Assessment of vibrational spectroscopy performance in geographical identification of virgin olive oils: A world level study.

Noelia Tena,¹ Ramón Aparicio,¹ Vincent Baeten,² Diego Luis García-González,^{*1} Juan Antonio Fernández-Pierna²

¹ Instituto de la Grasa (CSIC), Ctra. de Utrera, km. 1, Campus Universitario Pablo de Olavide - building 46, 41013 - Sevilla, Spain.

² Valorisation of Agricultural Products Department, Food and Feed Unit, Walloon Agricultural Research Centre (CRA-W), Henseval Building, Chaussée de Namur 24, 5030 Gembloux, Belgium.

*corresponding author: dlgarcia@ig.csic.es; Phone: +34954616790.

Abstract

In order to guarantee food integrity, testing methods should also include the tools needed to verify the geographical origin described on the label. In the particular case of virgin olive oil (VOO), regulatory bodies have yet to establish a standard for geographical identification. This manuscript described a procedure based on Raman spectroscopy to identify the provenance of these oils based on a classification criterion (European vs. non-European VOOs). The sample collection was considered a relevant step in which multiple factors were taken into account (cultivar distribution, complexity of production in each location, new agricultural practices, and new/old cultivars). A total amount of 78 virgin olive oils collected around the world allowed checking the classification accuracy of the model for identifying sample origin. The PLS-model built around the Raman spectroscopic data was validated with a blind set of samples, and the results were evaluated in terms of false negative and false positive quality parameters. This study provides an analytical application to detect mislabelling and fraudulent practices related to the geographical provenance of the virgin olive oils declared on the labels.

Practical applications: The results of this manuscript provide a robust mathematical model to assess the origin of virgin olive oils (EU vs. non-EU) based on spectroscopic data. Nowadays, the extensive knowledge of olive oil chemical composition has proven that this composition varies according to pedoclimatic conditions, while non-targeted methods can be proposed in geographical traceability for their ability to analyse the chemical profile of samples. However, it was deemed necessary to conduct a study through a collaborative work initiative (European FoodIntegrity project, grant agreement No. 613688). Thus, this study provides a strict evaluation of Raman spectroscopy through a defined strategy for evaluating non-targeted methods for geographical identification.

Keywords: Geographical authenticity, Raman spectroscopy, Traceability, Virgin olive oil, Chemometric analysis.

INTRODUCTION

Virgin olive oils (VOOs) have organoleptic and nutritional properties that give them access to a relatively highly priced market compared with most of the other vegetable oils. In recent years, the world's production and consumption of VOOs has been increasing due to greater consumer knowledge of the quality and nutritional advantages of this premium oil. Olive oil production has been traditionally associated to Mediterranean basin countries, but this association is not completely accurate today. Thus, new production areas have been set up in locations from two hemispheres with similar climates to those found in the Mediterranean basin. Thus, the total worldwide amount of productive olive trees is spread over more than 11 million hectares, including countries as far from its historical origin as Australia, USA, South Africa, Argentina, and Chile. In addition to the diversity of the geographical locations of the orchards, the enormous diversity of cultivars—more than 1,275 autochthonous olive cultivars have been identified—must be also considered as a source of variability.^[1] Furthermore, cultivars are sometimes associated to a particular geographical area and, therefore, to production specifications in terms of protected designation of origin (PDO) or protected geographical indication (PGI).^[2] Although these terms are only applicable in Europe, the geographical indications on labels is today a worldwide requirement demanded by consumers. Moreover, the oil provenance declared on the label has a major impact on the market price of the relevant virgin olive oils.^[3] In addition to cultivars, other factors such as climate, technological characteristics in production, and the application of new agricultural practices, make the challenge of characterizing the geographical provenance of a virgin olive oil (VOO) much more difficult.^[4,5] New high-density and super-high-density orchards (anywhere from 778 to 1,210 trees per acre) are planted worldwide (both in the Northern and Southern Hemispheres) with a few cultivars (e.g., Arbequina, Arbosana,

Koroneiki, Leccino),^[6] which means that the traditional association between cultivars and specific geographical regions is not so clear as it was in the past.

VOO composition depends on the above-mentioned variables: cultivars, technological characteristics in production, agricultural practices, climate and soils of orchards. In addition, it is well accepted that altitude and latitude, directly associated to climate, can affect the concentration of the chemical compounds of VOO.^[7] Previous studies have shown that the temperature in certain time frames during olive ripening and the presence of some metals in the soil significantly affect the activity of the enzymes responsible for olive oil chemical composition.^[4]

The importance of geographical origin of virgin olive oils is starting to be reflected in regulations. In this regard, European Regulation No 29/2012^[8] established that “the agricultural traditions, the local systems of extraction and the blending practices applied directly to produce marketable virgin olive oils may result in quite different taste and quality of virgin olive oil depending on their geographical origin”. This may also result in price differences within the same category that disturb the virgin olive oil market. In order to avoid this and ensure that consumers are not misled, by information concerning the geographical area in which olives are harvested and olive oil is obtained, accurate information on the oil provenance should be stated on the label of the package. In this connection, the European Union has established a mandatory regime for designations of origin, which should be restricted to extra virgin and virgin olive oils. Thus, regulation No. 29/2012 establishes that extra virgin olive oil and virgin olive oil shall bear a designation of origin on the labelling which consist of the mention of the Member State, the Union or the third country where the oil was originated.

The mandatory necessity of certifying the geographical origin of extra virgin olive oil and virgin olive oil requires developing adequate analytical methods capable of giving

scientific support to the European legislation thus verifying the information declared on the label. On the other hand, the geographical identification of olive oil is welcomed by olive oil producers who see this as a possibility of informing consumers about the genuineness of their products by guaranteeing the information declared on the label. Thus, incorrect product labelling, such as a false declaration of geographical origin, motivated by economic interests to increase profit, could result in a bad image for the product and in a loss of consumers' confidence. So, the development of an analytical procedure to determine/confirm VOO geographical origin is needed for consumer protection and authenticity. Despite the panoply of chemical parameters and analytical methods for virgin olive oil included in national and international regulations,^[9] no method has been established by regulatory bodies to be used as standard for geographical identification.

One of the requirements to develop new methods for geographical traceability is to define the objective -classification criterion- before optimizing sample collection, method set up and data treatment.^[10] Thus, a clear objective allows targeting the physical-chemical properties of the oils that have a clear relationship with the classification criterion under study. The actual meaning of geographical traceability can refer to multiple aspects as countries, regions, PDO or cultivars among others. Considering that 75% of the total production of this commodity around the world is European and taking into account the higher preference of consumers for European olive oils,^[11] this study has focused on developing models to identify European VOOs. This study aims to provide an analytical application to detect mislabelling and fraudulent practices related to the geographical provenance of virgin olive oils based on the label description.

Different approaches and strategies have been proposed to address the geographical origin characterization of VOOs. In the framework of the "SEXIA" project,^[12] an expert system was built to characterize the geographical origin of the most

representative European VOOs by means of a large database containing concentrations of major and minor chemical compounds. This project demonstrated that the geographical identification of samples can rely on chemical analyses and chemometric procedures.^[12,13] More recently, molecular markers, also known as DNA-based markers, have been explored for their suitability for geographic traceability of VOOs.^[14] The analytical approaches tested in previous works have shown that a considerable amount of chemical information, mostly obtained from chromatographic methods, is required for tackling geographical identification.^[14] Consequently, these approaches involve extensive laboratory work, which is usually time-consuming, destructive, and requires the use of reagents. Although different criteria have been studied for the geographical traceability of virgin olive oil in these studies, the specific criterion of European vs. non-European (EU vs. non-EU) has been scarcely studied^[15] and it concerns the European regulation demand.^[8] Bontempo et al.^[15] studied this classification criterion by Isotope Ratio Mass Spectrometry (IRMS) and IRMS coupled with GC. An alternative is the use of vibrational spectroscopies, which has the advantages of being a fast analytical technique working with almost no sample pre-treatment and solvents and being independent of the operator. The main drawback of vibrational spectroscopic methods is the fact that they are indirect techniques, i.e. they need to be calibrated according to some reference method to be used in routine. However, once they have been calibrated and are robust enough, vibrational spectroscopy can work at least as good as the classical wet chemistry methods. Vibration spectroscopies comprise near-infrared, mid infrared, and Raman spectral regions. Vibrational spectroscopies have been previously applied in authenticating the geographical origin, the variety/species, and the production process of food and feed products.^[18,16,17] In particular, FT-Raman spectroscopy has great potential for compositional analysis of oils and fats. Several studies have been carried out on the

adulteration of oils,^[19,20] quantitative analysis of major constituents,^[21,22] classification of edible oil and fat from different sources vegetable and animal,^[23,24] and determination of the total unsaturated fatty acids in different fats and oils.^[25,26,27,28] In this manuscript, a vibrational spectroscopic technique combined with accurate chemometric tools is presented as a rapid alternative to the chromatographic methods for tracing the geographical provenance of virgin olive oils. Raman spectroscopy has been combined with chemometric analysis to classify VOOs according to their geographical origin in the particular application of distinguishing European and non-European samples. In order to develop an exhaustive performance study, a protocol has been implemented to select the adequate samples that amply cover the variability of virgin olive oils produced inside and outside of the European Union. The large number of data obtained with these techniques has encouraged the use of chemometrics, whose outputs provide useful and easy to visualize information extracted from the spectroscopic data.

MATERIALS AND METHODS

Samples

To obtain satisfactory and relevant results from this study, total certainty of origin of the samples was considered a pre-requisite. A complete sampling strategy design had to be defined taking into account different variables that would yield a representative and traceable set of samples. Figure 1 shows all the countries and cultivars that have been considered in the sampling. Different criteria and important factors were considered to select the samples:

- (i) Freshness: Southern and Northern Hemispheres have different crop seasons, so the sample collection step was extended to cover both crops. Samples were stored at -20°C in dark flasks until analysis.

- (ii) Cultivar distribution and complexity of each location. The selection of the samples was done taking into account the diversity in cultivars. This diversity depends on each geographical area. Thus, for instance, in Italy, there is a large number of cultivars and the commercialised oils are blends of some of them.
- (iii) New agricultural practices as well as new/old cultivars: the implementation of new agricultural practices plays an important role in the chemical composition of oils.^[5,29] So these variables were also considered. Samples from Arbequina, Frantoio, Coratina, and Koroneki cultivars, originally from Spain, Italy, and Greece were included along with those other from geographical origins such as Argentina, California, Portugal, and Australia (Figure 1) where new agricultural practices are currently used.

The total amount of samples has been divided into three different sets: (i) modelling set of samples, which has been used to develop the chemometric models and select important variables (ii) verification set of samples, which is an independent set of samples selected to improve the classification method and includes samples from a second year crop and new geographical origins, and (iii) blind set of samples, which is used to determine the classification power of the model. Table 1 shows the division of the total amount of samples in these three sets. An equilibrium between the number of European and non-European samples was considered when distributing the samples between these sets. The total number of samples (78 virgin olive oils) were divided into two groups: 48 European oils and 30 non-European oils (henceforth EU and non-EU samples, respectively). Table 2 shows the provenance countries of the samples along with the total number of samples per country and information about the total production and consumption of each geographical origin. More information about these samples can be found in Table S1 (Supporting Information). The collection of samples was carried out

carefully and all the samples were directly provided by producers with total certainty of their origins. The blind set of samples (10 virgin olive oils, 5 EU samples and 5 non-EU samples) was used as validation set. The provenance of these samples was unknown during the analysis, spectra processing and chemometric analysis and their identity were revealed only after the work was completed. Therefore, these oils were considered ‘blind samples’. Details about these samples can be found in Table S2. The selection of producers was carefully done to avoid mixing with oils from other origins, particularly in those countries where the production is not enough to meet local consumption needs and must import olive oil to support their demands (e.g., USA, Australia, Uruguay).

Fatty acid composition

The fatty acid profile of all the samples has been determined by following the standard method COI/T20/Doc No 33.^[30] The quantification of the fatty acids methyl esters was expressed as a percentage by mass of methyl esters, by determining the percentage represented by the area of the corresponding peak relative to the sum of the areas of all the peaks.

Raman spectroscopy analysis

FT-Raman spectra were acquired on a Vertex 70 – RAM II Bruker FT-Raman spectrometer (Bruker, Optics, Germany). This instrument is equipped with a Nd:YAG laser (yttrium aluminium garnet crystal doped with triply ionised neodymium) with an output at 1064 nm (9398.5 cm^{-1}). The maximum level of laser power is 1.5 W. The measurement accessory is pre-aligned, only the Z-axis of the scattered light is adjusted to set the sample placed in a support tube of 4-5 ml in the appropriate position regarding the focal point. The RAM II spectrometer is equipped with a liquid-nitrogen cooled Ge

detector. FT-Raman spectra ($3600\text{--}50\text{ cm}^{-1}$) were collected with a resolution of 4 cm^{-1} by co-adding 128 scans for each spectrum. Each spectrum was collected in approximately 4 min. The analyses were carried out in quadruplicate and averaged. Laser power was set at 600 mW. The OPUS 6.0 software (Bruker Optics, Ettlingen, Germany) was used for spectral acquisition, manipulation, and transformation.

Statistical Analysis

Before applying any multivariate data analysis, all the Raman spectra (78 samples) were pre-processed using Multiplicative Scatter Correction (MSC)^[31] to reduce the effects caused by the scattering in the measurements mainly due to instrumental fluctuations, slight variation in the distance between samples and optics, state of the sample (linked to temperature), and the presence of fluorescence.

Two different methods have been applied for data reduction: (i) Principal Component Analysis (PCA)^[32] is used to reduce the data dimensionality, and the selected features were used as the input of the classifiers. PCA creates new orthogonal variables (scores or latent variables) that are linear combinations of the measured x-variables (e.g., absorbances at different wavelengths or scattering intensities); (ii) the Fisher Criterion (FC)^[33] describes the ratio of between-class /within-class variances a helpful datum to decide which original variables have significant discriminating powers.

For the discrimination analysis between European and non-European samples, Partial Least Squares Discriminant Analysis (PLS-DA)^[34] has been applied. It is a multivariate classification method based on the maximum separation between the classes (European vs. non-European samples) defined in the data. It is done by a PLS regression against the dummy matrix Y (± 1 for a two-class model) which describes the variation according to class. For comparison purposes, other two chemometric models were built:

Stepwise Linear Discriminant Analysis (SLDA)^[35] and Soft Independent Modelling of Class Analogy (SIMCA).^[36]

For the evaluation of the performance of the method with the blind samples, the following analytical quality parameters used for the validation of the binary response methods were calculated considering the number of false positive (FP, a non-European VOO classified as European) and false negative (FN, a European VOO classified as non-European) results as explained by Valcárcel et al.:^[37] Rate of FP, Rate of FN, Reliability, Sensitivity, Selectivity, Efficiency, Youden's index, and Likelihood ratio.

All computations, chemometric analyses, and graphics were carried out with programmes developed in Matlab (The Mathworks, Inc., Natick, MA, USA), the PLS Toolbox 8.1 (Eigenvector Research Inc.) and Statistica 8.0 (Stat soft. Inc., Tulsa, OK, USA).

RESULTS AND DISCUSSION

Following the sampling procedure explained in the Materials and Methods section, a representative and reliable set of samples was obtained. A initial step was to explore the predicted power of the fatty acid composition of the samples as a first approach to determine the variability of major compounds. The fatty acid composition has been used in the past to classify the oils according to their cultivars.^[3,38] Other studies have pointed out that fatty acids can only give basic information or separate clusters that are very different among them. It is well accepted that to carry out a proper characterization of virgin olive oils information on fatty acid is not enough and additional information on their minor chemical compounds is required.^[10] However, traditionally fatty acid profiles have been used to classify the geographical origin of samples.^[12]

The purpose of this study is to differentiate between two groups of samples with different geographical origins (European vs. non-European). To study the ability of the

fatty acid composition to classify VOOs as European or non-European samples, the fatty acid profile of the samples was determined. Table 3 shows the mean value, the standard deviation and the minimum and maximum of the percentage of each individual fatty acid obtained per country.

In order to determine the capacity of fatty acids for classifying the samples, a principal component analysis (PCA) was performed. The results of the PCA showed a complete overlap between samples (result not shown). Thus, the information provided by the fatty acids is not enough to properly separate the groups, despite the fact that fatty acids report more than 98% of the total chemical composition of VOO. These results confirm the need to add more chemical information about the minor chemical compounds present in the oils to carry out an appropriate characterization of the samples.^[13] In order to obtain more information about the potential of the fatty acid composition in the classification of the sample, a supervised discriminant analysis (PLS model) was performed on the entire set of samples (78 samples, 48 European and 30 non-European). The response variables were set at -1 for European samples and +1 for non-European samples. The results obtained by the discriminant analysis presented a misclassification in samples of both groups (Figure 2). However, only 5 non-European samples were classified as European and 9 European samples were classified as non-European. These results are in agreement with previous studies classifying oils by country or cultivar.^[7,10,13,39] The PLS model developed in this study only with the information of the fatty acid misclassified a total of 14 samples from a pool of 78. The variables selected for the PLS model were the percentage of C18:2, C18:1, C18:3, C22:0, C18:1*trans*, C17:1, C18:3, C14:0 and C20:1. In the model, the fatty acids with the highest contribution were C18:2 and C18:1. In the literature, the content of these fatty acids were already identified as markers of monovarietal virgin olive oils. Thus, the oleic/linoleic fatty acid content

ratio has been described as a reliable marker of the affinity between different cultivars.^[40,41] The linoleic acid content has also been identified as a marker, along with the content of other fatty acids, to characterize European virgin olive oils, in particular those produced in Italy, Spain and Portugal.^[42] Furthermore, the effect of latitude and altitude on the accumulation of linoleic acid to the detriment of oleic acid has also been demonstrated in other studies of geographical characterization of virgin olive oil.^[10] Thus, these results show that the information of fatty acid content can provide some information about geographical characterization of olive oil although this requires selecting fatty acids that have a closer relationship with location due to cultivar or climate features.

For a better classification of geographical origin, incorporating the information of minor components is necessary, this fact having been proven in previous studies.^[13] Thus, once the potential of the content of the fatty acid in the classification of European and non-European samples was studied, the next step was to evaluate the discriminant power of Raman spectroscopy as an alternative to chromatography data. -. After the MSC of the all spectra, a first exploratory analysis of the data of the modelling set of samples (50 virgin olive oils) was performed using the Fisher criterion (ratio of between-class variance/within-class variance) and the Principal Component Analysis. Figure 3 shows the spectra after MSC for the modelling data set (n=50) where no clear distinction can be made between European and non-European samples. Then, PCA is applied to visualize the data and estimate the difficulty of the classification problem for this data set. Figure 4 shows the PC1 vs. PC3 plot for the MSC data set. These two PCs have been selected as they show the best tendency to separate both European and non-European groups (EU and non-EU) while not achieving 100% efficiency. In this analysis 7 samples from Portugal had to be removed since they provided abnormal data, probably because they

were not filtered. Both EU and non-EU samples become clearly distinguishable except for one sample from Tunisia and a few from Italy.

Figure 5 shows the spectral bands with the largest discriminating ability between classes according to the results of the Fisher test after MSC preprocessing of the modelling set of samples. The Fisher criterion between EU and non-EU groups allowed selecting a few variables that have a high discriminating ability for this data set, i.e. 1658 cm^{-1} , 2850 cm^{-1} and 3014 cm^{-1} . These Raman shifts correspond mainly to the presence of Trilinolein ($\text{C}_{57}\text{H}_{98}\text{O}_6$),^[19,43] which consists of glycerol esterified with three fatty acids containing two double bonds. The intensity of the C=C stretching bands at 1658 cm^{-1} and 3014 cm^{-1} is high when compared to the other Raman bands in the spectra. The band around 2850 cm^{-1} corresponds to $-\text{CH}_2$ and $-\text{CH}_3$ groups. The trilinolein content of virgin olive oil is usually less than 0.3% and it can be also used as an indication of possible adulteration of oils of foreign origin.^[44] This result confirms the findings of studies based on HPLC that used the presence of trilinolein content as a first approach towards the development of geographical origin tracing models in olive oils.^[45,46]

A new PCA model was performed using exclusively the bands selected by the Fisher criterion. The results are shown in Figure 6. Similar conclusions can be drawn for Figure 4, i.e. both classes could be clearly distinguished except for few samples that were misclassified.

In order to obtain better classification results, different statistical methods were tested to check their performance in classifying EU-Non EU samples. Table S3 shows that the best classification results were obtained by PLS-DA model compared with the other two methods (SLDA, SIMCA). Thus, the modelling stage with a PLS-DA model was performed using the leave-one-out-cross-validation procedure to optimise the model,

i.e. to find the number of variables which could yield the best classification rate. The PLS-DA result was indicated in terms of the success rate:

$$successrate = \left(\sum_{i=1}^K \frac{correctly\,classified\,samples\,in\,class\,i}{total\,number\,of\,samples\,in\,class\,i} \right) / K$$

with K being the number of classes obtained for the data set with cross-validation.

For the whole data range, success rates of 82% and 80% of the objects correctly classified in class 1 (EU) and class 2 (non-EU), respectively, were obtained.

Figure 7 shows the predicted values using the whole spectral range where the misclassified samples are indicated. Again 5 samples from Italy and a single sample from Tunisia were misclassified, as observed when performing PCA. An Australian sample was also misclassified.

In order to improve the method, once the variables were selected and the classification procedure was established, 18 samples from the verification set of samples (Table S1) were additionally included in the model. The objective was to consider another crop year that could show possible variations and to include new geographical origins not considered in the first 50 samples (Greece, Morocco, USA, Peru). In all these samples, 72% of the EU oils, and 84% of the non-EU oils were correctly classified.

Finally, 10 unknown samples (blind set of samples, Table S2) were measured in quadruplicate by Raman spectroscopy under the same conditions used in the calibration samples. Table 4 shows the results obtained from the 10 blind samples. Table 5 shows the analytical quality parameters obtained with the VOO samples of the blind set computed from the number of false positive and false negative results. In this study, Raman spectroscopy was applied for a binary classification EU sample/non-EU sample mostly based on chemical profiling of the samples. The rate of false negative results (a

non-European VOO classified as European) was lower than the rate of false positive results (a European VOO classified as non-European), showing that the errors seem to be more related with the false positive results.

Conclusions

This study contributes both to reinforcing the usefulness of non-targeted techniques to solve current safety/quality problems in the olive oil industry and increase the confidence of the olive oil actors (mainly regulatory bodies, producers and consumers) on these techniques. In general terms, it can be concluded that a guideline for evaluating non-targeted methods for geographical identification should include a defined classification objective that allows evaluating the method with false positive and false negative errors. A lack of definition in the classification task does not allow applying this approach. Additionally, in this work we considered other relevant aspects: (i) the set of samples selected for the study is sufficiently representative of the universe of data (Table 2); (ii) in some cases, some cultivars were presented in both groups, EU and non-EU samples (Arbequina, Coratina, Frantoio, Picual, Koroneki), so this was a complex and realistic classification case; (iii) the chemometric analyses applied to the spectral data allowed selecting the right variables for the classification; (iv) the models developed have been tested with a proper validation procedure with external samples that were unknown by the analysts and chemometricians. The geographical identification of virgin olive oils is a challenge because it involves classifying samples with very similar composition. However, olive oil actors include among their current demands the development of an efficient method with this purpose. Further studies can be focused on testing the method at an industrial plant level in a routine basis in order to check variations between crop years and to characterise the climate influence in some areas.

Acknowledgments

The authors would like to thank Mr. Quentin Arnould from CRA-W for the assistance in FT-Raman analysis.

Funding

The project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 613688 (FoodIntegrity).

Conflict of interest

This manuscript does not have any financial or commercial conflict of interest.

References

- [1] P. Vossen, in *Handbook of Olive Oil: Analysis and Properties*, 2nd ed. (Eds: R. Aparicio, J. Harwood), Springer, USA **2013**, Ch. 2.
- [2] D.L. García-González, N. Tena, R. Aparicio, *Grasas Aceites*, **2012**, 63, 26.
- [3] P. Dais, in *Handbook of Olive Oil: Analysis and Properties*, 2nd ed. (Eds: R. Aparicio, J. Harwood), Springer, USA **2013**, Ch. 11.
- [4] D. L. García-González, C. Infante-Domínguez, R. Aparicio, in *Handbook of Olive Oil: Analysis and Properties*, 2nd ed. (Eds: R. Aparicio, J. Harwood), Springer, USA **2013**, Ch. 20.
- [5] N., Romero, J. Saavedra, F. Tapia, B.R. Sepúlveda, R. Aparicio, *J Sci Food Agric*, **2016**, 96, 583.
- [6] A. Larbi, M. Ayadi, A. Ben Dhiab, M. Msallem J.M. Caballero, *SJAR*. **2011**, 9 (4), 1279.
- [7] R. Aparicio, L.S. Conte, H.J. Fiebig, in *Handbook of Olive Oil: Analysis and Properties*. 2nd ed. (Eds: R. Aparicio, J. Harwood), Springer, USA **2013**, Ch. 16.
- [8] EU regulation 29/**2012** *Off. J. Eur. Union*. L 12/14, 14-1-2012.

- [9] D.L. García-González, N. Tena, I. Romero, R. Aparicio-Ruiz, M.T. Morales, R. Aparicio, *Grasas aceites*, **2017**, 68 (3), 1.
- [10] R. Aparicio, D.L. García-González, in *Handbook of Olive Oil: Analysis and Properties*, 2nd ed. (Eds: R. Aparicio, J. Harwood), Springer, USA **2013**, Ch. 12.
- [11] Ch. Emmanouilides, P. Fousekis, V. Grigoriadis, *SJAR*. **2014**, 12 (1), 3.
- [12] R. Aparicio, V. Alonso, *Prog Lipid Res*. **1994**, 33, 29.
- [13] D.L. García-González, G. Luna, M.T. Morales, R. Aparicio, *Eur. J. Lipid Sci. Technol.* **2009**, 111, 1003.
- [14] N. Tena, R. Aparicio-Ruiz, A. Koidis, D.L. García-González, in *Food Traceability and Authenticity Analytical Techniques*, 1st ed, (Eds. D. Montet, R.C. Ray), CRC Press, Boca Raton **2017**, Ch. 13.
- [15] L. Bontempo, M. Paolini, P. Franceschi, L. Ziller, D. L. García-González, F. Camin, *Food Chem.* **2019**, 276, 782.
- [16] P. Vermeulen, J.A. Fernández Pierna, O. Abbas, H. Rogez, F. Davrieux, V. Baeten, in: *Food Traceability and Authenticity: Analytical Techniques*, (Eds. D. Montet, R.C. Ray) Biology series, CRC press, USA **2017**, Ch.16.
- [17] V. Baeten, H. Rogez, J.A. Fernández Pierna, Ph. Vermeulen, P. Dardenne, in *Handbook of Food Analysis*, 3rd ed. Vol. 2 (Eds: L.M.L Nollet, F. Toldra), CRC Press, Boca Ratón **2015**, Ch 32.
- [18] D.L., García-González, V., Baeten, J.A. Fernández Pierna, N. Tena, in *Handbook of Olive Oil: Analysis and Properties*, 2nd ed. (Eds: R. Aparicio, J. Harwood), Springer, USA **2013**, Ch.10.

- [19] V. Baeten, M. Meurens, M.T. Morales, R. Aparicio, *J. Agric. Food Chem.* **1996**, *44*, 2225.
- [20] V. Baeten, J.A. Fernández Pierna, P. Dardenne, M. Meurens, D.L. García-González, R. Aparicio-Ruiz, *J. Agric. Food Chem.* **2005**, *53* (16), 6201.
- [21] H. Sadeghi-Jorabchi, R.H. Wilson, P.S. Belton, J.D. Edwards-Webb, D.T. Coxon, *Spectrochim Acta A Mol Biomol Spectrosc*, **1991**, *47 A* (9/10), 1449.
- [22] J.H. Giles, D.A. Gilmore, M.B. Denton, *J Raman Spectrosc.* **1999**, *30*, 767.
- [23] V. Baeten, P. Hourant, M.T. Morales, R. Aparicio, *Food Chem.* **1998**, *46* (7), 2638.
- [24] O. Abbas, J.A. Fernández Pierna, R. Codony, C. von Holst, V. Baeten, *J. Mol. Struct.* **2009**, *924/926*, 294.
- [25] H. Sadeghi-Jorabchi, H. Hendra, R.H. Wilson, P.S. Belton, *JOACS*, **1990**, *67* (8), 483.
- [26] R.C. Barthus, R.J. Poppi, *Vib. Spectrosc.* **2001**, *26* (1), 99.
- [27] E.F. Olsen, E.O. Rukke, A. Flåtten, T. Isaksson, *Meat Sci.* **2007**, *76*, 628.
- [28] F. Gao, L.J. Han, X. Liu, *Int J Agric & Biol Eng.* **2017**, *10* (3), 255.
- [29] M. Torres, P. Pierantozzi, P. Searles, M.C. Rousseaux, G. García-Inza, A.M. Miserere, R. Bodoira, C. Contreras, D. Maestri, *Front Plant Sci.* **2017**, *8* 1830, 1.
- [30] COI/T.20/Doc. No 33, *International Olive Council (IOC)*, **2015**, Madrid, Spain.
- [31] P. Geladi, D. MacDougall, H. Martens, *Appl. Spectrosc.* **1985**, *39*, 491.
- [32] D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte, L. Kaufman, *Chemometrics: a textbook*, Elsevier Science Publishers B. V., Amsterdam **1988**.

- [33] R.O. Duda, P.E. Hart, *Pattern classification and scene analysis*, Wiley, New York **1973**.
- [34] H. Martens, T. Naes, *Multivariate Calibration*, Wiley, New York **1989**.
- [35] C.M. Sánchez-Peña, G. Luna, D.L. García-González, R. Aparicio. *Meat Sci.*, **2005**, 69(4).
- [36] E. Guzmán, V. Baeten, J.A. Fernández-Pierna, J.A. García-Mesa, J. A., *Talanta*, **2012**, 93, 94.
- [37] M. Valcárcel, S. Cárdenas, D. Barceló, L. Buydens, K. Heydorn, B. Karlberg, K. Klemm, B. Lendl, B. Milman, B. Neidhart, A. Ríos, R. Stephany, A. Townshend, A. Zschunke, *Metrology of Qualitative Chemical Analysis. Final Report Contract G6MA-CT2000-01012*, Project report EUR 20605 EN. / BCR Information ed. Luxembourg : European Commission, **2002**, 166 p.
- <https://publications.europa.eu/en/publication-detail/-/publication/afa0233e-7b8e-497e-bce8-b3c17136cd53>
- [38] A. Sacco, M.A. Brescia, V. Liuzzi, F. Reniero, C. Guillou, S. Ghelli, P. van der Meer, *JAOCs*, **2000**, 77, 619.
- [39] M.J. Lerma-García, J.M., Herrero-Martínez, G., Ramis-Ramos, E.F. Simó-Alfonso, *Food Chem.* **2008**, 108, 1142.
- [40] M. Poiana, A. Mincione, *Grasas Aceites*, **2004**, 55 (3), 282.
- [41] M. Tsimidou, K.X. Karakostas, *J Sci Food Agric.* **1993**, 62, 253.
- [42] M.V. Alonso, and R. Aparicio, *Grasas Aceites*. **1993**, 44, 18.
- [43] J. De Gelder, K. De Gussem, P. Vandenabeele, L. Moens, *J. Raman Spectrosc.* **2007**, 38, 1133.

- [44] A. Sass-Kiss in *Modern techniques for food authentication*, (Ed. D.W. Sun), Elsevier, Ireland **2008**, Ch. 11
- [45] N. Araghipour, J. Colineau, A. Koot, W. Akkermans, J.M. Moreno Rojas, J. Beauchamp, A. Wisthaler, T.D. Märk, G. Downey, C. Guillou, L. Mannina, S. van Ruth, *Food Chem.* **2008**, *108 (1)*, 374.
- [46] A. Bajoub, S. Medina-Rodríguez, E. Hurtado-Fernández, E.A. Ajal, N. Ouazzani, A. Fernández-Gutiérrez, A. Carrasco-Pancorbo, *Eur. J. Lipid Sci. Technol.* **2016**, *118 (8)*, 1223.

FIGURE CAPTIONS

Figure 1. Map indicative of the countries and cultivars included in the study.

Figure 2. Predicted values obtained by applying the discriminant analysis based on fatty acid composition to classify 78 samples in European (48 samples) and non-European (30 samples) VOOs. Results shown with blue dots are European samples and with red squares are non-European samples. The filled blue dots and red squares indicate misclassified samples. The response variable was -1 for European samples and 1 for non-European samples.

Figure 3. FT-Raman spectra (MSC corrected spectra) of the modelling sample set included in the study.

Figure 4. Score plot obtained on the projection of PC1 vs. PC3 after MSC with the modelling set of samples.

Figure 5. Results of the Fisher test after MSC preprocessing of the modelling set of samples.

Figure 6. Score plot obtained on the projection of PC1 vs. PC3 after MSC and variable selection carried out with Fisher test.

Figure 7. Sample prediction by applying PLS-DA model performed using the leave-one-out-cross-validation. The misclassified samples are labelled.

Table 1: VOO samples used for the construction and test of the PLS-DA model for geographical origin discrimination. Distribution of the samples into different sets of samples.

Geographical origin	Set of samples		
	Modelling	Verification	Blind
European	35	8	5
Non-European	15	10	5
Total	50	18	10

Table 2: Characteristics of the samples included in the study. Geographical origin of virgin olive oil samples, number of samples, production, and consumption of olive oil (2017/18) for each geographical origin studied.

Country	Number of sample	Production (tonnes)	Consumption (tonnes)
Italy (EU)	18	320.000	557.000
Spain (EU)	15	1.090.500	470.000
Portugal (EU)	12	78.800	70.000
Australia	10	21.000	45.000
Turkey	5	287.000	170.000
California (USA)	4	15.000	315.000
Tunisia	4	220.000	35.000
Uruguay	3	1.000	2.000
Argentina	3	37.500	7.500
Greece (EU)	2	300.000	120.000
Morocco	1	140.000	120.000
Peru	1	--	--

Table 3: Mean of percentage of individual fatty acid in virgin olive oil studied. Standard deviation, minimum and maximum values.

Values per countries (%)	C16:0	C16:1	C17:0	C17:1	C18:0	C18:1	C18:2	C18:3	C20:0	C20:1	C22:0	C24:0
Italy (n=18)												
Minimum	9.69	0.10	0.04	0.07	1.86	65.03	6.43	0.63	0.39	0.30	0.12	0.04
Maximum	15.92	1.97	0.23	0.24	3.10	79.18	13.23	0.75	0.54	0.50	0.14	0.07
Mean	11.57	0.78	0.09	0.13	2.44	74.64	8.58	0.70	0.45	0.38	0.13	0.06
SD	1.71	0.44	0.05	0.07	0.43	3.69	1.82	0.04	0.04	0.06	0.01	0.01
Spain (n=15)												
Minimum	9.74	0.68	0.04	0.07	1.65	66.16	3.18	0.52	0.25	0.25	0.10	0.04
Maximum	17.00	1.82	0.16	0.28	3.93	81.04	12.40	1.11	0.52	0.42	0.15	0.17
Mean	12.80	1.18	0.08	0.15	2.61	74.65	6.90	0.73	0.40	0.30	0.12	0.07
SD	1.98	0.34	0.05	0.08	0.63	4.62	3.09	0.16	0.06	0.06	0.01	0.03
Portugal (n=12)												
Minimum	11.23	0.85	0.04	0.07	1.80	67.51	3.30	0.47	0.24	0.23	0.10	0.05
Maximum	16.04	2.37	0.16	0.31	3.92	80.27	11.06	0.85	0.51	0.39	0.18	0.22
Mean	13.42	1.39	0.11	0.22	2.46	74.27	6.60	0.65	0.42	0.29	0.13	0.08
SD	1.52	0.51	0.04	0.08	0.69	3.95	2.87	0.11	0.08	0.04	0.02	0.04
Australia (n=10)												
Minimum	8.20	0.41	0.05	0.07	2.04	71.01	4.04	0.55	0.39	0.28	0.12	0.05
Maximum	13.12	1.37	0.09	0.21	3.26	82.22	10.70	0.90	0.47	0.56	0.16	0.09
Mean	10.65	0.73	0.06	0.11	2.39	76.44	7.89	0.73	0.44	0.36	0.14	0.06
SD	1.60	0.33	0.01	0.05	0.41	3.55	2.04	0.10	0.03	0.09	0.01	0.01
Turkey (n=5)												
Minimum	10.60	0.75	0.05	0.06	2.66	70.38	8.95	0.68	0.34	0.25	0.11	0.04
Maximum	12.98	0.93	0.17	0.25	2.97	73.52	11.42	0.87	0.53	0.51	0.16	0.08
Mean	11.92	0.83	0.14	0.21	2.86	71.58	10.66	0.76	0.46	0.36	0.15	0.07
SD	0.86	0.07	0.05	0.08	0.12	1.45	1.01	0.08	0.08	0.09	0.02	0.02
California (n=4)												
Minimum	12.95	0.98	0.05	0.08	2.01	66.28	7.02	0.61	0.42	0.29	0.12	0.06
Maximum	15.66	1.55	0.17	0.32	2.60	74.69	12.56	0.73	0.47	0.30	0.17	0.08
Mean	14.75	1.36	0.12	0.23	2.26	69.64	10.01	0.66	0.45	0.30	0.14	0.07
SD	1.27	0.26	0.05	0.10	0.26	3.75	2.50	0.05	0.03	0.01	0.02	0.01
Tunisia (n=4)												
Minimum	12.79	0.54	0.06	0.07	2.51	59.43	14.86	0.72	0.45	0.27	0.12	0.06
Maximum	17.14	2.10	0.51	0.09	3.03	66.76	17.08	0.90	0.50	0.38	0.15	0.08
Mean	14.35	1.39	0.28	0.08	2.75	62.89	16.45	0.79	0.48	0.32	0.13	0.07
SD	1.95	0.65	0.25	0.01	0.23	3.03	1.07	0.08	0.02	0.05	0.01	0.01
Uruguay (n=3)												
Minimum	12.29	0.91	0.04	0.08	2.35	72.35	3.21	0.69	0.32	0.27	0.10	0.06
Maximum	14.79	1.67	0.14	0.26	2.71	79.25	6.38	0.76	0.45	0.40	0.13	0.08
Mean	13.20	1.27	0.08	0.14	2.47	75.90	5.32	0.72	0.38	0.32	0.11	0.07
SD	1.38	0.38	0.06	0.10	0.21	3.45	1.83	0.04	0.07	0.07	0.02	0.01
Argentina (n=3)												
Minimum	11.15	0.84	0.05	0.09	1.75	55.54	7.75	0.76	0.38	0.25	0.10	0.06
Maximum	19.00	2.25	0.10	0.20	2.51	75.95	19.51	0.88	0.42	0.33	0.13	0.07
Mean	14.48	1.37	0.07	0.13	2.18	67.92	12.15	0.81	0.41	0.29	0.12	0.06
SD	4.06	0.77	0.03	0.06	0.39	10.88	6.41	0.06	0.02	0.04	0.02	0.01
Greece (n=2)												
Minimum	9.56	0.50	0.04	0.07	2.74	76.82	6.38	0.76	0.41	0.31	0.13	0.06
Maximum	11.08	0.81	0.05	0.08	2.81	77.03	8.47	0.78	0.49	0.39	0.15	0.06
Mean	10.32	0.66	0.05	0.08	2.78	76.93	7.43	0.77	0.45	0.35	0.14	0.06
SD	1.07	0.22	0.01	0.01	0.05	0.15	1.48	0.01	0.06	0.06	0.01	0.00

7

Table 4: Prediction of the origin of unknown samples (blind set).

Sample Code (blind samples)	Predicted classification	Correct classification
A3Z	Non-European	European
1QV	Non-European	Non-European
X8T	European	Non-European
7KW	Non-European	Non-European
LPT	European	European
BC2	Non-European	Non-European
DEF	European	European
ZYR	European	European
SV1	European	Non-European
R5P	European	European

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

Table 5: Analytical quality parameters of the Raman spectroscopy method used to determine the geographical provenance of blind samples.

Rate of FP ^a	Rate of FN ^b	Reliability	Sensitivity	Selectivity	Efficiency	Youden's index	Likelihood ratio
0.4	0.2	40	0.8	0.6	0.7	40	2.0

Note: ^a, False positive; ^b, False negative

Figure 1.

Figure 2.

88 **Figure 3.**

89

90

91

Figure 4.

Figure 5.

Figure 6.

Figure 7.

Supporting Information

Assessment of vibrational spectroscopy performance in geographical identification of virgin olive oils: A world level study.

Noelia Tena, Ramón Aparicio, Vincent Baeten, Diego Luis García-González, Juan Antonio Fernández-Pierna.

Table S1. Code, provenance and cultivar of samples used to build the classification models.

Sample Code	Provenance	Cultivar
1	Uruguay	Picual.
2	Uruguay	Manzanilla.
3	Turkey (North Aegean)	Ayvalik.
4	Turkey (South Aegean)	Memecik.
5	Argentina (Catamarca)	Arbequina.
6	Australia (Western Australia)	Frantoio.
7	Australia (Victoria)	Frantoio.
8	Australia (Victoria)	Arbequina.
9	Australia (Victoria)	Picual.
10	Australia (Victoria)	Coratina.
11	Australia (Tasmania)	Frantoio.
12	Australia (Victoria)	Barnea.
13	Australia (Victoria)	Picual.
14	Australia (New South Wales)	Frantoio.
15	Italy (P.D.O. Val di Mazara, Sicily)	Biancolilla, Nocellara del Belice, Cerasuola.
16	Italy (P.D.O. Sardegna, Sardinia)	Bosana, Tonda di Cagliari, Nera di Villacidro, Semidana.
17	Italy (P.D.O. Riviera ligure-Riviera dei Fiori, Liguria)	Taggiasca.
18	Italy (P.D.O. Terra di Bari Bitonto, Puglia)	Cima di Bitonto, Coratina.
19	Italy	Coratina, Ogliarola, Carolea.
20	Italy (Oneglia, Liguria)	Non-specified.
21	Italy (P.G.I. Toscano)	Frantoio, Leccino, Moraiolo, Mignola, Maurino, Pendolino.
22	Italy (P.D.O. Terra di Bari Bitonto, Puglia)	Cima di Bitonto, Coratina.
23	Italy (P.D.O. Terra di Bari Bitonto, Puglia)	Cima di Bitonto, Coratina.
24	Italy (P.D.O. Umbria colli Martani, Umbria)	Moraiolo, Leccino, Frantoio.
25	Italy (P.D.O. Riviera ligure-Riviera dei Fiori, Liguria)	Taggiasca, Lavagnina, Pignola.
26	Italy (P.D.O. Val di Mazara, Sicily)	Nocellara del belice, Cerasuola, Biancolilla.
27	Italy (P.D.O. Val di Mazara, Sicily)	Biancolilla, Nocellara del Belice, Cerasuola.
28	Italy (P.D.O. Terra di Bari Castel del Monte, Puglia)	Coratina.
29	Italy	Frantoio.
30	Italy (P.D.O. Riviera ligure-Riviera dei Fiori, Liguria)	Taggiasca, Lavagnina, Pignola.
31	Spain (Les Garrigues, Catalonia)	Arbequina.
32	Tunisia	Non-specified.
33	Spain (Córdoba, Andalusia)	Non-specified.

Table S1. Code, provenance and cultivar of samples used to build the classification models. (cont.)

Sample Code	Provenance	Cultivar
34	Spain (Jaén, Andalusia)	Picual.
35	Spain (Jaén, Andalusia)	Picual.
36	Spain (P.D.O. Gata-Hurdes, Extremadura)	Manzanilla cacereña.
37	Spain (Almería, Andalusia)	Picual.
38	Spain (La Rioja)	Arbequina.
39	Spain (Toledo, Castilla-La Mancha)	Cornicabra.
40	Spain (Córdoba, Andalusia)	Picudo.
41	Spain (Catalonia)	Arbequina.
42	Spain (Sevilla, Andalusia)	Hojiblanca.
43	Portugal (Sta. Vitória, Alentejo)	Arbequina.
44	Portugal (Canhestro/Ferreira do Alentejo, Alentejo)	Arbequina.
45	Portugal (Ferreira do Alentejo, Alentejo)	Arbosana.
46	Portugal (Sta. Margarida do Sado/ Ferreira do Alentejo, Alentejo)	Cobrançosa.
47	Portugal (Alvito/Cuba, Alentejo)	Galega.
48	Portugal (Ferreira do Alentejo, Alentejo)	Koroneiki.
49	Portugal (Beja, Alentejo)	Manzanilla.
50	Portugal (Quintos, Alentejo)	Picual.
51	Portugal	Arbequina.
52	Portugal	Cobrançosa.
53	Portugal	Galega.
54	Portugal	-
55	Spain	Arbequina.
56	Spain	Picual.
57	Greece	-
58	Greece	-
59	Tunisia	-
60	Tunisia	-
61	Morocco	Picholine.
62	Peru	Unknown.
63	Turkey	Ayvalik.
64	Turkey	Gemlik.
65	USA (California)	Arbequina.
66	USA (California)	Koroneiki.
67	USA (California)	Arbosana.
68	USA (California)	Arbequina.

Note: Codes 1-50 represent the modelling set of samples; codes 51-68 represent the verification set of samples.

Table S2. Code, provenance and cultivar of samples used to validate the classification model (blind samples).

Sample Code (blind samples)	Provenance	Cultivar
A3Z	Spain	-
1QV	Turkey	Ayvalik.
X8T	Uruguay	-
7KW	Tunisia	-
LPT	Italy (Sicily)	PDO: val de Mazara.
BC2	Argentina	-
DEF	Portugal	-
ZYR	Spain (Andalusia, Jaén)	Picual.
SV1	Australia	-
RP5	Italy	-

Table S3. Percentage of correct classification for different statistical methods for discriminating EU vs. Non-EU samples.

Class	Statistical Procedure		
	PLS-DA	SLDA	SIMCA
European	82	74	50
Non-European	80	53	60

Note: Partial Least Square-Discriminant Analysis (PLS-DA), Stepwise Linear Discriminant Analysis (SLDA), Soft Independent Modelling by Class Analogy (SIMCA).