´ lambda parameter for the overall model after removing the selected
197 variable. Then, a forward stepwise algorithm is used to select the variables to be
198 included in the model. According to this algorithm, the F value is used as criterion for
199 inclusion or removal of variables in the model. Thus, Wilks´ lambda and F values were
200 used to check the significance of each predictor in LDA analysis accomplished in the
201 present work.
202 Partial least square discriminant analysis (PLS-DA) is a supervised pattern recognition
203 method based on searching an optimal set of latent variables (or components) to
204 discriminate between the previously defined categories. The PLS-DA method consists
205 of a classical PLS regression where the dependent variable is categorical and
206 represents the sample class membership, which can be conveniently used when the
207 number of objects is fewer than the number of variables. The principle of PLS is to find
208 the components in the matrix ***X*** (matrix of predictors) which describe, as much as
209 possible, the relevant variations in the input variables and, at the same time, have
210 maximal correlation with the target value in ***Y*** (matrix of responses), giving less weight
211 to the variations that are irrelevant or noisy (Berrueta et al., 2007). That is, PLS
212 searches for a set of components that performs simultaneous decomposition of X and
213 Y with the constraint that these components explain, as much as possible, the
214 covariance between X and Y. The optimal number of components is usually achieved

215 by cross-validation techniques, so that the overall quality of PLS-DA models is
216 described by R_X^2 and Q^2 values. R_X^2 is defined as the proportion of variance in the data
217 explained by the models and indicates goodness of fit, while Q^2 is a measurement of
218 the predictive ability of the model.
219 On the other hand, the support vector machine is a non-parametric machine learning
220 technique applicable for both classification and regression problems (Brereton, Richard
221 G., 2010). The SVM looks for an optimal separating hyperplane that will form the
222 boundary between classes. This optimal boundary is found by considering a subset of
223 samples, corresponding to the support vectors, which lie at the very border between
224 the classes so that the distance between them is maximum. This hyperplane is
225 optimized by means of an iterative algorithm that minimizes an error function.
226 According to the form of the error function, SVM models can be classified in two
227 groups: SVM type 1, also known as C-SVM, and SVM Type 2 or ν-SVM. These
228 parameters, C (capacity constant) and ν, can be estimated using a cross-validation
229 algorithm, thus controlling the complexity of the model by avoiding overtraining.
230 Furthermore, SVM can be applied to solve non-linear problems where the optimal
231 boundary between classes cannot be represented by a hyperplane. In these cases, it is
232 possible to linearly discriminate the original data projecting them in a new higher
233 dimensional space, using a set of mathematical functions (i.e. Kernel functions) in
234 which a linear solution is possible, being the Gaussian and polynomial functions the
235 most popular Kernel functions. Therefore, the optimization of training constants (i.e. C
236 or ν) and Kernel parameters (i.e. γ) is crucial for obtaining optimal SVM models suitable
237 to classify a new sample (Capron, Massart, & Smeyers-Verbeke, 2007; Zomer, Del
238 Nogal Sánchez, Brereton, & Pérez Pavón, 2004).

239 Finally, random forest (RF) is a non-parametric and non-linear classification and
240 regression algorithm based on a learning strategy called ensemble learning, which
241 operates by generating many decision trees on bootstrap samples taken from original
242 data and aggregating their results (Breiman, 2001). Trees are split to many nodes using
243 different subsets of randomly selected input variables (m). Thus, the main parameters
244 for RF model are the value of m and the number of decision trees.

245 **3. Results and discussion**

246 **3.1. Characteristic multi-element profile for Spanish olive oils**

247 In this study, the mineral composition of 125 EVOOs from multiple Spanish geographic
248 areas was determined by ICP-MS/OES analysis. Descriptive statistical analyses for the
249 55 elements determined in these olive oil samples are summarized in Table S1. This
250 method provided excellent intra- and inter-day precision, with relative standard
251 deviations below 6.4% and 5.7%, respectively. Furthermore, recovery rates were in the
252 range 82-110%.

253 Calcium, iron, magnesium, vanadium, chromium, copper, zirconium, antimony,
254 hafnium and lead were found in quantifiable amounts in all samples analyzed (Table
255 S1). In contrast, thallium and cesium could be only quantified in approximately half the
256 samples, while significant concentrations of titanium, beryllium and selenium were
257 detected in fewer than half. Calcium was the most abundant element (8.1 $\mu g\ kg^{-1}$),
258 accounting for 37% of the total mineral fraction, in agreement with results previously
259 published by Zeiner et al. (2010) and Karabagias et al. (2013) for Croatian and Greek
260 olive oils, respectively. Secondly, potassium and sodium represented 19.8% and 20.7%
261 of the olive oil element profile, with average contents of 4.3 and 4.5 $\mu g\ kg^{-1}$, values
262 slightly higher than those described by Camin et al. (2010) and Zeiner et al. (2010).

263 Chromium, aluminum, boron, iron, magnesium and vanadium were present at
264 moderate concentrations in olive oil samples, averaging 1.018, 0.749, 0.698, 0.682,
265 0.595 and 0.261 μg kg^{-1}, respectively, within the ranges reported in literature (Beltrán
266 et al., 2015; Camin, Larcher, Perini, et al., 2010; Karabagias et al., 2013; Zeiner,
267 Juranovic-Cindric, & Škevin, 2010). The remaining mineral elements determined in this
268 study showed low average values (ranging from 0.218 to 185.02 μg Kg^{-1}), thus
269 representing less than 4% of the total mineral content. In this context, mean levels of
270 essential elements manganese, selenium, cobalt and molybdenum were 100.5, 6.46,
271 1.95 and 44.33 μg Kg^{-1}, respectively. Mn and Co contents agreed with those found by
272 Camin et al. (2010), while Mo and Se concentrations differed from that depicted in
273 previous studies, probably as a consequence of the great influence of the geographical
274 origin of the investigated oil samples (Camin, Larcher, Perini, et al., 2010; Zeiner et al.,
275 2010). On the other hand, levels of toxic elements such as arsenic, cadmium, nickel
276 and lead were significantly lower than those reported in other works (Camin et al.,
277 2010; Zeiner et al., 2010; Karabagias et al., 2013). Moreover, it should be noted that
278 Pb, As and Cd concentrations were below the limits established by the World health
279 Organization in all samples analyzed (10 μg g^{-1}, 4 μg g^{-1} and 1 μg g^{-1} for Pb, As and Cd,
280 respectively). To conclude, other elements such as lithium, rubidium and cesium were
281 found in similar concentration ranges to that described in literature (Camin et al.,
282 2010), while lanthanum, cerium, samarium, europium ytterbium and uranium were
283 detected in greater quantities than those previously reported, thus evidencing the
284 particular geochemistry of the Spanish regions investigated in the present work.

285 **3.2 Exploratory data analysis**

As a first exploratory step, principal component analysis (PCA) and linear discriminant analysis (LDA) were applied for a preliminary evaluation of data quality. PCA is an unsupervised method that allows to get an overview of the data and to identify possible outliers and trends towards the grouping of samples. The application of this statistical tool to the data matrix generated by ICP-MS/OES analysis yielded two latent variables with eigenvalues higher than 1 (Kaiser criterion). Figure 2A shows the distribution of samples in the plane defined by these two principal components, which explained 41.2 and 15.8 % of the original variance, respectively. Thus, it can be clearly observed that samples were clustered in two groups, the first one comprising oil samples from the province of Huelva, located in the right side of the graph, while the second cluster showed a greater dispersion and included oils from other Spanish regions. Afterwards, linear discriminant analysis (LDA) was applied with exploratory purposes in order to find homogeneous groups of olive oil samples according to origin. Considering each investigated location as an independent group, a clear distinction among olive oil samples from the province of Huelva and the rest could be observed (Figure 2B). However, this second cluster of samples also showed two slightly differentiated groups corresponding to oils from Mediterranean coastal provinces and those from interior provinces (Figure 2B). Therefore, data were again analyzed by means of LDA using three grouping criteria as mentioned above: Huelva, Mediterranean coast and interior. Thereby, as can be seen in Figure 2C, a good separation of the three groups of oil samples was observed.

3.3. Differential element profile of olive oils according to origin

Considering results obtained after exploratory multivariate statistics, olive oil samples were grouped into three classes on the basis of their proximity to the sea. Then, non-

parametric ANOVA was applied with the aim to identify significant differences in the element profile according to the origin. Table 1 lists the mean concentration values for the 55 mineral elements determined in 125 Spanish extra virgin olive oils gathered according to the previously described criterion: Huelva, Mediterranean coast and interior provinces. Statistical differences between the three studied olive oil groups were investigated by using the Kruskal-Wallis non-parametric multiple comparison test. This test calculates the H parameter by comparison with the chi-squared distribution for n-1 degrees of freedom (being n the number of groups; n=3) and α=0.05. Then, when significant differences were detected, a post hoc comparison was used to highlight the pairs of groups responsible of those differences (Muth, 2014). Table 1 also shows the calculated H-statistic for the 55 monitored elements, as well as the post hoc analysis results. It is noteworthy that most of these elements were statistically significant, with calculated H-values exceeding the critical value of chi-squared, i.e. χ^2 (2, 0.05) = 5.99, except for Ga, Sr, Y, Nb, Tl and Bi. The lowest H values were found for Zn, Al and Ca, which presented significant differences between olive oil samples from Huelva and inland provinces. On the other hand, iron levels differentiated oil samples from inland provinces with respect to coastal ones; while cooper and titanium concentrations were statistically different between Mediterranean samples and those from the other two study areas. The remaining elements enabled the discrimination between samples from Huelva and the other two considered groups. The highest H values were observed for V, Cr, Ge, As, Se, Tm, Lu, Hf and Ta (H>80.0), which presented lower concentrations in oils coming from Huelva compared with samples collected in Mediterranean coastal regions and inland provinces. On the basis of these differences, it is noteworthy that the mineral content

could be considered as a suitable traceability marker in food research. In this sense, it has previously been demonstrated that rare earth elements are robust geographical origin tracers of olive oil since they are representative of the soil, and their concentrations do not change over time (Farmaki et al., 2012). However, it should be taken into account that many of these variations can also be influenced by other sources such as phytochemicals, fertirrigation and contamination deposition.

3.4. Comparison of pattern recognition techniques to classify olive oil samples

In view of results obtained by means of exploratory multivariate statistics as well as by Kruskal-Wallis multiple comparison, mineral compounds could be considered as good descriptors to build classification models with the aim to differentiate extra virgin olive oils according to their geographical origin. Accordingly, complementary multivariate classification approaches (i.e. PLS-DA, SVMs and Random Forest) were tested in the present study. A critical issue in any empirical modelling is to determine the correct complexity of the model by selecting the optimum number of variables to avoid overfitting (Cozzolino, Cynkar, Shah, & Smith, 2011; Roberts & Cozzolino, 2016). For this reason, validation procedures have to be applied in order to assess the predictive ability of the model. In the absence of a real validation set, cross-validation is a practical and reliable way to test how well the model predicts new data. In cross-validation, the data matrix is randomly divided in two sets, a training set used to construct the classification model and a test set applied to study the model performance, both of them containing the same percentage of samples within each class. Random division is applied in order to ensure that all samples have the same probability of belonging to the training set and, consequently, that the constructed model will take into account the within-class variability. This procedure, called k-fold

358 cross-validation, is repeated several times to avoid conclusions by chance. Then, the
359 performance of the model can be evaluated by computing its sensitivity (SENS) and
360 specificity (SPEC), where SENS refers to the percentage of cases belonging to a
361 determinate class correctly classified and SPEC refers to the percentage of cases not
362 belonging to a class correctly not classified in this class (Ceballos-Magaña et al., 2012).
363 In the present work, we first applied supervised PLS-DA in order to sharpen the
364 separation between study groups and to assess the influence of sample geographical
365 origin in the mineral content. Using the complete data matrix, a three-component
366 model was obtained with good quality of fit and validity (R^2 = 0.67) as well as
367 prediction (Q^2 = 0.646). Furthermore, as shown in Figure 3A, a good separation
368 between the three geographical groups was obtained by projecting the data in the
369 space defined by these three components. The first latent variable allowed to
370 discriminate between samples from Huelva (positioned on the right side of the plot,
371 with negative scores) and the other two groups (positioned on the left side of the plot,
372 with positive scores), which were slightly separated by the third component.
373 Thereafter, two-class comparisons were also performed in order to identify potential
374 markers for each origin. The first two models compare olive oils from Huelva with
375 samples from the other two study areas, Mediterranean coast and inland regions. In
376 both cases, satisfactory values for quality parameters R^2 and Q^2 were obtained, with
377 variance explained above 96% and variance predicted around 67%. Moreover, a third
378 model was also developed with data for Mediterranean and inland olive oils, rending
379 lower R^2 and Q^2 values (0.60 and 0.35, respectively), in accordance with the worst
380 separation observed between these two clusters of samples in the three-group PLS-DA
381 model (Fig. 3A). Figures 3B-D shows the scores plots obtained for these two-class

382 statistical models, demonstrating a satisfactory separation according to the origin. On
383 the other hand, the evaluation of PLS-DA loadings plots allowed a good understanding
384 of the variables that contributed to the discrimination of classes. In this sense, only
385 variables with VIP (variable importance for the projection) values higher than 1.0 were
386 selected as statistically significant. Thereby, the examination of these loadings plots
387 demonstrated that levels of elements such as Mg, Mn, K, B, Co and Ni were increased
388 in olive oil samples from Huelva compared to those from the Mediterranean coast and
389 inland regions, in accordance with data tabulated in Table 1. The loadings plot
390 comparing Huelva and Inland olive oils also showed an enrichment of iron in the first
391 group of samples, while high levels of Fe, Ti, Al, Hf, Mn, Ca and K were found in
392 samples from the Mediterranean coast compared to those from inland provinces.
393 Despite PLS-DA is the most well-known tool to perform statistical classification and
394 regression due to its capability to deal with multicollinearity and its robustness to
395 missing data and skew distributions, it can also provide misleading results due to a lack
396 of suitable statistical validation (Gromski et al., 2015). For this reason, in this study we
397 also employed other two complementary supervised learning methods, including
398 support vector machines and random forests. Multiple SVM classification models were
399 constructed by using 60% and 75% random division of the input data into training and
400 test set (Table 2). Furthermore, we compared C-SVM and ν-SVM modelling, as well as
401 four different kernel functions, including sigmoid, polynomial, linear and radial basis
402 functions (RBF). In order to control the complexity of the models and to avoid
403 overtraining, C, ν and γ values were optimized by cross-validation. The gamma value
404 was 0.018 in non-linear SVM models, while C and ν parameters are listed in Table 2.
405 Moreover, the accuracy rates in the training, testing and overall set were computed as

406 evaluation criteria for the goodness of classification. As can be seen in Table 2, the
407 selection of the proper kernel function has high influence on the performance of SVM
408 models. Thereby, best results were obtained by applying the linear kernel function,
409 with overall accuracy values between 95% and 98%. However, it should be noted that
410 the best prediction ability was provided by the C-SVM model using the RBF and a 1/4
411 of the inputs as test data, thus obtaining a 100% of correctly classified samples in the
412 test set.
413 Finally, random forest analysis was also applied because of its ability to provide a
414 measure of the predictor variables importance, high resistance to overfitting,
415 applicability in data matrices with low samples/features ratio, and the non-
416 requirement of scaling techniques prior to analysis (Gromski et al., 2015). Random
417 forest is performed by growing several decision trees on bootstrap samples and then
418 selecting the best split at each node among a random selection of predictor variables.
419 The number of predictors to be selected at each node (m) and the number of trees to
420 be grown are the most important parameters to be considered, but the use of default
421 values usually provide good results (Breiman, 2001). For constructing RF models with
422 data from our multi-element profiles, the number of trees was varied from 10 to 100,
423 and the number of selected features was set as the square root of the total number of
424 features. In order to examine the influence of the number of decision trees, the
425 misclassification rates in training and test sets were considered as evaluation criteria.
426 Accordingly, 35 decision trees yielded the lowest misclassification rate and allowed a
427 good classification of olive oils by using seven predictors for each node. Thereby, we
428 obtained excellent accuracy rates for both training and test sets (92.30% and 93.02%,
429 respectively). Furthermore, a list of discriminant variables ranked according to their

430 importance for classification in this RF model was also obtained, being Hf, K and Cs the
431 most important classifiers in agreement with results provided by PLS-DA.
432 To conclude, we compared the statistical performance of the three pattern recognition
433 tools here employed by computing the sensitivity (SENS) and specificity (SPEC)
434 parameters from confusion matrixes. Sensitivity and specificity mean values for each
435 study group (i.e. Huelva, Mediterranean coast, inland provinces) as well as for the
436 overall dataset are listed in Table 3. It is noteworthy that only olive oils from Huelva
437 showed SENS and SPEC values of 100% by using the three classification models. On the
438 other hand, the majority of misclassified samples were allocated to the inland class,
439 and for this reason the SPEC computed in this class was lower for the three models,
440 ranging from 93.4% to 98.0%. With regards to Mediterranean oil samples, SENS was
441 below 80% for all the models, while SPEC was 100% for PLS-DA and SVM models, and
442 98% for the RF model. Taking into account overall results, it could be concluded that RF
443 and SVM provide higher sensitivity than that observed in PLS-DA models, but
444 specificity was slightly decreased. Finally, it should be also noted that models built in
445 this work provided slightly higher classification rates than those reported in previously
446 published studies (ranging from 82.6-95%) (Camin, Larcher, Nicolini, et al., 2010;
447 Karabagias et al., 2013; Beltrán, Sánchez-Astudillo, Aparicio, & García-González, 2015),
448 especially for samples coming from Huelva and interior provinces (SENS=100%).

449 **4. Conclusions**

450 This work demonstrates the potential of multi-element fingerprinting to classify extra
451 Spanish virgin olive oils according to their origin. Samples from the Atlantic coast (i.e.
452 Huelva province) could be clearly differentiated from the rest of locations herein
453 studied, probably as a consequence of the characteristic geochemistry of this area. On

the other hand, olive oils produced in Mediterranean and inland provinces showed similar mineral profiles, and significant differences were only observed for iron and titanium levels. On this basis, several complementary data mining methods were subsequently applied to assess their utility in food traceability and authenticity research. Partial least squares discriminant analysis (PLS-DA), support vector machine (SVM) and random forest (RF) statistical models were developed and validated using the multi-element dataset, thus enabling to differentiate extra virgin olive oils according to their geographical origin.

Acknowledgements. This work was supported by the *Consejería de Innovación, Ciencia y Empresa, Junta de Andalucía* (grant number P10-FQM-6185). Authors also thank to Prof. Jesús de la Rosa for his selfless assistance in the interpretation of results.

Conflict of interest. Authors have no conflict of interest to declare.

References

Agiomyrgianaki, A., Petrakis, P. V., & Dais, P. (2012). Influence of harvest year, cultivar and geographical origin on Greek extra virgin olive oils composition: A study by NMR spectroscopy and biometric analysis. Food Chemistry, 135(4), 2561–2568. https://doi.org/10.1016/j.foodchem.2012.07.050

Aguilera, Y., Martin-Cabrejas, M. A., & González de Mejia, E. (2016). Phenolic compounds in fruits and beverages consumed as part of the mediterranean diet: their role in prevention of chronic diseases. Phytochemistry Reviews, 15(3), 405–423. https://doi.org/10.1007/s11101-015-9443-z

Alonso-Salces, R. M., Héberger, K., Holland, M. V., Moreno-Rojas, J. M., Mariani, C., Bellan, G., … Guillou, C. (2010). Multivariate analysis of NMR fingerprint of the unsaponifiable fraction of virgin olive oils for authentication purposes. Food Chemistry, 118(4), 956–965. https://doi.org/10.1016/j.foodchem.2008.09.061

Aparicio, R., & Harwood, J. (2013). Handbook of Olive Oil: Analysis and Properties. (R. Aparicio & J. Harwood, Eds.) (2nd ed.). New York: Springer New York LLC. https://doi.org/10.1007/978-1-4614-7777-8

Arslan, D., Karabekir, Y., & Schreiner, M. (2013). Variations of phenolic compounds, fatty acids and some qualitative characteristics of Sariulak olive oil as induced by growing area. Food Research International, 54(2), 1897–1906. https://doi.org/10.1016/j.foodres.2013.06.016

Batista, B. L., da Silva, L. R. S., Rocha, B. A., Rodrigues, J. L., Berretta-Silva, A. A., Bonates, T. O., … Barbosa, F. (2012). Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques. Food Research International, 49(1), 209–215. https://doi.org/10.1016/j.foodres.2012.07.015

Beltrán, M., Sánchez-Astudillo, M., Aparicio, R., & García-González, D. L. (2015). Geographical traceability of virgin olive oils from south-western Spain by their multi-elemental composition. Food Chemistry, 169, 350–357. https://doi.org/10.1016/j.foodchem.2014.07.104

Ben-Hassine, K., Taamalli, A., Ferchichi, S., Mlaouah, A., Benincasa, C., Romano, E., … Hammami, M. (2013). Physicochemical and sensory characteristics of virgin olive oils in relation to cultivar, extraction system and storage conditions. Food Research International, 54(2), 1915–1925. https://doi.org/10.1016/j.foodres.2013.09.007

501 Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern
502 recognition in food analysis. Journal of Chromatography A, 1158(1–2), 196–214.
503 https://doi.org/10.1016/j.chroma.2007.05.024
504 Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
505 https://doi.org/10.1023/A:1010933404324
506 Brereton, Richard G., G. R. L. (2010). Support Vector Machines for Classification and
507 Regression. The Analyst, 135(2), 230–267. https://doi.org/10.1039/B918972F
508 Camin, F., Larcher, R., Nicolini, G., Bontempo, L., Bertoldi, D., Perini, M., … Hoogewerff,
509 J. (2010a). Isotopic and elemental data for tracing the origin of European olive oils.
510 Journal of Agricultural and Food Chemistry, 58(1), 570–577.
511 https://doi.org/10.1021/jf902814s
512 Camin, F., Larcher, R., Perini, M., Bontempo, L., Bertoldi, D., Gagliano, G., … Versini, G.
513 (2010b). Characterisation of authentic Italian extra-virgin olive oils by stable isotope
514 ratios of C, O and H and mineral composition. Food Chemistry, 118(4), 901–909.
515 https://doi.org/10.1016/j.foodchem.2008.04.059
516 Capron, X., Massart, D. L., & Smeyers-Verbeke, J. (2007). Multivariate authentication of
517 the geographical origin of wines: A kernel SVM approach. European Food Research
518 and Technology, 225(3–4), 559–568. https://doi.org/10.1007/s00217-006-0454-2
519 Ceballos-Magaña, S. G., Jurado, J. M., Muñiz-Valencia, R., Alcázar, A., de Pablos, F., &
520 Martín, M. J. (2012). Geographical Authentication of Tequila According to its Mineral
521 Content by Means of Support Vector Machines. Food Analytical Methods, 5(2), 260–
522 265. https://doi.org/10.1007/s12161-011-9233-1
523 Cozzolino, D., Cynkar, W. U., Shah, N., & Smith, P. (2011). Multivariate data analysis
524 applied to spectroscopy: Potential application to juice and fruit quality. Food

525 Research International, 44(7), 1888–1896.
526 https://doi.org/10.1016/j.foodres.2011.01.041
527 Farmaki, E. G., Thomaidis, N. S., Minioti, K. S., Ioannou, E., Georgiou, C. A., &
528 Efstathiou, C. E. (2012). Geographical Characterization of Greek Olive Oils Using Rare
529 Earth Elements Content and Supervised Chemometric Techniques. Analytical Letters,
530 458(February 2017), 920–932. https://doi.org/10.1080/00032719.2012.655656
531 Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., & Goodacre,
532 R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant
533 analysis - a marriage of convenience or a shotgun wedding. Analytica Chimica Acta,
534 879, 10–23. https://doi.org/10.1016/j.aca.2015.02.012
535 Karabagias, I., Michos, C., Badeka, A., Kontakos, S., Stratis, I., & Kontominas, M. G.
536 (2013). Classification of Western Greek virgin olive oils according to geographical
537 origin based on chromatographic, spectroscopic, conventional and chemometric
538 analyses. Food Research International, 54(2), 1950–1958.
539 https://doi.org/10.1016/j.foodres.2013.09.023
540 Laroussi-Mezghani, S., Vanloot, P., Molinet, J., Dupuy, N., Hammami, M., Grati-
541 Kamoun, N., & Artaud, J. (2015). Authentication of Tunisian virgin olive oils by
542 chemometric analysis of fatty acid compositions and NIR spectra. Comparison with
543 Maghrebian and French virgin olive oils. Food Chemistry, 173, 122–132.
544 https://doi.org/10.1016/j.foodchem.2014.10.002
545 Longobardi, F., Ventrella, A., Casiello, G., Sacco, D., Catucci, L., Agostiano, A., &
546 Kontominas, M. G. (2012b). Instrumental and multivariate statistical analyses for the
547 characterisation of the geographical origin of Apulian virgin olive oils. Food
548 Chemistry, 133(2), 579–584. https://doi.org/10.1016/j.foodchem.2012.01.059

549 Longobardi, F., Ventrella, A., Casiello, G., Sacco, D., Tasioula-Margari, M., Kiritsakis, A.
550 K., & Kontominas, M. G. (2012a). Characterisation of the geographical origin of
551 Western Greek virgin olive oils based on instrumental and multivariate statistical
552 analysis. Food Chemistry, 133(1), 169–175.
553 https://doi.org/10.1016/j.foodchem.2011.09.130
554 Luykx, D. M. A. M., & van Ruth, S. M. (2008). An overview of analytical methods for
555 determining the geographical origin of food products. Food Chemistry, 107(2), 897–
556 911. https://doi.org/10.1016/j.foodchem.2007.09.038
557 Martínez-González, Miguel Angel; Sanchez-Villegas, A. (2004). The emerging role of
558 Mediterranean diets in cardiovascular ... European Journal of Epidemiology, 19, 9–13.
559 Mihailova, A., Abbado, D., Kelly, S. D., & Pedentchouk, N. (2015). The impact of
560 environmental factors on molecular and stable isotope compositions of n-alkanes in
561 Mediterranean extra virgin olive oils. Food Chemistry, 173, 114–121.
562 https://doi.org/10.1016/j.foodchem.2014.10.003
563 Muth, J. E. (2014). Basic Statistics and Pharmaceutical Statistical Applications. (J. E.
564 Muth, Ed.) (3rd ed.). CRC Press.
565 Palacios-Morillo, A., Jurado, J. M., Alcázar, A., & Pablos, F. (2016). Differentiation of
566 Spanish paprika from Protected Designation of Origin based on color measurements
567 and pattern recognition. Food Control, 62, 243–249.
568 https://doi.org/10.1016/j.foodcont.2015.10.045
569 Roberts, J. J., & Cozzolino, D. (2016). An Overview on the Application of Chemometrics
570 in Food Science and Technology-An Approach to Quantitative Data Analysis. Food
571 Analytical Methods, 9(12), 3258–3267. https://doi.org/10.1007/s12161-016-0574-7

572 Romero, N., Saavedra, J., Tapia, F., Sepúlveda, B., & Aparicio, R. (2016). Influence of
573 agroclimatic parameters on phenolic and volatile compounds of Chilean virgin olive
574 oils and characterization based on geographical origin, cultivar and ripening stage.
575 Journal of the Science of Food and Agriculture, 96(2), 583–592.
576 https://doi.org/10.1002/jsfa.7127
577 Zeiner, M., Juranovic-Cindric, I., & Škevin, D. (2010). Characterization of extra virgin
578 olive oils derived from the Croatian cultivar Oblica. European Journal of Lipid Science
579 and Technology, 112(11), 1248–1252. https://doi.org/10.1002/ejlt.201000006
580 Zomer, S., Del Nogal Sánchez, M., Brereton, R. G., & Pérez Pavón, J. L. (2004). Active
581 learning support vector machines for optimal sample selection in classification.
582 Journal of Chemometrics, 18(6), 294–305. https://doi.org/10.1002/cem.872
583
584
585 **Figure Captions**
586 **Figure 1.** Geographical origin and number of extra virgin olive oil samples studied in
587 this work.
588 **Figure 2.** Scores plots of the exploratory statistical models used to get a preliminary
589 overview of data, showing the distribution of samples in the plane defined by two first
590 principal components. (A) Principal component analysis (PCA); (B) linear discriminant
591 analysis (LDA) considering all the investigated locations as independent groups; (C) LDA
592 considering only three geographical groups: Huelva, Mediterranean coast and inland
593 provinces.
594 **Figure 3.** PLS-DA scores plots showing the distribution of samples in the plane defined
595 by three first principal components. (A) all samples; (B) Huelva vs Mediterranean coast

596 samples; (C) Huelva vs inland province samples; (D) Mediterranean coast vs inland

597 province samples.

598