Application of Fourier Transform Mid-Infra-Red Attenuated Total Reflectance (FT-MIR-ATR) for the authentication of Maltese extra virgin olive oil

Frederick LIA[1,*]
Marion ZAMMIT MANGION[2]
Claude FARRUGIA[3]

[1] Chemistry Department
University of Malta

[2] Department of Physiology and Biochemistry
University of Malta, Msida, Malta
e-mail: marion.zammit-mangion@um.edu.mt

[3] Department of Chemistry,
University of Malta, Msida, Malta
e-mail: claude.farrugia@um.edu.mt

(*) CORRESPONDING AUTHOR:
E-mail: frederick.lia.08@um.edu.mt
Tel.: +365 99018310

Received: August 28, 2020
Accepted: February 8, 2021

The price of extra virgin olive oil, a universally used natural product, depends on its botanical source and its production environment, causing extra virgin olive oil to be vulnerable for adulteration through mislabelling and inappropriate fraudulent production. The application of FT-MIR-ATR spectra in conjunction with several chemometric methods was found to provide a cheap, fast, and reliable way for the discrimination of Maltese EVOOs from non-Maltese EVOOs. Due to the high level of similarity and collinearity, the application of unsupervised PCA models was deemed to be unsatisfactory when it comes to discrimination of geographical origin. Application of supervised methods of classification namely PLS-DA, ANN, LDA and SVM, showed to be highly effective in classifying and discriminating local and non-local EVOOs samples. The use of variable selection methods significantly increased the effectiveness of PLS-DA models when compared to no variable selection. ANN, SVM and LDA models were also shown to offer similar classification rates to PLS-DA models, giving further confidence in the application of FT-MIR.

Keywords: Maltese Olive oil, Geographic discrimination, Multivariate data analysis, Machine learning, Food authenticity, FTIR, PCA, PLS-DA, Support vector machine, Neural networks.

1. INTRODUCTION

Vibrational molecular spectroscopy techniques, including FT-Raman, FT-IR and NIR are emerging analytical techniques which show great potential in the determination of adulterant concentrations of refined seed oils in extra virgin olive oils (EVOOs) [1]. The use of vibrational spectroscopy has also been extended for the determination of fatty acid and triacylglycerols composition [2, 3] as well as providing insights of the overall quality of olive oil including peroxide value [4], acidity [5] and genetic variety [6]. The application of such methods for the determination of geographical and botanical origins has also been studied, providing a cheaper, faster, and more reliable form of authentication [7-11]. To enhance their potential application, these techniques are used in conjunction with chemometric procedures. These methods involve the use of statistical and mathematical methods, designed to select an optimum number of variables to provide maximum chemical information. The traditional, analytical methodology relies on the identification of each compound followed by the quantification of the predefined specific chemical markers. However, this methodology has several disadvantages, as it involves the identification of an adequate number of chemical markers from an overly complex signal. Modern techniques in the field of chemometrics offer another solution to the problem, by attaining a more holistic view, the analysis of the complete set of unidentified and unquantified

markers yields a “fingerprint”. This allows the complete identification of without the prior need to identify and quantify the specific markers as specificity lies within the complexity of the signal obtained.
Both Raman and IR spectroscopy relies on the vibrational excitation of chemical bonds, but they differ in the measure of the nature of the bonds. In the case of IR, it is the change in the molecular dipole moment during vibration that is recorded, whilst in the case of Raman, it is the change in polarizability of the bond which occurs during the vibration that is recorded [10]. The peaks/bands in the IR and Raman spectra at a specific frequency/wavenumber are characteristic of chemical groups that constitute the components in the samples, thus such techniques not only enable quantitative information but also qualitative information giving insights on the different chemical structures and functional groups in the samples. In the case of near-infrared spectroscopy (NIR), several papers in the literature report the use of near [11], however, the use of the mid-infrared (MIR) region (4000-400 cm^{-1}) is preferred as narrower and sharper peaks are obtained due to fundamental vibrations of the molecules resulting in a higher signal/noise ratio and better resolution.
To protect and preserve the authenticity of EVOOs and other traditional foods, the establishment of specific production protection systems came into action in 1992. The European Commission introduced two types of accreditation namely the protected designation of origin (PDO) and protected geographical indication of origin (PGI) (EEC Regulation 2082/92 and later 510/2006). The EEC Regulation 510/2006 regarding labelling, production and commercialisation of olive oil were designed to protect the typical characteristics and authenticate food products, to discourage competition from similar replacement products [12, 13]. The Maltese olive oil industry Is an interesting case, as the industry has only recently been regenerated using indigenous olive stock. Considering the small state of the market, mislabelled EVOO originating from other countries sold as Maltese EVOO could severely impede the growth of the industry, with severe negative economic repercussions. Recent studies have shown that Maltese EVOOs have a significantly different phenolic and mineral composition furthermore spectrofluorimetric and NMR data have shown the possibility to authenticate Maltese EVOO [14-18]. The overall aim of the study was to determine the singularity of the Maltese olive oil, providing an opportunity for local producers to pursue the PDO certification.
This study aims to provide a quick, easy, and cost-saving authentication of the origin of Maltese extra virgin olive oils for the protected designation of origin certification, through the application of mid-infrared spectroscopy and chemometrics. Although the possibility of using mid- and near-infrared spectroscopy to authenticate the origin of extra virgin olive oil samples has been already described in literature [6-11], there are no current studies focusing on the actual EVOOs derived from the Maltese islands, furthermore only a few papers [9]) use the modelling approach to solve this classification problem, and this approach is compared to the discrimination in an even smaller number of papers. In this study, spectroscopic data were analysed both by a discriminant (PLS-DA, LDA) and modelling (SVM, ANN) chemometric tools so that the difference between the methodologies could be assessed. Moreover, the study aims to identify a clear understanding of which signal pre-treatment could be better for authentication purposes using different chemometric methods. Furthermore, the effect of the different spectral transformation on the final classification outcomes was also investigated in this study.

2. MATERIALS AND METHODS

2.1 SAMPLE COLLECTION

For this study, a total of 65 extra virgin olive oil samples were collected from the Maltese islands over 4 harvest seasons from 2013-2016 and from other neighbouring Mediterranean countries. The samples were all taken from different oil producers to cover a representative sample of the Maltese islands in terms of pedological and microclimatic conditions and of manufacturing techniques and the different presses employed. Foreign olive oils obtained were bought having a protected designation of origin to ensure the traceability of the product. All samples were stored at 4°C in the absence of light prior analysis. The samples were preheated to 35°C in a water bath for an hour and mixed to ensure homogeneity.

2.2 FT-MIR-ATR SPECTROSCOPY ACQUISITION

Spectra were acquired on the EVOO in the mid-infrared range at room temperature without any further sample pre-treatment step, through the use of an attenuated total reflection (ATR) cell made of a ZnSe crystal (10 reflections at 45° angles; PerkinElmer Inc., San Jose, CA). Spectra were recorded using a Shimadzu IRAffinity-1 FTIR spectrophotometer controlled by a PC running IRsolution software that accompanies the equipment. 10 µL of oil was deposited on the crystal and using the press tower of the ATR set at the constant height the layer of oil was uninformed throughout the cell. Spectra were then acquired between 4000 and 630 cm^{-1}, collecting 90 scans at a nominal resolution of 2 cm^{-1}. A background spectrum was recorded before each sample analysis. The crystal was cleaned after each analysis, using first hexane followed by chloroform and wiped dry using lens paper tissues. The spectra were exported as an ASCII file using the instrumental software Spectrum (PerkinEl-

mer Inc., Waltham, MA) and imported directly into The Unscrambler X 10.3 (CAMO Software Oslo, Norway) for all subsequent mathematical data processing.

2.3 SPECTRAL PRE-PROCESSING AND PRE-TREATMENT

Different spectroscopic signal processing techniques were evaluated and compared: ATR correction, 5-point smoothing, subtraction of a linear baseline, multiplicative scatter correction (MSC), orthogonal signal correction (OSC), Standard Normal Variate (SNV), Savitzky–Golay, first and second derivative. Furthermore, flat line regions of the signal (3200-4000 cm^{-1}) were eliminated before the statistical analysis. The effect of the different spectral transformations on the final classification outcomes as compared to those obtained without any signal processing. Following a basic ATR-correction, smoothing was the first transformation applied to the FTIR spectra. The smoothing parameters were first determined by trial and error to maximise the smoothness and minimising distortion but, at the same time, retaining enough information from the original signal. It was found that 5-point smoothing reached the optimal noise reduction and retained the maximum information from the spectra. The spectra were normalised, a transformation that put all spectra on the same scale, thus eliminating the fluctuations in intensities between spectra arising from slightly different sample concentrations. Both peak normalisation and area normalisation were carried out separately on the baseline-corrected spectrum. Normalisation was followed by detrending and deresolving procedures. The detrend transformation removes the effects of non-linear trends, showing only the absolute changes in values by removing the least-squares line of best fit from the data, thus focusing only on fluctuations between data. Deresolve is a noise-reducing transformation that operates by artificially lowering the resolution of the spectra. Other treatments applied to the baseline-corrected spectrum include multiplicative and orthogonal scatter corrections (MSC and OSC), and standard normal variate (SNV). MSC corrected for scaling effects by performing a regression of a spectrum against a reference spectrum, thereby correcting the spectrum using the slope of the fit obtained from the regression. OSC removes variance from the factors that is not related to the response, by finding directions in X that describe large variances while being orthogonal to Y and subtracting them from the data. The SNV transformation works similarly to MSC, however, it standardises each spectrum using data from the spectrum itself rather than data averaged from all the spectra. Several derivatising procedures (1st and 2nd derivatives, Savitzky-Golay) were also carried out. The 1st derivative removes baseline effects while the 2nd derivative also removes the slope of the spectrum by measuring the change in slope, thereby sharpening spectral features. The Savitzky-Golay derivative fits a low-degree polynomial to adjacent points in a spectrum, thereby smoothing the spectrum while minimally affecting the signal-to-noise ratio.

2.4 SUPERVISED AND UNSUPERVISED MULTIVARIATE STATISTICAL TECHNIQUES

A principal component analysis (PCA) was carried out using Unscrambler X 10.3 to identify any gross outliers and determine any preliminary clustering reflecting the geographical origin. An inspection of the PCA loadings was carried out to determine whether the loadings had a spectral shape indicating that observed variation was due to the FTIR spectra and not due to noise. PCA was carried out on all treated spectra to reduce all the spectral information down to 7 principal components (PCs) that retain the information of the original dataset. The first PC accounted for most of the variation in the dataset, with successive principal components accounting for decreasing amounts of the variation. The resulting PC-1 vs. PC-2 plots could be examined for any clustering that might arise from each spectral pre-treatment. Similarly, to PLS, PCA generates loading plots that indicate those x-values that are most responsible for the variability between the different spectra. The loading plots for the first two principal components (which explain most of the variability in the dataset) were used to determine which wavelength values have the largest influence on the separation of PC-1 and PC-2.

The main aim of this study is to develop methods that can predict whether an unknown olive oil sample comes from Maltese islands or not. This was done through the application of multivariate pattern recognition statistical method on the FTIR. The whole dataset consisted of two sets: the training and test sets (the former to build the model, the latter to validate it). To preserve the diversity in the training and test sets and to account for the fact that different pre-treatments had to be tested, a unique sample splitting scheme was used. The following method was adopted to cover such variation in the two sets and at the same time be able to compare the outcomes after the different pre-treatments. The Maltese and the non-Maltese samples were grouped in an ascending way so that the first 30 samples would represent Maltese EVOO's whilst the rest correspond to non–Maltese EVOO's. A Venetian blind cross-validation was applied, which selects every 5th sample from the data by making data splits such that all samples are left out exactly once (s = 5). This sampling method excluded 20% of the observation so that they would be retained as the testing set. The remaining 80% of the observation was used to build the training set.

In the case of PLS-DA, an inspection of the VIP scores was carried out. VIP is an index of how much a single variable contributes to the bilinear model. VIP larger

than 0.8 is significantly contributing to discrimination. Variables having a smaller VIP than 0.8, are an ideal candidate for deletion from the model [19]. An adjusted PLS model was repeated after removing these variables and the good the suitability for the adjusted model was evaluated. In SLC-DA a manual selection of chemical shifts was carried out based on the largest F-ratios and smallest p-values. The variables with a p-value smaller than 0.05 and with the highest F-ratio, as 0 obtained through a stepwise forward and backwards method, were selected since these represent the highest correlation with the response. The selected variables obtained in SLC-DA were arranged in ascending order in terms of their scoring coefficients. A smaller set of variables were selected which consisted of 20 variables which corresponded to 10 variables having most positive standardised scoring coefficients and 10 most negative standardised scoring coefficients. These variables were then subjected to a Fisher LDA.

2.5 SUPPORT VECTOR MACHINE AND ARTIFICIAL NEURAL NETWORKS

In the case of SVM, the models were built using different cross-validation forms. In training set which constituted 80% of the samples was used to build the SVM models. These models were segment validated whereby the training data set was first partitioned into 10 equally (or nearly equally) sized segments. Subsequently, 10 iterations of training and validation were performed so that, within each iteration, a different segment of the data was held out for validation while the remaining 9 folds used for learning. Data matrix was stratified before being split into segments to ensure each segment is a good representative of the whole data set. Once the model was fitted the % accuracy in the training and validation was derived. In this experiment, a second cross-validation method was employed whereby the models fitted on the 80% of the training set were tested on the remaining 20% of the data so that % predictability for each model is obtained. Although SVM models have no limit on the number of variables that can be used, it requires a computationally intensive grid search to optimise the C parameter and thus analysis was performed on SLC-DA selected variables. The C parameter is a trade-off parameter between complexity and risk. If the C parameter is exceptionally low, then the errors produced during the training stage become less important, thus risking the model to become overfitted.

To determine the suitability of the whole FTIR spectra to discriminate EVOOs of Maltese origin, an artificial neural network analysis was carried out. The main advantage of a neural network model is that it can efficiently model different response surfaces due to its nonlinearity, allowing a better fit to the data given enough hidden nodes and layer, providing an accurate prediction for kind of data. Unlike other modelling and discriminate methods (LDA, PLS) the main disadvantage of a neural network model is that the results are not easily interpretable, due to presence of intermediate hidden layers that direct path from the X variables to the Y variables, as in the case of regular regression but cannot be fully interpreted. In this experiment, 25 iterations were carried out using a TanH activation function as the standard neuron activation function in JMP software. TanH function transforms values to be between -1 and 1, acting like the centred and scaled version of the logistic function. In the case of ANN three different cross-validation techniques were employed to prevent model overfitting; the k-fold (CV-10), hold back (33.3%) and excluded rows (Venetian blinds).

2.6 ASSESSMENT OF MODEL PERFORMANCE

2.6.1 Goodness-of-fit

The goodness-of-fit of PLS was assessed from the % of variability explained in X and Y, and the sensitivity of the model on a specific number of latent variables was assessed by Predictive Residual Sum of Squares (PRESS). The Van der Voet T^2 and the Prob > van der Voet T^2 were also taken under consideration in which the null hypothesis of the test states that the squared residuals for both models have the same distribution, hence the same predictive ability.

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \widehat{y_{i,-i}})^2$$

Where for each observation i^{th} observation a regression model is fitted to the remaining observations N-1. The is model is used to predict y_i denoted by $\widehat{y_{i,-i}}$. The residual is defined by $(y_i - \widehat{y_{i,-i}})$, which is calculated for all the remaining observations.

2.6.2 Evaluation of model performance

The accuracy of training, testing and prediction by PLS-DA was determined as the numerical coordinate systems were rounded up to the nearest integer of either zero or one. A negative value was assigned a value of zero, whereas a value greater than one was assigned a value of one. The numerical output obtained was compared to the previously assigned value based on the known origin. In the case of LDA, the output obtained classified the sample as either zero or one, so there was no need for any manipulation. The % accuracy was determined by the following equation:

$$\% \ accuracy = 100 - \left(\left(\frac{K_m}{K_t}\right) \times 100\right)$$

Where K_m is the number of samples that are misclassified, K_t is the total number of samples used in training, testing or prediction model.

3. RESULTS AND DISCUSSION

Figure 1 shows a typical EVOO NIR spectrum, the maxima obtained at an absorbance of 3006 cm^{-1} is attributed to the stretching vibration of (=C–H) of oleic acid acyl groups and linoleic and linolenic acyl groups. The strong band absorptions observed in the region of 3000-2800 cm^{-1} corresponds to the (-C–H) stretching vibrations of methylene (–CH_2–) and methyl (–CH_3) groups observed at frequencies of 2922 and 2853 cm^{-1}, respectively. The bending vibrations of the methylene and methyl groups are observed at 1465 cm^{-1} and 1377 cm^{-1} that correspond to C=H scissors deformation vibration and bending vibration of CH_2 groups, respectively. The sharp intense peak around 1740 cm^{-1} is attributed to the presence of carbonyl groups that corresponds to the (-C=O) double bond stretching vibration of the ester carbonyl functional group of the triglycerides. The medium intensity peaks observed at 1160.74 cm^{-1} and 1236.86 cm^{-1} is assigned to the vibration of the C-O ester groups and CH_2 groups, respectively. Small intensity peaks observed at 1117 cm^{-1} are associated with the stretching vibration of the C-O ester group. The low-intensity peak observed at 722 cm^{-1} corresponds to the cis-CH=CH– bending out of plane [20-24] (Fig. 1).

Different kinds of spectral pre-treatments were tested and compared to overcome the instrumental limitation and account for scattering and other minor variations that would hinder the performance of the classification models. A total of 12 spectral pre-treatment methods were used, in each case, after pre-treatment, a principal component analysis was carried to dimensionally reduce the number of variables into a small set of principal components whilst retaining the information of the larger set. PCA enabled the preliminary identification of which pre-treatment method offered the highest variability and possible sample grouping based on the geographical origin. Figure 2 shows some of the different forms of spectral pre-treatments employed and the corresponding PCA plot for the first two principal components. From the % variability explained it was found that 5 points smoothing enhanced the variability explained by the 1st principal component when compared to the ATR correction. This was attributed to the improved signal to noise ratio especially in the 1550-1650 cm^{-1} region, thus this observation justified the use of smoothing before the other spectral pre-treatment methods. Whilst the other spectral pre-treatment methods displayed an improvement in the variability explained after smoothing, QN, MSC, SNV showed a lower % variability when compared to the basic ATR correction and smoothing. In the case of QN, this was expected as, unlike the other pre-treatment methods, it aims to achieve the same distribution of intensities of all spectra, making it not particularly useful when dealing with spectra of a continuous nature. None of the spectral methods managed to show any form of distinctive sample grouping reflecting the geographical origin of EVOOs (Fig. 2).

3.1. APPLICATION OF PLS-DA FOR THE DISCRIMINATION OF MALTESE EVOOs

After splitting the data according to the procedure described above, chemometric classification models were built and tested on all the MIR spectral pre-treatment using a PLS regression algorithm using JMP 10 and its inbuilt leave one out cross-validation method (LOOCV). Table I shows the number of latent variables extracted, the predicted root mean square error, for the different spectra pre-treatment methods. The % accuracy (correct classification), showed that, except for normalisation, all the other spectral treatments had the same effectiveness in correctly classifying the geographical origin of EVOOs. From the results obtained it was observed that the Savitzky–Golay, 1st derivative, MSC, OSC, and detrending, showed lower press

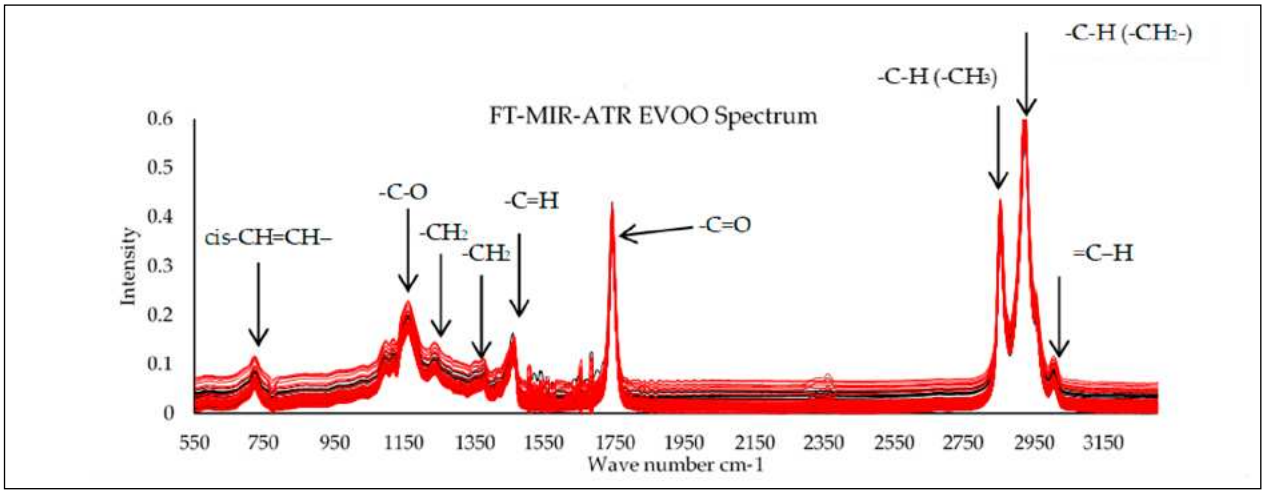

Figure 1 - The major peaks of interest obtained using FT-MIR-ATR for Maltese EVOOs (Black) and non-Maltese EVOOs (Red).

and a higher % accuracy and % predictability. On the other hand, normalisation seems to negatively affect the amount of variability of in fixed data set. To fully interpret the PLS models obtained, an inspection of the VIP scores was used to determine which predictors (variables) are mainly influencing the latent vectors obtained. VIP is an index of how much a single variable contributes to the bilinear model and it is scaled in such a way that indices having VIP larger than 0.8 is deemed as significantly contributing to discrimination (Tab.I).

As shown in Figure 3, the VIP > 0.8 identified relevant features in the spectra, particularly, stretching vibration of (=C–H) of acyl groups 3006 cm^{-1}, (-C=O) double bond stretching (around 1700 cm^{-1}) and C-H bending in the fingerprint region (650–750 cm^{-1}) appear to be the regions contributing the most to the bilinear model. Additionally, C-H stretching (2800–3100 cm^{-1}) and C-O single bond stretching (1100 cm^{-1}) also show a VIP value significantly larger than 0.8. The next step was to build another PLS model this time using only variables with a VIP score > 0.8. Table I shows the results obtained by using the adjusted PLS model. Comparing the models obtained using variable selection to the one previously obtained without any variable selection, no noticeable difference was observed in % accuracy and predictability of the model. Nonetheless, a lower PRESS was observed for most of the pre-treatments. This observation suggests that, in the case of ATR-FTIR data, no need for extensive variable selection is required to obtain an exceptionally good classification method with the use of PLS models (Fig. 3).

3.2. APPLICATION OF SLC-DA AND LDA METHODS FOR DISCRIMINATING MALTESE EVOOs

To obtain a more robust method of classification with the use of a smaller number of variables, the VIP data

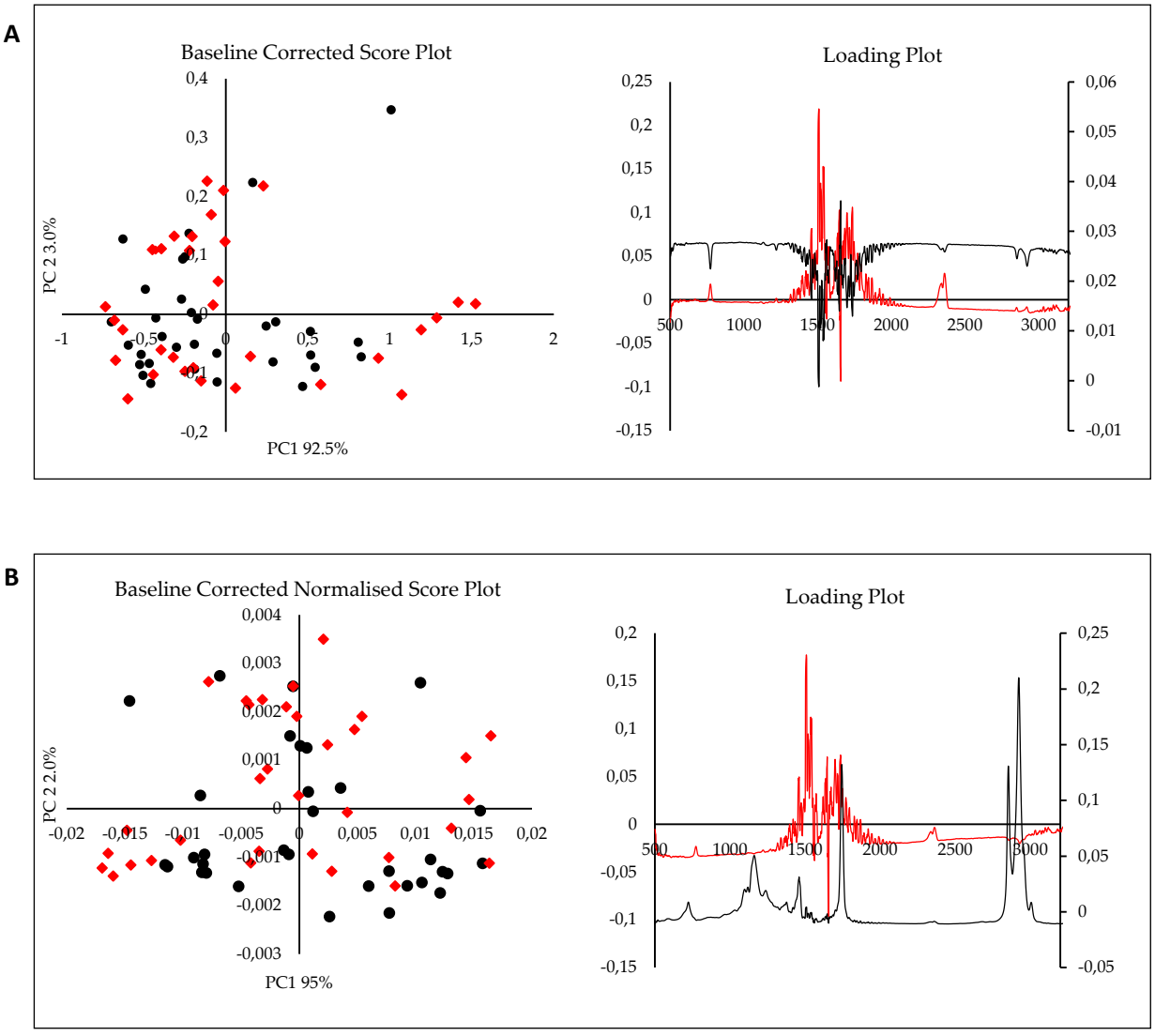

Figure 2 - PCA score plots (black dots represent the Maltese EVOOs whilst red diamonds represent the non-Maltese EVOOs) and loading plots for PC1 (black line) and PC2 (red line) for the baseline-corrected spectra (A) and baseline corrected spectra after normalisation (B).

set obtained from the previous PLS-DA analysis was subjected to a stepwise linear canonical discriminate analysis SLC-DA. SLC-DA was performed on the MIR data from all the pre-treatment methods to extract only a small amount of highly discriminate variables which would enable easier and faster discrimination between the origins of EVOOs. This strategy involved a substantial reduction of the dimensionality of the data. To further reduce the number of variables selected from the SLC-DA analysis, a minimum of 14 variables was selected to carry out a conventional LDA. During the SLC-DA the variables chosen by applying a forward stepwise variable selection algorithm using JMP 10 using a Wilks' Lambda as a selection criterion and an F-statistic factor to determine the significance of the changes in Lambda when the influence of a new variable is evaluated. The most significant variables were then extracted, and their canonical scoring coefficients were plotted as shown in Figure 4. The main advantage of using SLC-DA over the convention LDA is the ability to perform a feature selection. Only those variables which helped to improve classification performance were used whereas variables without discriminant information were discarded. Furthermore, LDA is greatly affected by the normality distribution of predictor variables and their homogeneity. In this experiment, most of the selected variables for LDA were found to be non-normally distributed and thus a Fisher LDA rather than a Bayesian LDA model was built. Furthermore, the application of Fisher LDA model avoided the need for homogeneity since the normality assumption overrides the need for homogeneity.

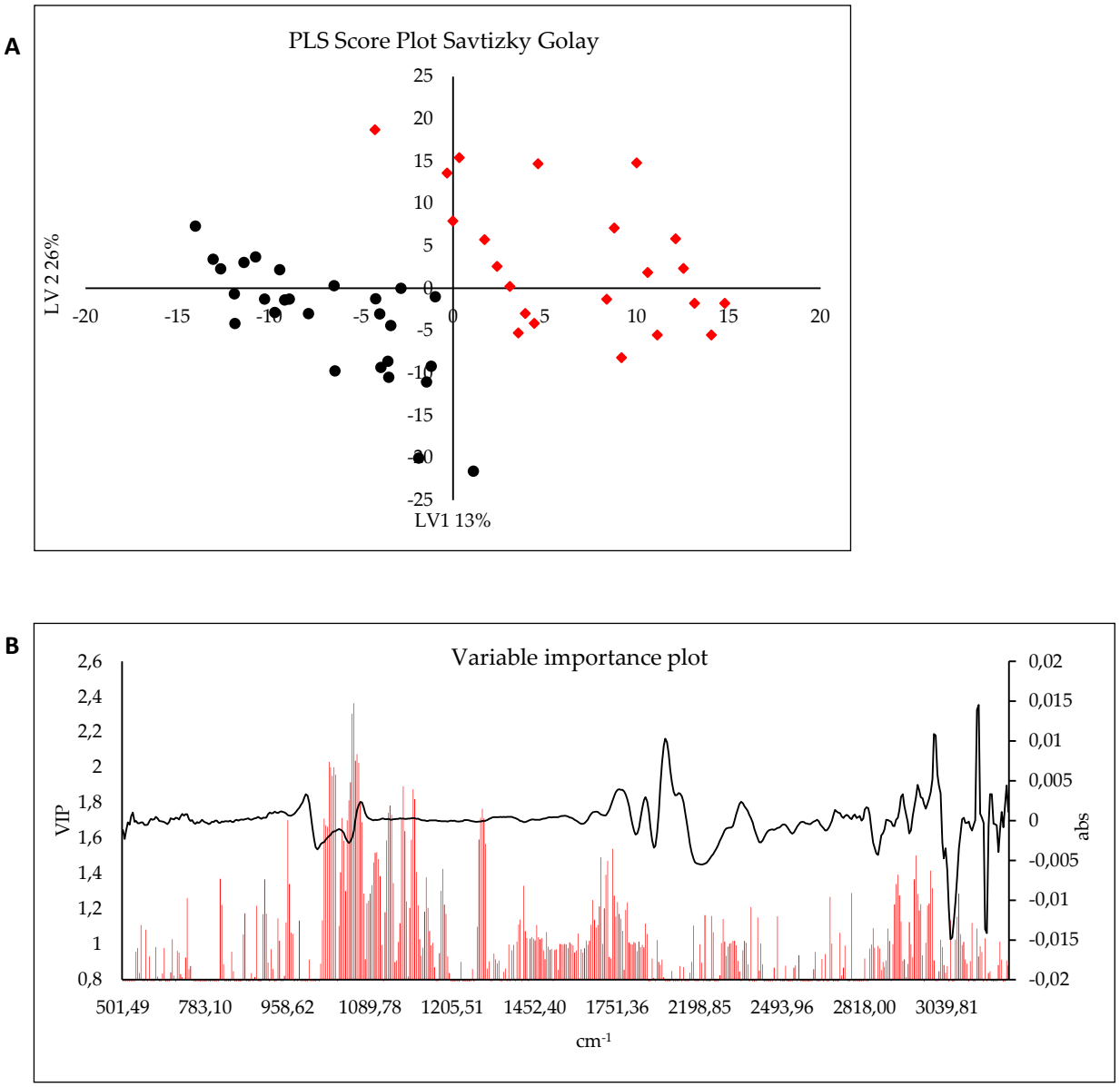

Figure 3 - (A) PLS score plots (black dots represent the Maltese EVOOs whilst red diamonds represent the non-Maltese EVOOs) for Savitzky–Golay derivatised spectra. (B) superimposed Savitzky–Golay spectrum (black line) and variables having a VIP score > 0.8.

Table I - Results from PLS-DA models applied to spectral transformations of ATR-FT-MIR spectra and variable selection procedures. (MSC = Multiplicative Scatter Correction, OSC = Orthogonal Signal Correction, Q Norm= Quantile Normalise, SNV = Standard Normal Variate, SGD = Savitzky–Golay derivatised spectra, 1st = first deriviative specta, 2nd = second derivistised spectra)

	Whole Spectrum				VIP > 0.8			
Pretreatment	Latent Variables	Press	% Accuarcy	% Predicatablity	Latent Variables	PRESS	% Accuarcy	% Predicatablity
Raw	5	0.79	92.65	100.00	5	0.85	97.06	100.00
Smoothing	15	0.73	91.18	100.00	15	0.72	91.18	100.00
Baseline	8	0.75	89.71	84.62	1	1.01	88.25	84.62
Norm	1	1.02	50.00	61.54	5	0.79	50.00	61.54
QNorm	5	0.88	85.29	92.31	10	0.76	86.76	92.31
Detrend	15	0.49	94.12	100.00	14	0.46	94.12	100.00
Deresolve	15	0.71	92.65	100.00	15	0.72	92.65	100.00
SNV	9	0.68	91.18	100.00	13	0.66	91.18	100.00
MSC	15	0.65	91.12	100.00	13	0.66	92.65	100.00
OSC	15	0.56	92.65	100.00	15	0.19	100.00	100.00
SGD	9	0.65	97.06	100.00	5	0.43	98.56	100.00
1st Der.	9	0.65	97.06	100.00	5	0.43	97.06	100.00
2nd Der.	13	0.80	92.65	100.00	8	0.53	98.53	100.00

Table II - Summary of LDA and SVM Model performance

Pre-treatment	LDA		SVM			
			Kernel Type Linear		Kernel Type Radial	
	% Accuracy	% Predictability	% Accuracy	% Predictability	% Accuracy	% Predictability
Raw	97.92	95.00	100.00	70.00	100.00	80.00
Smoothing	87.50	70.00	100.00	100.00	89.58	80.00
Normalized	100.00	95.00	97.92	90.00	95.83	75.00
Q Norm	81.25	65.00	91.67	75.00	97.92	70.00
Baseline	100.00	85.00	100.00	60.00	100.00	65.00
Detrend	91.67	85.00	100.00	70.00	93.75	70.00
Deresolve	100.00	95.00	97.92	65.00	91.67	100.00
SNV	97.92	85.00	100.00	95.00	100.00	60.00
MSC	100.00	95.00	100.00	95.00	95.83	90.00
OSC	83.33	30.00	100.00	80.00	79.17	60.00
SGD	100.00	100.00	100.00	100.00	100.00	95.00
1st	100.00	100.00	100.00	100.00	100.00	95.00
2nd	100.00	100.00	16.67	35.00	100.00	95.00

Figure 4 shows that for the variables selected during SLC-DA for 1st derivatised spectra, it was shown that the selected variables were mainly concentrated in the 1100-500 cm^{-1} range which corresponds to the C-H bending in the fingerprint region, whilst shouldering peaks next to the stretching vibration of (=C–H) of acyl groups 3006 cm^{-1}, (-C=O) double bond stretching (around 1700 cm^{-1}) also appear to be the regions contributing the bilinear model. Table IV shows the results obtained during the training and testing the phase of the LDA models obtained for all the pre-treatments ranged from 81-100% accuracy. The classification model obtained was then tested and except for OSC and quantile normalisation, the validation accuracy ranged from 70-100%. From the results obtained it was shown that 1st, 2nd and Savitzky–Golay derivatisation of the spectra had the highest % accuracy in both the training and validation dataset (100% for both). Figure 4 shows the biplot obtained for the 1st derivatised spectral pre-treatments, it was clear that there was complete discrimination between the EVOOs of Maltese and non-Maltese origin (Fig. 4).

3.3. APPLICATION OF SUPPORT VECTOR MACHINE CLASSIFICATION FOR DISCRIMINATING MALTESE EVOOs

Results obtained from SVM classification are present in Table II high rates of accuracy and predictability were obtained for the majority of the spectral pre-treatments further validating that SVM classification is highly adaptable to the kind of data used. In the case of linear SVM, the best classification was obtained

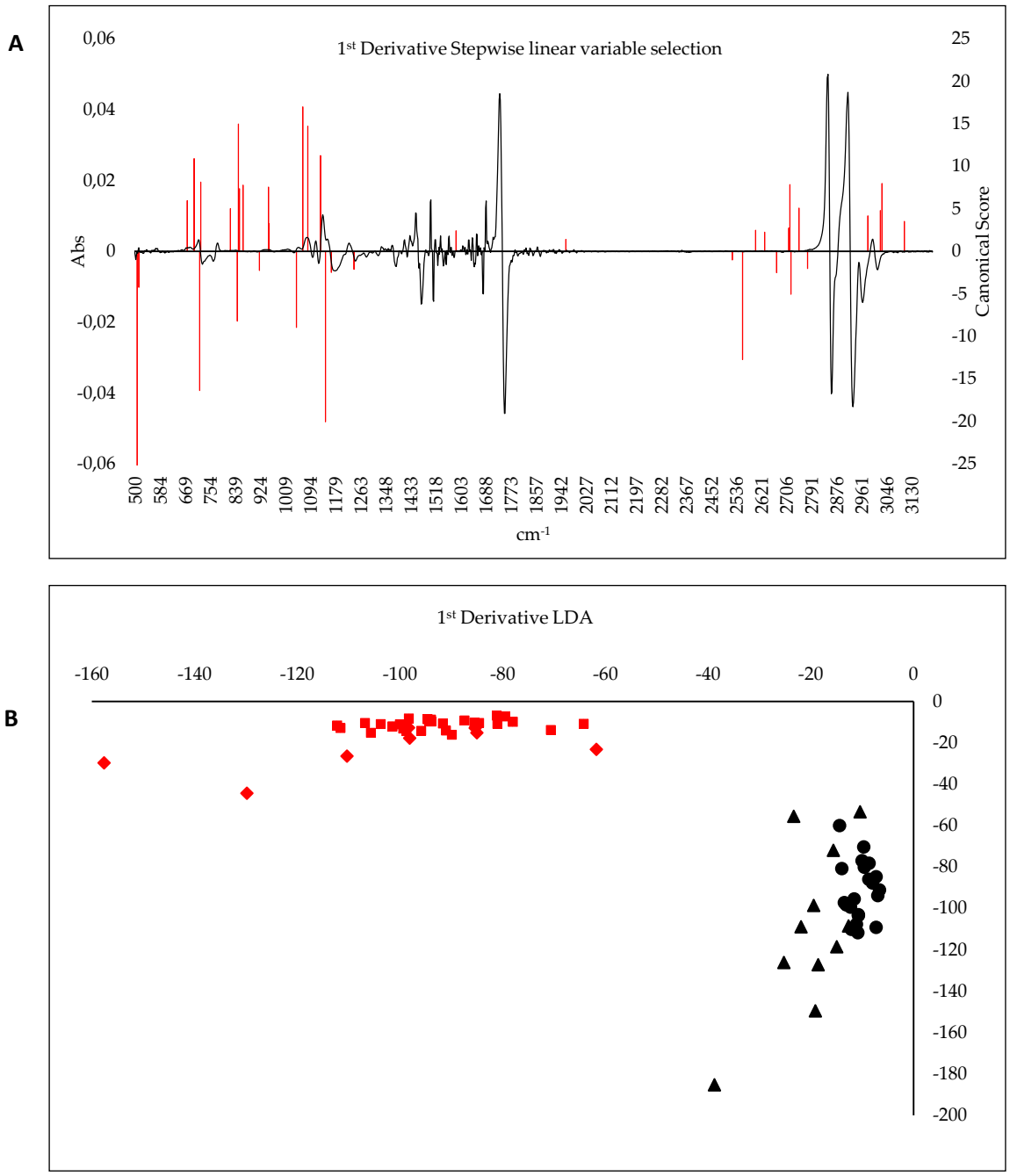

Figure 4 - (A) 1st order derivatise spectrum showing variables selected during stepwise linear discriminate analysis and the canonical scores obtained for the selected variables (B) LDA plot using the most discriminated variables showing Maltese EVOOs in black and non-Maltese EVOOs in red. (▲) Maltese EVOOS samples used in the training set (●) Maltese EVOOs samples used in the validation set (■) Foreign EVOO samples used in the training set (◆) Foreign EVOOs samples used in the validation set.

by using 5-point smoothing, 1st and Savitzky–Golay derivatisation techniques as 100% accuracy and predictability were obtained. Unlike what was observed in PLS-DA, and LDA, spectra pre-treated with 2nd order derivatisation had the lowest % accuracy in the training set (16.17%) and % predictability (35%) when compared to the rest of the spectral pre-treatments under the linear type SVM. Unlike the rest of the other spectral pre-treatments, the 2nd order derivatisation showed an increase in accuracy and predictability on the use of a radial type Kernel function. The 2nd order derivatisation reaches a 100% accuracy and a 95% predictability on the use of a radial type Kernel function, suggesting that the group projected in the higher dimensional space cannot be separated using a linear hyperplane but through a spherical hyperplane formed from the use of the radial Kernel type function (Tab II).

Table III - Summary of ANN Model performance with no variable selection on the 'whole' spectrum using different cross-validation methods

CV Type	Hold back		K-fold		Excluded Row	
Pre-treatment	% Accuracy	% Predictability	% Accuracy	% Predictability	% Accuracy	% Predictability
Raw	100.00	100.00	85.51	92.31	85.51	92.31
Smoothing	97.10	92.31	97.10	92.31	94.20	76.92
Normalized	81.16	76.92	91.30	92.31	53.62	61.54
Q Norm	92.75	92.31	100.00	100.00	91.30	92.31
Baseline	63.77	69.23	98.55	100.00	68.12	69.23
Detrend	100.00	100.00	98.55	92.31	98.55	92.31
Deresolve	97.10	97.10	94.20	97.10	97.10	94.20
SNV	94.20	100.00	100.00	100.00	82.61	84.62
MSC	98.55	100.00	100.00	100.00	84.06	76.92
OSC	92.75	61.54	92.75	61.54	95.65	92.31
SGD	95.65	92.31	100.00	100.00	91.30	100.00
1st	97.10	100.00	98.55	92.31	98.55	92.31
2nd	92.75	92.31	100.00	100.00	97.10	84.62

3.4. WHOLE FTIR MODELLING USING FEED-FORWARD PREDICTIVE ARTIFICIAL NEURAL NETWORKS

Artificial neural network (ANN) is a mathematical algorithm with the capability of relating large amounts of the input and output parameters. The main advantages of ANN are their nonlinearity, allowing the better fit to the data; noise insensitivity, providing accurate predictions. For these reasons, the application of AANs was computed on either FTIR spectrum without any form of variable selection methods. The performance of ANNs was compared to the PLS-DA models obtained using either spectrum as shown in Table I. In the case of ANN using three different forms of cross-validation, namely 33.3% of data holdback, CV-10 k-fold and excluded row validation were employed. In comparison to the PLS-DA models, ANNs had a lower performance especially when it comes to the testing phase as shown in Table III. The lower % precision recorded in the ANN is coherent to several other studies which showed that PLS-DA has a higher sensitivity and performance [23, 24]. ANNs work better if they deal with non-linear dependence between input and output vectors and generally, are more efficient in modelling classes separated with non-linear boundaries, however from the experimental data we have shown that FTIR data for the different EVOO origins attain a more linearly discrimination as shown by SVM, and LDA results. Nonetheless, ANN can provide a substantially good corroboration of PLS-DA without the excessive need for variable selection (Tab. III).

4. CONCLUSION

In conclusion, it has been shown that FT-MIR-ATR spectra, in conjunction with several chemometric methods, provided a cheap, fast, and reliable way for the determination of the geographical origin of EVOOs, especially when it comes to discrimination of Maltese EVOOs from non-Maltese EVOOs. From the preliminary assessment using only unsupervised PCA models, no significant clustering was observed. This was attributed to the high levels of similarity between the two classes of EVOOs studied, such method was deemed to be unsatisfactory when it comes to discrimination of geographical origin. Application of supervised methods of classification namely PLS-DA, ANN, LDA and SVM showed to be highly effective in discriminating Maltese EVOOs. The use of variable selection methods significantly increased the effectiveness of PLS-DA models when compared to those in which either spectrum was used. ANN, SVM and LDA models were also shown to offer similar classification rates to PLS-DA models and thus corroborate the results obtained from the PLS-DA models and put confidence in the use of FT-MIR-ATR methods in conjunction with spectral transformation for the classification of Maltese and foreign EVOOs samples. It was further highlighted that the use of Savitsky Golay 1st and 2nd derived FTIR spectra greatly improve the potential application of multivariate analysis in discriminating and predicting Maltese EVOOs.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research was funded by the Malta Government Scholarships Post-Graduate Scheme for 2014 (MGSS-PG 2014).

REFERENCES

[1] A. Rohman, Y.B. Che Man, Quantification and Classification of Corn and Sunflower Oils as Adulterants in Olive Oil Using Chemometrics and FTIR Spectra. The Scientific World Journal, 1-6, (2012).

[2] O. Galtier, O. Abbas, Y.Le Dréau, C. Rebufa, J. Kister, J. Artaud, N. Dupuy, 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, (2011)

[3] A.M. Inarejos-Carcía, A. Gómez-Alonso, G. Fregapane, M.D. Salvador, Evaluation of minor components, sensory characteristics and quality of virgin olive oil by near infrared (NIR) spectroscopy. Food Res. 50, 250-258, (2013).

[4] A. Bendini, L. Cerretani, F. Di Virgilio, P. Belloni, M. Bonoli-Carbognin, G.Lercker, Preliminary evaluation of the application of the FTIR spectroscopy to control the geographic origin and quality of virgin olive oils. J. of Food Quality 30, 424-437, (2007).

[5] F.A. Iñón, J.M. Garrigues, S.Garrigues, A.Molina, M. de la Guardia, Selection of calibration set samples in determination of olive oil acidity by partial least squares-attenuated total reflectance-Fourier transform infrared spectroscopy. Anal. Chim. Acta 489, 59-75, (2003).

[6] V. Concha-Herrera, M.J.Lerma-García, J.M. Herrero-Martínez, E.F. Simó-Alfonso, Prediction of the genetic variety of extra virgin olive oils produced at La Comunitat Valenciana, Spain, by fourier transform infrared spectroscopy. J. Agric. Food Chem 57(21), 9985-9989, (2009).

[7] H.S.Tapp, M. Defernez, E.K. Kemsley, FTIR Spectroscopy and Multivariate Analysis Can Distinguish the Geographic Origin of Extra Virgin Olive Oils. J. Agric. Food Chem. 51(21), 6110-6115, (2003).

[8] M. De Luca, W. Terouzi, G. Ioele, F. Kzaiber, A. Oussama, F. Oliverio, G. Ragno, Derivative FTIR spectroscopy for cluster analysis and classification of morocco olive oils. Food Chem. 124(3), 1113-1118, (2011).

[9] M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: A case study. Anal. Chim. Acta 717, 39-51, (2012).

[10] F.P. Capote, J.R. Jim´enez, M.D.L. de Castro Sequential, (step-by-step) detection, identification and quantitation of extra virgin olive oil adulteration by chemometric treatment of chromatographic profiles. Anal. Bioanal. Chem. 388(8), 1859-1865, (2007).

[11] N. Sinelli, M. Casale, V. Di Egidio, P.Oliveri, D. Bassi, D. Tura, Varietal discrimination of extra virgin olive oils by near and mid infrared spectroscopy. Int. Food Res. J. 43, 2126-2131, (2010).

[12] Council Regulation (EEC) No 2082/92 of 14th July 1992 on certificates of specific character for agricultural products and foodstuffs.

[13] Council Regulation (EC) No 510/2006 of 20th March 2006 on the protection of geographical indications and designations of origin for agricultural products and foodstuffs

[14] F. Lia, A. Morote Castellano, M. Zammit-Mangion, C. Farrugia, Application of fluorescence spectroscopy and chemometric models for the detection of vegetable oil adulterants in Maltese virgin olive oils. Int. J. Food Sci. Technol. 55(6), 2143-2151, (2018).

[15] F. Lia, M. Zammit-Mangion, C. Farrugia, A first description of the phenolic profile of EVOOs from the Maltese islands using SPE and HPLC: Pedo-climatic conditions modulate genetic factors. Agriculture 9(5), (2019)

[16] F. Lia, M.Z. Mangion, C. Farrugia, Application of elemental analysis via energy dispersive x-ray fluorescence (ED-XRF) for the authentication of maltese extra virgin olive oil. Agriculture 10(3), (2020).

[17] F. Lia, J.P. Formosa, M. Zammit-Mangion, C. Farrugia, The first identification of the uniqueness and authentication of Maltese extra virgin olive oil using 3D-fluorescence spectroscopy coupled with multi-way data analysis. Foods 9(4), (2020)

[18] F. Lia, B. Vella, M.Z. Mangion, C. Farrugia, Application of 1H and 13C NMR Fingerprinting as a Tool for the Authentication of Maltese Extra Virgin Olive Oil, Foods 9(6), 689 (2020)

[19] S. Wold, PLS for Multivariate Linear Modeling, QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry (1994).

[20] H. Yang, J. Irudayaraj, Comparison of Near-Infrared, Fourier Transform-Infrared,and Fourier Transform-Raman Methods for Determining olive Pomace Oil Adulteration in Extra Virgin Olive Oil, JOAC, 78(9), 889-895, (2001).

[21] M.D. Guillen, N. Cabo, Usefulness of the Frequencies of Some Fourier Transform Infrared Spectroscopic Bands for Evaluating the Composition of Edible Oil Mixtures, Fett-Lipid 101, 71-76, (1999)

[22] M.J. Lerma-García, E.F. Simo-Alfonso, A. Bendini, L. Cerretani, Rapid evaluation of oxidised fatty acid concentration in virgin olive oil using

Fourier-transform infrared spectroscopy and multiple linear regression. Food Chem. 124, 679-684, (2011)

[23] M. Khanmohammadi, N. Dallali, A. Bagheri Garmarudi, M. Zarnegar, K. Ghasemi, Comparison of partial least squares and artificial neural network chemometric techniques in determination of sulfamethoxazole and trimethoprim in pharmaceutical suspension by ATR-FTIR spectrometry. Spectroscopy 26(2), 105-114, (2011)

[24] E. Z.Panagou, F.R. Mohareb, A.A. Argyri, C.M. Bessant, G.J.E. Nychas, A comparison of artificial neural networks and partial least squares modelling for the rapid detection of the microbial spoilage of beef fillets based on Fourier transform infrared spectral fingerprints. Food Microbiology 28(4), 782-790, (2011)