ONE INPUT-CLASS AND TWO INPUT-CLASS CLASSIFICATIONS FOR DIFFERENTIATING OLIVE OIL FROM OTHER EDIBLE VEGETABLE OILS BY USE OF THE NORMAL-PHASE LIQUID CHROMATOGRAPHY FINGERPRINT OF THE METHYL-TRANSESTERIFIED FRACTION

Ana M. JIMÉNEZ-CARVELO [a,✉], Estefanía PÉREZ-CASTAÑO[a], Antonio GONZÁLEZ-CASADO [a], Luis CUADROS-RODRÍGUEZ [a]

[a] Department of Analytical Chemistry, University of Granada, c/ Fuentenueva, s.n. E-18071 Granada, Spain.

Abstract

A new method for differentiation of olive oil (independently of the quality category) from other vegetable oils (canola, safflower, corn, peanut, seeds, grapeseed, palm, linseed, sesame and soybean) has been developed. The analytical procedure for chromatographic fingerprinting of the methyl-transesterified fraction of each vegetable oil, using normal-phase liquid chromatography, is described and the chemometric strategies applied and discussed. Some chemometric methods, such as k-nearest neighbours (kNN), partial least squared-discriminant analysis (PLS-DA), support vector machine classification analysis (SVM-C), and soft independent modelling of class analogies (SIMCA), were applied to build classification models. Performance of the classification was evaluated and ranked using several classification quality metrics. The discriminant analysis, based on the use of one input-class, (plus a dummy class) was applied for the first time in this study.

Keywords

Olive oil authentication, methyl-transesterified fraction, chromatographic fingerprinting, one input-class and two input-class classification

✉ Corresponding author: phone: +34 958240797; fax: +34 958243328; email: amariajc@ugr.es

1. INTRODUCTION

Edible vegetable oils are important worldwide products, which are used as raw materials and/or ingredients in several foodstuffs. Although most vegetable oils are extracted from oilseeds, some are obtained directly from the fruit as a juice. This is the case for virgin olive oil, which is collected directly from olive fruits by mechanical procedures (grinding followed by centrifugation and/or decantation). Furthermore, in contrast to other vegetable oils, virgin olive oil is not refined for human consumption. Extra virgin olive oil is more expensive than other vegetable oils, owing to the specific process required for extraction [Jabeur, Zribi, Makni, Rebai, Abdelheidi, & Bouaziz, 2014]. For this reason, olive oils are susceptible to adulteration, with cheaper vegetable oils, to achieve an illicit profit. Unauthorized blends or adulteration of olive oil of any quality category with oils obtained from seeds is a particular problem in Spain, as well as other Mediterranean countries, which have specific legislation prohibiting the marketing of such blends. Therefore, it is desirable to develop rapid and simple methods to monitor the authenticity of olive oil.

The analytical methodologies applied to authenticate the olive oil are, generally, based on the quantification of certain chemical markers, which constitute a characteristic fraction of the oils [Arvanitoyannis & Vlachos, 2007; Aparicio, Morales, Aparicio-Ruiz, Tena & García-González, 2013]. Thus, families of compound, such as fatty acids, triacylglycerols (or triglycerides) or sterols, have been proposed. Other chemical fractions, such as volatile compounds or phenols, have also been used but they are not stable enough to give reliable results.

Triglycerides represent 95-99% of the chemical composition of vegetable oils. The compositional characterization of these compounds, determined by gas chromatography (GC) or high-performance liquid chromatography (HPLC), has been proposed for the detection of other oils due to their specific compositional profiles [Aparicio & Aparicio-Ruiz, 2000; Ruiz-Samblás, Marini, Cuadros-Rodríguez, & González-Casado, 2012; Lerma-Garcia, Simó-Alfonso, Méndez, Lliberia & Herrero-Martínez, 2011]. The content of some fatty acids, such as linolenic and oleic acids, has also been used to detect blending in olive oils [Aparicio & Aparicio-Ruiz, 2000].

Fatty acids are quantified using GC, following derivatization to increase the volatility of the compounds, as necessary [Sanchéz de Medina, El Riachy, Priego-Capote & Luque de Castro, 2014; Fernandes, Fernandes, Simas, Barrera-Arellano, Eberlin, & Alberici, 2013]. Moreover, sterols are applied as markers of authenticity in vegetable oils. In order to characterize the compositional profile of these compounds, firstly, it is necessary to carry out saponification of the oil followed by isolation of free sterols by means preparative chromatography or solid phase extraction, and silanization. Then, GC analysis is performed

[Gázquez-Evangelista, Pérez-Castaño, Sánchez-Viñas & Bagur-González, 2013]. Consequently, this methodology is difficult, tedious and time-consuming.

In 1993, Bierdermann et al. [Biedermann, Grob & Mariani, 1993] developed a new strategy that replaced the conventional saponification/ isolation process with a methyl-transesterification reaction. This approach, which inexplicably has been underused by the analytical community, requires less vegetable oil and facilitates extraction since soaps are not produced and the process is faster. The breakdown of molecules during transesterification leads to the formation of methyl esters from fatty acids and the liberation of sterols. Two fractions are obtained during this process: (1) the water soluble fraction, which contains the polar compounds, and (2) the organic fraction (transesterified fraction) in which fatty acid methyl esters, sterols, alcohol, monoglycerides, diglycerides and other molecules can be found. In the latter fraction, Bierdermann et al. [1993] identified methyl sterols, dimethyl sterols and linear alcohols. The methyl-transesterified fraction can be analysed by liquid chromatography to obtain a characteristic fingerprint of each vegetable oil, which might also be used to detect potential adulteration. The fingerprinting methodology is based on treating the entire or a part of the chromatogram as a whole, without identifying or quantifying each compound [Ellis et al., 2012; Cuadros-Rodríguez, Ruiz-Samblás, Valverde-Som, Pérez-Castaño, & González-Casado, 2016]. Effective implementation of fingerprinting requires the use of chemometric tools. Chromatograms are exported as data vectors and treated with pattern recognition methods to develop multivariate classification or regression models, which are suitable to differentiate among vegetable oils.

The chemometric methods fall in to two groups: supervised and non-supervised [Naes, Isaksson, Fearn & Davies, 2002; Marini, 2010]. In the first group, the category or class membership of each data vector is known and used to build the multivariate model. In contrast, the model from non-supervised methods does not consider this information [Correira & Ferreira, 2007]. Supervised classification methods are used to categorize objects (samples) in two or more classes according to a set of characteristic features of each class. Such features are extracted previously from information supplied for standard objects and selected during the model-training step. In order to perform the classification process, two approaches could be applied: discriminant analysis methods and class-modelling methods [Bevilacqua, Nescatelli, Bucci, Magrì, Magrì & Marini, 2014]. A discriminant method works by finding the borders between groups of objects from different classes, while a class-modelling method defines a particular enclosed space for all the objects from the same class.

Sometimes class-modelling methods are described as 'one-class classifier' (e.g. SIMCA) where each class is modelled independently [Brereton, 2011] and as many model as classes are built. Classification is performed considering all the models simultaneously. In our opinion, however, this term should not be used as synonym for class-modelling since the two might be confused.

This study proposes a multivariate method to differentiate olive oil from other edible vegetable oils. For this, the methyl-transesterified fraction from each oil class (olive and non-olive) was analysed using normal-phase conventional high-performance liquid chromatography. The chromatograms (chromatographic fingerprints), acquired by means of a corona charged aerosol detector (CAD), were used as a source of analytical information to set up the classification models. Some common and well-established classification methods were applied, such as k-nearest neighbours (kNN), partial least squares discriminant analysis (PLS-DA), support vector machine classification (SVM-C) and soft independent modelling of class analogies (SIMCA). Two classification strategies were tried for each classification method according to the number of class used for model training: two input-class and one input-class classifications. In addition, the use of a 'dummy' class was proposed for applying discrimination methods with a one input-class strategy. The classification results from each method and strategy were compared and ranked on the basis of several classification performance metrics [Cuadros-Rodríguez, Pérez-Castaño & Ruiz-Samblás, 2016].

2. MATERIALS AND METHODS

2.1. Chemicals

All solvents used were HPLC grade. Isopropanol, n-hexane, methanol and tert-butyl methyl ether (TBME) were provided by the VWR International Eurolab, S.L. (Barcelona, Spain). Sodium methoxide (MeONa), citric acid monohydrate, and anhydride sodium sulphate were purchased from Merck (Darmstadt, Germany). The nitrogen (99.9999 %) used was provided by Air Liquid (Madrid, Spain).

2.2. Chromatography

The analyses were carried out with an Agilent 1100 series liquid chromatograph (Santa Clara, USA) equipped with a column thermostat (Eppendorf CH30), a quaternary pump and degasser auto sampler. Detection was performed with a corona charged aerosol detector (CAD) (ESA Biosciences Inc., Chemlsford, MA, USA). Agilent ChemStation software (rev. B.02.01-SR1) for LC systems was used to collect and process data.

The HPLC analysis was carried out on a (250 × 4 mm i.d, 5 µm) column Lichrospher® 100 CN. The column temperature was set at 30 ºC during the entire operation. The composition of the mobile phase was n-hexane/isopropanol (96:4, v/v) at a flow rate of 1.2 mL min^{-1}. The injection volume was 20 µL and the run time was only 8 min.

2.3. Samples

A total of 127 vegetable oil samples of different types were analysed. The samples were obtained directly from local providers. More specifically, 66 samples were different categories of marketed olive oil (virgin extra, virgin, refined+virgin, and pomace+virgin), and the other 61 were canola, safflower, corn, peanut, sunflower, (no-specified) seed, grapeseed, palm, linseed, sesame, and soybean oils. Table 1 summarizes the different vegetable oils and the number of samples analysed for each.

TABLE 1

2.4. Sample preparation

Previous to the chromatographic analysis, a transesterification reaction was applied. A modification of the procedure described by Biedermann et al [Biedermann, Grob & Mariani, 1993] was used. For this, 0.1 g of oil was weighed into a centrifuge tube. 1 mL of extracting agent (MeONa at 10 % in methanol in TBME, 4:6 (v/v)) was added and mixed with the oil. The mixture was stirred for 20 s and then allowed to stand for 20 min. This step was repeated twice. Then, 1 mL of water and 8 mL of hexane was added, and the mixture centrifuged for 3 min at 1,500 g. The aqueous phase was removed with a Pasteur pipette and 1 mL of 1 % citric acid in water added to the residual. Again, the aqueous phase was eliminated before 2 g of anhydrous sodium sulphate added and the mixture allowed to stand for 20 min. The methyl-transesterified organic fraction was passed through a polytetrafluoroethylene (PTFE) membrane syringe filter (0.22 µm) and the solution stored at –20ºC until analysis. For the chromatographic analysis, 200 µL of transesterified solution was added to a 2 mL HPLC vial before 450 µL of n-hexane was added and 20 µL injected.

2.5. Chemometrics

The raw data files from each chromatogram were obtained in a CSV file and exported to MATLAB (version R2013a). In this way, a data vector composed of 839 variables defined each chromatogram. The data pre-processing was done with a home-programmed MATLAB function, "Medina" (version 10) [Pérez Castaño et al., 2015]. This function implemented several algorithms from the MATLAB Bioinformatics Toolbox^{TM} and 'icoshift' (*interval correlation optimized shifting*) algorithm [Tomasi, Savorani & Engelsen, 2011] to align the peaks of the chromatograms. The steps for pre-processing the data were: (1) raw chromatograms data grouping and overlay; (2) selection of interval of interest in chromatograms; (3) filtered of the raw chromatograms data to eliminate noise of signal analytical; (4) correction of the baseline using the 'msbackadj' function (included in the Bioinformatics Toolbox^{TM}); (5) alignment of the peaks with the function 'icoshift'; and finally (6) mean centring of the data set.

The original dataset was divided in two groups: (1) the training set, which was made up of 84 oil samples (44 olive oil, 40 non-olive oil), and (2) the validation (or test) set composed of the remaining oil samples (25 olive oil, 18 non-olive oil). Selection was carried out ensuring that a sample from each class of oil was allocated to one vegetable oil group or the other. Within each group, the samples were selected randomly.

Classification of the vegetable oils was achieved using multivariate chemometric pattern recognition in the PLS_Toolbox (version 7.5.2, Eigenvector Research, Wenatchee, WA).

Principal Component Analysis (PCA)

The main aim of PCA is to reduce the number of variables to evaluate which components contain essential information. Each principal component (PC) is a lineal combination between original variables (chromatographic intensities) of each object, which are described as: $X=T \times P^T$ where X is the original data matrix, T is the score matrix and P is the transposed loading matrix [Bro & Smilde, 2014].

k-Nearest Neighbours (kNN)

kNN is a based-similarity classification method that uses distance measures between objects. The classification is carried out as follows: first, a multidimensional hyperspace is defined with the training set and, then, the prediction is performed. The assigned class of each new object will be one where the number of k-neighbours is largest [Correira & Ferreira, 2007; Alsberg, Goodacre, Rowland & Kell, 1997] and k is an odd integer that could be selected previously. Each sample is classified based on the most represented classes of the k-nearest samples.

Partial Least Squares Regression-Discriminant Analysis (PLS-DA)

PLS-DA is a latent variable-based method that builds a PLS regression model on latent variables (LV) to establish limits of the class and, then, carries out a discriminant analysis (DA) to classify the samples [Bevilacqua, Nescatelli, Bucci, Magrì, Magrì & Marini, 2014; Ballabio & Consonni, 2013]. In order to develop the best PLS model, it is necessary to optimize the number of LVs to be used in advance.

Support Vector Machine Classification (SVM-C)

SVM is a based-machine learning method. As with PLS-DA, SVM-C works by carrying out a SVM regression model for building hyperplanes in a multidimensional space that separates the different classes of objects [Xu, Zomer, Brereton, 2006; Luts, Ojeda, Van de Plas, De Moor, Huffel & Suykends, 2010]. SVM can be optimized with 'nu' and 'C' parameters. The former optimizes a model with an adjustable parameter Nu [0 → 1], which indicates the upper boundary for the number of misclassifications allowed, and the latter optimizes a model with an adjustable cost function C [0 → ∞], which indicates how strongly misclassifications should be penalized [SVM Function Settings, Eigenvector Documentation wiki. URL http://wiki.eigenvector.com/index.php?title=SVM_Function_Settings. Accessed 29.06.15].

Soft Independent Modelling of Class Analogies (SIMCA)

This chemometric technique performs as many principal component (PC) models as input-classes in study and, then, the classification is carried out from the distance of the object to the centre of each principal component score space [Bevilacqua, Nescatelli, Bucci, Magrì, Magrì & Marini, 2014]. The assignment of each unknown sample to a particular class is based on the nearest distance to the corresponding regions established by the PC model.

Two input-class (2iC) and one input-class (1iC) classification

Usually a two-class classification method (or more properly, two output-class classification) requires using two input-classes, the target class and the non-target class (in this paper, olive and non-olive classes). The term 'output' is related to the classes to which objects or samples will be assigned as result of the classification while the term 'input' refers to the class that is used to train the classification model [Cuadros-Rodríguez, Pérez-Castaño, & Ruiz-Samblás, 2016]. It is also possible to perform the same classification method by training the model with a single input-class, *i.e.* the target class.

Working with one input-class classification has significant advantages. For example, in food authentication, the model can be built with data from only genuine foods (target class) and it is not necessary to have other foods (non-target class) to train the model. Consequently, the necessary experimental work is halved. When this model is applied on unknown foods, only those recognized by the model will be declared as "true" whereas the remaining food will be refused and they are candidate to be considered as "false". The greater the training set of genuine representative samples, the better the quality classification performance. Obviously, this strategy can be applied to differentiate olive oils from other edible vegetable oils.

This is a very easy task when a class-modelling method is applied because each class is modelled independently. This approach has been used already with SIMCA [López, Trullos, Callao & Ruisanchez, 2014]. However, the discriminant methods, such as PLS-DA or SVM-C, usually require two input-classes to define the discrimination model. Although some proposals have been reported as one-class PLS (OCPLS) [Xu, Yan, Cai & Yu, 2013], in fact, this is a class-modelling method. To resolve this drawback, a fictitious class or 'dummy' class could be used as a substitute for the second class (the non-target class). The dummy class should be defined from inactive objects that do not have analytical information of interest for the target class, e.g. analytical blank.

In this study, both 2iC and 1iC strategies were applied to devise a classification model for differentiating olive oil from non-olive oil. When the 1iC was applied, a dummy class was from the dataset provided using 30 chromatograms for the solvent blank.

3. RESULTS AND DISCUSSION

A chromatogram was recorded for each vegetable oil sample. Figure 1 shows the superposed chromatograms for all vegetable oil samples. Two regions could be easily differentiated: (1) region A shows a major peak, which was essentially composed of methyl esters of fatty chains derived from triglycerides, phospholipids, waxes, esterified sterols and free fatty acids, and (2) region B that was composed of several minor peaks and contained information about the families of free sterols and terpenic alcohols.

FIGURE 1

Exploratory Analysis

A principal component analysis (PCA) was carried out considering the dataset composed of the whole chromatogram from each vegetable oil sample. Four PCs were enough to explain

87.16% of the variance. Figure 2a shows the biplot for scores on the PC2-PC1 plane. PC1 and PC2 explained 56.2% and 17.3% of the variance, respectively. Three groups of vegetable oils could be distinguished easily, which corresponded with olive oil (centre left), palm oil (top left) and other vegetable oils (right).

Two additional PCA were carried out, one for each of the regions of the chromatograms to check if both regions grouped the oil samples in the same way. Figure 2b and 2c show the biplot for scores on the PC2-PC1 plane, corresponding to the data subset from regions A and B, respectively.

FIGURE 2

The three scores biplots allowed differentiation in similar ways to the three sample groups and, in principle, there was no conclusive reason –from a chemometric point of view– to select one dataset or the others. However, looking the chromatographic retention time, region A was preferred to minimize the analysis time.

Two input-class (2iC) classification

In order to differentiate olive oils from other vegetable oils, a two input-class (2iC) classification strategy was applied where the target class was 'olive oil' and the alternative class was, generally, denoted as 'non-olive oil'. Four well-established classification methods were tried: kNN, PLS-DA, SVM-C and SIMCA.

To differentiate the two vegetable oils classes, k=3 was enough to decide the neighbour distance in the kNN model. The olive class was defined by a class predicted probability value equal to 1, while the non-olive class was defined by a probability of 0. Classification of the samples contained in the validation set was carried out directly by the software. All of olive oil samples were well classified (probability=1) and the non-olive oil samples were also classified correctly (probability=0), with exception of palm oil samples, which had an assigned probability of 0.5; in this case, we also classified these samples as non-olive oil.

The PLS-DA model was built using four LVs, with 92.91% of the variance explained. Each class was characterized by a predicted value around 1 for olive oil and 0 for non-olive oil. The classification threshold established by the software from the corresponding probability curves was a predicted value of 0.6 for the olive oil class.

The SVM-C model was optimized with 'C-svc' and 'nu-svc' parameters, and the results obtained in both cases were similar. As in the kNN method, the olive class was assigned to samples with a predicted probability value equal to 1 and the non-olive class was defined by samples with a probability of 0. The software also carried out the class assignment for the validation samples. Both olive and non-olive oil samples were classified correctly.

Figure 3 (a) and (b) show the classification plots obtained from both 2iC PLS-DA and 2iC SVM-C methods.

FIGURE 3 (a) (b)

The application of SIMCA implies building of two PC models. The number of PCs chosen for each model was four for 'olive oil' and five for 'non-olive oil'. The software carried out classification of the validation dataset based on the Q-residual values for each olive oil sample. Samples with a normalized Q-residual (95% confidence) value less than $\sqrt{2}$ were classified as olive oil.

Table 2 shows the different quality performance features for the 2iC classification method, calculated according to the olive oil samples classification. These show that, in this classification scenario, 2iC kNN and SVM-C were faultless and, in contrast, 2iC SIMCA performed poorly.

TABLE 2

One input-class (1iC) classification

Since the aim of this study was differentiation of olive oil from other vegetable oils, the classification model could be trained using objects from the olive oil class. In this way, the objects recognized by the model should be assigned as olive oil whereas the remainder, regardless of their botanical origin, should be classified as non-olive oils. The same classification methods, kNN, PLS-DA, SVM-C and SIMCA, were applied. For each, a confidence interval-based classification criterion was established because the default classification threshold defined by the software was not applicable.

The kNN model conformed with k=3, but did not generate good results and all the non-olive oil samples were misclassified because they were considered to be "nearest neighbours" to the target class (olive oil). Thus, the 1iC strategy was not applicable for the kNN method.

Two strategies were applied for 1iC PLS. In a first step, a PLS-DA classification with dummy class was performed using the PLS_Toolbox. Next, a one-class PLS without dummy class (OCPLS) was performed using software provided by Xu [Xu, Yan, Cai, & Yu, 2013].

A conventional PLS-DA was built with only two LVs explaining 99.74% of the variance. A confidence interval was established centred on 1, which was the value assigned for the olive oil class. The width of the interval was calculated as plus/minus 2.33-times the standard deviation (s) from the predicted values for the olive oil samples in the training set. The expression 2.33×s is an "ad-hoc" application, recommended by the EC for estimating the decision limit (DL), formally termed as CCα, concerning the performance of analytical methods in the case of substances for which no permitted limit has been established [EU Commission Decision, 2002]. This decision limit defines the limit at and above which it can be concluded with an error probability of α that a sample is non-compliant. Strictly speaking, the correct expression would be: DL = $1.645\times\sqrt{2}$ s, where 1.645 is the critical value for the standardized normal distribution (α = 1%) and s the whiting-batch standard distribution of the difference between the predicted values of both the target and the non-target samples, which are considered equal, and consequently: $s = \sqrt{s^2(\mathrm{targ}) + s^2(\mathrm{non{-}targ})} = \sqrt{2}$ s(targ). The coefficient 2.33 is the result of multiplying 1.645 $\times\sqrt{2}$ (or 1.414). The confidence band is calculated from an estimated standard deviation of 0.026.

Figures 4(a) and 4(b) show the classification plots obtained from the 1iC PLS-DA method.

FIGURE 4 (a) (b) (c) (d)

Most olive oil samples were included within the confidence interval while the no-olive oils were not. However, samples in the non-olive oil class were separate into two subclasses on both sides of the interval. The seed oils were located in the upper region whereas the palm oils were in the lower region. This surprising outcome implies the classification scenario is suitable for implementing a three output-class classification (olive oil, palm oil, and generically seed oil) from a one input-class strategy, making it possible to distinguish palm oil from a classification model trained only with olive oils. Currently, the authors are working to develop and apply this approach.

OCPLS was built with seven LVs. For classification purpose, the regions pre-established by the software were used. The results are showed in Table 3.

SVM-C classification was carried out by optimization of 'C-svc' and 'nu-svc' parameters, and the results obtained in both cases were similar. All the oil samples were assigned to a predicted probability close to 1 and always distant from 0, which was assigned to the dummy class. Specifically, the probability value was ca. 0.98 for the olive oil class and less but always greater than 0.92 for the non-olive class. The confidence interval was determined by means of a probability interval centred on the average olive oil class probability calculated from the training set. The width of the interval was also calculated as plus/minus 2.33 times the standard deviation from the predicted class probability. The estimated value of the probability standard deviation was 0.0015. Figures 4 (c) and (d) show the classification plots obtained from the 1iC SVM-C method.

Finally, the SIMCA method was also applied. Since SIMCA is a class-modelling method, two options were applied: i) a double PCA model using both the olive oil and dummy classes; and ii) a single model from the olive oil class. In both cases, five PCs were used to build the olive oil model. In both cases, a sample oil was classified as olive oil when the normalized Q-residual (95% confidence) value was less than $\sqrt{2}$.

Table 3 shows the quality performance features of the different 1iC classification methods. In contrast with the 2iC classification method, the 1iC PLS-DA provided the best classification performance and 1iC SIMCA (without dummy class) was, again, the worst.

TABLE 3

4. CONCLUSIONS

In this study, several classification methods were applied and the application strategy has been discussed. Four well-established classification methods were used, namely kNN, PLS-DA, SVM-C and SIMCA. Each was applied using two classification strategies designated as two input-class (2iC) and one input-class (1iC) classifications. This is the first time a dummy class has been used to perform discriminant analysis methods with a single input-class. This new approach does not require having and analysing samples from the non-target class (non-olive vegetable oil) in order to train the classification model. In order to assess and rank the different classification methods and strategies, several quality classification metrics were calculated. kNN and SVM-C, on the one hand, and PLS-DA, on the other, proved to be the best when 2iC or 1iC classification strategies were applied, respectively. Furthermore, the proposed analytical method consumed less time in sample treatment (transesterification reaction, 60 min) and chromatographic elution (8 min) than previous methods (saponification, 120 min) and chromatographic analysis (40 min).

REFERENCES

Alsberg, B.K., Goodacre, R., Rowland, J.J., & Kell, D.B. (1997). Classification of pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, K-nearest neighbour, neural and decision-tree methods. *Analytica Chimica Acta, 348,* 389–407.

Aparicio, R., & Aparicio Ruíz, R. (2000). Authentication of vegetable oils by chromatographic techniques. *Journal of Chromatography A*, *881*, 93–104.

Aparicio, R., Morales, M.T., Aparicio-Ruiz, R., Tena N., & García-González D.L. (2013). Authenticity of olive oil: Mapping and comparing official methods and promising alternatives. *Food Research International, 54*, 2025–2038.

Arvanitoyannis, I.S., & Vlachos, A. (2007). Implementation of physicochemical and sensory analysis in conjunction with multivariate analysis towards assessing olive oil authentication/adulteration. *Critical Reviews in Food Science and Nutrition, 47*,441–498.

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods, 5,* 3790–3798.

Bevilacqua, M., Nescatelli, R., Bucci, R., Magrì, A.D., Magrì, A.L., & Marini, F. (2014). Chemometric Classification Techniques as a Tool for Solving Problems in Analytical Chemistry. *Journal of AOAC International, 97,* 19–28.

Biedermann, M., Grob, K., & Mariani, C. (1993). Transesterification and on-line LC-GC for determining the sum of free and esterified sterols in edible oils and fats. *Fett Wissenschaft Technologie - Fat Science Technology*, *95*(4), 127–133.

Brereton, R.G. (2011). One-class classifiers. *Journal of Chemometrics, 25,* 225–246.

Bro, R., & Smilde, A. (2014). Principal component analysis. *Analytical Methods, 6,* 2812–2831.

Correira, M.M., & Ferreira, P.R.M. (2007). Non-supervised pattern recognition methods: Exploring chemometrical procedures for evaluating analytical data. *Quimica Nova, 30,* 481–487.

Cuadros Rodríguez, L., Pérez Castaño, E., & Ruiz Samblás, C. (2016). Quality performance metrics in multivariate classification methods for qualitative analysis. *Trends in Analytical Chemistry, 80*, 612–624.

Cuadros Rodríguez, L., Ruiz Samblás, C., Valverde Som, L., Pérez Castaño, E., & González Casado, A. (2016). Chromatographic fingerprinting: An innovative approach for food 'identitation' and food authentication - A tutorial. *Analytica Chimica Acta, 909*, 9–23.

Ellis, D.I., Brewster, V.L., Dunn, W.B., Allwood, J.W., Golovanov, A.P., & Goodacre R. (2012). Fingerprinting food: current technologies for the detection of food adulteration and contamination. *Chemical Society Reviews*, *41*, 5706-5727.

EU Commission Decission (2002/657/EC) implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results, Official Journal of the European Communities, L 221/8-36.

Fernandes, A.M.A.P., Fernandes, G.D., Simas, R.C., Barrera-Arellano, D., Eberlin, M.N., & Alberici, R.M. (2013). Quantitation of triacylglycerols in vegetable oils and fats by easy ambient sonic-spray ionization mass spectrometry. *Analytical Methods, 5,* 6969-6975.

Gázquez Evangelista, D., Pérez Castaño, E., Sánchez Viñas, M., & Bagur González, M.G. (2013). Using offline HPLC-GC-FID 4-Desmethylsterols Concentration Profiles, Combined with Chemometric Tools, to Discriminate Different Vegetable Oils. *Food Analytical Method*, *7*, 912–925.

Jabeur, H., Zribi, A., Makni, J., Rebai, A., Abdelheidi, R., & Bouaziz, M. (2014). Detection of chemlali extra-virgin olive oil adulteration mixed with soybean oil, corn oil, and sunflower oil by using GC and HPLC. *Journal of Agricultural and Food Chemistry, 62,* 4893–4904.

Lerma García, M.J., Simó Alfonso, E.F., Méndez, A., Lliberia, J.L., & Herrero Martínez, J.M. (2011). Classification of extra virgin olive oils according to their genetic variety using linear discriminant analysis of sterol profiles established by ultra-performance liquid chromatography with mass spectrometry detection. *Food Research International, 44,* 103–108.

López, M.I., Trullos, E., Callao, M.P., & Ruisánchez, I. (2014). Multivariate screening in food adulteration: untargeted versus targeted modelling. *Food Chemistry, 147,* 177-181.

Luts, J., Ojeda, F., Van de Plas, R., De Moor, B., Van Huffel, S., & Suykens, J.A.K.(2010). A tutorial on support vector machine-based methods for classification problems in chemometrics. *Analytica Chimica Acta, 665,*129–145.

Marini, F. (2010). Classification Methods in Chemometrics. *Current Analytical Chemistry*, *6*, 72–79.

Naes, T., Isaksson, T., Fearn T., & Davies, T. (2002). A user-friendly guide to multivarate calibration and classification, NIR Publicationes, Chichester.

Pérez Castaño, E., Ruiz Samblás, C., Medina Rodríguez, S., Quirós Rodríguez, V., Jiménez Carvelo, A.M., Valverde Som, L., González Casado, A., & Cuadros Rodríguez, L. (2015) Comparison of different analytical classification scenarios: application for the geographical origin of edible palm oil by sterolic (NP) HPLC fingerprinting. *Analytical Methods, 7,* 4192–4201.

Ruiz Samblás, C., Marini, F., Cuadros Rodríguez, L., & González Casado, A. (2012). Quantification of blending of olive oils and edible vegetable oils by triacylglycerol fingerprint gas chromatography and chemometric tools. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, 910*, 71–77.

Sánchez de Medina, V., El Riachy, M., Priego-Capote, F., & Luque de Castro, M.D.(2014). Composition of fatty acids in virgin olive oils from cross breeding segregating populations by gas chromatography separation with flame ionization detection. *Journal of the Science of Food and Agriculture, In press*, doi:10.1002/jsfa.7030.

SVM Function Settings, Eigenvector Documentation Wiki [accessed 09 November 2015], http://wiki.eigenvector.com/index.php?title=SVM_Function_Settings

Tomasi, G., Savorani, F., & Engelsen, S.B. (2011). Icoshift: An effective tool for the alignment of chromatographic data. *Journal of Chromatography A, 1218*, 7832–7840.

Xu, L., Yan, S-M., Cai, C-B., Yu, X-P. (2013). One-class partial least squares (OCPLS) classifier. *Chemometrics and Intelligent Laboratory Systems, 126*, 1-5.

Xu, Y., Zomer, S., & Brereton, R.G. (2006). Support Vector Machines: A recent method for classification in chemometrics. *Critical Reviews in Analytical Chemistry, 36,* 177–188.

Table 1. Class and types of vegetable oils analysed.

Class	Category/type	Nº samples
Olive oil (66 samples)	Virgin extra	50
	Virgin	4
	"Refined" [a]	6
	"Pomace" [b]	6
Non-olive oil (61 samples)	Canola	4
	Safflower	4
	Corn	5
	Peanut	5
	Sunflower [c]	13
	Seeds	6
	Grapeseed	4
	Palm	7
	Linseed	3
	Sesame	3
	Soybean	7

[a] A marketed blend of refined and virgin olive oil (5-10 %).

[b] A marketed blend of pomace and virgin olive oil (5-10 %).

[c] Two samples of high-oleic sunflower oils are included.

Table 2. Values of the quality performance features of the different 2iC classification methods.

Performance features	**kNN**	**PLS-DA**	**SVM-C**	**SIMCA**
Sensibility (or Recall)	1.00	1.00	1.00	0.48
Specificity	1.00	0.94	1.00	1.00
Positive predictive value (Precision)	1.00	0.96	1.00	1.00
Negative predictive value	1.00	1.00	1.00	0.58
Youden index	1.00	0.94	1.00	0.48
Positive likelihood rate	–	18.00	–	–
Negative likelihood rate	0.00	0.00	0.00	0.52
F-measure	1.00	0.98	1.00	0.65
Discriminant power	–	–	–	–
Efficiency (or Accuracy)	1.00	0.98	1.00	0.70
AUC (Correctly classified rate)	1.00	0.97	1.00	0.74
Matthews correlation coefficient	1.00	0.95	1.00	0.53
Kappa coefficient	1.00	0.95	1.00	0.44

The hyphen "–" is signifying that the performance feature cannot be determined

Table 3. Values of the quality performance features of the different 1iC classification methods.

Performance features	With a dummy class				Without dummy class	
	kNN	**PLS-DA**	**SVM-C**	**SIMCA**	**OCPLS**	**SIMCA**
Sensibility (or Recall)	1.00	0.96	0.88	0.88	0.80	0.80
Specificity	0.00	1.00	1.00	0.83	0.89	1.00
Positive predictive value (Precision)	0.58	1.00	1.00	0.88	0.91	1.00
Negative predictive value	–	0.95	0.86	0.83	0.76	0.78
Youden index	0.00	0.96	0.88	0.71	0.69	0.80
Positive likelihood rate	1.00	–	–	5.28	7.20	–
Negative likelihood rate	–	0.04	0.12	0.14	0.23	0.20
F-measure	0.74	0.98	0.94	0.88	0.85	0.89
Discriminant power	–	–	–	0.86	0.83	–
Efficiency (or Accuracy)	0.58	0.98	0.93	0.86	0.84	0.88
AUC (Correctly classified rate)	0.50	0.98	0.94	0.86	0.84	0.90
Matthews correlation coefficient	–	0.95	0.87	0.71	0.68	0.79
Kappa coefficient	0.00	0.95	0.86	0.71	0.67	0.77

The hyphen "–" is signifying that the performance feature cannot be determined

FIGURE CAPTIONS

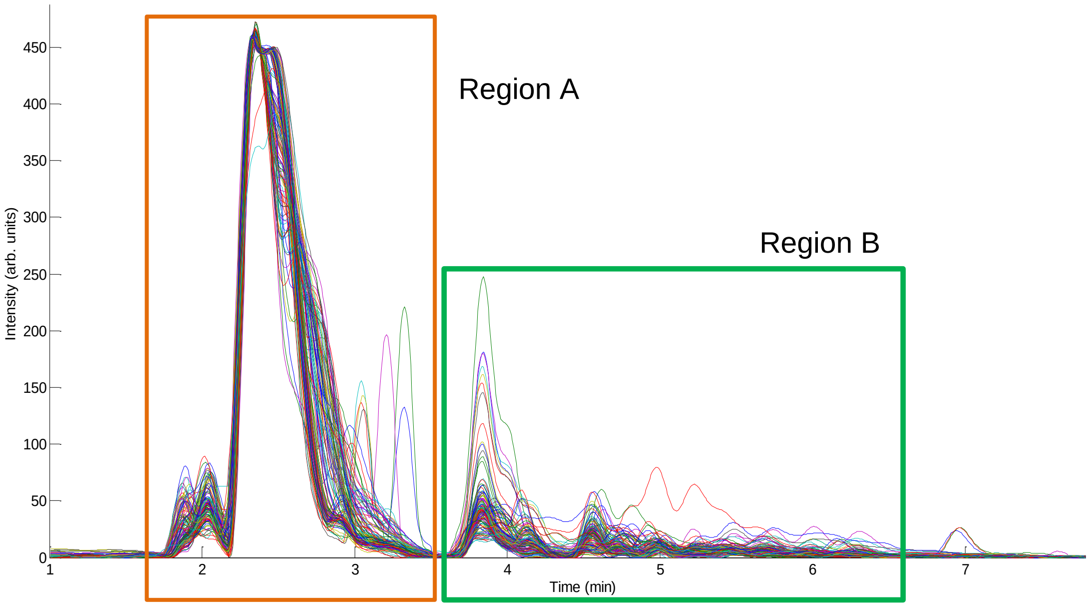
Figure 1. Superposed chromatograms of the 127 vegetable oil samples showing the two characteristic regions (see text for additional explanations). The chromatograms have been previously pre-processed with the exception of the mean centring step.

Figure 2. PCA scores biplot obtained from the fingerprint data of the methyl-transesterified fraction of the 127 vegetable oil samples: **(a)** PC2-PC1 plane of the whole chromatogram; **(b)** PC2-PC1 plane from region A; **(c)** PC2-PC1 plane from region B.

Figure 3. Classification plots on the 2iC classification strategy: **(a)** PLS-DA; **(b)** SVM-C.

Figure 4. Classification plots on the 1iC classification strategy: **(a)** and **(b)** PLS-DA full plot and zoomed plot, respectively; **(c)** and **(d)** SVM-C full plot and zoomed plot, respectively. In addition, the confidence bands are superposed on (**b**) and (**d**) plots.

Figure 1



Region A
Region B
Intensity (arb. units)
450
400
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
Time (min)

Figure 2

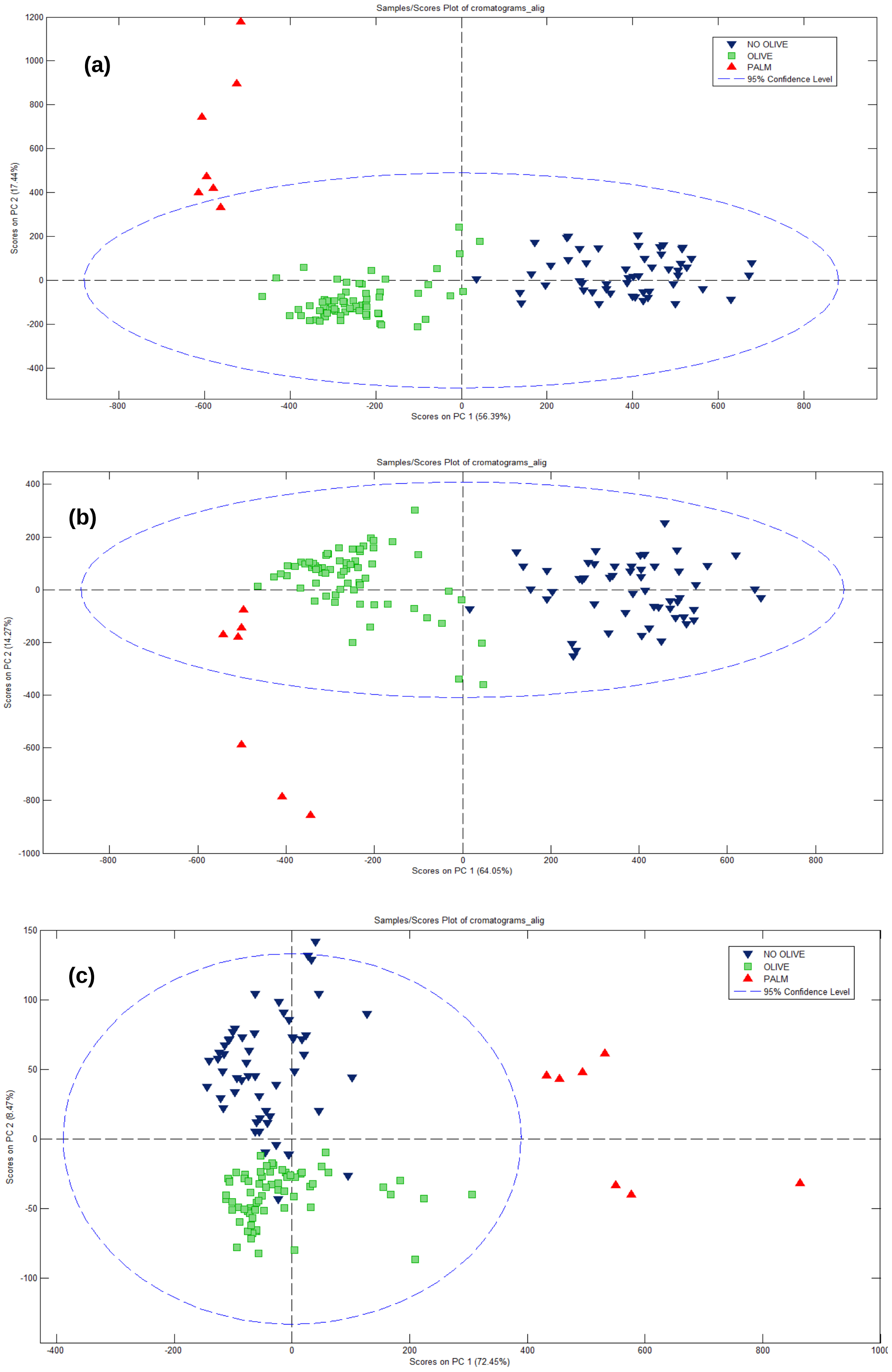

Samples/Scores Plot of cromatograms_alig
(a)
NO OLIVE
OLIVE
PALM
95% Confidence Level
Scores on PC 2 (17.44%)
Scores on PC 1 (56.39%)
1200
1000
800
600
400
200
0
-200
-400
-800
-600
-400
-200
0
200
400
600
800
Samples/Scores Plot of cromatograms_alig
(b)
Scores on PC 2 (14.27%)
Scores on PC 1 (64.05%)
400
200
0
-200
-400
-600
-800
-1000
-800
-600
-400
-200
0
200
400
600
800
Samples/Scores Plot of cromatograms_alig
(c)
NO OLIVE
OLIVE
PALM
95% Confidence Level
Scores on PC 2 (8.47%)
Scores on PC 1 (72.45%)
150
100
50
0
-50
-100
-400
-200
0
200
400
600
800
100(

Figure 3

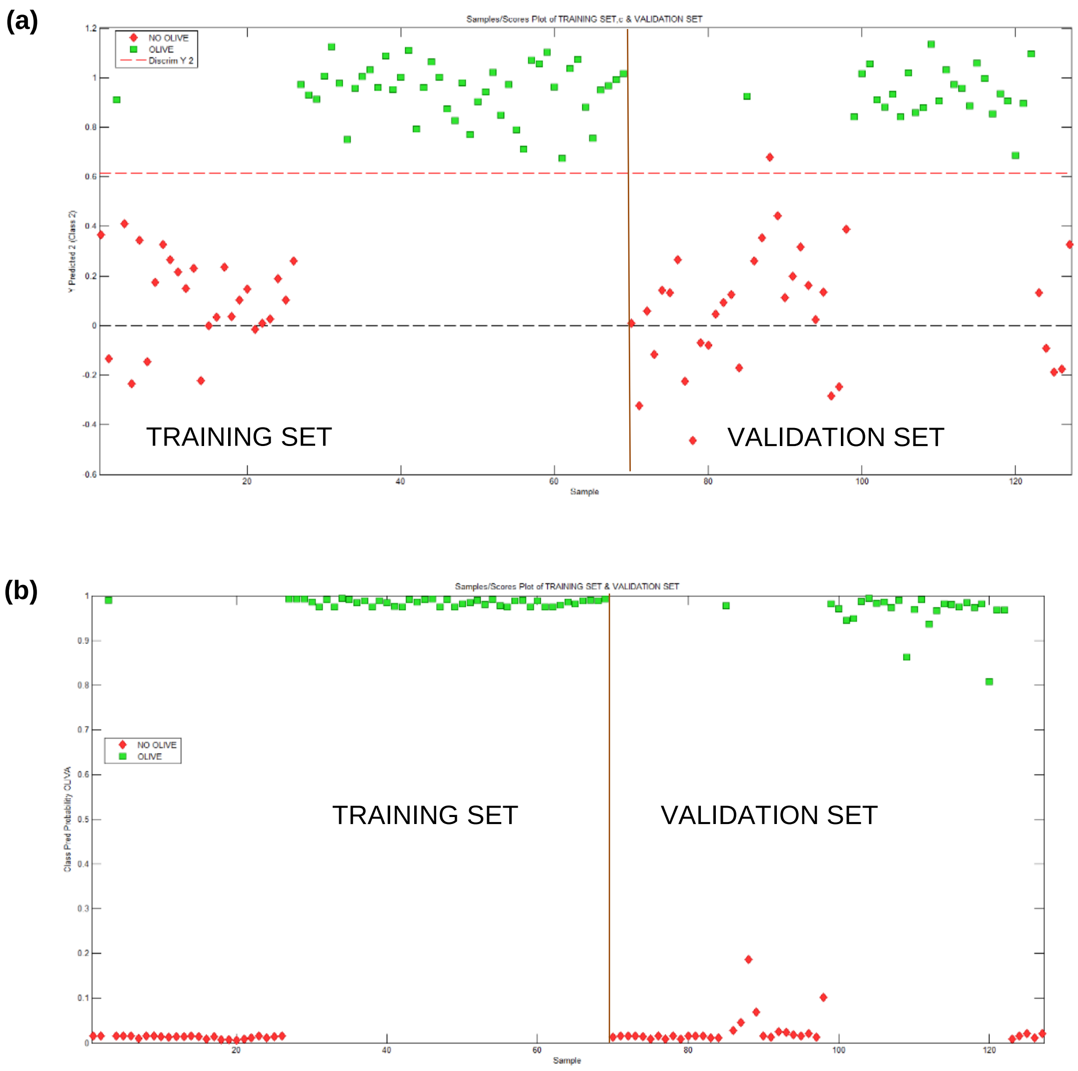
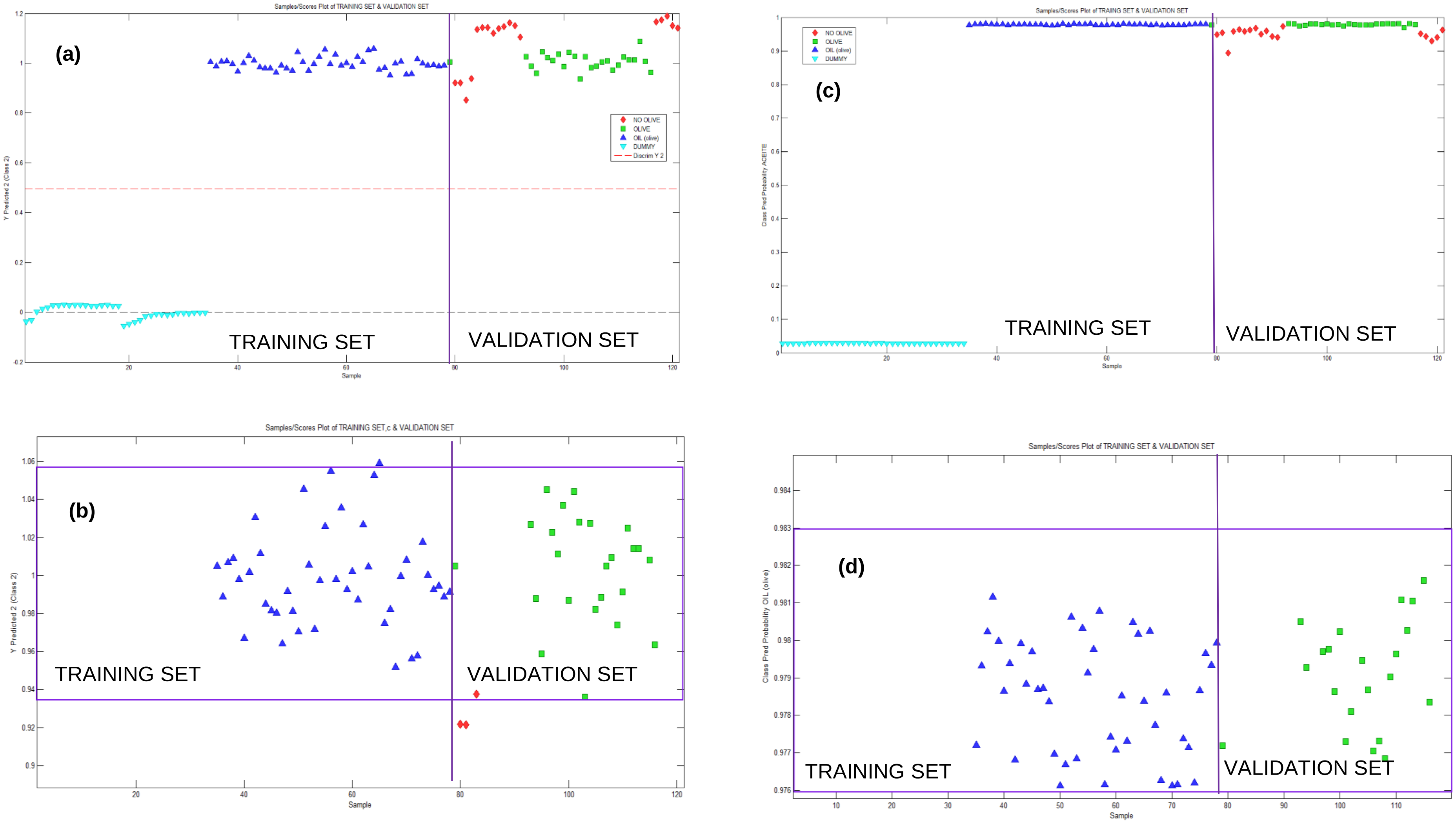

(a)
Samples/Scores Plot of TRAINING SET,c & VALIDATION SET
NO OLIVE
OLIVE
Discrim Y 2
Y Predicted 2 (Class 2)
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
20
40
60
80
100
120
Sample
TRAINING SET
VALIDATION SET
(b)
Samples/Scores Plot of TRAINING SET & VALIDATION SET
NO OLIVE
OLIVE
Class Pred Probability OLIVA
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
40
60
80
100
120
Sample
TRAINING SET
VALIDATION SET

Figure 4


(a)
Samples/Scores Plot of TRAINING SET & VALIDATION SET
Y Predicted 2 (Class 2)
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
20
40
60
80
100
120
Sample
NO OLIVE
OLIVE
OIL (olive)
DUMMY
Discrim Y 2
TRAINING SET
VALIDATION SET
(c)
Samples/Scores Plot of TRAIING SET & VALIDATION SET
Class Pred Probability ACEITE
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
40
60
80
100
120
Sample
NO OLIVE
OLIVE
OIL (olive)
DUMMY
TRAINING SET
VALIDATION SET
(b)
Samples/Scores Plot of TRAINING SET,c & VALIDATION SET
Y Predicted 2 (Class 2)
1.06
1.04
1.02
1
0.98
0.96
0.94
0.92
0.9
20
40
60
80
100
120
Sample
TRAINING SET
VALIDATION SET
(d)
Samples/Scores Plot of TRAINING SET & VALIDATION SET
Class Pred Probability OIL (olive)
0.984
0.983
0.982
0.981
0.98
0.979
0.978
0.977
0.976
10
20
30
40
50
60
70
80
90
100
110
Sample
TRAINING SET
VALIDATION SET