RESEARCH ARTICLE

European Journal of
Lipid Science and Technology
www.ejlst.com

# A Systematic Chemometric Approach to Identify the Geographical Origin of Olive Oils

*Christian Gertz, Alexander Gertz, Bertrand Matthäus, and Ina Willenberg**

**The verification of the geographical origin of olive oils by analytical techniques is still a challenge. The goal of this work is to explore the application and accuracy of different chemometric tools combined with near infrared spectroscopy (NIR) based analytical methods in the field of geographical authenticity of olive oils. As olive oils associated with different geographical origins are mainly characterized by different fatty acid (FA) and triacylglycerol (TAG) compositions, NIR methods for the fast and reliable determination of these parameters are developed. Next, these NIR methods are used to characterize a comprehensive set of olive oils ($n > 5000$) derived from 19 different countries. This set of data is used to build a statistical workflow, which allows the determination of the geographical origin of unknown olive oil samples. First of all, the untreated data set is pretreated by *k*-means clustering and the selection of the relevant analytical variables by principal component analysis (PCA) and linear discriminant analysis (LDA) and min/max normalization of all parameters. Subsequently, classification is performed with a reduced sample set of the 200 most similar samples identified by *k*-nearest neighbor tool (kNN). For classification purpose kNN, LDA, naïve Bayes classifier, and logit regression are applied.**

***Practical Applications*: The established statistical workflow can be used to verify the geographical origin of olive oils. The application and usage of up to four different statistical models for classification purpose results in a superior probability of the predicted origin in comparison to the application of only one single statistical classification test. As standardized methods are used as reference methods for building the NIR methods, the FA and TAG composition and the iodine value can be either determined by the standard methods or by the described NIR method. The presented statistical approach will help to build up a system for the verification of the geographical origin of olive oils.**

## 1. Introduction

Information on the geographical origin of olive oil has the most important influence on olive oil consumer choices.[1] When discussing authenticity of olive oils one of the main issues is non-compliance with origin stated on the label. In 1992, Council Regulation (EEC) No 2081/92 came into force providing a system for the protection of regional foods by the introduction of the "PGI" (Protected Geographic Indication) and "PDO" (Protected Designation of Origin) labels (today: Regulation (EU) No 1151/2012). The aims of this legislation were to support diversity in agricultural production, to protect consumers by giving them information on the specific characteristic of the product and to protect product names against fraud and imitation.

At present, the geographical origin of extra virgin olive oils can only be ensured by documented traceability although chemical analysis may also be able to contribute to the verification of the geographical origin. Therefore, the search for methods which enable to verify geographical origin and authenticity of olive oils has been the object of numerous studies in the past few years. Different targeted and non-targeted approaches as well as different analytical techniques in combination with multivariate data analysis were applied for this purpose: While some studies focused on the analysis of specific compounds like sterols, carotinoids, tocopherols, isotope ratios, volatiles, and phenolic compounds, other studies used the so called "chemical fingerprints" of olive oils analyzed by gas and liquid chromatography, mass spectrometry, spectroscopic techniques (NMR, NIR, MIR, and Raman fluorescence) or potentiometric electronic nose.[2–12] The combination of the determination of phenolic compounds, sterols or other minor compounds with fatty acid (FA) patterns is also proposed.[13] The analytical advantages and drawbacks of these methods have been highlighted by different authors.[11,14,15] Some of the analytical parameters may affect the quality of classification of the geographical origin because they are altered during production and storage, by climatic changes or by grade of ripeness of the olives. Several authors propose analysis of the geographical origin based on $^1$H NMR spectra of the phenolic extract of olive oils.[16,17] However, minor

Dr. C. Gertz, A. Gertz
Maxfry GmbH
Hagen 58095, Germany
Dr. B. Matthäus, Dr. I. Willenberg
Max Rubner-Institut (MRI)
Department of Safety and Quality of Cereals
Working Group for Lipid Research
Schützenberg 12,, 32756, Detmold, Germany
E-mail: ina.willenberg@mri.bund.de



**DOI: 10.1002/ejlt.201900281**

compounds like phenolic compounds are changed in their whole or individually by hydrolysis and/or oxidation reactions during production and storage.[18] Another interesting tool is the application of DNA based markers since it is independent from environmental factors.[19]

The main drawback of all papers is the limited number of samples (<400) derived from only a few geographical origins. There is a lack of systematic studies on the chemical composition of virgin olive oils, which are not limited to specific regions or a few countries and which also include influences of different harvest periods, varieties or climatic changes. There is a strong need for simple and statistically approved methods to establish the authenticity confirmation of olive oils.

Olive oils are complex matrices with high chemical variability due to genetic and changing environmental factors, different states of ripening of the fruit, ages, different ways of harvesting and extraction prepared with different varieties of different geographical origins resulting in different organoleptic properties and different chemical patterns.[20–23] Thus, each olive oil has its unique fingerprint characterized by changing contents and types of metabolites such as FA and triacylglycerols (TAG), sterols, *n*-alkanes, volatile compounds, carotinoides, tocopherols, or stable isotopes.[24,25] Especially, the possibilities of the variations in the FA pattern and their combination in TAG molecules lead to enormous complexity in all vegetable fats and oils. The same FA distribution can lead to different patterns of the individual tri-acylglycerols. For this reason, many studies use the TAG-profile and the FA-distribution to characterize fats and oils.[26–28]

In this regard, the study of compositional differences among olive oils from different geographical regions has been the basis for different statistical approaches. Official methods are not available for this purpose. Conte et al.[29] revealed the gaps of the existing official legal standards and requested more efficient analytical solutions to overcome drawbacks and limitations of the official methods in Europe. For the identification of the geographical origin of olive oils a large data set of samples including olive oils from non-European countries such as Tunisia, Turkey, or Egypt harvested in different years and at different grade of ripeness is needed to cover the great possible variations of composition.

For this purpose, the analytical methods must be quick, less expensive, and simple in order to increase the number of analyses per day for an effective fight against olive oil fraud. Modern instrumental methods provide a large number of reproducible data in one run and thereby increase the information of the collected data per analysis. The combination of such methods with proper chemometric methods helps to extract the maximum of relevant information and to develop a new analytical approach. The classical approach is often erroneous, time-consuming and determines only one factor at a time to prove the validity of a previously designed model.

NIR shows a great potential to provide the complete FA and TAG composition of olive oil even if NIR has a much lower sensitivity in comparison to chromatographic methods such as gas chromatography or high-performance liquid chromatography (HPLC). The advantage of NIR in comparison to other methods based on vibrational spectroscopic techniques or chromatographic methods lies in the simple sample handling without any pretreatment,[30] which reduces also systematic errors due to different handling of the samples. Furthermore, the reevaluation of spectra scanned years before under same measurement conditions is possible and guarantee the homogeneity of all data analyzed over years.

The first part of this project was the development of NIR methods for the analysis of FA and TAG composition as well as iodine value (IV) of olive oils. Next, these methods were applied to analyze a high number of olive oils elucidating the different effects of olive varieties, climatic and geographical conditions.[31] The last part deals with comprehensive data analysis of a set of more than 4000 olive oils with known geographical origin. Different strategies for data reduction were established and a workflow comprising different statistical models was developed for the prediction of the geographical origin of olive oils.

## 2. Experimental Section

### 2.1. Samples

Different olive oils covering a wide range of variations of FA- and TAG-profiles determined by GC analyzed in 2009 to 2011 according DGF standard methods were used for the NIR calibration and validation measurement (**Table 1**).

To verify the geographical origin of olive oil, since 2011, olive oil samples have been collected from various sellers, producers, fillers, im- and exporters, and organizers of competitions worldwide. The origin of these samples was reported on the packaging or guaranteed from the producers. Finally, more than 14 000 NIR spectra of olive oil samples from different geographical origins (Table S1, Supporting Information) covering a wide range of different varieties, crop years and sensory qualities were selected for building the statistical models.

### 2.2. FT-NIR Spectroscopy

Olive oils were filled in 8 mm disposable glass vials and submitted for FT-NIR analysis after being thermally equilibrated to 50 °C for 5 min. All spectra were recorded in triplicate. Spectra were obtained in transmission mode from 11 500 to 4000 $cm^{-1}$. Each spectrum was time-averaged based on 32 scans at a resolution of 8 $cm^{-1}$ using a Bruker MPA-FT-NIR spectrometer (Bruker Optik GmbH, Ettlingen, Germany) equipped with OPUS software version 7.8.

Methyl esters of the FA were analyzed according to method DGF C-VI 10 (13) in combination with DGF C-VI 11d (98) and for the individual TAG method DGF C-VI 14 (08) was used.[32] IV was determined as described in method DGF C-V 11d (14). These standard methods are technically equivalent to the international standards (ISO) and provide the needed precision data for repeatability and reproducibility. No precision data is given in the corresponding EU methods.[33] The official European standard for analyzing TAG uses isocratic non-aqueous reversed-phase HPLC with refractive index detector neglecting the advantages of GC over HPLC such as better separation efficiencies, reproducibility of retention data, and the availability of an universal detector. The TAG are separated by HPLC according to their equivalent carbon number (ECN). This chromatographic system does

**Table 1.** NIR specifications for the different methods.

| Parameter | Repeatability sdv [sR] | Set | Number of samples | Sample | Range | Wavelength regions [1/cm] | Data treatment | Rank | $R^2$ [Regression model = PLS2] | RMSEP | MD limit |
|---|---|---|---|---|---|---|---|---|---|---|---|
| C16:0 | 0.09 | Calibration | 131 | Olive oil | 6.0–18.6 | 5384.7–5292.1; 5110.8–4833.1 | 1st Deriv. and vector normalization | 18 | 98.2 | | |
| | 0.49 | Validation | 130 | | 6.0–18.6 | 4655.7–4470.5 | | | 96.3 | 0.256 | 0.59 |
| C16:1 [n-7] | 0.04 | Calibration | 302 | Olive oil | 0–1.7 | 5384.7–5199.5; | 1st Deriv. | 17 | 95.6 | | |
| | 0.49 | Validation | 290 | | 0.1–1.8 | 4929.5–4833.1 | | | 95.6 | 0.0733 | 0.27 |
| C18:0 | 0.04 | Calibration | 218 | Olive oil | 1.0–3.7 | 9114.6–7718.3; 6792.6–5863; | 1st Deriv. and vector normalization | 18 | 94.5 | | |
| | 0.25 | Validation | 214 | | 1.6–3.8 | 5400.1–4933.4 | | | 89.2 | 0.154 | 0.50 |
| C18:1 [n-9] | 0.08 | Calibration | 358 | Olive oil/ | 3.3–78.2 | 6838.9–6241; | vector normalization | 15 | 99.8 | | |
| | 0.88 | Validation | 351 | vegetable oils | 3.3–77.6 | 5654–5060.7 | | | 99.7 | 0.635 | 0.39 |
| C18:1 [n-11] | Not available | Calibration | 129 | Olive oil | 0.9–3.7 | 9878.4–9199.5; 7857.2–5836; | 1st Deriv. and vector normalization | 15 | 96.8 | | |
| | Not available | Validation | 127 | | 1.3–3.8 | 5168.7–4493.7 | | | 94.0 | 0.108 | 0.46 |
| C18:2 [n-6] | 0.12 | Calibration | 341 | Olive oil/ | 3.0–63.2 | 10 383.7–8609.3; | 1st Deriv. and vector | 10 | 99.9 | | |
| | 0.62 | Validation | 337 | vegetable oils | 3.0–63.2 | 5654.7–5060.7 | normalization | | 99.8 | 0.411 | 0.35 |
| C18:3 [n-3] | 0.04 | Calibration | 299 | Olive oil/ | 0–52.1 | 9793.5–7425.2; | 1st Deriv. and vector | 18 | 99.9 | | |
| | 0.22 | Validation | 294 | vegetable oils | 0.1–55.0 | 5654.7–5060.7 | normalization | | 99.9 | 0.082 | 0.47 |
| Iodine value | 0.07 | Calibration | 144 | Vegetable oils | 30.6–133.2 | 8971.2–8323.3; | vector normalization | 12 | 99.9 | | |
| [IV] | 0.13 | Validation | 105 | | 33.7–133.2 | 8138.13–8045.5 | | | 99.8 | 0.030 | 0.23 |
| POP | 0.16 | Calibration | 105 | Olive oil/ vegetable oils | 3.1–7.1 | 10 383.7–3789.6; 9203.3–8015.3; | 1st Deriv. and vector normalization | 11 | 95.6 | | |
| | 0.32 | Validation | 96 | | 3.1–7.1 | 7429–6835 | | | 91.5 | 0.287 | 0.39 |
| POO | 0.33 | Calibration | 259 | Olive oil/ | 1.7–30.7 | 5465.6–5361.5; | vector-normalization | 16 | 98.1 | | |
| | 1.32 | Validation | 194 | vegetable oils | 1.7–30.7 | 5060.6–4651.8 | | | 97.8 | 1.15 | 0.21 |
| PLP | 0.10 | Calibration | 304 | Olive oil | 0.1–3.0 | 9002.6–7442; | 1st Deriv. and vector | 17 | 93.3 | | |
| | 0.18 | Validation | 274 | | 0.1–3.0 | 5554.3–5214.9 | normalization | | 90.7 | 0.122 | 0.25 |
| PLO | 0.21 | Calibration | 112 | Olive oil | 3.7–17.1 | 10 383.9–9199.5; 8019.2–6241; | 1st Deriv. and vector normalization | 16 | 99.1 | | |
| | 0.38 | Validation | 100 | | 4.2–13.1 | 5656.7–5060.7 | | | 98.3 | 0.241 | 0.51 |
| OOO | 0.57 | Calibration | 109 | Olive oil | 22.1–53.2 | 10 383.7–3241; | 1st Deriv. and vector | 17 | 98.9 | | |
| | 1.78 | Validation | 103 | | 22.4–49.9 | 5064.5–4470.5 | normalization | | 98.9 | 1.14 | 0.56 |
| OLO | 0.15 | Calibration | 112 | Olive oil | 3.7–16.3 | 8812.9–8015.1; | 1st Deriv. and vector | 11 | 95.3 | | |
| | 0.44 | Validation | 100 | | 3.0–16.1 | 5654.5–5060.5 | normalization | | 93.0 | 0.562 | 0.34 |
| OLL | 0.51 | Calibration | 110 | Olive oil | 0.5–3.8 | 8613.2–7425.2; | vector normalization | 15 | 95.3 | | |
| | 1.86 | Validation | 105 | | 0.5–3.7 | 5064.5–4470.5 | | | 94.4 | 0.157 | 0.47 |
| LLL | 0.30 | Calibration | 97 | Olive oil | 0–22.3 | 10 383.7–8015.3 | 1st Deriv. and vector | 13 | 99.9 | | |
| | 0.85 | Validation | 96 | | 0.0–22.3 | | normalizatio | | 99.9 | 0.041 | 0.50 |
| PLL | 0.09 | Calibration | 60 | Olive oil | 0.6–2.7 | 8018.9–7424.9; | 1st Deriv. and vector | 7 | 91.1 | | |
| | 0.25 | Validation | 51 | | 0.0–2.6 | 5064.4–4470.4 | normalization | | 87.7 | 0.187 | 0.42 |

Reference methods: FA: DGF C-VI 10 (13) in combination with DGF C-VI 11d (98); TAG: DGF C-VI 14 (08); IV: DGF C-V 11d (14).

not allow to determine the relevant individual TAG because of many co-eluting TAG.[34]

### 2.3. Statistical Analyses

Reported data were expressed in terms of the means and SD. The XLSTAT software (version 2019.1.3. Addinsoft Deutschland, Andernach, Germany) was applied for the Kolmogoroff test for normal and experimental distribution, outlier tests according Dixon, descriptive statistics, *k*-means clustering, nearest neighbor (kNN), logit regression (LR), linear discriminant analysis (LDA), and naïve Bayes test.

### 2.4. Development of NIR Methods

NIR methods for the determination of the FA and TAG composition were built by calibration of the NIR spectra against the reference method using a set of samples which was analyzed by both techniques. Test set calibration was applied. It is important that the calibration and validation set covers a wide range of variability representing the product variation.[35] The scanned NIR spectra were statistically evaluated by validated calibration software to develop the multivariate equations using partial least squares (PLS2) algorithm. Mathematical data treatment within the NIR calibration process was conducted with OPUS/Quant 2 (Bruker Optik GmbH, Ettlingen, Germany). For the calculation 30–50%

of all samples were selected as test samples per random access. All methods were built with 155 up to 746 samples for the calibration and the validation (Table 1). In general, more than 150 samples are needed for calibration and test in test validation.[35] Wavelength range and data treatment (first derivative, vector normalization or a combination of both) were optimized individually for each parameter with the aim to generate the most suitable calibration with a low prediction error (RMSEP) and high regression factors ($R^2$).

## 3. Results and Discussion

The FA composition is often used as identity criteria for edible oils and fats in official standards. Moreover, FA composition is also proposed to detect adulteration with foreign oils or to classify olive varieties.[36,37] Some promising attempts have been made to confirm authenticity of vegetable oils based on their TAG and FA profiles because the composition of any vegetable or animal fat or oil is generally defined in terms of the nature and distribution of the FA present in the TAG.[38] The information provided by FA pattern is much enhanced by the combination with data on the TAG pattern. Moreover, the analytical approach measuring the individual FA and TAG and finally combine both is strong, since not only a single analyte (marker) is taken into consideration but several in combination.[24] In addition the statistical combination of several markers makes manipulation more difficult, since it is not possible to manipulate several substances by dilution or removal whose contents also influence each other.

Apart from the traditionally popular Mediterranean basin, the cultivation of the olive tree is spreading worldwide to other countries like the United States, Australia, South America, and Middle Eastern countries. Actually still about 80% of the total world production of olive oil is made in Europe. Consequently, a large data set including also olive oil from non-European countries is needed to construct statistical prediction models for the geographical authenticity of olive oils worldwide.

NIR has become one of the most used analytical techniques in routine analysis of food because it provides quick and economic analysis which does not need skilled staff and no sample preparation. The structural features of the TAG molecules with different FA attached at different position of the glycerol backbone produce different characteristic absorptions in the spectrum. Therefore, NIR methods for the most relevant FA (C16:0, C16:1 [n-7], C18:1 [n-9], C18:1 [n-11], C18:2 [n-6], C18:3 [n-3]) and TAG (POP, POO, PLP, PLO, OOO, OLO, OLL, LLL, PLL) and the IV had been developed. The number of samples used for building the NIR methods ranged from 60 (calibration set) and 51 (validation set) samples for PLL up to 358 (calibration set) and 351 (validation set) samples used for the calibration of oleic acid (Table 1). The concentration range, wavelength area and the data treatment used for the calibration of the single FA and TAG can be found in Table 1. Root mean standard error of prediction (RMSEP) is often used to compare the accuracy and correlation between the reference method and the NIR method. The values obtained for the optimized NIR methods are shown in Table 1. They were found to be in a similar range as the repeatability standard deviation of the reference methods or even lower because many systematic errors such as changing operator, instrumentation and sample preparation and derivatization which occur within the traditional analysis process must not be considered.[39,40]

The Mahalanobis distance (MD) indicates if an unknown sample fits to the population of the calibration or if its composition is too different. The MD for the test set samples were all within an acceptable range, also indicating that the calibration set comprised a suitable set of oils. The results for RMSEP, $R^2$, and MD limit demonstrate that the developed NIR analysis is a suitable alternative for the time-consuming and laborious reference methods traditionally used for analyzing FA, TAG, and IV. Another advantage of NIR methods is the usually given stability of the measurements over time. For chromatographic methods it is often harder to get this stability of the measurement over time because columns and instrument might change their performance with time so that it is more difficult to get identical results over a period of several months. All findings based on the NIR analysis can be repeated also with other NIR instruments or traditional analytical methods in a laboratory because all reference data of the conventional methods for calibrating and validating the FT-NIR methods were obtained using international standards. However, the calibration and validation of the developed NIR methods cannot be simply transferred from one unit to a unit of another producer due to small differences in the optical systems.

The developed NIR methods for FA, TAG, and IV were applied to analyze FA compositions and the TAG profiles of olive oil samples from 19 countries derived from different olive varieties and produced within the last 10 years. Details about number of samples per geographical origin and the varieties are given in Table S1, Supporting Information. This comprehensive data set was used to establish a statistical workflow which allows determining the geographical origin of olive oils form all over the world.

### 3.1. Statistical Evaluation

Many analytical methods or techniques concerning the authenticity of olive oil have been developed only with a limited number of samples or just to confirm the feasibility of the method without elucidating the effects on olive composition besides geographical origin. In most cases, only one statistical tool (LDA) is used to evaluate the analytical data neglecting different assumptions to apply the statistical test or the structure and variation of the data set. The main effects on olive oil composition are the different cultivars and their blend, soil, climate, ripeness, technology of production, storage time, and geographical position (altitude). Each factor may have a different impact on the FA and TAG profile. Consequently, the requisite of any authentication method is to have available a large number of reference data of olive oils from many sources of variation which may have an influence on the oil composition. Samples have to represent most of the variability which might occur due to the conditions during cultivation and production mentioned above. Therefore, a good sampling representing most of the variability is necessary to produce accurate results in the chemometric evaluation.

It is also a fact of statistics, that the number of samples and the distribution of the different classes (i.e., countries) can have a significant impact on the final result of the statistical evaluation.[40]

Some statistical methods like so-called parametric tests like LDA require that the number of objects in each class of the training set should be approximately equal otherwise the class with most representatives will always be selected. Other (non-parametric) methods like LR or kNN make obviously no assumptions concerning a normal distribution of the objects in a class.

The first step after analyzing an adequate number of samples is to make a data cleaning of the raw data which is absolutely necessary for successful data mining. The aim of any data reduction is to obtain a reduced training set without a significant loss in classification accuracy and to generate a new, more balanced spectrum of representatives.[42,43] Erroneous values caused by measurement errors have to be removed. Duplicate data is another source of error as it increases the relevance of this multiple samples. All these kind of data were detected by applying *k*-means clustering as a statistical tool.[44] *k*-means clustering is an iterative method which, wherever it starts from, converges on a solution with a target within class variance of less than 2% and a much higher variance compared to the other classes. Applying this technique, distinct patterns are evaluated in order to group similar or duplicate objects together. They are classified stepwise iteratively into *k* number of clusters in which each observation belongs to the cluster with nearest mean and lowest variance. The algorithm continues until no observation (instance) change the cluster membership. Instead of taking the mean value of the variables in a cluster as a reference point, the most centrally located object is used which represents now all samples in the group (cluster). By this treatment the data sets were reduced for more than 20% from 5177 to 4093 datasets (i.e., Italy 1473 to 1213; Spain 1232 to 1062; Greece 1310 to 960; Table S1, Supporting Information).

### 3.2. Selecting Principal Variables

Principal component analysis (PCA) is one of the most frequently used multivariate data analysis methods to reduce the number of variables and to create new, completely independent composite variables from the variables of the raw database. In addition to the different FA, TAG, and the IV which were determined by NIR, the list of variables has been extended with ratios of some variables as proposed by Rossell[45] to improve the discriminating power (Table S3, Supporting Information). Only those ratios were used which produce the highest differences. All variables including ratios were statistically independent and do not correlate which other variables.

The analytical parameters provide a global characterization of the samples and finally the classes to be differentiated. The availability of large sets of data does not mean a maximum of relevant information. The main goal of a preselection is to remove variation within the data that does not pertain to the analytical information needed for a specific unknow sample. There are no rules about which data or variable reduction strategy is optimal for a given problem so that it still contains the information of the large set. PCA looks at the data set as a whole and select the components that describe the majority of the variance.[46] But often there are data available where the separation is not based on the highest variance and the use of the most important components of PCA will not work because of the lack of normal distributed values. To exclude the influence of inhomogeneous data sets (number and distribution), the data of Italy, Spain and Greece were selected to find out the variables which may be relevant for the differentiation of the origin. The number of samples per group of these three countries is almost equal.

Applying PCA in XLSTAT program, using the Kaiser–Meyer–Olkin criterion[47] with 0.87 (=very good) the following variables were selected: C16:0, C16:1, C18:1 (n-9), C18:1 (n-11), POP, PLP, PLO, OOO, OLO, OLL, OOO/OLO, POO/PLO, OLO/POP. These parameters selected by PCA were used to perform a LDA analysis. A total correctness of prediction of only 61.7% (ESP: 72.7%, GRE: 60.4%, ITA: 51.7%) was achieved. Additionally, the discriminating effect of the variables was examined by another test which gives a more significant improvement of the fit. Hereby, each variable is stepwise eliminated to evaluate the influence of the selected variables on the prediction accuracy determined by LDA. This process is repeated until no further variables can be deleted without a statistically significant loss of fit. Table S4, Supporting Information, shows the changing correctness of the predictions during this process determined by applying LDA. Starting with 24 different variables the number of parameters was reduced to 14 (C16:0, C16:1, C18:0, C18:1 (n-9), C18:2 (n-6), IV, POP, POO, PLP, OOO, OLO, OLL, the ratios C16:0/C18:2 (n-6), and C18:1(n-9)/C18:2 (n-6) without a loss of the initial prediction accuracy (Table S4, Supporting Information).

Many publications generally reported better results with 90% or higher. This may be the case if sampling is restricted to a few countries or regions. In addition, often only monovarietal olive oils were analyzed. In the present study, the origin and the olive varieties were known only to a maximum of 50%. The rest of the olive oils were oils that have been mixed from different regions in a country using different varieties. After all, the total number of analyzed olive oils in this study is approximately ten times higher.

The descriptive statistics (minimum, maximum, and median) for the 14 selected variables applied to the training set are very similar. No significant differences could be seen visually when comparing the results for the countries (Table S2, Supporting Information) except Tunisia and Lebanon.

It is believed that the FA and TAG composition is also influenced by the variety and not only by the geographical origin because monovarietal olive oils have specific flavor characteristics related to the olive variety from which they are elaborated. Some varieties like Arbequina and Koroneiki are planted all over the world. The analysis of some oils extracted from Arbequina and Koroneiki olives (Tables S5 and S6, Supporting Information) produced in different countries demonstrates that a typical pattern for a special olive variety could not be detected. However, when comparing samples from different topographical locations in Spain or Greece changes can be observed in FA and TAG distributions as a function of north to south or island or mainland. Obviously, the FA composition of olive oil is more influenced by climate factors and geographical origin than by cultivar, maturation stage of fruit or harvest year. This observation is encouraging to develop a chemometric model based on the FA and TAG composition to identify the geographical origin of olive oils with a better prediction power.

### 3.3. Generating a Training Set with Preselected Data

The absolute values of the individual parameters vary greatly from one parameter to the other (Table S3, Supporting Information). For stearic acid values of up to 4.4% were determined, whereas the linoleic acid values varied from 2.5 to 17%. Therefore, database normalization is needed to improve data integrity.[48] Min–max normalization is one of the most common tools to normalize data. Applying the min–max normalization all numeric ranges of the individual feature such as steric acid or linoleic acid were reduced to a scale between 0 and 1.[49]

An advantage of min–max rescaling is that the ranges of different parameters are equalized, which allows a better differentiation and improves the data integrity. Within the data set derived from more than 4000 samples from 19 countries the numbers of samples per country vary widely between 3 and about 1200. Due to this inhomogeneity the whole training set cannot be used for a statistical evaluation, that is, classification. It is necessary to start with another preselection of those data sets which are more similar to the sample to be identified and reduce the number of countries. kNN has long been used in pattern recognition and data analysis. However is has also been used[50] for the similarity search in large databases.

The advantage of the kNN tool is that it does not scale or it is not specific to certain similarity measures. It is a simple method to assign those samples which are the nearest to the query sample. kNN measures the distance for continuous variables for instance as Euclidean distances. Deriving the Euclidean distance between two data points involves computing the square root of the sum of the squares of the differences between corresponding values.

After the min–max normalization of all variables, kNN algorithm measuring Euclidean distances is used to select the first 200 samples nearest to the query sample for the final statistical evaluation. The data are ordered by increasing differences of distance. It seems to be an efficient method for obtaining a ranking of all objects in approximate order of similarity to the reference object.[51] For the further statistical evaluation the data of these 200 preselected objects are taken. Table S7, Supporting Information, demonstrates that variances and ranges of the different variables have been drastically reduced. The number of query countries is considerably reduced, too. The composition and ranking of the individual datasets in this preselection of 200 data sets change with each new sample since kNN with the basic data in the training set is performed again for each sample.

A complication in applying LDA or also other classification tools to real data occurs also[52] when the number of analytical parameters measured in a sample exceeds the number of records of the same class ("overfitting"). Therefore, the number of objects in every class must be larger than 14 for all of the preselected countries. It means that in the 200 preselected data sets a country must be represented 14 times or more otherwise it will not be considered for the test. In this case, the number of data sets for the final evaluation is lower than 200.

In a next step, the reduced data set which is very similar to the sample composition (Table S7, Supporting Information) can be used to predict the origin of the sample by applying different statistical strategies (e.g., kNN, LDA, LR, and naïve Bayes test), if the data meets the assumptions of the applied statistical tool. The combination of different statistical tools may provide additional information, which might not be available when only using one single method. However, it has been observed that different statistical tools provide not always the same results and different geographical origins are indicated. When the assumptions of the statistical test method are not fulfilled, the results of the analysis can be misleading or wrong. Many statistical tests like LDA requires normally distributed data. Such kind of tests are called parametric test. In case of non-parametric data non-parametric tests like LR can be used instead as such tests do not rely on a specific probability distribution function. Other tests require that the different classes comprise an almost equal number of objects. For instance, we are running a classification model with a training set consisting of 95 records, A, and 5 records, B. The classification model simply predicts A and achieves 95% classification accuracy which is not correct. For these two reasons, a balancing of the training data set is recommended for classification and set aside a number of non-rare records to obtain a number of about 15–25% rare records. For instance, to achieve a 20% balance within a sample set comprising 5 rare records and 95 non-rare records in the unbalanced data, the number of the non-rare records have to be reduced the number of 25. The balancing proportion can be relatively low (e.g., 10%) if the analyst is confident that the rare group comprise sufficiently rich variety of records. However, the balancing proportion should be higher, for example, 20%, if the analyst is not so confident about this circumstance.[53] The reduction is performed by a randomized selection of data of the country from the training set.

Due to different assumptions of statistical tests, large differences in the number of objects per country in the final training set have to be compensated by offsetting the proportions of data sets of the most frequently represented countries in favor of the less present ones.

### 3.4. Validation of the Statistical Model

As mentioned above, LDA is expected to work well if the numbers in each class of the training set is approximately equal and approximately normally distributed. Consequently, LDA does not work correctly if the datasets are not balanced and the number of objects of each country is highly different. Furthermore, LDA is not applicable (inferior) for non-linear (binary) problems. LDA is often used to differentiate properties of samples whereas the binary logit model is often applied to model the impact of properties on a binary phenomenon (e.g., YES or NO; ESP or ITA). Therefore, in cases with less than three classes, it is necessary to apply another appropriate statistical method. Applying LR the dependent variable must be a dichotomy (i.e., two categories), that means a binary variable coded as 0 and 1. In the regression analysis, the metric dependent variable Y is directly estimated, whereas the LR only tries to calculate the probability of the occurrence of values of the dependent variable as a function of various influencing variables such as TAG or FA. LDA is a more appropriate method when the explanatory variables are normally distributed but fails when number of categories is really small (<4). Therefore, it is recommended to apply different tests to get a verification of the results. Another tool is naïve Bayes classifier. The

naïve Bayes classifier is a supervised machine learning algorithm that allows classifying a set of observations according to a set of rules determined by the algorithm itself. Naïve Bayes works quite well with low amounts of data. Presence of one particular feature does not affect the other (="naïve"). Based on all objects in the database it is the aim to find the class with a maximum of probability.

The kNN tool is another possible statistical tool. This supervised classification method is based on the distance of the objects in a multidimensional space, defined by the variables. To classify an unknown sample the distance between the unknown sample and a set of samples with known class membership is calculated. Then, the predicted class is assigned as the class of the k samples nearest to it. This conceptually simple approach works well in many situations, but it is important to realize the limitations. The numbers in each class of the training set should be approximately equal otherwise the "votes" will be biased toward the class with most representatives. Another problem of kNN is that the tool does not learn anything from the training data and uses the training data itself for classification without any filtering of noisy data or neglecting bad data sets. In literature there are many discussions about the advantages and disadvantages of all these statistical tools.[41]

The goal of classification is to build a model to predict the outcome of a new observation based on observable predictors using the training set. The number of the observations in the 19 classes (countries) is varying from 3 to more than 1200 per country. The majority of classifiers such as naïve Bayes, LDA, kNN, and LR are sensitive to different proportions of the classes.[54] These algorithms tend to favor the class accuracy in the imbalanced data set due to the effect of the majority class. Several classification tools can be used (KNN, LR, LDA, naïve Bayes classifier) and the different assumptions for their application can be ignored to a certain extent if the number of objects in the classes are balanced. **Tables 2–4** shows that the probabilities ($p$) of correct predictions calculated by the tests increased if the database is more balanced. The accuracy of the different test results including misclassification rate can be compared with a confusion matrix of the training set.

It is necessary to develop applicable concepts for identifying the geographical origin of olive oils in routine. The general workflow presented in this study is shown in **Figure 1**. It is recommended to start with kNN. kNN is one of the simplest classification algorithms. To measure the distance between test data and each row of the training set data different functions can be used (Euclidian, Manhattan, Minkowski, Tanimoto, Jaccard, Mahalanobis, Chebyshew, cosine, etc.). Practical experiences show that the Manhattan distances seem to perform better in terms of lowest RMSE (=root-mean-square error) over various values of k. Manhattan distance is less sensitive to outliers and more sensitive to small scale behavior than the Euclidian distance function.

kNN is directly applied to analyze the entire unbalanced data set ($n$ = 200) because no assumptions have to be considered. However, the results can be still influenced by the chosen $k$-value. One method to validate the number of clusters is the Elbow method.[55] If $k$ = 5 is chosen as the appropriate number in this dataset with 200 clusters a correctness of prediction (85–90%) can be expected. kNN can only be used to get first information

**Table 2.** Validation of one olive oil sample from Italy. (Sicily, var. Nocellara del Belice).

| Training set | Sample | Number of samples | C16:0 | C16:1 [n-7] | C18:0 | C18:1 [n-9] | C18:2 [n-6] | IV | POP | POO | PLP | OOO | OLO | OLL | kNN | LDA | | Naïve Bayes | | LR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | Confusion matrix – training set [% correct] | Prediction [$p$ = probability] | Confusion matrix – Training set[% correct] | Prediction [$p$ = probability] | Confusion matrix – Training set [% correct] | Prediction [$p$ = probability] |
| Unbalanced | ESP | 35 | 12.9 | 1.1 | 3.1 | 69.8 | 8.7 | 80.9 | 5.6 | 27.5 | 1.4 | 40.4 | 9.0 | 1.6 | ITA- | 66.7 | 0.000 | 33.3 | 0.000 | | |
| | FRA | 3 | 13.1 | 1.1 | 2.9 | 69.1 | 8.8 | 81.5 | 5.7 | 28.5 | 1.5 | 39.5 | 8.9 | 1.4 | ITA- | 83.3 | 0.000 | 80.1 | 0.000 | | |
| | GRE | 18 | 12.5 | 1.0 | 3.2 | 70.7 | 8.4 | 80.3 | 5.1 | 27.6 | 1.2 | 41.5 | 9.1 | 1.3 | ITA- | 66.7 | 0.000 | 0.00 | 0.000 | | |
| | ISR | 3 | 13.1 | 1.1 | 3.4 | 68.8 | 9.0 | 80.8 | 5.6 | 27.9 | 1.7 | 37.0 | 9.4 | 1.6 | POR- | 86.6 | **0.983** | 66.7 | **0.995** | | |
| | **ITA** | 82 | 13.1 | 1.0 | 3.6 | 68.9 | 8.7 | 81.2 | 5.6 | 28.7 | 1.6 | 38.3 | 8.8 | 1.4 | TUR- | 80.0 | 0.000 | 48.6 | 0.000 | | |
| | PAL | 5 | 13.2 | 0.9 | 3.9 | 68.0 | 9.1 | 80.4 | 6.0 | 28.9 | 1.3 | 38.4 | 9.4 | 1.7 | | 70.0 | 0.000 | 47.2 | 0.000 | | |
| | POR | 10 | 13.2 | 1.0 | 3.0 | 69.1 | 8.7 | 81.0 | 5.8 | 28.7 | 1.5 | 36.7 | 8.9 | 1.5 | | 72.2 | 0.011 | 63.4 | 0.004 | | |
| | TUR | 36 | 13.0 | 1.0 | 3.4 | 69.0 | 9.1 | 80.9 | 5.7 | 28.0 | 1.5 | 38.4 | 9.0 | 1.5 | | 84.4 | 0.000 | 29.8 | 0.000 | | |
| Balanced | ESP | 34 | 12.6 | 1.1 | 3.2 | 69.9 | 8.7 | 81.0 | 5.6 | 27.4 | 1.4 | 40.6 | 9.0 | 1.6 | | 72.4 | **0.963** | 7.7 | **0.994** | 88.2 | **0.987** |
| | **ITA** | 29 | 12.9 | 1.0 | 3.5 | 69.0 | 8.8 | 81.3 | 5.6 | 28.6 | 1.5 | 38.9 | 8.9 | 1.4 | | 80.0 | 0.037 | 12.5 | 0.006 | 92.5 | |
| | TUR | 37 | 13.0 | 1.0 | 3.4 | 69.1 | 9.0 | 80.9 | 5.7 | 28.0 | 1.5 | 38.5 | 9.0 | 1.5 | | | | | | | |

**Table 3.** Validation of one olive oil sample from Spain (Alicante, var. Changlot Real, Picual).

| Training set | | Number of samples | C16:0 | C16:1 [n-7] | C18:0 | C18:1 [n-9] | C18:2 [n-6] | IV | POP | POO | PLP | OOO | OLO | OLL | kNN | LDA | | Naïve Bayes | | LR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | Confusion matrix – Training set [% correct] | Prediction [$p$ = probability] | Confusion matrix – Training set [% correct] | Prediction [$p$ = probability] | Confusion matrix – Training set[% correct] | Prediction [$p$ = probability] |
| Sample | | | 11.8 | 0.9 | 3.1 | 73.8 | 5.9 | 79.6 | 5.1 | 28.7 | 1.1 | 43.5 | 6.5 | 0.9 | | | | | | | |
| Unbalanced | **ESP** | 76 | 11.4 | 0.8 | 3.1 | 73.7 | 6.2 | 80.3 | 4.8 | 27.0 | 0.9 | 46.7 | 7.5 | 1.0 | POR- | 86.7 | **0.864** | 61.8 | 0.044 | | |
| | GRE | 101 | 11.6 | 0.8 | 3.0 | 73.7 | 6.3 | 79.8 | 4.7 | 27.8 | 0.9 | 45.8 | 7.6 | 0.9 | ESP- | 92.9 | 0.011 | 68.3 | 0.090 | | |
| | ITA | 8 | 11.6 | 0.8 | 3.0 | 73.7 | 6.4 | 79.7 | 4.6 | 28.5 | 0.9 | 44.3 | 7.2 | 0.9 | GRE- | 37.5 | 0.001 | 12.5 | 0.808 | | |
| | POR | 8 | 11.5 | 0.8 | 3.0 | 73.4 | 6.1 | 80.5 | 4.9 | 27.4 | 0.8 | 45.2 | 7.3 | 1.0 | GRE- | 12.5 | 0.125 | 0.0 | 0.058 | | |
| Balanced | **ESP** | 22 | 11.5 | 0.8 | 3.1 | 73.4 | 6.3 | 80.3 | 4.9 | 27.4 | 0.9 | 45.9 | 7.5 | 1.0 | GRE | 95.5 | **1.00** | 76.0 | **0.692** | 100.0 | **1.00** |
| | GRE | 17 | 11.6 | 0.8 | 3.1 | 73.5 | 6.3 | 79.7 | 4.7 | 27.7 | 0.9 | 45.4 | 7.5 | 0.9 | | 100.0 | 0.000 | 50.0 | 0.308 | 100.0 | |

**Table 4.** Validation of an olive oil sample from Turkey (Izmir, var. Venos).

| Training set | | $N$ | C16:0 | C16:1 [n-7] | C18:0 | C18:1 [n-9] | C18:2 [n-6] | IV | POP | POO | PLP | OOO | OLO | OLL | kNN | LDA | | Naïve Bayes | | LR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | Confusion matrix – Training set [% correct] | Prediction [$p$ = probability] | Confusion matrix – Training set [% correct] | Prediction [$p$ = probability] | Confusion matrix – Training set [% correct] | Prediction [$p$ = probability] |
| Sample | | | 12.7 | 0.6 | 3.3 | 72.0 | 6.4 | 79.1 | 5.5 | 26.6 | 0.9 | 37.8 | 7.3 | 0.6 | | | | | | | |
| Unbalanced | ESP | 54 | 11.5 | 0.8 | 3.1 | 73.0 | 6.6 | 80.8 | 4.9 | 27.0 | 0.9 | 44.4 | 7.8 | 1.0 | TUR- | 74.5 | 0.000 | 63.0 | 0.000 | | |
| | GRE | 97 | 11.7 | 0.8 | 2.9 | 73.1 | 6.7 | 79.8 | 4.8 | 27.5 | 0.9 | 43.5 | 8.1 | 0.9 | TUR- | 90.6 | 0.000 | 53.5 | 0.000 | | |
| | IST | 4 | 11.8 | 0.9 | 2.6 | 73.5 | 6.4 | 80.7 | 4.9 | 27.5 | 0.8 | 42.1 | 7.5 | 0.8 | TUR- | 100.0 | 0.000 | 75.0 | 0.000 | | |
| | ITA | 18 | 11.8 | 0.8 | 3.0 | 72.3 | 7.1 | 80.1 | 5.0 | 27.9 | 1.0 | 40.3 | 8.2 | 0.9 | TUR- | 38.9 | 0.000 | 61.1 | 0.000 | | |
| | POR | 8 | 11.5 | 0.7 | 2.9 | 73.4 | 6.4 | 80.5 | 4.9 | 27.2 | 0.8 | 44.3 | 7.7 | 0.9 | TUR- | 0 | 0.000 | 25.0 | 0.000 | | |
| | **TUR** | 14 | 12.6 | 0.7 | 3.2 | 72.4 | 6.7 | 79.4 | 5.5 | 27.7 | 1.0 | 39.0 | 7.3 | 0.7 | | 78.5 | **1.000** | 78.5 | **1.000** | | |
| Balanced | ESP | 33 | 11.5 | 0.8 | 3.1 | 72.9 | 6.5 | 80.7 | 4.9 | 27.0 | 0.9 | 44.2 | 7.8 | 1.0 | | 87.9 | 0.000 | 76.5 | 0.000 | | |
| | GRE | 35 | 11.8 | 0.8 | 3.0 | 73.1 | 6.7 | 79.9 | 4.7 | 27.7 | 0.9 | 43.4 | 8.1 | 1.0 | | 88.6 | 0.000 | 45.7 | 0.000 | | |
| | ITA | 15 | 11.8 | 0.8 | 3.0 | 72.2 | 7.1 | 80.2 | 5.1 | 27.8 | 1.0 | 40.6 | 8.3 | 1.0 | | 53.3 | 0.000 | 61.1 | 0.000 | 100.0 | |
| | **TUR** | 154 | 12.6 | 0.7 | 3.2 | 72.4 | 6.7 | 79.4 | 5.5 | 27.7 | 1.0 | 39.0 | 7.3 | 0.7 | | 85.7 | **1.000** | 78.6 | **1.000** | 100.0 | **1.00** |

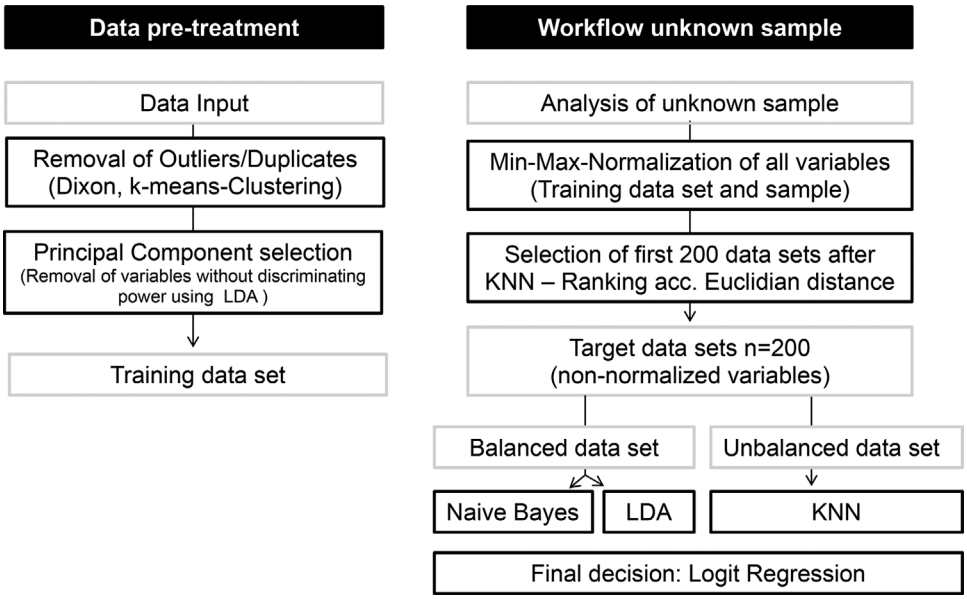


**Figure 1.** Overview showing the different steps of data pretreatment of the comprehensive data set and the proposed statistical workflow for the determination of the origin of an unknown sample (data sets are bordered in gray; mathematical operations are indicated by a black box).

(Table 2a–c). Next, LDA and naïve Bayes can be applied. The results for LDA and naïve Bayes are better when balanced data are used. For the final decision between two countries, LR is used. When a binary outcome variable is modeled using LR, it is assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables. If the probability of an event is 0.8, the probability of failure is 0.2 (1–0.8). The odds of success are defined as the ratio $0.8/0.2 = 4$, then the odds of success are 4:1. If the probability of success is 0.5 then the odds of success is 1:1. The odds are transformed into probability with $p = \text{odds}/(1+\text{odds})$. $p$ is defined only as the relative probability that this sample represents this country, which means a value of $p = 0.5$ does not correspond to a 50/50 blend of the tested two countries. The LR analysis is used in this study to make a final decision between two countries which were also proposed as alternatives by kNN, LDA, and naïve Bayes (Figure 1). The training set for the balanced data set usually showed a sufficient correctness (>80%). If this percentage of correctness will not be achieved, it can be assumed that the training set might not be suitable and care has to be taken into account when making a statement about the verification of the labeled origin of this sample.

If an oil is a blend of oil from two or more countries of origin this results in a new FA- and TAG-profile of the sample, which will in most cases result in a lower probability for the labeled country. However, it might also happen that the FA- and TAG profile is more similar to another country and thus results in a high probability for a country not related to the sample. For this reason, the proposed model should only be used for the verification of a given origin and it is not possible to predict the percentages of blends from different origin.

After identification of the country is done, the origin can even be specified to a specific region. This is much easier because the observed changes in the individual parameters are more characteristic, if only the geographical region in a country has to be identified. The olive varieties grown here and the topological location (altitude, north-south gradient or island location) influence characteristic changes in the TAG-FA pattern. For all well-known regions in, Greece, Italy, Spain, and Portugal samples with verified region (see Figure 1a–c) were collected and analyzed. Therefore, a preselection to reduce the records does not seem to be necessary. It is sufficient to remove duplicates by applying *k*-means clustering (see above) and to adjust the number of objects in each class in order to have a balanced sample set. In almost all cases, LDA can be successfully used to identify the region. Even for Toledo, Cordoba or Jaen, which are all parts of Andalusia, the more specific classification is possible, and they are not simply grouped to Andalusia. The centroids of the different important regions in Spain, in Greece such as Peloponnes and Crete, Portugal, and many regions in Italy are well separated (**Figure 2**) and might be used for a chemometric evaluation of the different regions within one country.

Nevertheless, the results illustrate that a single statistical test is not sufficient to make a correct statement about the origin of an olive oil. The changing compositions of the reference data and the possible different assumptions for the applicability of a statistical test require the verification of the results by different statistical tests. However, it is not possible to directly determine the geographic region of a sample unless the country has been verified in advance. It could happen that a Greek olive oil would then be identified as Italian oil from Apulia. The same would happen if the database contains only three countries such as Italy, Spain, and Greece. In that case, a Turkish olive oil might to be assigned as a Spanish olive oil due to the statistics missing reference data. An extensive data set that is not limited to a few countries is therefore an important prerequisite for a more reliable statement.

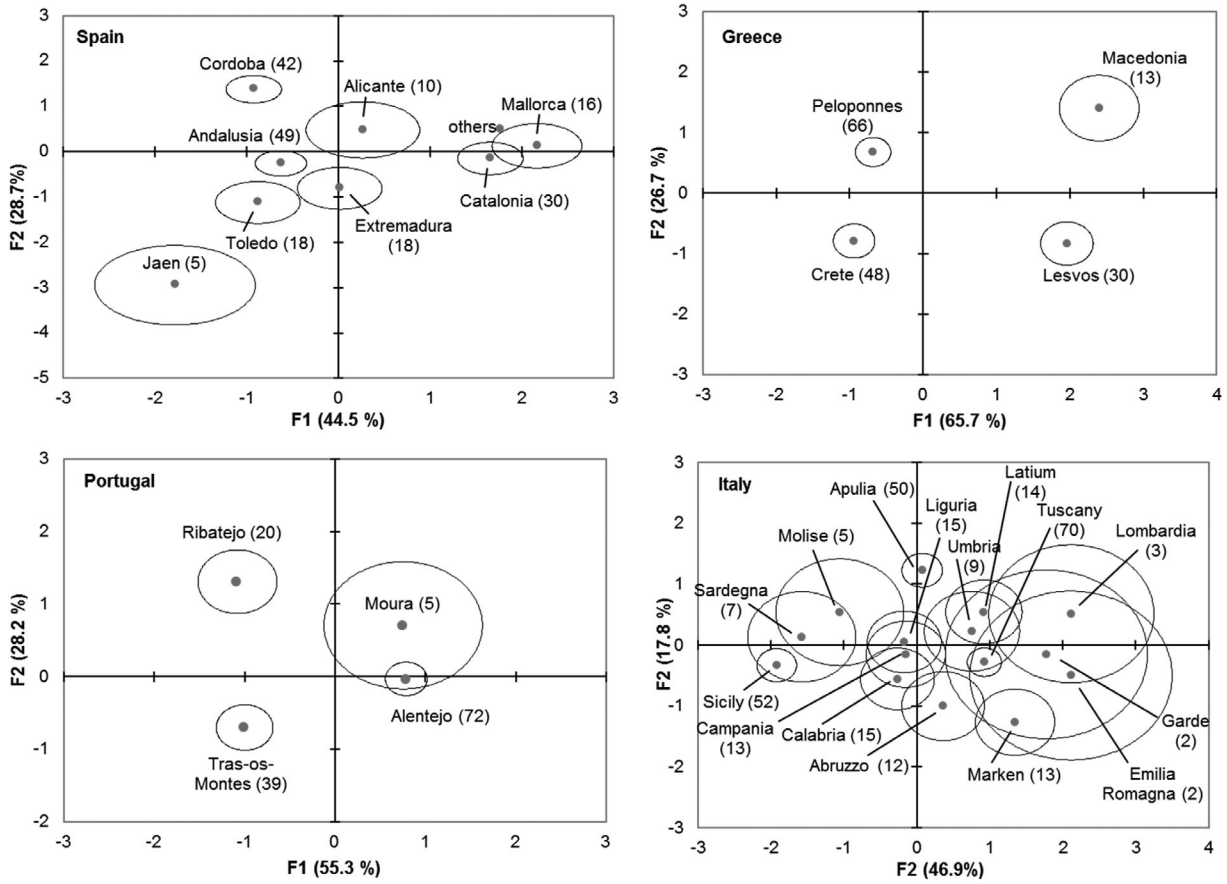


**Figure 2.** Classification of olive oils according to the distinct regions of Spain, Greece, Portugal, and Italy by LDA. Shown are centroids and 95% confidence intervals (total sample number: Spain $n = 188$, Greece $n = 157$, Portugal $n = 136$, Italy $n = 282$).

## 4. Conclusion

The authentication of olive oil samples requires usually the use of sophisticated and time-consuming analytical techniques. NIR spectroscopy analysis associated to chemometric tools is a fast and efficient method to ensure the traceability of olive oils. It could be demonstrated that FA and TAG composition are appropriate analytical parameters to identify the origin of olive oils because they seemed to be independent of the variety and quality of the olives before harvest, the extraction process of oils, and environmental conditions. A preselection of data with a reduction to the most similar samples is necessary to apply several different statistical methods for classification. Furthermore, the number of the different classes in the training data set has to be balanced to avoid misclassifications due to different parametric assumptions of the applied statistical tools. The calibration and validation of NIR methods is based on reference data obtained using standard GC methods. The high correlation for all analytical parameters allows this statistical approach to be applied to data not only provided by NIR but also by GC analysis.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

The authors are very grateful to the members of the German and Swiss Olive Oil Panel for providing a huge number of samples.

## Conflict of Interest

The authors declare no conflict of interest.

## Keywords

authenticity, chemometrics, FT-NIR, geographical origin, olive oil

Received: July 9, 2019
Revised: August 7, 2019
Published online:

[1] T. Del Giudice, C. Cavallo, F. Caracciolo, G. Cicia, *Agric. Food Econ.* **2015**, *3*, 20.
[2] S. Portarena, C. Baldacchini, E. Brugnoli, *Food Chem.* **2017**, *215*, 1.
[3] M. Casale, C. Casolino, P. Oliveri, M. Forina, *Food Chem.* **2010**, *118*, 163.

[4] R. Korifi, Y. Le Dreau, J. Molinet, J. Artaud, N. Dupuy, *J. Raman Spectrosc.* **2011**, *42*, 1540.
[5] L. Mannina, M. Patumi, N. Proietti, D. Bassi, *J. Agric. Food Chem.* **2001**, *49*, 2687.
[6] H. S. Tapp, M. Defernez, E. K. Kemsley, *J. Agric. Food Chem.* **2003**, *51*, 6110.
[7] M. S. Cosio, D. Ballabio, S. Benedetti, C. Gigliotti, *Anal. Chim. Acta* **2006**, *567*, 202.
[8] F. Souayah, NR. Rodrigues, A. Veloso, LG. Dias, J. A. Pereira, S. Oueslati, AM. Peres, *J. Am. Oil Chem. Soc.* **2017**, *94*, 1417.
[9] F. Longobardi, A. Ventrella, G. Caseillo, D. Sacco, L. Catucci, A. Agostiano, M. G. Kontominas, *Food Chem.* **2012**, *133*, 579.
[10] C. Montealegre, M. L Alegre, C. Garcia-Ruiz, *J. Agric. Food Chem.* **2010**, *58*, 28.
[11] P. Dais, E. Hatzakis, *Anal. Chim. Acta* **2013**, *765*, 1.
[12] A. Bajoub, A. Bendini, A. Fernández-Gutiérrez, A. Carrasco-Pancorbo, *Crit. Rev. Food Sci. Nutr.* **2018**, *58*, 832.
[13] N. G. Kamoun, W. Zarrouk, *Int. J. Food Sci. Technol.* **2012**, *47*, 1496.
[14] A. M. Gómez-Caravaca, R. M. Maggio, L. Cerretani, *Anal. Chim. Acta* **2016**, *913*, 1.
[15] R. Aparicio, M. T. Morales, R. Aparicio-Ruiz, N. Tena, D. L. García-Gonzáleza, *Food Res. Int.* **2013**, *54*, 2025.
[16] A. Sacco, M. A. Brescia, V. Liuzzi, F. Reniero, G. Guillou, St. Ghelli, P. van der Meer, *J. Am. Oil Chem. Soc.* **2000**, *77*, 619.
[17] R. M. Alonso-Salces, N. Segebarth, S. Garmón-Lobato, M. V. Holland, J. M. Moreno-Rojas, J. A. Fernández-Pierna, V. Baeten, S. R. Fuselli, B. Gallo, L. A. Berrueta, F. Reniero, C. Guillou, K. Héberger, *Eur. J. Lipid Sci. Technol.* **2015**, *117*, 1991.
[18] S. Gómez-Alonso, V. Mancebo-Campos, M. Desampaardos Salvador, G. Fregapane, *Food Chem.* **2007**, *100*, 36.
[19] R. Ben-Ayed, N. Kamoun-Grati, A. Rebai, *Compr. Rev. Food Sci. Food Saf.* **2013**, *12*, 218.
[20] F. Caponio, T. Gomes, A. Pasqualone, *Eur. Food Res. Technol.* **2001**, *212*, 329.
[21] R. Boggia, F. Evangelisti, N. Rossi, P. Salvadeo, P. Zunin, *Grasas y Aceites* **2005**, *56*, 276.
[22] M. A. Shaker, A. A. Azza, *Int. Food Res. J.* **2013**, *20*, 197.
[23] C. Montealegre, M. L. M. Alegre, C. Garcia-Ruiz, *J. Agric. Food Chem.* **2010**, *58*, 28.
[24] N. Semmar, S. Laroussi-Mezghani, N. Grati-Kamoun, M. Hammami, J. Artaud, *Food Chem.* **2016**, *208*, 150.
[25] A. Mihailova, D. Abbado, S. D. Kelly, N. Pedentchouk, *Food Chem.* **2015**, *173*, 114.
[26] N. K. Andrikopoulos, I. G. Giannakis, V. Tzamtzis, *J. Chromatogr. Sci.* **2001**, *39*, 137.
[27] E. Christopoulou, M. Lazaraki, M. Komaitis, K. Kaselimis, *Food Chem.* **2004**, *84*, 463.
[28] F. Ulberth, M. Buchgraber, *Eur. J. Lipid Sci. Technol.* **2000**, *102*, 687.
[29] L. S. Conte, A. Bendini, E. Valli, P. Luccia, S. Moret, A. Maquet, F. Cacoste, P. Brereton, D. L. Garcia-Gonzáles, W. Moreda, T. G. Toschi, *Trends Food Sci. Tech.* [In press]. https://doi.org/10.1016/j.tifs.2019.02.025
[30] H. Azizian, J. K. G. Kramer, *Lipids*. **2005**, *40*, 855.
[31] S. Vichi, L. Pizzale, L. S. Conte, *Eur. J. Lipid Sci. Technol.* **2007**, *109*, 72.
[32] *DGF Deutsche Einheitsmethoden zur Untersuchung von Fetten: Fettprodukten. Tensiden und verwandten Stoffen*. Wissenschaftliche Verlagsgesellschaft, Stuttgart, Germany **2013**.
[33] European Commission Regulation. *Off. J. Eur. Union* **2007**, *L161*, 11.
[34] W. Moreda, M. C. Pérez-Camino, A. Cert, *Grasas y Aceites* **2003**, *54*, 175.
[35] M. Casale, R. Simonetti, *J. Near Infrared Spectrosc.* **2014**, *22*, 59.
[36] W. Zarrouk, B. Baccouri, W. Taamali, A. Triguli, D. Daoud, M. Zarrouk, *Grasas y Aceites* **2009**, *60*, 500.
[37] F. M. Haddada, H. Manai, I. Oueslati, D. Daoud, J. Sánchez, E. Osario, M. Zarrouk, *J. Agric. Food Chem.* **2007**, *55*, 10941.
[38] M. Buchgraber, F. Ulberth, H. Emons, E. Anklam, *Eur. J. Lipid Sci. Technol.* **2004**, *106*, 621.
[39] N. M. Faber, B. R. Kowalski, *J. Chemom.* **1997**, *11*, 181.
[40] P. Williams, P. Dardenne, P. Flinn, *J. Near Infrared Spectrosc.* **2017**, *25*, 85.
[41] R. G. Brereton, *Chemom. Intell. Lab. Syst.* **2015**, *149*, 90.
[42] D. R. Wilson, T. R. Martinez, *Mach. Learn.* **2000**, *38*, 257
[43] W. Towler, *Analytics* **2015**, *Nov/Dec*, *58*.
[44] *Improving k-Means by Outlier Removal* (Eds: V. Hautamäki, S. Cherednichenko, I. Kärkkäinen, T. Kinnunen P. Fränti in H. Kalviainen) Springer, Berlin **2005**.
[45] J. B. Rossell, *Fat Sci. Technol.* **1991**, *93*, 526.
[46] K. J. Parsins, W. J. Cooper, R. C. Albertson, *PLoS One*, **2009**, *4*, e7957.
[47] H. F. Kaiser, *Psychometrika* **1974**, *39*, 31.
[48] E. F. Codd, *IBM Research Report*, IBM, San Jose, CA, **1971**.
[49] I. B. Mohamad, D. Usman, *Res. J. Appl. Sci., Eng. Technol.* **2013**, *6*, 3299.
[50] N. S. Altman, *Am. Stat.* **1992**, *46*, 175.
[51] V. Athitsos, J. Alon, S. Sclaroff, G. Kollios, *CVPR* **2004**, *2*, 268.
[52] A. M. Martinez, A. C. Kak, *TPMAI*, **2001**, *23*, 228.
[53] D. T. Larose, C. D. Larose, *Data Mining and Predictive Analytics* (2nd ed.), Wiley, New York **2015**, Ch. 7.
[54] A. Santacruz, Why it is important to work with a balanced classification dataset, http://amsantac.co/blog/en/2016/09/20/balanced-image-classification-r.html (accessed: August 2019).
[55] D. Ketchen, C. Shook, *SMJ* **1996**, *17*, 441.