

Differentiation of Edible Oils by Type Using Raman Spectroscopy and Pattern Recognition Methods

Francis Kwofie¹, Barry K. Lavine¹, Joshua Ottaway², and Karl Booksh²

¹Department of Chemistry, Oklahoma State University, Stillwater, OK 74078

²Department of Chemistry and Biochemistry, University of Delaware, Newark, DE 19716

Corresponding author email: bklab@chem.okstate.edu

Abstract

The application of Raman spectroscopy and pattern recognition methods to the problem of discriminating edible oils by type was investigated. Two-hundred and eight six Raman spectra obtained from 53 samples spanning 15 varieties of edible oils were collected for 90 seconds at 2 cm^{-1} resolution. Employing a Whittaker filter, all Raman spectra were baseline corrected after removing the high intensity fluorescent background in each spectrum. The Raman spectral data were then examined using the three major types of pattern recognition methodology: mapping and display, discriminant development and clustering. The 15 varieties of edible oils could be partitioned into five distinct groups based on their degree of saturation and the ratio of polyunsaturated fatty acids to monounsaturated fatty acids. Edible oils assigned to one group could be readily differentiated from those assigned to other groups, whereas Raman spectra within the same group more closely resembled each other and therefore would be more difficult to classify by type.

Keywords: Raman spectroscopy, baseline correction, classification of edible oils, pattern recognition, variable selection, genetic algorithms

Introduction

Edible oils are mixtures of triglycerides that differ in their relative composition of fatty acids (e.g., palmitic, oleic, steric and linoleic). Because of their high nutritional value, edible oils are an important component of the human diet.¹ They are a source of essential fatty acids and are a carrier of fat soluble vitamins.^{2,3} Edible oils are used in cooking and are also ingredients in many processed or precooked foods because of their sensory characteristics. Most edible oils sold commercially are derived from plants, e.g., olive, corn, canola, sunflower, peanut, and safflower oil, although some are derived from animals, e.g., tallow and lard.

Adulteration of more expensive edible oils, for example, extra virgin olive oil or sesame oil by substitution or by blending with less expensive oils (such as corn or canola oil), is a problem that is of concern to government and regulatory officials.⁴ Analysis of edible oils for purposes of classification or authentication is usually carried out by gas chromatography/mass spectrometry (GC–MS).⁵ However, GC–MS analysis of edible oils can be laborious and time consuming. Approximately twenty years ago, analysis of edible oils (e.g., olive, sesame, and canola) for authentication,⁶ adulteration,⁷ and classification⁸ was demonstrated using Raman spectroscopy, which allowed spectra to be collected for samples in a short time period without the need for sample preparation. More recently, multivariate classification and calibration methods have been applied to Raman and IR spectra of edible oils to improve classification success rates^{9–15} and lower detection limits for adulterants.^{16–19} Generally, classification success rates of around 90% for edible oils have been reported in the literature. However, these studies, which used the 900 cm^{-1} to 1800 cm^{-1} region, were typically limited to 20 samples spanning five or six edible oils using either PLS or linear discriminant analysis to perform a flat (nonhierarchical) classification of the spectral data. Furthermore, the edible oils investigated were represented by samples obtained from a single brand within a limited production year range.

Raman spectra of edible oils usually contain weak narrow bands superimposed over a broad, high intensity fluorescence background (often referred to as baseline) that may distort the Raman bands of the components characteristic of the sample.²⁰ In these circumstances, baseline correction is necessary to ensure successful numerical processing of the Raman spectra using multivariate methods of data analysis such as principal component analysis (PCA) or partial least squares.²¹ In this study, 286 Raman spectra of 15 varieties of edible oils obtained from 53 samples from different vendors (possibly the same company, but a different batch and from a different manufacturing plant) were collected at relatively short integration times to test the robustness of the analysis to noise. The relatively large number of varieties (i.e., classes), samples, and spectra (i.e., replicates) were necessary to build better statistical distributions of expected in-class variance to determine classification performance when developing discriminants from training sets and to have sufficient number of samples to construct independent training and validation sets.

All Raman spectra used in this study were baseline corrected with the high intensity fluorescent background in each spectrum removed. The filtered spectral data were investigated using the three major types of pattern recognition methodology: mapping and display, discriminant development and clustering. The fifteen varieties of edible oils investigated were partitioned into 5 distinct groups based on their degree of saturation. Edible oils assigned to one group could be readily differentiated from those assigned to another group, whereas Raman spectra in the same group more closely resembled each other and therefore were more difficult to discriminate by type.

Experimental

Edible oils used in this study were purchased from supermarkets in the Newark, DE metropolitan area. The 53 samples, purchased over three years, spanned 15 varieties of edible oils (Table I). Each sample represented a different brand of edible oil or the same brand but a different production year. A QE65000 Raman spectrograph (Ocean Optics) equipped with a Hamamatsu TE cooled CCD detector (Hamamatsu) and a Inphotonics fiber optic probe for sampling was used to collect 286 Raman spectra of the 15 edible oils/edible oil blends at 785 nm. Each Raman spectrum ($2000\text{--}50\text{ cm}^{-1}$), which was collected for 90 seconds integration at 2 cm^{-1} resolution, consisted of 1044 points.

Data Preprocessing and Pattern Recognition Analysis

For each Raman spectrum (Figure 1a), noise and background were separated from the signal by a three-step procedure performed using the PLS Toolbox 8.6 (Eigenvector Technology). First, a Whittaker filter²² was applied to each Raman spectrum. Next, the baseline corrected Raman spectra were smoothed using a Savitzky–Golay linear filter with a 15-point window, and the spectra were then normalized to unit length. Figure 1b shows a representative Raman spectrum of corn oil that was baseline corrected, smoothed, and normalized to unit length. The bands at 2200 cm^{-1} and 300 cm^{-1} were not informative as they were an artifact of applying the Whittaker filter to the entire Raman spectrum (Figure 1a). This artifact in the filtering process occurred due to large changes in scattering intensity in these regions. The Raman spectra of edible oils in the region 1000 cm^{-1} to 500 cm^{-1} do not have active Raman bands.²³ If one looks at the signal to noise of all the Raman spectra for this region, it is very low. For these reasons, the baseline

corrected, smoothed and normalized spectra (shown in Figure 1b) were truncated to 1772.6 cm^{-1} – 1127.6 cm^{-1} (Figure 1c), which corresponded to 361 points. Below this region, the spectra were too noisy, and above this region there did not appear to be any information present about edible oil type. Table II lists the wavenumbers of the five Raman active bands in this region characteristic of the major components (i.e., triglycerides) found in the edible oils.²³

The 286 Raman spectra were divided into a training set of 257 spectra and a validation set of 29 spectra. Every tenth Raman spectrum was assigned to the validation set, with some substitutions to avoid having too few training set spectra for some edible oils. For pattern recognition analysis, each Raman spectrum was initially represented as a 361 dimensional data vector $\mathbf{x} = (x_1, x_2, x_3, \dots, x_j, \dots, x_{361})$ where x_j is the Raman intensity of the j th point. Thus, each spectrum is considered as a point in a 361 dimensional measurement space. A basic assumption inherent in this analysis is that distances between pairs of points (i.e., Raman spectra of edible oil samples) in this measurement space are inversely related to their degree of similarity. Points representing samples from one class of edible oils (e.g., Raman spectra of corn oil) will cluster in a limited region of this space distant from the points corresponding to the other class (e.g., Raman spectra of olive oil). Pattern recognition is a set of methods for investigating data represented in this manner in order to assess its general structure, which is defined as the overall relation of each Raman spectrum to every other spectrum in the data set.

In this study, PCA^{24,25} was used to analyze the 257 Raman spectra comprising the training set. PCA attempts to reduce the dimensionality of the data by computing a new set of basis vectors that are better at conveying the information present in the data (e.g., edible oil type) than axes defined by the original measurement variables. These axes, which represent the directions of greatest variance in the data, are called the principal components of the data. Each principal component is a linear combination of the original measurement variables. Often, only the two or three largest principal components are necessary to explain the information content of a data set when there are a large number of correlated measurement variables. By analyzing a data set using principal component (PC) plots, it is possible to find relationships in the data, e.g., identify similarities and differences among varieties of edible oils and to detect the presence of outliers (i.e., discordant observations) in the data.

To identify wavelengths characteristic of each class (i.e., the variety of an edible oil), a genetic algorithm (GA) for pattern recognition analysis^{26–30} was applied to the Raman spectra. The

pattern recognition GA identifies those wavelengths that optimize the separation of the classes in a plot of the two or three largest principal components of the data. Because principal components capture the bulk of the variance in the data, the wavelengths selected by the pattern recognition GA contain information primarily about differences between the varieties of edible oils in this data set. The PCA routine embedded in the fitness function of the pattern recognition GA acts as an information filter, significantly reducing the size of the search space since it restricts the search to wavelengths whose PC plots show clustering of the Raman spectra on the basis of their class label. In addition, the pattern recognition GA focuses on those classes (i.e., varieties of edible oils) and/or samples that are difficult to classify as it trains by boosting the value of the class and sample weights. Raman spectra that are classified correctly are not as heavily weighted as those spectra which are difficult to classify. Over time, the pattern recognition GA learns the optimal values of its parameters in a manner similar to that of a neural network.

To track and score the PC plots generated by the pattern recognition GA during each generation, class and sample weights are computed (Eqs. 1 and 2), where $CW(c)$ is the weight of class c (with c varying from 1 to the total number of classes in the data set), and $SW_c(s)$ is the weight of sample s in class c . Class weights sum to 100, and the sample weights for spectra from a particular class sum to a value equal to its class weight.

<AQ>Please check the equations carefully. NOTE: It is journal format that ALL Greek characters should appear in normal face (i.e., Roman, not italic). All variables represented by a Latin letter should be in italics. All constants and designators (i.e., λ_{max}) should be in normal (Roman) face. Vectors should be lowercase bold and matrices should be uppercase bold (both in normal, Roman, face).</AQ>

$$CW(c) = 100 \frac{CW(c)}{\sum_c CW(c)} \quad (1)$$

$$SW(s) = CW(c) \frac{SW(s)}{\sum_{s \in c} SW(s)} \quad (2)$$

Each PC plot generated for each feature subset after extracting the features from its chromosome is scored using the K-nearest neighbor classification algorithm.³¹ For each spectrum in the training set, Euclidean distances are computed between it and the other spectra which are also represented as points in the PC plot. These distances are arranged from the smallest to the largest. A poll is then taken of the K_c nearest neighbors of the spectrum. For the most rigorous classification of the data, K_c is assigned a value corresponding to the number of Raman spectra in the class to which the spectrum is a member. The number of K_c nearest neighbors with the same class label as the data point (i.e., spectrum) in question, the so-called sample hit count, $SHC(s)$, is computed ($0 < SHC(s) < K_c$) for each spectrum. It is then a simple matter to score a PC plot (Eq. 3). For class 1, SHC for each spectrum is divided by K_c and multiplied by $SW(s)$. This product/quotient for all class spectra is summed to yield the contribution of the class to the overall fitness score. This same calculation is repeated for the other classes with the fitness score of each class (i.e., variety of edible oil) summed to yield the overall fitness score, $F(d)$.

The fitness function of the pattern recognition GA is able to focus on those Raman spectra and/or specific varieties of edible oils that are difficult to classify by boosting sample and class weights over successive generations. To boost sample and class weights, it is necessary to compute both the sample-hit rate (SHR), which is the mean value of SHC/K_c over all feature

$$F(d) = \sum_c \sum_{s \in c} \frac{1}{K_c} \times SHC(s) \times SW(s) \quad (3)$$

subsets (which comprise the population of solutions) generated in a particular generation (Eq. 4), and the class-hit rate (CHR), which is the mean sample hit rate for all samples in a particular class (Eq. 5). The ϕ in Eq. 4 is the number of chromosomes in the population, and AVG in Eq. 5 refers to the average or mean value of SHR for the spectra comprising class g .

$$SHR(s) = \frac{1}{\phi} \sum_{i=1}^{\phi} \frac{SHC_i(s)}{K_c} \quad (4)$$

During each generation, class and sample weights are adjusted by a perceptron algorithm (Eqs. 6 and 7) with the momentum, P , set by the user and with $g + 1$ being the current generation

$$CHR_g(c) = AVG(SHR_g(s): \forall_{s \in c}) \quad (5)$$

and g being the previous generation. Classes (i.e., edible oils) with a lower class hit rate are boosted more heavily than classes that score well.

$$CW_{g+1}(s) = CW_g(s) + P(1 - CHR_g(s)) \quad (6)$$

$$SW_{g+1}(s) = SW_g(s) + P(1 - SHR_g(s)) \quad (7)$$

Boosting is crucial to the successful operation of the pattern recognition GA as it modifies the fitness landscape by adjusting the values of both the class and sample weights in each generation. This allows the GA to learn and as a result helps it to minimize the problem of premature convergence to a local optimum. Thus, the fitness function of the pattern recognition GA is continually changing as the population is evolving towards an optimal solution.

Results and Discussion

Outliers can obscure relationships present in multivariate data. For this reason, outlier analysis was performed on each variety of edible oil in the training set using PCA prior to classification of the data. One corn oil spectrum (spectrum 296 in Figure S1) and one extra virgin olive oil spectrum (spectrum 149 in Figure S2) were judged to be outliers based on the PC plot of each oil and a comparison of the average Raman spectrum of corn oil or extra virgin olive oil to the Raman spectrum of the suspected outlier (Figures 2 and 3). For this reason, these two Raman spectra were removed from the data set. In all likelihood, these two outliers were the result of small changes that occurred in the alignment of the fiber probe or the fiber optic connection to the monochromator of the Raman microscope. The spectra of the other replicates for these two samples were very similar to the average Raman spectrum of corn or extra virgin olive oil.

Because of the large number of classes in the data set, a hierarchical classification scheme was employed to discriminate the Raman spectra of the 15 varieties of the edible oils comprising the training set. To implement this scheme, the average Raman spectrum of each variety of edible oil was computed. The fifteen average Raman spectra were then analyzed using both PCA and cluster analysis.³² For both PCA and hierarchical cluster analysis (Ward's method), the spectral data were mean centered. Both the dendrogram and the PC plot of the average Raman spectra (Figure 2) were in good agreement, and each indicated that dividing the 15 classes of edible oils into five distinct groups is appropriate (Table III). A visual comparison

of the average Raman spectra revealed that spectra in the same oil group were more similar to each other than spectra in different oil groups. For this reason and because of the agreement in the results obtained from the PC score plot and the dendrogram, we chose to partition the edible oils into five groups.

The dataset consisted of 257 Raman spectra of 361 points each, which comprised the training set (Table IV), which was subjected to PCA. Figure 3 shows a plot of the two largest PCs of the 361 features obtained from the 257 Raman spectra comprising the mean-centered training set. Each spectrum is represented as a point in the PC plot. The five groups previously detected by cluster analysis are separated from each other in the PC plot. Since this projection is made without the use of information about the class assignment of each spectrum, the resulting separation is, therefore, a strong indication of real differences in the Raman spectral profile of these five edible oil groups.

A validation set of 27 Raman spectra (Table IV) was employed to assess the predictive capability of the PC plot developed from the training set data. The 27 Raman spectra were directly mapped onto the PC plot defined by the 257 spectra and 361 spectral features comprising the training set. Figure 4 shows the validation set samples projected onto the PC plot of the training set data. Each projected validation set spectrum lies in a region of the plot with Raman spectra from the same edible oil group. This suggests that the approach taken for discriminating these 15 edible oils by first dividing them into five groups is supported by the data.

Three different edible oils are well separated from the other twelve oils in the PC plot (Figures 5 and 6): Coconut oil (Group 4), flaxseed oil (Group 3) and the blend of avocado/flaxseed/olive oils (Group 5). The first principal component functionally distinguishes the oils by their degree of saturation. Of all the oils in this study (Table V), coconut oil has significantly higher percentage of saturated fats (~90%) than the other oils while flax seed oil has one of the highest percentages of polyunsaturated fats (~70%). As for the remaining 12 edible oils, they were assigned to Groups 1 or 2. Group 1 edible oils (extra virgin olive, extra light olive, pure olive, peanut, avocado, safflower, and hazelnut oils) tend to have a greater degree of unsaturation than those from Group 2 (corn, vegetable, canola, sesame, and grapeseed oils). This could manifest through a combination of a lower percentage of saturated fatty acids and a larger ratio of polyunsaturated fatty acids to monounsaturated fatty acids.

The next step was to investigate edible oils from Groups 1 and 2 individually using variable selection. The pattern recognition GA was applied to the 71 Raman spectra of Group 2 (Table VI) to identify spectral features that were discriminatory as the spectral profiles of the edible oils comprising this Group were similar. For this study, the mutation rate of the GA was set at 0.4 and the number of chromosomes at 10,000. After two-hundred generations, the pattern recognition GA identified 23 spectral features that were correlated with edible oil type. These spectral features were identified by sampling key feature subsets, scoring their PC plots, and tracking those edible oils and/or Raman spectra that were difficult to classify. The boosting routine of the pattern recognition GA used this information to steer the population to an optimal solution. Figure 5 shows a plot of the two largest principal components of the 23 spectral features identified by the pattern recognition GA. The eight validation set spectra assigned to Group 2 by the PC plot used to discriminate edible oils by their oil group (Figure 4) were projected onto the PC plot of the 71 Raman training set spectra from Group 2 and the 23 spectral features identified by the pattern recognition GA. All eight Raman validation set spectra (see Table VI) were correctly classified, i.e., they were projected in a region of the PC plot (Figure 6) with Raman spectra of the same type of edible oil.

Sesame, vegetable, and grapeseed oils were separated from each other and from the other edible oils in the plot (Figures 5 and 6), whereas canola and corn oil overlap. In other studies on the application of vibration spectroscopy to discrimination of edible oils, many authors have reported that corn oil can be readily discriminated from canola oil.³³ In these studies, the authors only addressed within supplier sample variation in making their best case assessment for classification. In this study, both sample to sample variation within a supplier and between suppliers were considered for classification, which would explain the differences in our reported results for corn and canola oil from those reported in previously published studies.

The within class variability associated with each variety of edible oil in the PC plot is correlated to the number of samples representing the edible oil, not the number of Raman spectra comprising the class. Both vegetable and sesame oil yielded compact and well separated clusters in the PC plot as each oil is represented by only a single brand (i.e., only within supplier variation for sesame oil but seasonal variation within supplier variation for vegetable oil). Grapeseed, corn, and canola oils do not form compact clusters in the PC plot as each edible oil is

represented by five or seven samples representing both within supplier including seasonal variation and between supplier variation.

The 174 Raman spectra comprising the training set for Group 1 (Table VII) were also investigated using the pattern recognition GA. Extra virgin olive oil, extra light olive oil, and pure olive oil could not be differentiated, which can probably be attributed to their triglyceride fraction being comparable.³⁴ For this reason, these three edible oils were merged into a single class. Variable selection was performed using the pattern recognition GA. After 200 generations, the pattern recognition GA identified 20 spectral features. Figure 7 shows a plot of the two largest principal components of the 174 Raman spectra comprising the training set and the 20 spectral features identified by the pattern recognition GA. Extra virgin olive oil, extra light olive oil, and pure olive oil are separated from the other edible oils in the PC plot (Figure 7). Safflower, hazelnut, and avocado oil form compact and well-defined clusters. Each contains spectra from only a single source (Table VI). By comparison, peanut oil spectra in the plot are divided into two clusters. The three peanut oil samples (obtained from two different suppliers) capture both within-supplier and between-supplier variability. When accounting for this second source of variability, it is plausible that peanut oil may not be able to be reliably differentiated from safflower oil using Raman spectroscopy.

Figure 8 shows the 18 validation-set Raman spectra assigned to Group 1 by the PC plot used to discriminate edible oils by group (Figure 6) projected onto the PC plot of the 174 Raman spectra comprising the training set for Group 1 and the 20 spectral features identified by the pattern recognition GA. The seventeen extra virgin olive oil, extra light olive oil, and pure olive oil spectra (Table VII) were assigned to the correct class in the PC plot. The Raman spectrum of safflower in the validation set was also correctly classified based on its projected location in the PC plot.

Conclusion

A hierarchical classification scheme was developed to differentiate Raman spectra of edible oils by type. The fifteen edible oils investigated in this study were divided into five groups using cluster analysis. Raman spectra in the same group were more similar to each other than Raman spectra from different edible oil groups. Edible oils assigned to Group 1 could be readily differentiated from those assigned to Groups 2, 3, 4, and 5, whereas Raman spectra of edible oils

within the same group more closely resembled each other and therefore were more difficult to classify by type. For example, corn and canola oil (Group 2), which previous workers have shown can be readily detected as adulterants in extra virgin olive oil (Group 1 oil) by Raman spectroscopy,^{35,36} can be readily differentiated from extra virgin olive oil using the hierarchical based classification scheme described in this study. This is clear from the PC plots shown in Figures 5 and 6.

Adulteration of extra virgin olive oil by less expensive grades of olive oil such as pure olive oil,⁴ which is in the same oil group as extra virgin olive oil, cannot be detected by Raman spectroscopy because of the similarity of the triglyceride profile of extra virgin olive oil and that of the adulterant. Furthermore, adulteration of extra virgin olive oil by hazelnut oil (both of which are in Group 1) is a difficult analysis problem as the Raman spectra of these two edible oils are very similar. Although Goodacre³⁷ has shown that PLS using all the wavelengths in the spectrum has the potential to detect hazelnut adulteration in extra virgin olive oil, his training set was restricted to a single source of extra virgin olive oil and hazelnut oil. The reported PLS results may represent a best-case scenario as the training set does not account for within-supplier and between-supplier variability. Clearly, the hierarchical classification scheme formulated in this study affords the analyst the means to realistically assess the difficulty associated with detecting the presence of a particular adulterant in a specific variety of edible oil using Raman spectroscopy.

These analyses also highlight the need for edible oil classification and adulteration studies that span multiple sources within each oil type. The supplier to supplier variation for edible oils (and possibly the seasonal variation within a supplier) is often greater than the within supplier variation. This work also shows that previous studies (which rely on only one source for each type of edible oil) provide an overly optimistic estimate of the ability to spectroscopically classify edible oils as to type or to detect low levels of adulterants in edible oils.

Acknowledgments

BKL and FK acknowledge the financial support of the National Science Foundation under award CHE-1506915. KB and JO acknowledge the financial support of the National Science Foundation under award CHE-1506853.

Supplemental Material

All supplemental material available in the text, consisting of Figures S1 and S2, is available in the on-line version of the journal.

References

1. P. Siri-Tarino, S. Chiu, N. Bergeron, R.M. Krauss. “Saturated Fats Versus Polyunsaturated Fats Versus Carbohydrates for Cardiovascular Disease Prevention and Treatment”. *Annu. Rev. Nutr.* 2015. 35: 517–543.
2. J. Orsavova, L. Misurcova, J.V. Ambrozova, R. Vicha, et al. “Fatty Acids Composition of Vegetable Oils and its Contribution to Dietary Energy Intake and Dependence of Cardiovascular Mortality on Dietary Intake of Fatty Acids”. *Int. J. Mol. Sci.* 2015. 16(6):12871–12890.
3. K. Chowdhury, L.A. Banu, S. Khan, A. Latif. “Studies on the Fatty Acid Composition of Edible Oil”. *Bangladesh J. Sci. Ind. Res.* 2007. 42(3): 311–316.
4. J.C. Moore, J. Spink, M. Lipp. “Development and Application of a Database of Food Ingredient Fraud and Economically Motivated Adulteration from 1980 to 2010”. *J. Food Sci.* 2012. 77(4): R118–R126.
5. T. Rezanka, P. Mares. “Determination of Plant Triacylglycerols Using Capillary Gas Chromatography, High-Performance Liquid Chromatography and Mass Spectrometry”. *J. Chromatogr.* 1991. 542: 145–159.
6. N.A. Marigheto, E.K. Kemsley, M. Defernez, R.H. Wilson. “A Comparison of Mid-Infrared and Raman Spectroscopies for the Authentication of Edible Oils”. *J. Amer. Oil Chem. Soc.* 1998. 75(8): 987–992.
7. V. Baeten, M. Meurens, M.T. Morales, R. Aparicio. “Detection of Virgin Olive Oil Adulteration by Fourier Transform Raman Spectroscopy”. *J. Agric. Food Chem.* 1996. 44(8): 2225–2230.
8. V.P. Baeten, M.T. Hourant, M.T. Morales, R. Aparicio. “Oil and Fat Classification by FT-Raman Spectroscopy”. *J. Agric. Food Chem.* 1998. 46(7): 2638–2646.
9. M.J. Lerma-Garcia, G. Ramis-Ramos, J.M. Herrero-Martinez, E.F. Simo-Alfonso. “Authentication of Extra Virgin Olive Oils by Fourier-Transform Infrared Spectroscopy”. *Food Chem.* 2010. 118(1): 78–83.

10. O.O. Abbas Galtier, Y. Le Dreau, C. Rebufa, J. Kister, et al. "Comparison of PLS1-DA, PLS2-DA and SIMCA for Classification by Origin of Crude Petroleum Oils by MIR and Virgin Olive Oils by NIR for Different Spectral Regions". *Vib. Spectrosc.* 2011. 55(1): 132–140.
11. R. Korifi, Y. Le Dreau, J. Molinet, J. Artaud, et al. "Composition and Authentication of Virgin Olive Oil from French PDO Regions by Chemometric Treatment of Raman Spectra". *J. Raman Spectrosc.* 2011. 42(7): 1540–1547.
12. Y. Saucedo-Hernandez, M.J. Lerma-Garcia, J.M. Herrero-Martinez, G. Ramis-Ramos, et al. "Classification of Pumpkin Seed Oils According to Their Species and Genetic Variety by Attenuated Total Reflection Fourier-Transform Infrared Spectroscopy". *J. Agric. Food Chem.* 2011. 59(8): 4125–4129.
13. X.F. Zhang, X.H. Qi, M.Q. Zou, F. Liu. "Rapid Authentication of Olive Oil by Raman Spectroscopy Using Principal Component Analysis". *Anal. Lett.* 2011. 44(12): 2209–2220.
14. S.F. Li, X.R. Zhu, J.H. Zhang, G.Y. Li, et al. "Authentication of Pure Camellia Oil by Using Near Infrared Spectroscopy and Pattern Recognition Techniques". *J. Food Sci.* 2012. 77(4): C374–C380.
15. P. Samyn, D. Van Nieuwerkerke, G. Schoukens, L. Vonck, et al. "Quality and Statistical Classification of Brazilian Vegetable Oils Using Mid-Infrared and Raman Spectroscopy". *Appl. Spectrosc.* 2012. 66(5): 552–565.
16. R. El-Abassy, M.P. Donfack, A. Materny. "Visible Raman Spectroscopy for the Discrimination of Olive Oils from Different Vegetable Oils and the Detection of Adulteration". *J. Raman Spectrosc.* 2009. 40(9): 1284–1289.
17. G. Gurdeniz, B. Ozen. "Detection of Adulteration of Extra-Virgin Olive Oil by Chemometric Analysis of Mid-Infrared Spectral Data". *Food Chem.* 2009. 116(2): 519–525.
18. W. Dong, Y.Q. Zhang, B. Zhang, X.P. Wang. "Quantitative Analysis of Adulteration of Extra Virgin Olive Oil using Raman Spectroscopy Improved by Bayesian Framework Least Squares Support Vector Machines". *Anal. Meth.* 2012. 4(9): 2772–2777.
19. X.F. Zhang, M.Q. Zou, X.H. Qi, F. Liu, et al. "Quantitative Detection of Adulterated Olive Oil by Raman Spectroscopy and Chemometrics". *J. Raman Spectrosc.* 2011. 42(9): 1784–1788.

20. G. Schulze, A. Jirasek, M.M.L. Yu, A. Lim, et al. "Investigation of Selected Baseline Removal Techniques as Candidates for Automated Implementation". *Appl. Spectrosc.* 2005. 59(5): 545–574.
21. M.A. Sharaf, D.L. Illman, B.R. Kowalski. *Chemometrics*. New York: John Wiley and Sons, 1986.
22. P.H.C. Eilers. "A Perfect Smoother". *Anal. Chem.* 2003. 75(14): 3631–3636.
23. C. Jimenez-Sanchidrian, J.R. Ruiz. "Use of Raman Spectroscopy for Analyzing Edible Vegetable Oils". *Appl. Spectrosc. Rev.* 2016. 51(5): 417–430.
24. I.T. Jolliffe. *Principal Component Analysis*. New York: Springer-Verlag, 1986.
25. J.E. Jackson. *A User's Guide to Principal Components*. New York: John Wiley and Sons, 1991.
26. B.K. Lavine, C.G. White, T. Ding, M.M. Gaye, et al. "Wavelet based Classification of MALDI-IMS-MS Spectra of Serum N-Linked Glycans from Normal Controls and Patients Diagnosed with Barrett's Esophagus, High Grade Dysplasia, and Esophageal Adenocarcinoma". *Chemom. Intell. Lab. Syst.* 2018. 176(1): 74–81.
27. M.M. Gaye, T. Ding, H. Shion, A. Hussein, et al. "Delineation of Disease Phenotypes Associated with Esophageal Adenocarcinoma by MALDI-IMS-MS Analysis of Serum N-Linked Glycans". *Analyst.* 2017. 142(9): 1525–1535.
28. A. Fasasi, N. Mirjankar, R.-I. Stoian, C. White, et al. Pattern Recognition Assisted Infrared Library Searching of Automotive Clear Coats". *Appl. Spectrosc.* 2015. 69(1): 84–94.
29. B.K. Lavine, A. Fasasi, N. Mirjankar, M. Sandercock, et al. "Search Prefilters for Mid-IR Spectra of Clear Coat Automotive Paint Smears Using Stacked and Linear Classifiers". *J. Chemom.* 2014. 28(5): 385–394.
30. B.K. Lavine, C. White, C.M. Sundling, C. Breneman. "Odor-Structure Relationship Studies of Tetralin and Indan Musks". *Chem. Senses.* 2012. 37(8): 723–736.
31. J. Tou, R.C. Gonzalez. *Pattern Recognition Principles*. Reading, Mass.: Addison-Wesley Publishing Company, 1974.
32. D.L. Massart, L. Kaufman. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*. New York: John Wiley and Sons, 1983.
33. C. Jimenez-Sanchidrian, J. Rafel Ruiz. "Use of Raman Spectroscopy for Analyzing Edible Vegetable Oils". *Appl. Spectrosc. Rev.* 2016. 51(5): 417–430.

34. F. Guimet, J. Ferre, R. Boque. “Rapid Detection of Olive-Pomace Oil Adulteration in Extra Virgin Olive Oils from the Protected Denomination of Origin Siurana Using Excitation-Emission Fluorescence Spectroscopy and Three-Way Methods of Analysis”. *Anal. Chim. Acta*. 2005. 544(1–2): 143–152.
35. Y. Rohman, A. Ismail, P. Hashim. "FTIR Spectroscopy Coupled with Chemometrics of Multivariate Calibration and Discriminant Analysis for Authentication of Extra Virgin Olive Oil". *Int. J. Food Prop.* 2017. 20(1): S1173–S1181.
36. C. Farley, A. Kassu, N. Bose, A. Jackson-Davis, et al. “Short Distance Standoff Raman Detection of Extra Virgin Olive Oil Adulterated with Canola and Grapeseed Oil”. *Appl. Spectrosc.* 2017. 71(6): 1340–1347.
37. E.C. Lopez-Diez, G. Bianchi, R. Goodacre, “Rapid Quantitative Assessment of the Adulteration of Virgin Olive Oils with Hazelnut Oils Using Raman Spectroscopy and Chemometrics”. *J. Agric. Food Chem.* 2003. 51(21): 6145–6150.

Table I. Edible Oil Data Set

Oil type	Oil ID	Number of samples	Number of spectra
Extra virgin olive oil (EVOO)	1	12	63
Extra light olive oil (ELOO)	2	4	31
Pure olive oil	3	6	75
Coconut oil	4	1	4
Avocado oil	5	1	4
Peanut oil	6	3	10
Corn oil	7	5	25
Grapeseed oil	8	5	29
Safflower oil	9	1	5
Hazelnut oil	10	3	5
Flaxseed oil	11	1	5
Canola oil	13	7	11
Avocado/flaxseed/olive oil	14	1	4
Sesame oil	15	1	10

Vegetable oil	17	2	5
Total	–	53	286

Table II. Raman shift assignments.

Wavenumber	Assignment
1270 cm ⁻¹	=CH deformation in an unconjugated cis C=C
1305 cm ⁻¹	In phase methylene twisting
1440 cm ⁻¹	CH ₂ scissoring deformation
1660 cm ⁻¹	C=C stretching (cis)
1750 cm ⁻¹	C=O stretching in an ester

Table III. Edible oil group assignments.

Edible oil	Edible oil ID	Edible oil group
Extra virgin olive oil	1	1
Extra light olive oil	2	1
Pure olive oil	3	1
Coconut oil	4	4
Avocado oil	5	1
Peanut oil	6	1
Corn oil	7	2
Grapeseed oil	8	2
Safflower oil	9	1
Hazelnut oil	10	1
Flaxseed oil	11	3
Canola oil	13	2
Avocado/flaxseed/olive oil	14	5
Sesame oil	15	2
Vegetable oil	17	2

Table IV. Training and validation set for edible oil group.

Edible oil	Group ID	Oil ID	Training set	Validation set
Extra virgin olive oil	1	1	54	8
Extra light olive oil	1	2	29	2
Pure olive oil	1	3	68	7
Coconut oil	4	4	4	0
Avocado oil	1	5	4	0
Peanut oil	1	6	10	0
Corn oil	2	7	21	3
Grapeseed oil	2	8	26	3
Safflower oil	1	9	4	1
Hazelnut oil	1	10	5	0
Flaxseed oil	3	11	5	0
Canola oil	2	13	11	0
Avocado/flaxseed/olive oil	5	14	3	1
Sesame oil	2	15	9	1
Vegetable oil	2	17	4	1
Total	—	—	257 spectra	27 spectra

Table V. Amounts of saturated and unsaturated fats in edible oils.

Edible oil	Saturated	Monounsaturated	Polyunsaturated
Olive oil	13%	74%	8%
Hazelnut oil	7%	80%	11%
Avocado oil	10%	70%	20%
Peanut oil	17%	49%	33%
Safflower oil	9%	12%	74%
Grapeseed oil	12%	17%	71%
Corn oil	13%	24%	59%
Sesame oil	14%	40%	46%

Canola oil	7%	58%	32%
Vegetable oil	15%	24%	61%
Flaxseed oil	10%	19%	68%
Coconut oil	87%	6%	2%

Table VI. Training and validation set for Group 2 edible oils.

Edible oil	Oil ID	Number of samples	Training set spectra	Validation set spectra
Corn oil	7	5	21	3
Grapeseed oil	8	5	26	3
Canola oil	13	7	11	0
Sesame oil	15	1	9	1
Vegetable oil	17	2	4	1
Total	—	20	71	8

Table VII. Training and validation set for Group 1 edible oils.

Edible oil	Oil ID	Number of samples	Training set spectra	Validation set spectra
Extra virgin olive oil	1	12	54	8
Extra light olive oil	2	4	29	2
Pure olive oil	3	6	68	7
Avocado oil	5	1	4	0
Peanut oil	6	3	10	0
Safflower oil	9	1	4	1
Hazelnut oil	10	3	5	0
Total	—	30	174	18

Figures

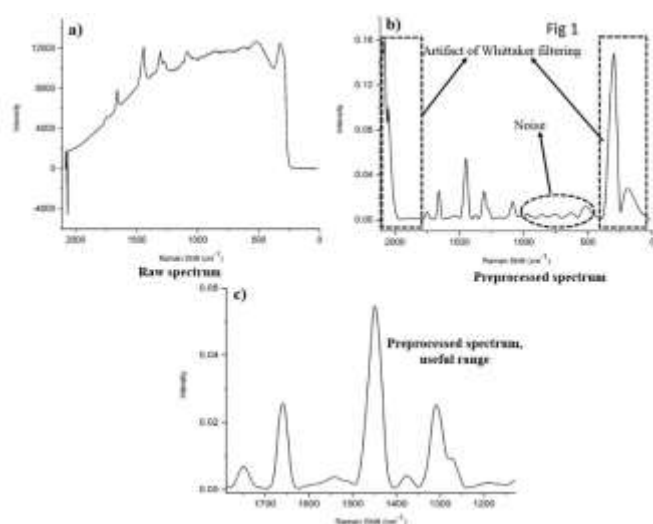


Figure 1. Representative Raman spectrum of corn oil: (a) before baseline correction, smoothing, and normalization to unit length, (b) after baseline correction, smoothing and normalization to unit length, and (c) truncation of the uninformative regions to yield the spectral range (1772.6 cm^{-1} – 1127.6 cm^{-1}) used for pattern recognition analysis.

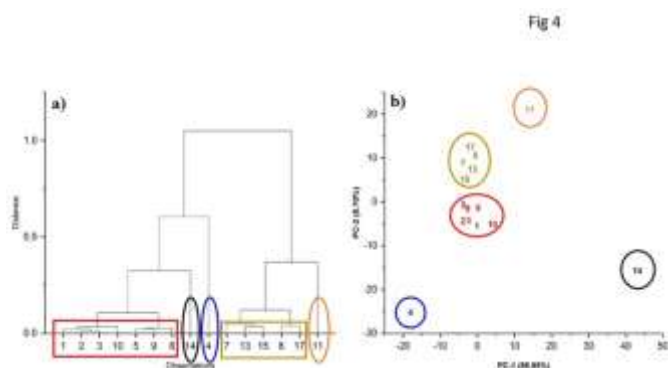


Figure 2. (a) Dendrogram (Ward's method) and (b) plot of the two largest principal components of the average Raman spectra of the 15 edible oils. The two plots agree and each indicates that the 15 varieties of edible oils investigated in this study can be divided into five distinct groups: Group 1: 1 = extra virgin olive oil; 2 = extra light olive oil; 3 = pure olive oil; 6 = peanut oil; 9 = safflower oil; 10 = hazelnut oil. Group 2: 7 = corn oil; 8 = grapeseed oil; 13 = canola oil; 15 = sesame oil; 17 = vegetable oil. Group 3: 11 = flaxseed oil. Group 4: 4 = coconut oil. Group 5: avocado/flaxseed/olive oil.

Fig 5

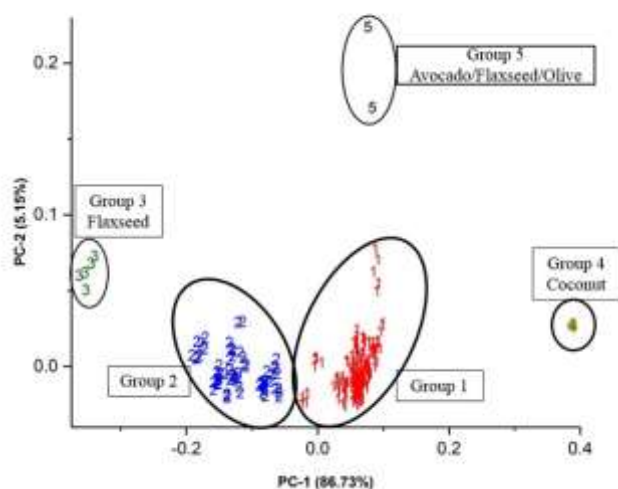


Figure 3. Plot of the two largest principal components of the 361 features obtained from the 257 Raman spectra comprising the mean-centered training set. 1 = Group 1; 2 = Group 2; 3 = Group 3; 4 = Group 4; and 5 = Group 5.

Fig 6

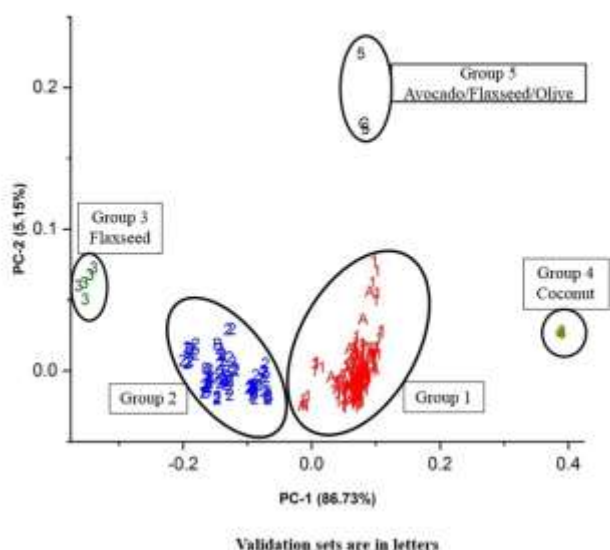


Figure 4. Projection of the 27 Raman spectra comprising the validation set onto the plot of the two largest principal components of the 361 features obtained from the 257 Raman spectra comprising the mean-centered training set. Training set: 1 = Group 1; 2 = Group 2; 3 = Group 3; 4 = Group 4; and 5 = Group 5. Validation set: A = Group 1; B = Group 2; C = Group 5.

Fig 7

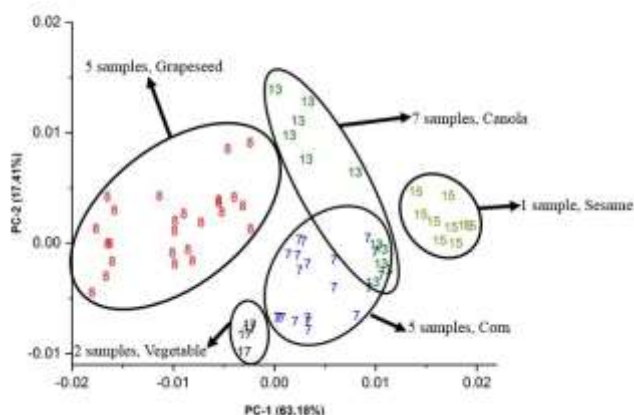


Figure 5. Plot of the two largest principal components of the 71 Raman spectra comprising the mean-centered training set for Group 2 and the 23 features identified by the pattern recognition GA: 7 = corn oil; 8 = grapeseed oil; 13 = canola oil; 15 = sesame oil; 17 = vegetable oil.

Fig 8

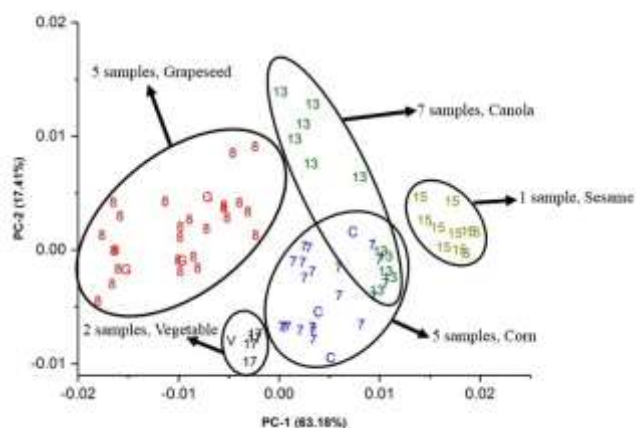


Figure 6. Validation set spectra projected onto the PC plot of the 71 Raman spectra comprising the training set for Group 2 and the 23 spectral features identified by the pattern recognition GA. Training set: 7 = Corn oil; 8 = grapeseed oil; 13 = canola oil; 15 = sesame oil; and 17 = vegetable oil. Validation set: C = corn oil; G = grapeseed oil; S = sesame oil; and V = vegetable oil.

Fig 9

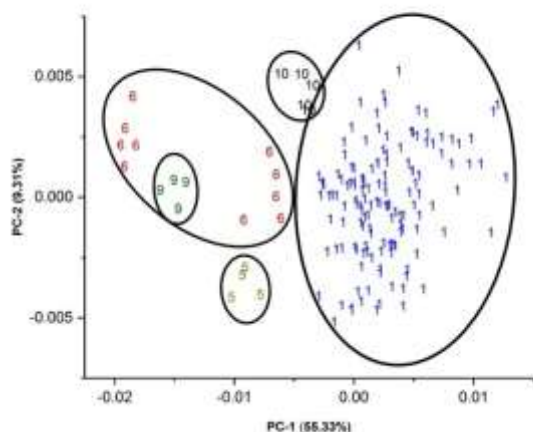


Figure 7. Plot of the two largest principal components of the 174 Raman spectra comprising the training set for Group 1 and the 20 spectral features identified by GA pattern recognition. Jackknifed estimates for the uncertainty of each training set sample are depicted in gray. 1 = extra virgin olive oil, extra light olive oil, and pure olive oil; 5 = avocado oil; 6 = peanut oil; 9 = safflower oil; 10 = hazelnut oil.

Fig 10

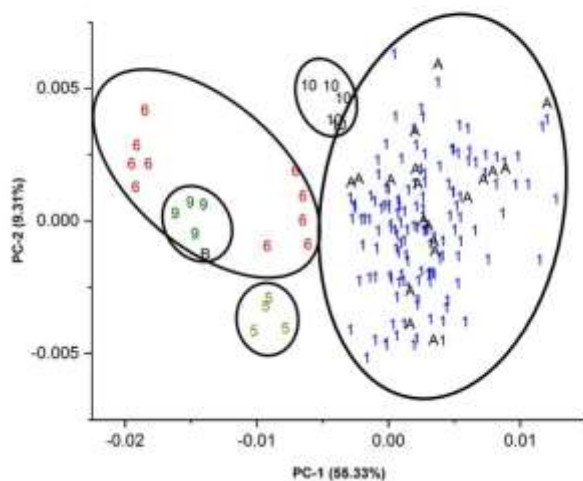


Figure 8. Validation set spectra projected onto the PC plot of the 174 Raman spectra comprising the training set for Group 1 and the 20 spectral features identified by the pattern recognition GA. Jackknifed estimates for the uncertainty of each training set sample are depicted in gray.

Training set: 1 = extra virgin olive oil, extra light olive oil, and pure olive oil; 5 = avocado oil; 6 = peanut oil; 9 = safflower oil; 10 = hazelnut oil. Validation set: O = olive oils; S = safflower oil.

Supplemental Material

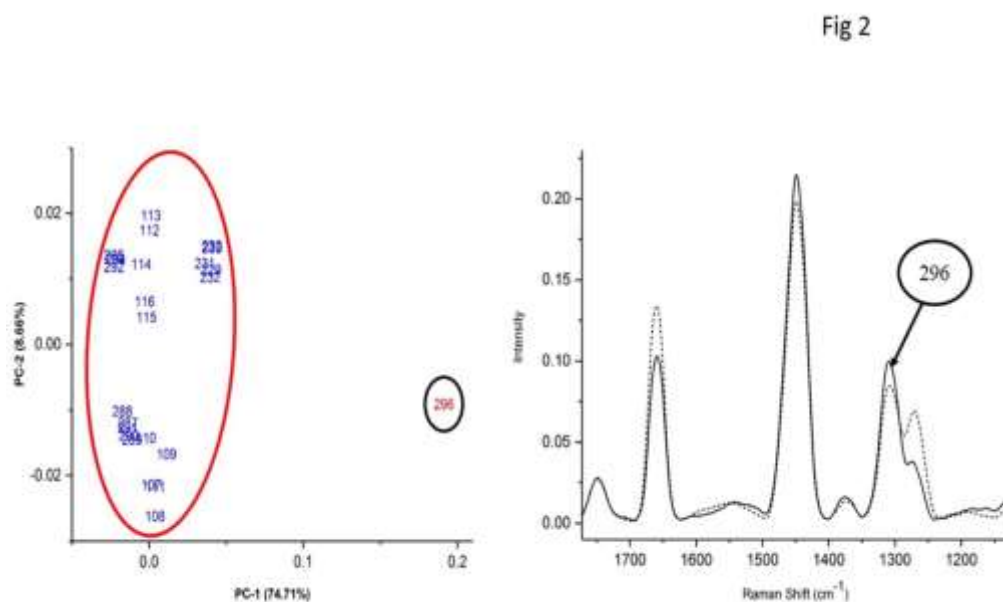


Figure S1. (a) Plot of the two largest principal components of the Raman spectra of corn oil. One corn oil spectrum (Spectrum ID #296) appears as an outlier in the PC plot. (b) Average Raman spectrum of corn oil (dashed line) and the Raman spectrum of the suspected outlier (solid line).

Fig 3

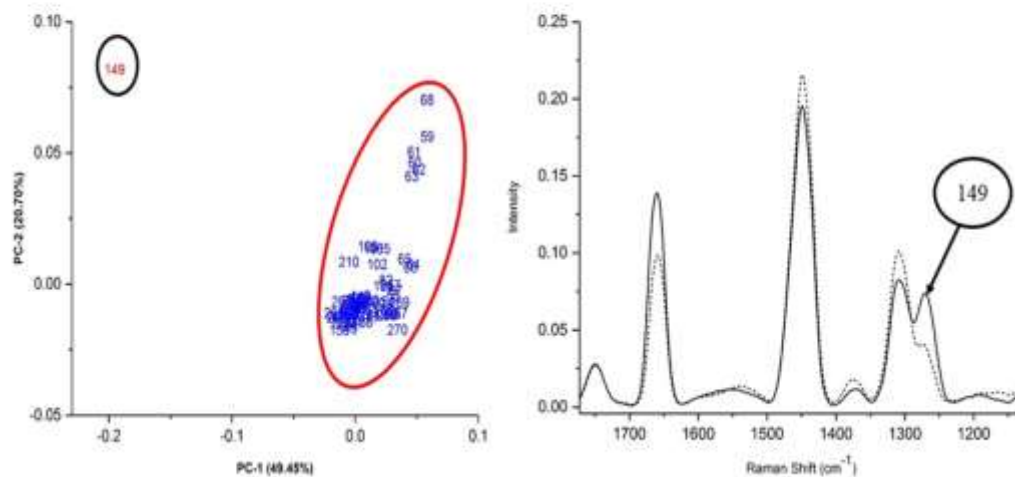


Figure S2. (a) Plot of the two largest principal components of the Raman spectra of extra virgin olive oil. One extra virgin olive oil spectrum (Spectrum ID #149) appears as an outlier in the PC plot. (b) Average Raman spectrum of extra virgin olive oil (dashed line) and the Raman spectrum of the suspected outlier (solid line).