SORT 38 (2) July-December 2014, 231-250

New approaches in the chemometric analysis of infrared spectra of extra-virgin olive oils

María Isabel Sánchez-Rodríguez[1,*], Elena M. Sánchez-López[2],
Alberto Marinas[2], José Mª Caridad[1], Francisco José Urbano[2]
and José Mª Marinas[2]

Abstract

The aim of this paper is to apply new chemometric approaches to obtain quantitative information from near and mid infrared spectra of Andalusian extra-virgin olive oils, using gas chromatography as a classical reference analytical technique. Estimations of the content in saturated, monounsaturated and polyunsaturated fatty acids are given using partial least squares regression from the near and mid infrared data matrices as well as their concatenated matrix. The different estimations are evaluated in terms of goodness of fit (calibration) and prediction (validation), as a function of the number of partial least squares factors in the regression model and the used matrix of data. Furthermore, the nature, systematic or random, of the prediction errors is studied by a decomposition of their mean squared error. Finally, procedures of cross-validation are implemented in order to generalize the previous results.

MSC: 62H25, 62J05, 62Q99.

Keywords: Extra-virgin olive oil, infrared spectroscopy, partial least squares regression, cross-validation.

1. Introduction

Extra-virgin olive oil is an edible oil very much appreciated by its taste and benefits for health. Mediterranean countries (Spain, Italy, Greece, Turkey, Tunisia and Morocco) and Portugal cover 90% of the world production, Spain and Italy being the major producers and consumers. In Spain, Andalusia produces 80% of the national product.

[*] Corresponding author e-mail: td1sarom@uco.es

[1] Dep. Estadística, Econometría, I.O., Org. Empresas y Ec. Aplicada. University of Córdoba

[2] Dep. Química Orgánica. University of Córdoba

Received: September 2013
Accepted: June 2014

The composition of olive oil depends on the type and the distribution of the fatty acids present in the triglycerides and on the positions in which these fatty acids are esterified to hydroxyl groups in glycerol backbone. The principal fatty acids of vegetable oils are oleic, linoleic, linolenic, myristic, palmitic and estearic. The last three types are classified as saturated (SAFA), the oleic is monounsaturated (MUFA) and the linoleic and linolenic acids are polyunsaturated (PUFA).

Extra-virgin oil is by definition obtained only from the olive, using solely mechanical or other physical means, in conditions, particularly thermal conditions, which do not alter the oil in any way. It presents a high price of commercialization, which makes it susceptible to adulteration with other cheaper oils, such as hazelnut, sunflower, soybean, maize or refined olive oils (see, for example, Baeten et al. (2005), Gurdeniz and Ozen (2009) and Öztürk et al. (2010)), considerably modifying its quality indices. This makes it necessary to provide fast, reliable and cost-effective analytical procedures which require no or little sample manipulation. In this sense, for several years we have elaborated an extra-virgin olive oils database using diverse spectroscopic techniques such as near and mid infrared (NIR and MIR, respectively). IR techniques provide continuous information (spectra) that is rich in both isolated and overlapping bands and not so obvious to analyse as in the case of gas chromatography (GC). Nevertheless, the application of multivariate statistics to the above-mentioned spectra allows to obtain quantitative information (as the content of oil in diverse compounds) or qualitative (as the geographical origin or the protected designation of origin, PDO) about the olive oil.

There are in the literature diverse examples of application of NIR, MIR or concatenated NIR-MIR spectroscopic techniques to the quantitative and qualitative analysis of olive oils. Thus, for example, Bertran et al. (2000) and Galtier et al. (2007, 2011) classify several olive oils according to different geographical zones and determine the composition in fatty acids and triacylglycerols by using NIR spectra. Baeten et al. (2005) and Gurdeniz and Ozen (2009) study the possible adulteration of olive oils with lower quality oils (such as hazelnut, sunflower or maize) by MIR spectroscopy. Dupuy et al. (2010a, 2010b) and Sinelli et al. (2008) use NIR, MIR and concatenated NIR-MIR spectra to develop quantitative and qualitative studies of olive oil. Sinelli et al. (2010) apply NIR and MIR spectroscopies as a rapid tool to classify extra virgin olive oil on the basis of fruity attribute intensity. Casale et al. (2012) use NIR and MIR spectroscopical data, individually and jointly, to characterize olive oils from an Italian protected designation of origin.

Some of these works determine, using correlation, specially significant IR spectral bands to fit regression models to predict some components of olive oil (see Guillén and Cabo (1997) or Zhang et al. (2012)). Other works, such as Maggio et al. (2011), use partial least squares (PLS) regression to avoid the presence of multicollinearity in the model. PLS regression summarizes the information of a spectral band in some components or latent factors being orthogonal among them and so avoiding multicollinearity, incompatible with the hypothesis of uncorrelation in the general linear model. Other authors, such as Casale et al. (2012) or Dupuy et al. (2010a, 2010b), extract the PLS components from the complete NIR, MIR or concatenated NIR-MIR spectra, not from the previously selected spectral bands.

The aims of this paper are to revisit the procedures used in the literature to obtain quantitative information of olive oil from the near and mid zones of the infrared spectra and propose new approaches. The goal is to determine the profile in SAFA, MUFA and PUFA fatty acids of diverse extra-virgin olive oils by using the information provided by the NIR, MIR and concatenated NIR-MIR matrices of data, using the values obtained from GC as a reference . The estimations are provided by partial least squares regression models and are compared in terms of goodness of fit (calibration) and prediction (validation), that is, measuring errors that correspond to data used or not used to train the regression model. In addition, a decomposition of the mean squared error of prediction is provided to evaluate the character, systematic or random, of prediction errors (see Sánchez-Rodríguez et al. (2013)) . The obtained results are generalized using procedures of cross-validation, based on the design of repetitive algorithms that, for each iteration, modify the partition of the available data set in subsets of calibration and validation. Finally, three-dimensional scatterplots give a global vision for the three types of fatty acids and matrices of data, simultaneously.

The computer programs commonly used in Chemometrics have internally implemented a stopping criterion to determine the number of PLS latent factors to retain in the regression model. But, in this work, the PLS factors are progressively introduced in the model, with the aim of determining the evolution of calibration and validation errors as a function of the number of factors and the estimated fatty acid type. The chemometric software has also cross-validation procedures implemented that change, at each iteration, the learning and the test data sets, and provide a global mean of the calibration and validation errors. On the contrary, in the present work, the procedures of cross-validation have been programmed and show the results corresponding to each iteration. Therefore, the evolution and the variability of the fit and prediction errors can be studied for the successive iterations.

2. Acquisition of data

The studied samples include 128 Andalusian extra-virgin olive oils, collected for four consecutive seasons (from 2007 to 2011) with a ripeness index of 3. The varieties studied are, mainly, 'Arbequina', 'Hojiblanca', 'Picual', 'Lechín', 'Manzanilla', 'Picudo' and 'Royal'. Olive oil was extracted by the producers through a two-phase centrifugation system. The data for the subsequent statistical treatment have been provided by the following analytical chemical procedures:

- **Gas chromatography**. Classical separation technique that leads to discrete information including several usually well-defined, separated peaks from which, on proper integration, the content of various chemical components (for example, SAFA, MUFA and PUFA fatty acids) can be determined. It will be considered as reference technique in the next studies.

- **Spectroscopical techniques**. Infrared techniques, such as NIR and MIR, generate continuous information, rich in both isolated and overlapping bands attributed to vibration of chemical bonds in different molecules. The use of mathematical and statistical procedures allows us to extract the maximum useful information from data (Berrueta et al. (2007)).

2.1. Gas Chromatography (GC)

The determinations of fatty acid composition by GC-FID, according to the official methods for olive and pomace oil control in the European Union, EU (2011) and the International Olive Council, COI (2001a, 2001b), were performed by the staff of Organic Chemistry of University of Córdoba, using an Agilent 7890A gas chromatograph with a capillary column (SGE FORTE BPX-70 de 50 m $\times$ 220 μm $\times$ 0.25 μm). The conditions of analysis were as follows: 250 °C of injector temperature, 2 μL of injection volume, 260 °C of detector temperature. The oven temperature was programmed to remain at 180 °C for 15 min and then raised to 240 °C at a rate of 4 °C/min and maintained at this temperature for 5 min.

The triacylglycerol samples (olive oil samples), were initially submitted to a cold transesterefication procedure to convert the triacylglycerol into fatty acid methyl esters. This method is indicated for edible oils with acidity index lower than 3.3°. In this process, 0.1 g of olive oil are transferred into a 5 mL volumetric flask. Next, 2 mL *n*-heptane and 0,2 mL of a 2N KOH solution in methanol were added and reaction mixture was vigorously stirred. Finally, the methyl esters were extracted and subject to GC analyses.

2.2. NIR and MIR spectra

NIR and MIR spectra were obtained by the staff of the Organic Chemistry Department of the University of Córdoba within 15 days after reception of the samples which where kept in the fridge so that properties were not modified [Baeten et al. (2003)]. The instruments employed for spectra collection were available at the Central Service of Analyses (SCAI) at the University of Córdoba.

As for NIR instrument, it consisted in a Spectrum One NTS FT-NIR spectrophotometer (Perkin Elmer LLC, Shelton, USA) equipped with an integrating sphere module. Samples were analyzed by transflectance by using a glass petri dish and a hexagonal reflector with a total transflectance pathlength of approximately 0.5 mm. A diffuse reflecting stainless steel surface placed at the bottom of the cup reflected the radiation back through the sample to the reflectance detector. The spectra were collected by us-

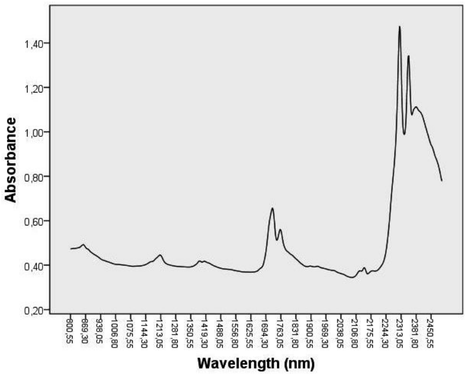


Figure 1: *NIR spectrum of an extra-virgin olive oil.*

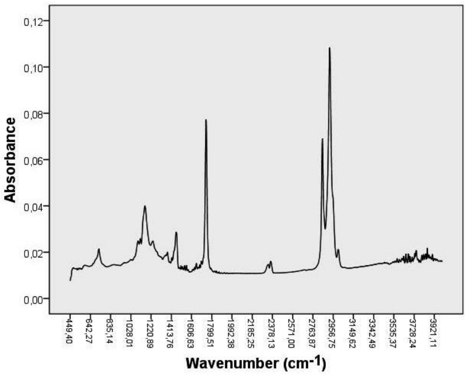


Figure 2: *MIR spectrum of an extra-virgin olive oil.*

ing Spectrum Software 5.0.1 (Perkin Elmer LLC, Shelton, USA). The reflectance ($\log 1/R$) spectra were collected with two different reflectors. Data correspond to the average of results with both reflectors, thus ruling out the influence of them on variability of the obtained results. Moreover, spectra were subsequently smoothed using the Savitzky-Golay technique, which performs a local polynomial least squares regression in order to reduce the random noise of the instrumental signal. Once pre-treated, NIR data of 1237 measurements for each case (representing energy absorbed by olive oil sample at 1237 different wavelengths, from 800.62 to 2499.64 nm) were supplied to the Department of Statistics (University of Córdoba) in order to be analysed.

Regarding MIR spectra of olive oil samples, they contain both well-resolved (3100-1721 cm^{-1}) and overlapping peaks (1500-700 cm^{-1}). Spectra were registered at room temperature in the 600 to 4000 cm^{-1} range on a Tensor 27 FTIR Spectrometer (Bruker Optics, Milano, Italy) coupled to an ATR (Attenuated total reflectance) device consisting in several reflection crystals (ZnSe). Software used was OPUS r. 5,0 (Bruker Optics), the resolution 2 cm^{-1} and 50 scan per sample. The number of measurements for each case was 1843, which were supplied to the Department of Statistics (University of Córdoba) for analysis.

3. Multivariate data analysis

3.1. Selection criteria of regression models

The purpose of this work is to use statistical regression models to determine the profile in fatty acids SAFA, MUFA and PUFA of extra-virgin olive oils obtained by gas chromatography (classical technique used as reference) from the information provided by the IR spectroscopy technique. The regression models are evaluated in terms of goodness-of-fit and predictive capability, using the following measures.

Let $y_1, y_2, \ldots, y_n$ be the observations of a dependent variable, Y, and the corresponding predictions, $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n$, of a regression model. The *mean squared error of calibration*, MSEC$= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 / n$, takes values nearer to 0 for a good fit (calibration). But MSEC is not dimensionless, that is, it depends on the units of measurement of the variable, hence it is useful. It is useful in comparisons of models with variables measured in the same units but not as an absolute goodness of fit measure.

Given the predictions for the future t observations, $\widehat{y}_{n+1}, \widehat{y}_{n+2}, \ldots, \widehat{y}_{n+t}$, of a certain regression model, the *mean squared error of the prediction*, MSEP$= \sum_{j=1}^{t} (y_{n+j} - \widehat{y}_{n+j})^2 / t$, evaluates the predictive capability (validation) of a regression model. The predictive capability of a model is obviously better as MSEP approaches 0, taking into account that MSEC has no upper bound and depends on the measurement units.

As indicated by Berrueta et al. (2007), the ideal situation in the evaluation of the predictive capability of a model is when there are enough data available to create separate test sets completely independent from the model building process (this validation procedure is known as *external validation*). When an independent test set is not available (e.g. because cost or time constraints make it difficult to increase the sample size), MSEP has to be estimated from the data used to train the regression. For this reason, as validation set, part of the original data set is used, avoiding the bias associated to the fact that the same data are used to the fit of the regression model and the evaluation of the predictions). The *cross-validation* procedures are designed to modify the selections repeatedly, using an algorithm that, for each iteration, changes the partition of the original data set into calibration and validation sets.

Besides, in line with the approach introduced by Fisher around 1920 relative to analysis of variance, Sánchez-Rodríguez et al. (2013) described new insights into evaluation of regression models through a decomposition of MSEP to analyse more in depth the causes of the prediction errors. Let $\overline{y}$ and $\overline{\widehat{y}}$ be the means of the t future observations and their predictions, s_Y and $s_{\widehat{Y}}$ are the corresponding standard deviations and $s_{Y\widehat{Y}}$ represents the covariance. Therefore, MSEP can be expressed as

$$\text{MSEP} = \frac{1}{t} \sum_{j=1}^{t} (y_{n+j} - \widehat{y}_{n+j})^2 = \left(\overline{y} - \overline{\widehat{y}}\right)^2 + \left(s_Y - s_{\widehat{Y}}\right)^2 + 2\left(s_Y s_{\widehat{Y}} - s_{Y\widehat{Y}}\right) = E_B + E_V + E_R,$$

or, equivalently, with the identity

$$1 = \frac{E_B}{\text{MSEP}} + \frac{E_V}{\text{MSEP}} + \frac{E_R}{\text{MSEP}} = U_B + U_V + U_R,$$

where U_B is the part of MSEP corresponding to the bias due to the systematic prediction errors; U_V indicates the difference between the variability of the real values and the variability of the predicted values; finally, U_R shows the random variability in the prediction errors.

A model is obviously better as MSEP approaches 0 (taking into account that MSEP is not upper bounded and depends on the unit of measurement). But, using the proposed decomposition, if MSEP shows a great percentage attributable to systematic errors, this aspect indicates that there is some detectable cause causing these deviations in the predictions. This cause must be detected in order to eliminate systematic errors. Thus, a great percentage of MSEP attributable to systematic prediction errors indicates that the model can be improved in some sense. Nevertheless, this improvement is difficult if the predictions generated by a model have a random nature because random errors, with a white noise appearance, are usually inherent to a process.

Definitively, the ideal situation for evaluating the predictive capability of a model is presented when MSEP has a value as close as possible to 0 and besides $U_B = 0$, that is, systematic errors do not exist in the prediction; $U_V = 0$, which indicates that the variability of the real values is the same as that of the predictions; and $U_R = 1$, which corresponds to prediction errors with random nature.

3.2. New methodological approaches in the chemometric analysis of IR spectra

Now, the procedures used in the literature to extract information of olive oil from IR spectra (NIR, MIR and concatenated NIR-MIR) are revised. The different contributions to each technique are conveniently motivated and justified.

1. **Extraction of the information from the complete IR spectra versus the analysis of some particular IR bands**. There are in the literature many references in which the analysis of IR spectra is made based on the detection of highly informative bands. One such example is the work by Guillén and Cabo (1997), who relate IR spectral bands of edible oils with some chemical functional groups. This approach is based on the *Lambert-Beer Law*, which states that the intensities of the spectral bands are proportional to the concentration of their respective functional groups. The frequencies of some bands, fundamentally the ones associated to the so-called *fingerprint region*[1], are highly correlated to the composition of olive oil. Guillén and Cabo (1997) successfully obtained regression equations to predict the content in SAFA, MUFA and PUFA fatty acids of olive oil from the frequencies of some bands in the fingerprint region (see Table 1). A follow-up study (Guillén and Cabo, 1998) generalized the previous results by regression models that provide relationships between the composition in SAFA, MUFA and PUFA of edible oils and the ratio of absorbance of specific bands of the IR spectra, not necessarily

[1] The region 1500-700 cm^{-1} of the MIR spectra is named *fingerprint region* because this region is highly characteristic of a specific compound. Little changes in the molecular structure frequently cause significant changes in the absorption peaks of this region.

associated to the fingerprint region. Besides, Guillén and Cabo (1999) used the previous regression equations to determine the composition of mixtures of olive oil and other low quality oils (such as sunflower or peanut), using gas chromatography values as references.

There are other works which analyse IR spectra by determining relevant frequency bands. Vlachos et al. (2006) establish the relation between the frequency 3009 cm^{-1} of the IR spectra and the percentage of adulteration of olive oil with low quality oils. Rohman and Man (2010) use PCA and PLS components extracted from the fingerprint region 1500-1000 cm^{-1} (MIR spectra) to quantitatively and qualitatively analyse extra-virgin olive oils, to detect possible adulteration with palm oil. Nicoletta et al. (2010) select some regions from NIR and MIR spectra to classify, by discriminant analysis, diverse extra-virgin olive oils based on their fruity attribute intensity. Zhang et al. (2012) divide the IR spectra in regions, attending to the absorbance peaks, to establish linear regression equations to detect possible adulteration of vegetables oils with used frying oils.

All the previously cited works determine, by using correlation, highly informative IR spectral regions to predict the composition of olive oil. In general, the determined zones are localized in the mid infrared spectral region (MIR), where the spectral fingerprint is localized.

Our previous study (Sanchez-Rodriguez et al., 2013), from NIR spectral data, compares the estimation results obtained by extracting information from the whole spectra with those provided by some specific NIR bands (either determined by cluster analysis or associated to certain spectral peaks). The best calibration and validation results are obtained from the whole spectra. This is the reason why the present work uses the whole NIR, MIR and concatenated NIR-MIR spectra to

Table 1: *Coefficients for IR equations, Frequency* $= a + b\%M$ $(\%P$ *o* $\%S)$, *and linear correlation coefficients*[a] *(Guillén and Cabo, 1997).*

Percentage	a	$b(10^{-2})$	r
M	3010.40	-7.24	0.9853
P	3004.85	+6.10	0.9492
M	1394.90	+9.90	0.9910
P	1402.61	-8.43	0.9223
M	1100.46	-4.87	0.9908
P	1096.68	+4.43	0.9176
S	2926.04	-6.28	0.8504
S	2855.07	-6.59	0.9565
S	1122.65	-22.65	0.9510
S	724.30	-9.93	0.9924
S	1238.11	+2.33	0.8408
S	1160.20	24.44	0.9989

[a] %M, %P and %S represents the percentage of MUFA, PUFA and SAFA, respectively.

predict the content of olive oil in SAFA, MUFA and PUFA fatty acids. As subsequently shown, although in the analysis of spectral bands the most informative zones are localized in the MIR spectral region, the NIR spectra provides better estimations in certain situations when the whole spectral information is used.

2. **PLS regression versus general linear regression or PCA regression**. The analysis of IR spectra from the detection of relevant wavelength bands to obtain quantitative information (such as the prediction of the content of olive oil in some specific compounds) is based on the matching of some wavelengths with high correlation with the response variable. Then linear regression equations are fit to predict the percentage of the compound as a function of the wavelengths (see, for example, Guillén and Cabo (1997, 1998, 1999), Vlachos et al. (2006), Rohman and Che Man (2010), Zhang et al. (2012)). The selection of a single wavelength could lead to a waste of useful statistical information. But the selection of many wavelengths highly correlated with the dependent variable could cause the presence of multicollinearity among the explanatory variables, incompatible with the hypothesis of uncorrelation in the general linear model. This is why the use of principal component regression (PCA regression) or partial least squares regression (PLS regression) is more interesting. Both methodologies summarize the information of the whole IR spectrum in some latent factors or components, orthogonal among themselves, thus avoiding the possible multicollinearity in the model. These factors are obtained as linear combinations of the independent variables in both methodologies. However, the factors are obtained by maximizing the covariances (or correlations) among the explanatory variables, in PCA regression, and the covariances (or correlations) between the explanatory variables and the dependent one, in PLS regression.

   In this work, PLS regression has been selected because a previous one (Sánchez-Rodríguez et al. (2013)) highlights the benefits of PLS regression versus PCA regression in the determination of quantitative information from NIR data. There are also other works indicating that PLS regression is better than PCA regression in the multivariate analysis of NIR or MIR spectral data (see, for example, Frank and Friedman (1993) or Maggio et al. (2011)).

3. **Progressive introduction of PLS factors in the regression model**. The estimation algorithms that compare the MSEC and MSEP values obtained for PCA and PLS regression or for PCA or PLS discriminant analysis (PCA-DA or PLS-DA) have established, in the chemometric software, a stopping internal criterion to determine the number of factors to retain. This is the case, for example, of the article by Dupuy et al. (2010a) and (2010b), which uses UNSCRAMBLER software version 9.8 from CAMO (Computer Aided Modelling, Trondheim, Norway) and Matlab software from MathWorks in the analysis of NIR, MIR and concatenated NIR-MIR spectra.

The criteria determining the number of factors to retain in PLS regression are diverse. For example, in PCA, the *Kaiser criterion* is the default in most statistical software. It suggests that those principal components with eigenvalues equal to or higher than 1 should be retained, as each eigenvalue represents the variance of the corresponding factor. In PLS analysis, the *criterion of the first increase of the mean squared error of prediction* is considered: the number of latent factors taken into account is

$$h^* = min\{h > 1 : \mathrm{MSEP}(h+1) - \mathrm{MSEP}(h) > 0\},$$

where $\mathrm{MSEP}(h)$ is the mean squared error of prediction of the regression model with h factors. Gowen et al. (2010) present some measures for preventing the overfitting in PLS calibration models of near-infrared (NIR) spectroscopy data and investigate the simultaneous use of both model bias and variance in the selection of the number of latent factors to include in the model.

The cited criteria have an empirical character and are not unanimously applied. Therefore, the present work does not fix the number of factors. This number is progressively incremented in the PLS regression model and the associated MSEC and MSEP values are compared. As subsequently confirmed, the NIR matrix of data provides better estimations of the content in fatty acids of olive oil than MIR matrix for a lower number of factors whereas the opposite holds true for a higher number of latent factors.

4. **Procedures of cross-validation**. The procedures of cross-validation are aimed at avoiding the bias associated to the case of using the same data to fit the regression model and to evaluate the corresponding predictions. They are repetitive algorithms that, at each iteration, subdivide the set of original data in the calibration and the validation subsets. The calibration (or fit) set, formed by 80% of the data approximately, is used in the fit of the model and provides the MSEC value as a measure of goodness-of-fit. But the validation (or prediction) set, formed by the remaining 20% of the data, is reserved in the training of the model and so can be used to evaluate its predictive capability with the MSEP value.

   The computer programs frequently used in Chemometrics have some cross-validation procedures implemented. They iteratively repeat the selection of calibration and validation sets and provide and average of the MSEC and MSEP values obtained for each iteration. This is the case, for example, of the paper by Rohman and Man (2010), that uses the software TQ Analyst$^{\mathrm{TM}}$ Version 6 (Thermo Electron Corporation, Madison, WI); Sinelli et al. (2010), which uses the V-PARVUS package (Forina et al. (2008)); Dupuy et al. (2010b), that uses the UNSCRAMBLER software version 9.8 from CAMO.

But the present work calculates and represents the MSEC and MSEP values obtained for different random selections of the calibration and validation sets, from the NIR, MIR and concatenated NIR-MIR matrices of data. The algorithm has been programmed by using the Matlab software from MathWorks. The graphical representations permit to compare not only the mean MSEC and MSEP values but also their variability in the successive selections. The three-dimensional graphics permit to compare the results for the three type of acids and matrices of data simultaneously.

5. **Decomposition of the mean squared error of prediction**. With the aim of analyzing the nature of the prediction errors, this work uses a decomposition of the MSEP value in the terms E_B, E_V and E_R, attributable to systematic errors, the difference in variability among the real and the predicted values and random errors, respectively (Section 3.1, Sánchez-Rodríguez et al. (2013)). This decomposition is presented for the predictions for each type of fatty acid (SAFA, MUFA and PUFA) and spectral zone (NIR, MIR and concatenated NIR-MIR), as a function of the number of PLS factors in the regression model. Besides, in the context of cross-validation, this decomposition is also presented for the successive selections of calibration and validation sets. The ideal situation for evaluating the predictive capability of a model is presented when MSEP has a value nearer to 0 and the great percentage of this value is associated to the randomness and the lowest percentage is attributable to systematic errors. This work afterwards highlights that these percentages depend on the type of fatty acid to estimate, the IR spectral zone used for the estimation (NIR, MIR or NIR-MIR) and the number of PLS factors in the regression model.

6. **Treatment of the spectra in the context of functional data analysis**. A line of future research (see Sánchez-Rodríguez and Caridad (2014)) could consider IR spectra as so-called *data objects* in *object-oriented data analysis* (OODA). This is the particular case of functional data analysis (FDA), in which the data objects of OODA are curves (see the overview by Marron and Alonso (2014)). In this context, multivariate techniques such as PCA or PLS regression, have been extended to the functional case. For example, Aguilera et al. (2010) apply functional PLS and PCA regressions to simulated and spectrometric data, comparing the results with the corresponding discrete ones and concluding that functional PLS regression provides better estimations of the parameter function than functional PCA regression and similar predictions. Preda and Saporta (2005) apply functional regression models to predict the behaviour of shares and conclude that the functional PLS regression model provides the best forecasts evaluating the global model quality by the sum of squared errors. Finally, also the classification techniques, such as logit regression or discriminant analysis, have been successfully extended to the functional case (see, respectively, Escabias et al. (2007) or Preda et al. (2007)).

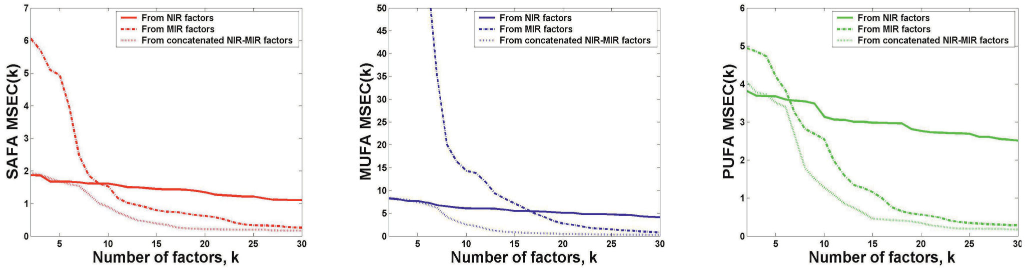
4. Results and discussion

Initially, Section 4.1 deals with the estimation of SAFA, MUFA and PUFA fatty acids of olive oils by PLS regression from the NIR, MIR and concatenated NIR-MIR matrices of data. The results of calibration and validation depend on the number of PLS factors in the regression model. These results are compared for the three matrices of data and types of fatty acids. The randomness of the prediction errors is analysed by a decomposition of MSEP. Subsequently, in Section 4.2, the previous results are generalized by using cross-validation procedures, that is, changing iteratively the training and the test data sets. Besides, three-dimensional scatterplots permit to obtain conclusion simultaneously for the three matrices of data or types of fatty acids.

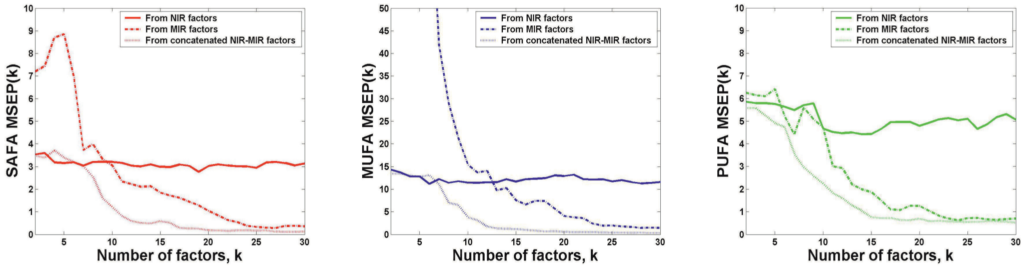
4.1. Chemometrics from IR data: progressive introduction of PLS factors in the regression model

NIR and MIR spectroscopies provide $n \times p$ data matrices whose rows refer to an olive oil ($n = 128$, in total) and each column is associated to a wavelength of the spectrum ($p_{\mathrm{NIR}} = 1237$ and $p_{\mathrm{MIR}} = 1843$). The information given by NIR and MIR data is summarized by using PLS regression. Sánchez-Rodríguez et al. (2013) pointed out that this technique, applied directly to the matrix of NIR data, provides a potential methodology to predict the content in fatty acids of olive oil. This paper shows that the results obtained from the whole matrix of data, being considered as a “black box”, are better than the ones obtained with the selection of some spectral peaks or spectral regions by cluster analysis. Besides, PLS regression considerably improves, in this context, the results obtained for PCA regression. As stated above, both PCA and PLS methodologies provide components or factors orthogonal among themselves, thus avoiding the possible presence of multicollinearity in the regression model.

With regard to the rows of the NIR and MIR data matrices, 80% of them, randomly selected, will be used for calibration and the remaining 20%, for prediction or validation. Initially, the NIR matrix of data is considered and PLS components are extracted. Those components will be progressively introduced in the PLS regression models that consider the content in SAFA, MUFA and PUFA fatty acids as explained variables, respectively. For each number of introduced components, the mean squared error of calibration and prediction, MSEC and MSEP, are calculated. The same process is repeated considering, secondly, the MIR data matrix and, finally, the concatenated NIR-MIR data matrix. In addition, with the purpose of determining the character, systematic or random, of the prediction errors, a decomposition of MSEP obtained by the PLS regression models on NIR, MIR and NIR-MIR matrices (in the U_B, U_V and U_R components) is obtained for each fatty acid.



Figures 3, 4 and 5: *MSEC in the estimation of SAFA, MUFA and PUFA from PLS components of NIR, MIR and concatenated NIR-MIR matrices.*

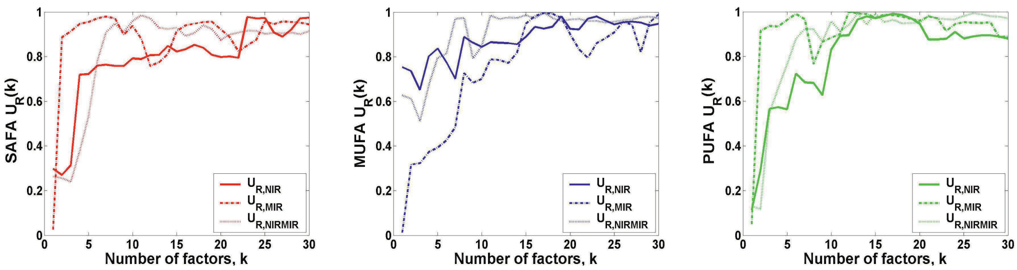


Figures 6, 7 and 8: *MSEP in the estimation of SAFA, MUFA and PUFA from PLS components of NIR, MIR and concatenated NIR-MIR matrices.*

Figures 3, 4 and 5 represent the MSEC in the estimation of SAFA, MUFA and PUFA acids obtained by successively introducing components in the regression model. These components are obtained on the NIR, MIR and concatenated NIR-MIR matrices of data. These figures show that, for the three types of fatty acids, the NIR (and concatenated NIR-MIR) matrix of data provides better calibration results in regression models with a lower number of PLS factors. But MIR (and concatenated NIR-MIR) matrix supplies better estimations for models with a higher number of factors.

Then, Figures 6, 7 and 8 represent, in the same context, the respective MSEP values. MSEP evaluates the predictive capability of a model, taking into account that the estimations are calculated by using observations that are not included in the fit or calibration of the model. The conclusions in prediction are similar to the ones obtained in calibration: the MSEP values obtained from MIR data are lower than the ones obtained from NIR data when the number of PLS factors in the model is sufficiently high, but not for low values.

Figures 9, 10 and 11 show the U_R term in the decomposition of MSEP for SAFA, MUFA and PUFA acids, respectively. This term corresponds to random prediction errors and, as in the previous graphics, is expressed as a function of the number of PLS components in the model. The figures evidence that the U_R term represents the great



Figures 9, 10 and 11: *U_R term of MSEP in the estimation of SAFA, MUFA and PUFA from PLS components of NIR, MIR and concatenated NIR-MIR matrices.*

percentage for each case, as this ratio is near to 1. With respect to the comparison of the techniques, there are differences depending on the used NIR, MIR and concatenated NIR-MIR matrices of data and the estimated fatty acid. For a higher number of PLS factors in the model, the three NIR, MIR and NIR-MIR U_R terms are very close to one. But for a lower number of factors, the MIR U_R term is closer to one in the estimation of SAFA and PUFA acids but it is farther from one in the estimation of MUFA acids. In this last case, the NIR matrix of data provides better results.

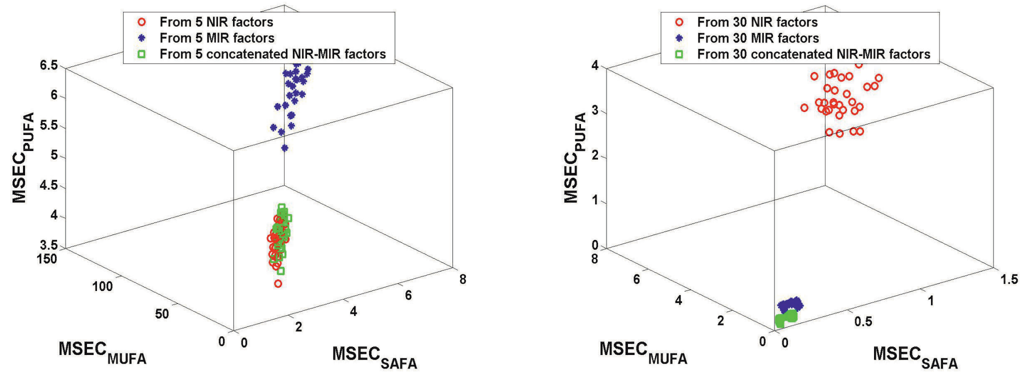
These results suggest that, under our experimental conditions, a more accurate estimation in calibration and validation of SAFA, MUFA and PUFA content in extra-virgin olive oil (taking GC as the reference technique) is obtained from the NIR matrix for a lower number of PLS factors. For a greater number of PLS factors, the MIR matrix provides the best results. The previous considerations are important as usual chemometric computer programs have internally implemented a stopping criterion to retain a concrete number of PLS factors in the regression model. It is interesting to identify the range of variation of this number to determine the region of the IR spectra, NIR or MIR, that provides better estimations of the different fatty acids. Then, analysing the nature of the prediction errors, the percentage of them attributable to random causes also depends on the region of the IR spectra and the type of acid. The MIR matrix provides, in general, better results in the estimation of SAFA and PUFA acids, irrespective of the number of PLS factors in the model. On the contrary, in the estimation of MUFA acids, NIR matrix supplies better results, also independently of the number of factors.

4.2. Cross-validation: generalization of the previous results

In the last subsection, the original data have been subdivided in a single calibration set (containing the 80% of the original data, specifically, 102 out of 128 data) and a single validation set (with the 20%, that is, 26 data). The calibration set is used to train the regression model. The validation set is used to test the model, using data reserved in the fit of the model. With the goal of generalizing the previously obtained results, procedures of cross-validation are used in this section. They are implemented by a repetitive algorithm that, for each iteration, modifies the partition in calibration and validation subsets of the original data set. For each iteration, MSEC and MSEP are calculated for evaluating, respectively, the goodness-of-fit and the predictive capability of the corresponding model.

More specifically, the cross-validation algorithm has been implemented for 30 iterations, randomly selecting, for each one, the sets considered for calibration and validation. Besides, since the previous section highlights differences depending on the number of PLS factors in the regression model, this section compares the results for a low number of factors, 5, and also for a high number of factors, 30.

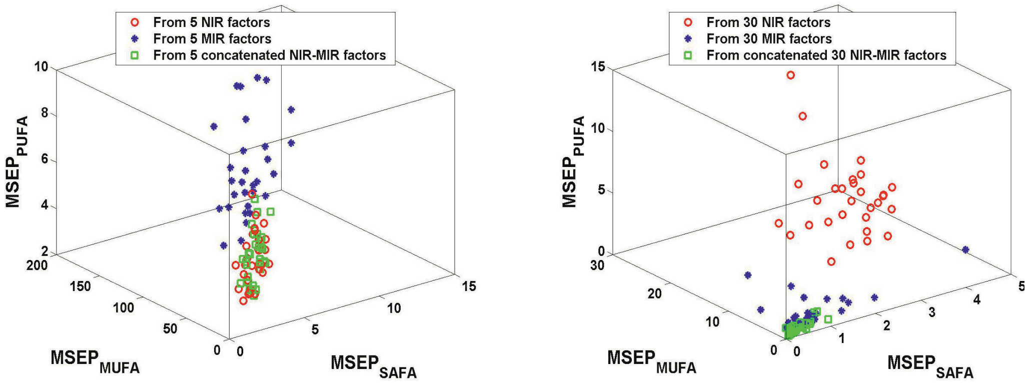
Figures 12 and 13 draw three-dimensional scatterplots for goodness-of-fit or calibration in cross-validation. The point clouds are associated to models with 5 and 30 PLS factors, respectively, representing $\mathrm{MSEC_{SAFA}}$, $\mathrm{MSEC_{MUFA}}$ and $\mathrm{MSEC_{PUFA}}$ in x, y, z axes. Unlike the previously represented figures, these graphics permit to compare, in a global manner, the results obtained for the three types of fatty acids simultaneously. For 5 PLS factors (Figure 12), the MIR point cloud is farther from the origin (0,0,0) than the corresponding to the NIR (and NIR-MIR) data. For 30 PLS factors (Figure 13), the conclusions are the opposite: in this case, the estimations from NIR data are associated with the high MSEC values.



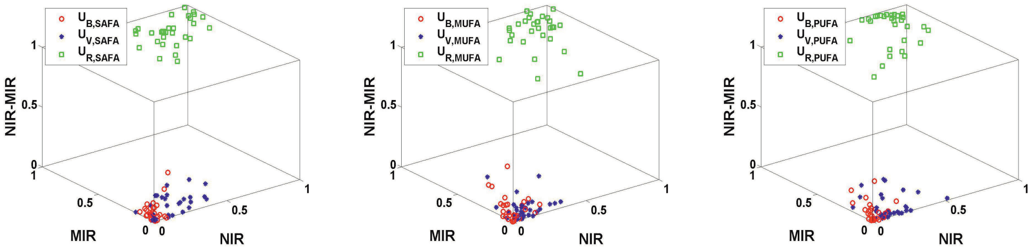
Figures 12 and 13: *MSEC obtained by the cross-validation algorithm from NIR, MIR and concatenated NIR-MIR data (for SAFA, MUFA and PUFA) for 5 and 30 factors, respectively.*

The same conclusions are obtained for the three fatty acid types, if the models are compared in validation or prediction terms (using MSEP, see Figures 14 and 15). Besides, the variability existing among the MSEP values is higher for the MIR than for the NIR estimations in the models with 5 PLS factors and lower for the models with 30 PLS factors.

Figures 16-18 represent, for each acid type, the decomposition of MSEP in the terms U_B, U_V and U_R obtained, by the cross-validation algorithm, for each iteration. The



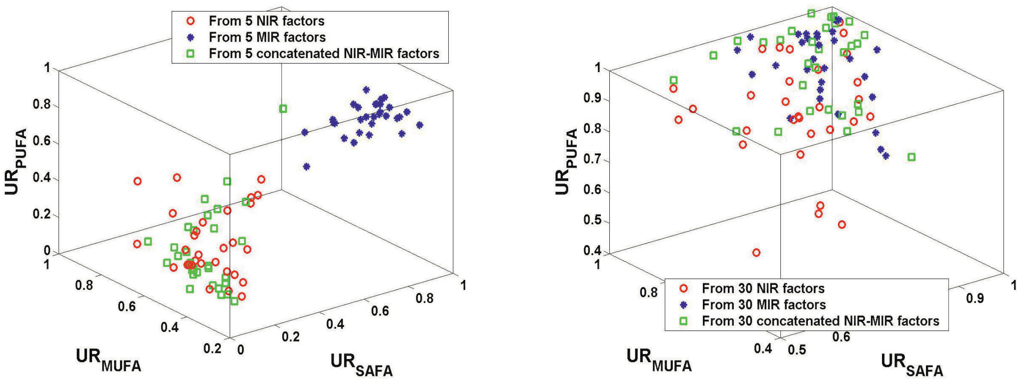
Figures 14 and 15: *MSEP obtained by the cross-validation algorithm from NIR, MIR and concatenated NIR-MIR data (for SAFA, MUFA and PUFA) for 5 and 30 factors, respectively.*



Figures 16, 17 and 18: *Decomposition of MSEP obtained by the cross-validation algorithm in the estimation of SAFA, MUFA and PUFA from PLS components of NIR, MIR and concatenated NIR-MIR matrices, respectively.*

aim is to determine the nature, random or systematic, of the prediction errors. The point clouds represent the values corresponding to NIR, MIR and NIR-MIR matrices in *x*, *y*, *z* axes, respectively. It is evident that, for each case, the component corresponding to random prediction errors, U_R, is associated to the great percentage, as this ratio is near to 1 for NIR, MIR and NIR-MIR axes. This is the suitable situation in the evaluation of the predictive character of a model.

Finally, with the aim to confirm the differences detected previously depending on the number of the PLS factors in the model, Figures 19 and 20 depict the U_R term associated with the models with 5 and 30 factors. The scatterplots represent the values corresponding to $U_{R,\text{SAFA}}$, $U_{R,\text{MUFA}}$ and $U_{R,\text{PUFA}}$ in *x*, *y*, *z* axes, respectively. The results show that the U_R term obtained from the NIR, MIR and concatenated NIR-MIR matrices of data is close to 1 in models with a relatively high number of factors (30). But, in model with a low number of factors, the NIR and NIR-MIR U_R terms are lower, clearly discriminated from the one obtained from the MIR data.



Figures 19 and 20: *U_R term of MSEP obtained by the cross-validation algorithm from NIR, MIR and concatenated NIR-MIR data (for SAFA, MUFA and PUFA) for 5 and 30 factors, respectively.*

5. Conclusions

In recent years, procedures which permit to determine in a fast and efficient manner the profile of olive oils in different components have been generalized, specially aiming at evaluating quality indexes. In this sense, spectroscopic techniques have been extended. In parallel, multivariate statistics has emerged as a powerful tool to identify and extract the information contained in spectra.

In this work, Chemometrics is applied to data obtained from IR spectra, in the near (NIR) and mid (MIR) zones, and using GC data as a reference. PLS regression models to predict the content in SAFA, MUFA and PUFA fatty acids of olive oil are proposed, using the three NIR, MIR and concatenated NIR-MIR matrices of data. The final conclusion is that the best estimation of calibration or fit and validation or prediction are obtained from the NIR data for lower numbers of PLS factors and from the MIR data for higher numbers of factors. This is important to be taken into account since, usually, chemometric computer programs have a stopping criterion implemented to determine the number of PLS factors to be retained.

These conclusions are generalized via cross-validation procedures. They compare estimations in terms of goodness-of-fit and prediction for different calibration and validation subsets and evidence the desirable main random nature of the estimation errors. Three-dimensional scatterplots confirm the differences among the three fatty acid types and matrices simultaneously.

Then, this study analyses the prediction errors to determine their nature, systematic or random. Also in this case the conclusions depend on the number of PLS latent factors, the type of fatty acid and the matrix of data. These differences are detected by the U_R term, that represents the percentage of randomness in the prediction errors. In general, irrespective of the number of factors in the regression model, the MIR zone provides a higher value in the estimation of SAFA and PUFA acids. But, in the estimation of MUFA acids, the NIR matrix gives better estimations. In the three-dimensional representation of the U_R term for the three acids and IR zones, this term is always close to 1 for a high number of PLS factors. But, for a low number of factors, the NIR and NIR-MIR U_R terms are clearly lower than the associated to the MIR data.

Acknowledgements

The authors thank the financial support by 'Junta de Andalucía' (Project P08-FQM-3931) and FEDER funds.

References

Aguilera, A. M., Escabias, M., Preda, C. and Saporta, G. (2010). Using basis expansions for estimating functional PLS regression. Applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems*, 104, 289–305.

Baeten, V., Aparicio, R., Marigheto, N. and Wilson, R. (2003). *Manual del aceite de oliva*. AMV ediciones, Mundi-Prensa.

Baeten, V., Fernández Pierna, J. A., Dardenne, P., Meurens, M., García-González, D. L. and Aparicio-Ruiz, R. (2005). Detection of the presence of hazelnut oil in olive oil by FT-Raman and FT-MIR spectroscopy. *Journal of agricultural and food chemistry*, 53(16), 6201–6206.

Berrueta, L. A., Alonso-Salces, R. M. and Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography A*, 1158, 196–214.

Bertran, E., Blanco, M., Coello, J., Iturriaga, H., Maspoch, S. and Montoliu, I. (2000). Near infrared spectrometry and pattern recognition as screening methods for the authentication of virgin olive oils of very close geographical origins. *Journal of Near Infrared Spectroscopy*, 8, 45.

Casale, M., Oliveri, P., Casolino, C., Sinelli, N., Zunin, P., Armanino, C., Forina, M. and Lanteri, S. (2012). Characterization of PDO olive oil Chianti Classico by non-selective (UV-visible, NIR and MIR spectroscopy) and selective (fatty acid composition) analytical techniques. *Analytical Chimica Acta*, 712, 56–63.

Commission regulation (EU) No 61/2011 of 24 January 2011 amending Regulation (EEC) No 2568/91 on the characteristics of olive oil and olive-residue oil and on the relevant methods of analysis.

D'Imperio, M., Mannina, L., Capitani, D., Bidet, O., Rossi, E., Bucarelli, F. M., Quaglia, G. B. and Segre, A. (2007). NMR and statistical study of olive oils from Lazio: a geographical, ecological and agronomic characterization. *Food chemistry*, 105(3), 1256–1267.

Dupuy, N., Galtier, O., Ollivier, D., Vanloot, P. and Artaud, J. (2010a). Comparison between NIR, MIR, concatenated NIR and MIR analysis and hierarchical PLS model. Application to virgin olive oil analysis. *Analytica chimica acta*, 666(1), 23–31.

Dupuy, N., Galtier, O., Le Dréau, Y., Pinatel, C., Kister, J. and Artaud, J. (2010b). Chemometric analysis of combined NIR and MIR spectra to characterize French olives. *European Journal of Lipid Science and Technology*, 112(4), 463–475.

Escabias, M., Aguilera, A. M. and Valderrama, M. J. (2007). Functional PLS logit regression model. *Computational Statistics and Data Analysis*, 51, 4891–4902.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109–135.

Forina, M., Lanteri, S., Armanino, C., Casolino, C., Casale, M. and Oliveri, P. (2008). *P-PARVUS. Dip. Chimica e Tecnologie Farmaceutiche e Alimentari*, University of Genova, http://www.parvus.unige.it.

Galtier, O., Abbas, O., Le Dréau, Y., Rebufa, C., Kister, J., Artaud, J. and Dupuy, N. (2011). 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. *Vibrational Spectroscopy*, 55(1), 132–140.

Galtier, O., Dupuy, N., Le Dréau, Y., Ollivier, D., Pinatel, C., Kister, J. and Artaud, J. (2007). Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra. *Analytica chimica acta*, 595(1), 136–144.

Gowen, A. A., Downewy, G., Esquerre, C. and O'Donnell, C. P. (2010). Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. *Journal of Chemometrics*, 25, 375–381.

Guillén, M. D. and Cabo, N. (1997). Characterization of edible oils and lard by Fourier transform infrared spectroscopy. Relationships between composition and frequency of concrete bands in the fingerprint region. *Journal of the American Oil Chemists' Society*, 74(10), 1281–1286.

Guillén, M. D. and Cabo, N. (1998). Relationships between the composition of edible oils and lard and the ratio of the absorbance of specific bands of their Fourier transform infrared spectra. Role of some bands of the fingerprint region. *Journal of Agricultural and Food Chemistry*, 46, 1788–1793.

Guillén, M. D. and Cabo, N. (1999). Usefulness of the frequencies of some Fourier transform infrared spectroscopic bands for evaluating the composition of edible oil mixtures. *European Journal of Lipid Sciences and Technology*, 1, 71–76.

Guillén, M. D. and Ruiz, A. (2003). Edible oils: discrimination by ^{1}H nuclear magnetic resonance. *Journal of the Science of Food and Agriculture*, 83(4), 338–346. Relationships between composition and frequency of concrete bands in the fingerprint region. *Journal of the American Oil Chemists' Society*, 74(10), 1281–1286.

Gurdeniz, G. and Ozen, B. (2009). Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. *Food chemistry*, 116(2), 519–525.

International Olive Oil Council, 200 (COI / T.20 / Doc. no. 24 / 2001a). Preparation of the fatty acid methyl esters from olive oil and olive-pomace oil.

International Olive Oil Council, 200 (COI / T.20 / Doc. no. 17 / 2001b). Determination of trans unsaturated fatty acids by capillary column gas chromatography.

Maggio, R. M., Valli, E., Bendini, A., Gómez-Caravaca, A. M., Toschi, T. G. and Cerretani, L. (2011). A spectroscopic and chemometric study of virgin olive oils subjected to thermal stress. *Food Chemistry*, 127, 216–221.

Marron, J. S. and Alonso, A. M. (2014). Overview of object oriented data analysis. *Biometrical Journal*, 56, doi: 10.1002/bimj.201300072.

Öztürk, B., Yalçin, A. and Özdemir, D. (2010). Determination of olive oil adulteration with vegetable oils by near infrared spectroscopy coupled with multivariate calibration. *Journal of Near Infrared Spectroscopy*, 18, 191–201.

Preda, C. and Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, 48, 149–158.

Preda, C., Saporta, G. and Lévéder, C. (2007). PLS classification of functional data. *Computational Statistics*, 22, 223–235.

Rezzi, S., Axelson, D. E., Héberger, K., Reniero, F., Mariani, C. and Guillou, C. (2005). Classification of olive oils using high throughput flow^{1}H NMR fingerprinting with principal component analysis, linear discriminant analysis and probabilistic neural networks. *Analytica Chimica Acta*, 552(1), 13-.24.

Rohman, A. and Che Man, Y. B. (2010). Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulteration with palm oil. *Food research international*, 43, 886–892.

Sánchez-Rodríguez, M. I., Sánchez-López, E., Caridad, J. M., Marinas, A., Marinas, J. M. and Urbano, F. J. (2013). New insights into evaluation of regression models through a descompositon of the prediction errors: application to near-infrared spectral data. *Statistics and Operations Research Transactions (SORT)*, 37(1), 57–78.

Sánchez-Rodríguez, M. I. and Caridad, J. M. (2014). Modelling and partial least squares approaches in OODA. *Biometrical Journal*, 37(1), 771–773.

Sinelli, N., Casiraghi, E., Tura, D. and Downey, G. (2008). Characterisation and classification of Italian virgin olive oils by near-and mid-infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 16, 335–342.

Sinelly, N., Cerretani, L., Di Egidio, V., Bendini, A. and Casiraghi, E. (2010). Application fo near (NIR) infrared and mid (MIR) infrared spectroscopy as a rapid tool to classify extra virgir olive oil on the basis of fruity attribute intensity. Food research international, 43, 369–375.

Vlachos, N., Skopelitis, Y., Psaroudaki, M., Konstantinidou, V., Chatzilazarou, A. and Tegou, E. (2006). Applications of Fourier transform-infrared spectroscopy to edible oils. *Analytica Chimica Acta*, 573–574, 459–465.

Zhang, Q., Liu, C., Sun, Z., Hu, X., Shen, Q. and Wu, J. (2012). Authentication of edible vegetable oils adulteration with used frying oil by Fourier transform infrared spectroscopy. *Food Chemistry*, 132, 1607–1613.