SORT 42 (2) July-December 2018, 209-236 DOI: 10.2436/20.8080.02.75

Effect of agro-climatic conditions on near infrared spectra of extra virgin olive oils*

M. I. Sánchez-Rodríguez*,[1], E. M. Sánchez-López[2], J. M. Caridad[1], A. Marinas[2] and F. J. Urbano[2]

Abstract

Authentication of extra virgin olive oil requires fast and cost-effective analytical procedures, such as near infrared spectroscopy. Multivariate analysis and chemometrics have been successfully applied in several papers to gather qualitative and quantitative information of extra virgin olive oils from near infrared spectra. Moreover, there are many examples in the literature analysing the effect of agro-climatic conditions on food content, in general, and in olive oil components, in particular. But the majority of these studies considered a factor, a non-numerical variable, containing this meteorological information. The present work uses all the agro-climatic data with the aim of highlighting the linear relationships between them and the near infrared spectra. The study begins with a graphical motivation, continues with a bivariate analysis and, finally, applies redundancy analysis to extend and confirm the previous conclusions.

MSC: 62H20, 62Pxx, 82-08, 62-09

Keywords: Extra virgin olive oil, infrared spectroscopy, agro-climatic data, linear correlations, redundancy analysis

1. Introduction

Spain is the first worldwide producer of extra virgin olive oil (EVOO), where Andalusia encompasses 80% of the national production. EVOO is an edible oil very much appreciated by its flavour and benefits for health. Its high quality could be affected by frauds in marketing, such as adulteration with other cheaper oils (for example, palm, corn, hazelnut or refined olive oil) or with the indication of a false geographical origin. These practices considerably modify its quality indexes. Therefore, authentication of EVOO requires fast, reliable and cost-effective analytical procedures which require no or little sample manipulation, such as near infrared spectroscopy (NIR). Contrary to classical separation techniques (for example, gas chromatography), NIR spectra provide

* *Corresponding author:* td1sarom@uco.es. Avda. Puerta Nueva, s/n. 14071. Córdoba

[1] Department of Statistics and Business. University of Cordova.

[2] Department of Organic Chemistry. University of Cordova.

Received: March 2018
Accepted: October 2018

continuous information rich in both isolated and overlapping bands and their analysis requires the application of multivariate statistics (see Öztürk, Yalçin and Özdemir, 2010).

There are in the literature many examples of the application of chemometrics to determine qualitative and quantitative information of EVOO from NIR spectra, specially, with the aim of its authentication. For instance, Bertran et al. (2000) apply NIR and pattern recognition as screening methods for the authentication of EVOO of very close geographical origins. Mailer (2004) shows a rapid evaluation of olive oil quality by NIR reflectance spectroscopy. Galtier et al. (2007) determine geographic origins and compositions of EVOO by chemometric analysis of NIR spectra. Woodcock, Downey and O'Donnell (2008) show a confirmation of declared provenance of European EVOO samples by NIR spectroscopy. Casale et al. (2012) present a characterization of Protected Designations of Origin (PDO) olive oil Chianti Classico by non-selective (UV-visible, NIR and MIR (mid-infrared) spectroscopy) and selective (fatty acid composition) analytical techniques. Finally, some previous papers of our research group (see Sánchez-Rodríguez et al. (2013) and Sánchez-Rodríguez et al. (2014)) show new chemometric approaches to empathize the potential of NIR and MIR spectra to determine the fatty acid profile of EVOO, the fatty acids being its major components and considered as a quality parameter in order to its authentication. Therefore, NIR and MIR spectra contain valuable and diverse information about EVOO.

Moreover, there are in the literature many works analysing the influence of weather, agro-climatic or meteorological[1] conditions on food content, in general, or in EVOO components, in particular. Thus, for example, Martínez-Herrera et al. (2006) analyse the chemical composition of *Jatropha curcas L.*, a multipurpose shrub of significant economic importance because of its several potential industrial and medicinal uses, from different agro-climatic regions of Mexico. Jarvis et al. (2008) and Khokhar et al. (2017) study the influence of agro-climatic conditions on wheat in western Canada and India, respectively. Zheng et al. (2012) show the effects of latitude and weather conditions on the contents of black currant, while Yang et al. (2017) analyse the same effects on Finnish berries. Falasca, Ulberich and Ulberich (2012) develop an agro-climatic zoning model to determine potential production areas for castor bean. Luciano et al. (2013) treat the effects of the weather and the soil on the composition of grapes. Rymbai et al. (2014) study the physiological characteristics of mango in different agro-climatic regions of India. Edmunds et al. (2015) analyse the relationships of preharvest weather conditions and soil factors to susceptibility of sweetpotato. Dorey et al. (2016) model sugar content of pineapple under agro-climatic conditions on Reunion Island. Finally, there are many papers treating the effect of weather and agro-climatic conditions on oils (such as Leskinen, Suomela and Kallio, 2009a and Leskinen et al., 2009b), especially the numerous studies of olive oils: for example, Sacco et al. (2000), D'Imperio et al.

1. Climatology deals with the scientific study of climate, that is, the processes and phenomena of the atmosphere over relatively long periods of time. However, Meteorology studies the characteristics of the atmosphere over a short period of time, especially as a means of forecasting the weather. The agro-prefix placed before both terms refers to the interrelationship between Climatology and Meteorology with the processes of agricultural production.

(2007), Cornejo, Bueno and Gines (2012), Awan (2014), Alowaiesh, Singh and Kailis (2016), Ozdemir (2016), Veizi, Peçi and Lazaj (2016), Zaied and Zouabi (2016) and Merchak et al. (2017). But there are few studies considering NIR data to study this agro-climatic influence on oils or other food products.

Regarding the multivariate statistical technique being applied, the majority of the studies included in the literature consider a single factor, a non-numerical variable, to establish different meteorological or agro-climatic zones – see, for example, Alowaiesh et al. (2016), Cornejo et al. (2012), Leskinen et al. (2009a) and Leskinen et al. (2009b), Merchak et al. (2017) or Zheng et al. (2012). If this factor is used as an independent variable in a statistical model, ANOVA (or MANOVA) and a post-hoc test can be used to compare the means corresponding to the defined zones in a numeric variable. The agro- climatic factor can also be used as a dependent variable in the linear discriminant analysis (LDA), where the high dimensionality of the independent variables can be reduced by previously applying principal component analysis (PCA) or partial least squares (PLS). However, the present study rather uses the complete agro- climatic data base obtained from the official webpage of the Automatic Weather Stations (AWEs) of Andalusia. In particular, the historical daily information from 2005 to 2010 has been downloaded for the following variables: temperature, humidity, wind speed, radiation, precipitation and evapotranspiration.

In this case, the agro-climatic data are aggregated in different ways and associated to the EVOO (taking into account the nearest AWE) by using computational programs designed by the powerful free software R-project (R Core Team (2018)). The aim of the study is to explore the linear relationships between agro-climatic and EVOO NIR data: firstly, by using bivariate correlation analysis and, then, generalizing the procedures to multivariate analysis with the application of Redundancy Analysis (RDA).

In particular, Section 2 describes the process of acquisition of NIR and agro-climatic data, the statistical bivariate and multivariate methodology and the computational implementation. Section 3 shows the results and discussion: firstly, the graphical analysis of NIR (original and derivative) spectra and the series of agro-climatic data; secondly, the results of the correlation analysis between the agro-climatic measurements and the spectral absorbance are shown; thirdly, some of the previous conclusions are confirmed and extended by the application of the multivariate technique of RDA. Finally, Section 4 includes the main conclusions of the work.

2. Materials and methods

2.1. Data

2.1.1. NIR data

Olive oil was extracted by the producers through a two-phase centrifugation system. Information from 222 Andalusian extra virgin olive oils, collected from consecutive

harvests from 2005-06 to 2010-11 (denoted H1, H2,..., H6, respectively), is available. The chemical data from each EVOO have been provided by near-infrared (NIR) spectroscopy by the staff of the Organic Chemistry Department of the University of Cordova (Spain). The instruments employed for spectra collection were available at Central Service of Analyses (SCAI) and included a Spectrum One NTS FT-NIR spectrophotometer (Perkin Elmer LLC, Shelton, USA) equipped with an integrating sphere module. Samples were analysed 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 using 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 (Savitzky and Golay (1964)). Once pre-treated, NIR data of 1237 measurements for each case (representing energy absorbed by olive the oil sample at 1237 different wavelengths, from 800.62 to 2499.64 nm) were supplied to the Department of Statistics (University of Cordova) in order to be analysed (Figure 1).

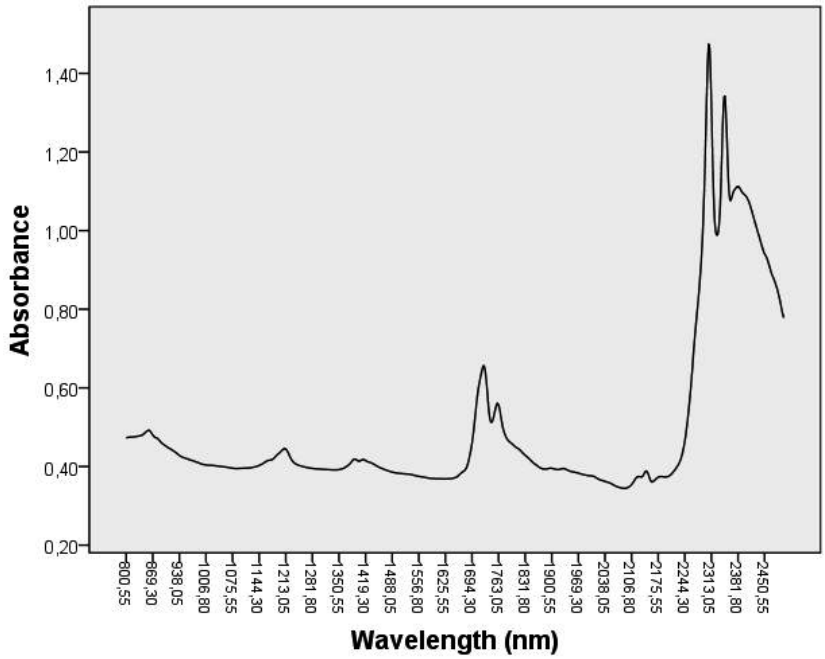


Figure 1: *NIR spectrum of an extra virgin olive oil.*

2.1.2. Agro-climatic data

The agro-climatic data used in the work has been obtained of the official website of the Andalusian Institute of Agricultural, Fisheries, Agrifood and Organic Production Research and Training (IFAPA). In this webpage, the long-run information registered in the Automatic Weather Stations (AWEs) can be accessed[2]. These stations have a suitable plan of maintenance and an exhaustive review of the records that supply the sensors. There are approximately 120 AWEs in all the Andalusian provinces, though in this work only the historical daily information corresponding to the 28 AWEs specified in Table 1 (see Appendix A), for the period 2005-2010 (years before the considered harvest years), has been downloaded. These AWEs have been selected due to their proximity with the point of extraction of the available oils.

Information about the following variables has been considered in this study:

- *Temp*: Daily average temperature, in °C. The temperature is measured by a sensor Pt1000 whose functioning is based on the variation of the resistance of the platinum element by the temperature.
- *Hum*: Daily average relative humidity, in %. The measurement of the relative humidity is realized by a capacitive device of solid condition: sensor HUMICAP 180, plastic polymer that tends to absorb humidity. The sensor changes its electrical characteristics by the variations of humidity, in such a way that diminishes its electrical capacity by the absorption of dampness.
- *WSpe*: Daily average wind speed, in meters per second. Its measurement is realized by a weather vane, in which the rotation of a propeller produces an electrical sign in alternating current, of frequency proportional to the wind speed.
- *Rad*: Daily average radiation, in MJ per m^2. The measurement is realized by a pyrometer constituted by a photoelectric cell of silicon being sensitive to the radiation from 350 to 1100 nm, orientated in a southerly direction and ensuring that another sensor or accessory of the tripod does not cast shade on it.
- *Precip*: Daily precipitation, in mm. The AWE has a device of swinging small containers to measure the volume of rainfall, that is measured by the number of contacts with a tab of the device (each one equivalent to 0.20 mm) that are produced by the overturning of the rain water from one container to the other.
- ET_0. The potential evapotranspiration (PET) is the loss of dampness (in mm per day) of a surface for direct evaporation together with the water loss for perspiration of the vegetation. PET represents the maximum quantity of water that can evaporate from a soil completely covered with vegetation, which develops in ideal conditions and supposing that there are no limitations in the availability of water. ET_0, denoted here as ETo for purposes of labelling, is similar to the ETP though

2. The link is https://www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/servlet/FrontController, where the historical data can be downloaded by clicking on the name of the station and selecting the agro-climatic measurements and the start and end dates [accessed on 02 October 2018].

it is applied to a specific or standard cultivation, habitually cereals or alfalfa, from 8 to 15 cm of uniform height, of active growth, totally covering the soil and not being submitted to water deficit.

2.2. Methodology

2.2.1. Bivariate analysis

Pearson's linear correlation coefficient, r, determines the degree of linear association existing between two numerical variables, being higher as the coefficient is closer to 1 in absolute value. Assuming bivariate normality of the variables, and under the null hypothesis of zero correlation, the statistic $t = r\sqrt{(n-2)/(1-r^2)}$ has t-Student distribution with $n-2$ degrees of freedom, where n is the sample size, equal to 222 in this study. Using a significance level of $\alpha = 0.05$, values of r such as $-0.1317 < r < 0.1317$ show no statistical evidence for rejecting the hyothesis of zero correlation.

2.2.2. Redundancy analysis

Canonical redundancy analysis (RDA) and canonical correspondence analysis (CCA) are two forms of asymmetric canonical analysis, where asymmetric means that the matrices involved in the analysis, $\mathbf{X}$ and $\mathbf{Y}$, do not play the same role: $\mathbf{Y}$ is a matrix of response variables – in this case, containing the spectral information – and $\mathbf{X}$ is the matrix of explanatory variables – in this study, the agro-climatic measurements. This aspect contrasts with canonical correlation analysis where the two matrices play the same role in the analysis and so can be interchanged. $\mathbf{X}$ is used to explain the variation in $\mathbf{Y}$, as in regression analysis, in two steps[3] [4]:

1. Multivariate regression of $\mathbf{Y}$ on $\mathbf{X}$, which is equivalent to a series of multiple linear regressions of the individuals variables of $\mathbf{Y}$ on $\mathbf{X}$ and produces a matrix of fitted values $\hat{\mathbf{Y}}$.
2. Principal component analysis (PCA) of $\hat{\mathbf{Y}}$ in order to reduce its dimension. PCA components of $\hat{\mathbf{Y}}$, called RDA components or *redundancy axes*, are obtained as a reduced number of linear combinations of the variables of $\hat{\mathbf{Y}}$, orthogonal among themselves, explaining a maximum percentage of their variability.

Therefore, in RDA the variability of the variables of $\mathbf{Y}$ are explained from PCA components (factors or latent variables) depending on the variables of $\mathbf{X}$ and so RDA can be seen as a constrained version of PCA.

3. $\mathbf{X}$ and $\mathbf{Y}$ are generally standardized to eliminate the effect of the measurement units.

4. The main assumptions of the data are linearity between the variables of matrix $\mathbf{Y}$ and the variables of the matrix $\mathbf{X}$ and the variance homogeneity of each set of data.

Each eigenvalue of the correlation matrix of the variables of $\hat{\mathbf{Y}}$, λ_j for $j = 1, \ldots, g$, represents the variance of each redundancy axis, whose direction is calculated from the corresponding eigenvector. The proportion of the total variance of $\mathbf{Y}$ explained by a redundancy axis k, $k = 1, \ldots, g$, is given by:

$$\frac{\lambda_k}{\sum_{j=1}^{g} \lambda_j}.$$

The *redundancy index* of the model (similar to a coefficient of determination) is defined by:

$$R_m^2 = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{g} \lambda_j},$$

being m the number of redundancy axes (among the possible g RDA components) to retain.

The results of the applications of RDA analysis are usually shown by representing both matrices, $\mathbf{X}$ and $\mathbf{Y}$, in a space of reduced dimension: the two or three-dimensional space formed by the first RDA components. Variables or cases with the highest coordinates (scores) in a RDA component or redundancy axis are very useful to interpret it, showing the variables and/or cases that are discriminated by this RDA component. Besides, the proximity between variables, cases or RDA components represents the high association between them.

Redundancy analysis as an alternative for canonical correlation analysis was presented by authors such as Rao (1964) and van den Wollenberg (1977). More recently, Legendre, Oksanen and ter Braak (2011), test the significance of the redundancy axes in RDA.

2.2.3. Functional data analysis

For some years, the computing applied to different areas has caused a major technological change due to the addition of faster and more precise measuring equipments. This fact affects one of the paradigms on classical statistics: the number of data should be greater than the number of variables. Currently, large databases corresponding to observations of random variables taken over a continuous interval (or increasingly extensive discretizations of this continuous interval). This kind of data, named *functional data*, appear in a natural way in fields such as the spectrometry, where the measurement result is a curve, a spectrum (see, for example, Aguilera et al. (2010) or Saeys, De Ketelaere and Darius, 2008).

Moreover, in chemometrics, the treatment of a spectrum in the context of functional data analysis, as a continuous function, enables the obtaining of the spectral derivatives as any differentiable function must be continuous at every point in its domain. Many studies of different fields, in particular, of olive oil have proven that the first or second derivative of NIR spectra provide valuable qualitative or quantitative information about

oil that, however, the original spectra do not show (see, for example, Chen et al. (2015) or Woodcock et al. (2008)). Although the original spectral curves overlap, sometimes those ones associated to a high content in a concrete compound or having the effect of an external factor show show higher variability. Therefore, these variations or discrepancies are appreciated more clearly in the first derivative of the spectra than in the original spectra.

2.2.4. Computational implementation

The agro-climatic data corresponding to the year previous to the olive harvest and to the nearest AWE (or the average of the nearest AWEs) are associated to each oil sample. A procedure has been programmed, using R, that permits to select the considered agro-climatic variable (*Temp*, *Hum*, *WSpe*, *Rad*, *Precip*, *ETo*) and accumulates the daily measurements corresponding to several days or months. In particular, the following function has been defined:

AGR-CLIM-function(*station*, *harvest*, *month1*, *month2*, *agro-climatic measurement*),

with the following arguments:

- *station*: among the 28 observed AWEs, the case has associated the code of the nearest geographically (see Table 1),
- *harvest* years, from 1 (2005-06) to 6 (2010-11),
- given the station and the harvest, the period of time (from *month1* to *month2*) can be selected to aggregate the daily agro-climatic measurements,
- *agro-climatic measurement*, distinguishing among the 6 previously described: *Temp*, *Hum*, *WSpe*, *Rad*, *Precip*, *ETo*,

The function returns as value the aggregated agro-climatic measurement according to the selected months and the established meteorological criterion.

Having extracted the data, Pearson's linear correlation is computed between the different agro-climatic measurement, aggregated for different months, and some spectral values of absorbance for the original spectra or their (first or second) derivatives. The graphical procedures will mark in all cases the correlation coefficients which are (or not) statistically different from zero (with $\alpha = 0.05$). As stated above, for the sample size $n = 222$ they are the values outside the range $(-0.1317, 0.1317)$.

The packages of R-project 'fda' (Ramsay et al. (2017)) and 'fda.usc' (Febrero-Bande and Oviedo de la Fuente, 2012) have been used to obtain the spectral derivatives and the multivariate analysis of RDA has been developed by using the package 'vegan' (Oksanen et al. (2018)). Detailed information of the code of the programs designed to read the agro-climatic and chemical data, including the above-mentioned function, and to obtain the diverse range of graphics considered in the study can be seen in the Supplementary Material.

3. Results and discussion

3.1. Graphical analysis

3.1.1. Analysis of NIR spectra

NIR spectra are the representation of the *absorbance*, that is, the quantity of energy absorbed by an oil at each wavelength (from 800.62 to 2499.64 nm, 1237 measures in total). As indicated above, the continuous treatment of a spectrum, instead of an extensive discretization, permits the obtaining of its derivatives that, in occasions, contain valuable information about olive oil compositions.

Thus, in Figure 2 the original spectra as well as their two first derivatives are represented, where the spectra are grouped in the same colour corresponding to a same harvest. The visual analysis highlights the separation or divergence of some spectra, especially those corresponding to the last harvest (H6, depicted in pink). This discrepancy is more pronounced in some ranges of wavelengths of derivative spectra, whose detail is represented in Figure 3 (where the points of maximum discrepancy are denoted by P_1, $P_2, \ldots, P_{10}$ for future analysis).

In addition, in Figure 4 the transposes of the original spectra and their two first derivatives are shown, i.e., the curves are represented as a function of the case. This graphic also highlights the structural change corresponding to the last harvest, H6; this change is especially evident by the view of the derivative spectra.

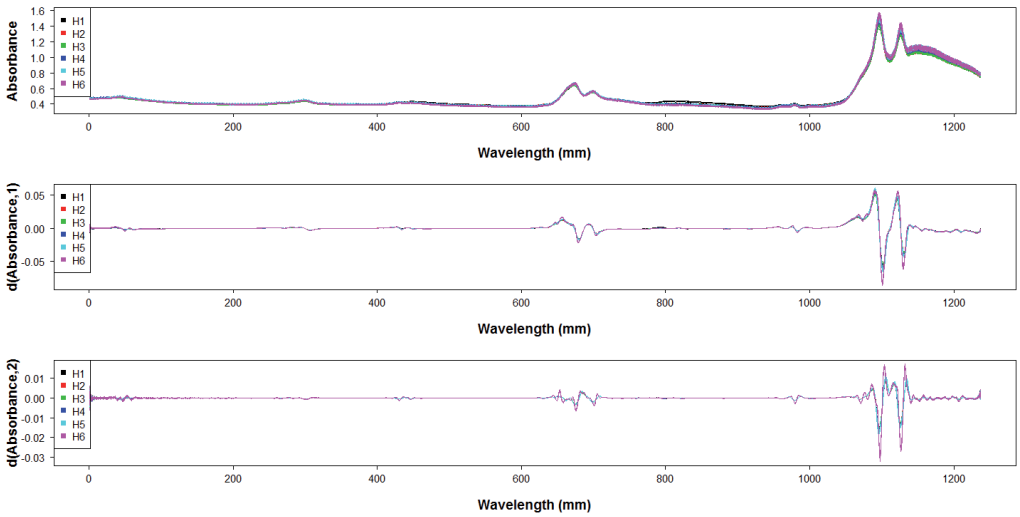


Figure 2: *NIR spectra and their first and second derivatives.*

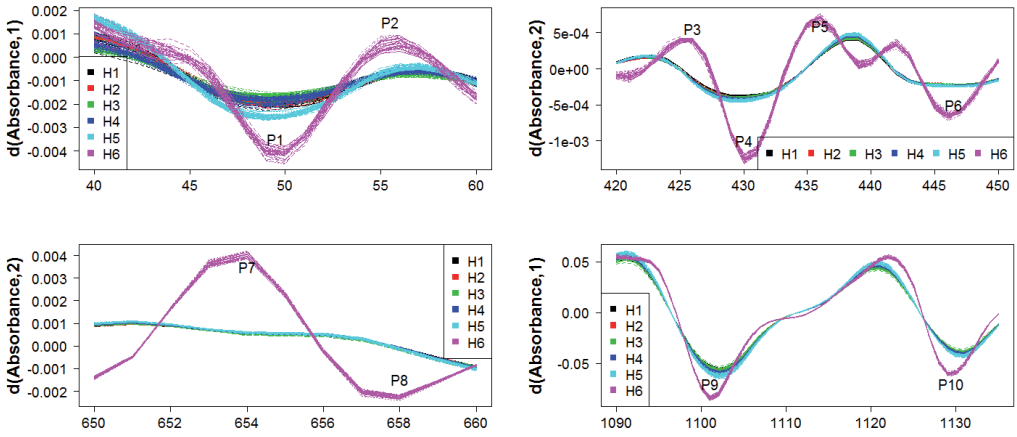


Figure 3: *Spectral details corresponding to the maximum discrepancies.*

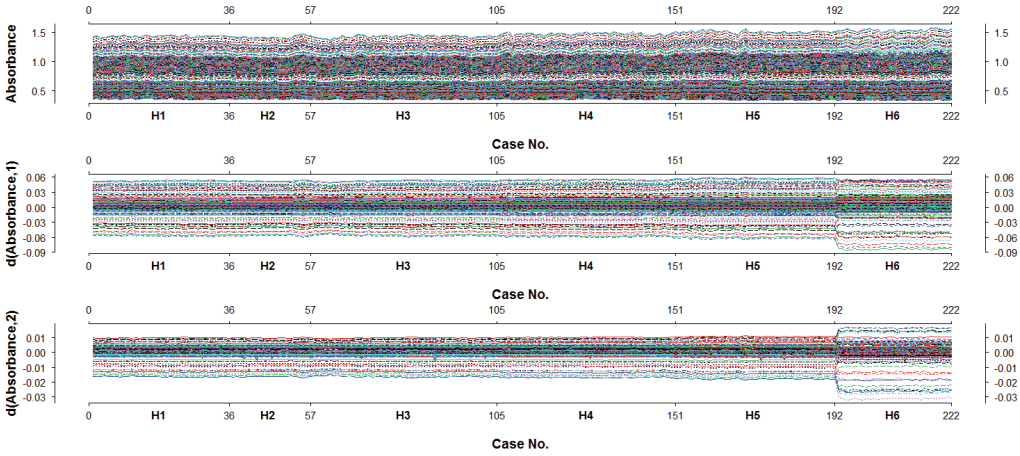


Figure 4: *Spectral and derivative NIR values, as a function of the case.*

3.1.2. Analysis of series of agro-climatic data

In this section, the series of the six agro-climatic measurements (*Temp*, *Hum*, *WSpe*, *Rad*, *Precip*, *Eto*) are represented for the six harvests. The daily values are accumulated for each month (afterwards, the reason is explained) and then standardized in order to eliminate the effect of the measurement units of each variable. So, dimensionless series are obtained that can be represented and compared in the same graphic. These standardized values are represented in Figure 5 which shows a cyclical tendency for all the considered variables. In general, the proximity of the trajectories of evolution of the variables *Temp*, *WSpe*, *Rad* and *Eto*, on the one hand, and *Hum* and *Precip*, on the other, is observed, noting also the symmetry among them. With regard to the relation between precipitation and radiation, Bradley et al. (2011) use cross-spectral analysis

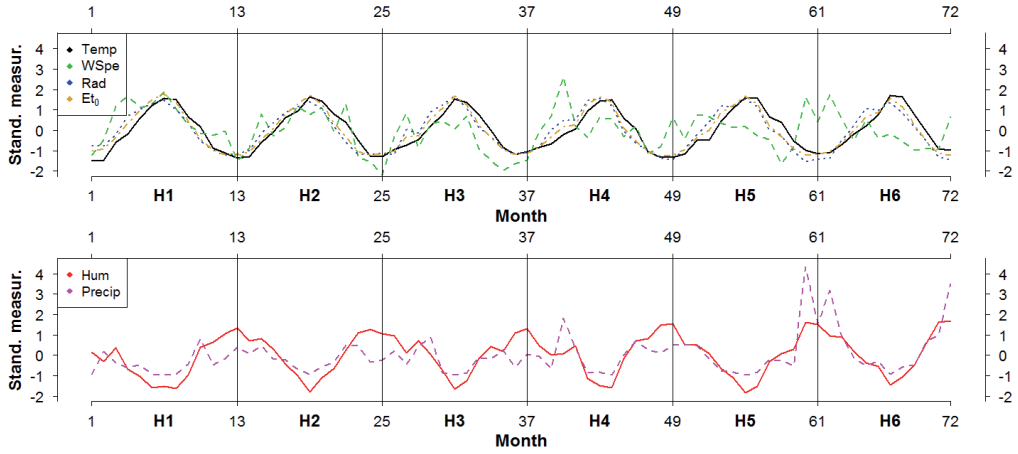
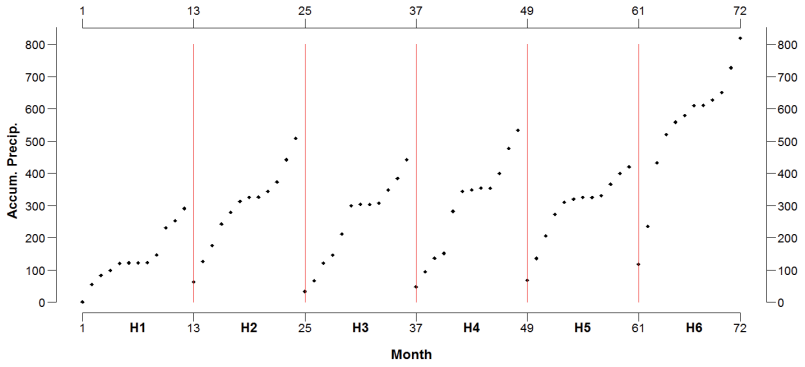


Figure 5: *Monthly accumulated and standardized agro-climatic measurements.*



Year	Ann. prec. (in mm.)
2005	290.65
2006	508.97
2007	442.13
2008	533.80
2009	419.85
2010	819.05

Figure 6: *Monthly accumulated precipitation (in mm.), for each harvest.*

to show that precipitation has a role to play in the maintenance of phenology cycles because it maintains constant vegetation growth reducing so the seasonal impact of the solar radiation.

As fundamental irregularity of Figure 5, the especially high values of the variables that represent the wind speed (*WSPe*, in green) and the volume of precipitation (*Precip*, in pink) at the beginning of the 4th harvest and at the beginning and the end of the 6th harvest (H6) can be highlighted. This fact corroborated the work of Back and Bretherton (2005) which studied the relationship between wind speed and precipitation in the Pacific and found a significant correlation between these variables. The specially irregular behaviour of the *Precip* variable in H6, whose accumulated mean values are specially high, can also be deduced from the observation of Figure 6.

Therefore, the anomalous accumulated precipitation (or wind speed) values corresponding to the sixth harvest together with the anomalous derivative NIR spectra corresponding to the same harvest justify the formulation of the following question: What

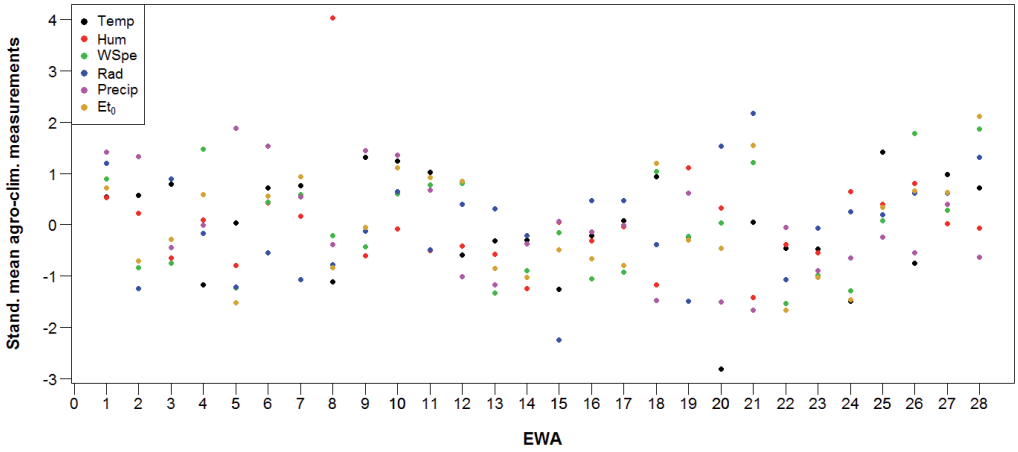


Figure 7: *Standardized mean agro-climatic measurements, for each AWE.*

is the effect of the precipitation or the wind speed, in particular, or the agro-climatic conditions, in general, on NIR spectra or on the chemical compounds of EVOO?

Finally, Figure 7 depicts the standardized mean values for the six agro-climatic measurements for the 28 automatic weather stations. The obvious discrepancies among the mean values corresponding to the different AWEs makes reasonable the assignation the agro-climatic measurements associated to the nearest AWE to each olive oil (case).

3.2. Bivariate analysis

The following function:

$$\textit{AGR-CLIM-function}(\textit{station}, \textit{harvest}, \textit{month1}, \textit{month2}, \textit{agro-climatic measurement}),$$

described in Section 2.2 (Methodology) and whose code is included in the Supplementary Material, has been applied to each EVOO (222, in total), considering the nearest AWE (station) and the corresponding harvest. The six agro-climatic measurements previously downloaded (*Temp*, *Hum*, *WSpe*, *Rad*, *Precip* and *Eto*) have been accumulated for each month, from January to December. Therefore, a list of 12 matrices of dimension 222×6, $[\mathbf{X}_1|\mathbf{X}_2|\cdots|\mathbf{X}_{12}]$, is available. Moreover, $\mathbf{Y}$ is the matrix of dimension 222×10 whose columns contain the absorbance associated to the 10 peaks of maximum discrepancy (P_1, P_2, ..., P_{10}) represented in Figure 3.

The aim of aggregating the agro-climatic measurements has been to relate them more adequately to the phenological cycle of the olive grove, which will directly influence the composition of the oil. As shown in Figure 8, this cycle is not distributed equally, and in this way the months of interest in each case could be studied independently. In the bibliography, authors such as Orlandi et al. (2012), in the study of the influence of

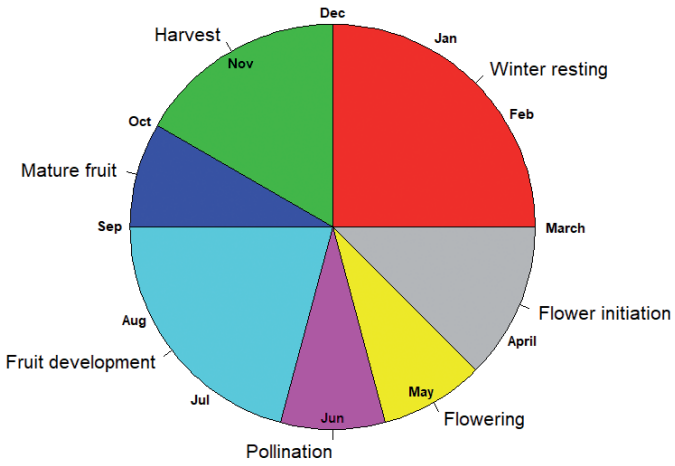


Figure 8: *Phenological stages of olive.*

climate data on oil production in southern Italy, also consider meteorological variables on a monthly basis.

3.2.1. Correlations between the agro-climatic measurements and the discrepancy spectral peaks

In this section, Pearson's linear correlation coefficients are calculated between each of the six agro-climatic measurements, accumulated for each month, and the discrepancy spectral peaks denoted in Figure 3. The results are shown in Figures B.1-B.4 (in Appendix B), where the light grey lines of points mark the correlations -0.5 and 0.5 and the dark grey lines of points show the frontier between the values being different (or not) statistically from zero for $\alpha = 0.05$.

The following fundamental conclusions can be deduced from the observation of Figures B.1-B.4:

- There are many high correlations, next to -1 or 1, specially for the accumulated agro-climatic measurements corresponding to January, February, March, June and November. Therefore, the lowest correlations between the discrepancy spectral peaks and the aggregate agro-climatic measurements appear for the months of the phenological stages corresponding to the development and the maturation of olives (see Figure 8). And so it may be interpreted that the highest effect of the meteorological conditions (in particular, of the precipitation), reflected in NIR spectra, takes place not on the fruit but on the tree.
- From the observation of the different agro-climatic measurements, the precipitation (*Precip*, in pink) and the radiation (*Rad*, in blue) are the variables showing, in general, the highest (positive or negative) correlations, having opposite sign.

As shown in Bradley et al. (2011), precipitation and radiation have negative linear correlation and now Figures B.1-B.4 highlight that both agro-climatic measurements have a contrary effect on the discrepancy spectral peaks. Besides, the sign of the pairwise correlations between *Precip-Rad* and the peaks P_2, P_3, P_5, P_7 are the same, and the opposite of the sign of the correlations between the rest of the peaks. By coincidence, these peaks are the relative maxima of the derivative NIR spectra while the other peaks are the relative minima.

- Some agro-climatic variables are almost uncorrelated between the discrepancy spectral peaks for many months but, nevertheless, shown values closer to 1 (in absolute terms) for a concrete month. These are the case, for example, of the evapotranspiration (*ETo*, in yellow) or the humidity (*Hum*, in red) in March or November, whose influence on the spectral peaks is the contrary. The negative or inverse correlation between both variables can be intuited from the observation of Figure 5. Besides, in March and November, the standardized values for *ETo* and *Hum* are quite similar and, however, the effect on the discrepancy spectral peaks is the highest.

3.2.2. Correlations between the agro-climatic measurements and the spectral absorbance

In this section, Pearson's linear correlation coefficients between the monthly accumulated agro-climatic measurements and the spectral absorbance are calculated. The results, that coincide with the ones obtained from Figures B.1-B.4, are shown in Figures C.1-C.4 (in Appendix C).

The following general conclusions can be obtained:

- January and December are the months showing, in general, the highest correlations (in absolute terms) and April is the one with the correlation values nearest to zero. This fact confirms, newly, that the highest correlations appear in the phenological stage of winter resting of olive tree (see Figure 8).
- Taking into account the different agro-climatic measurements, the precipitation (*Precip*, in pink) and the radiation (*Rad*, in blue) are the variables showing the highest correlations, being the opposite the sign of their linear correlation. In general, the sign of the correlation for the radiation and the evapotranspiration (*ETo*, in yellow) is the same, and the opposite to the sign of the correlation for all other variables.

3.3. Multivariate analysis

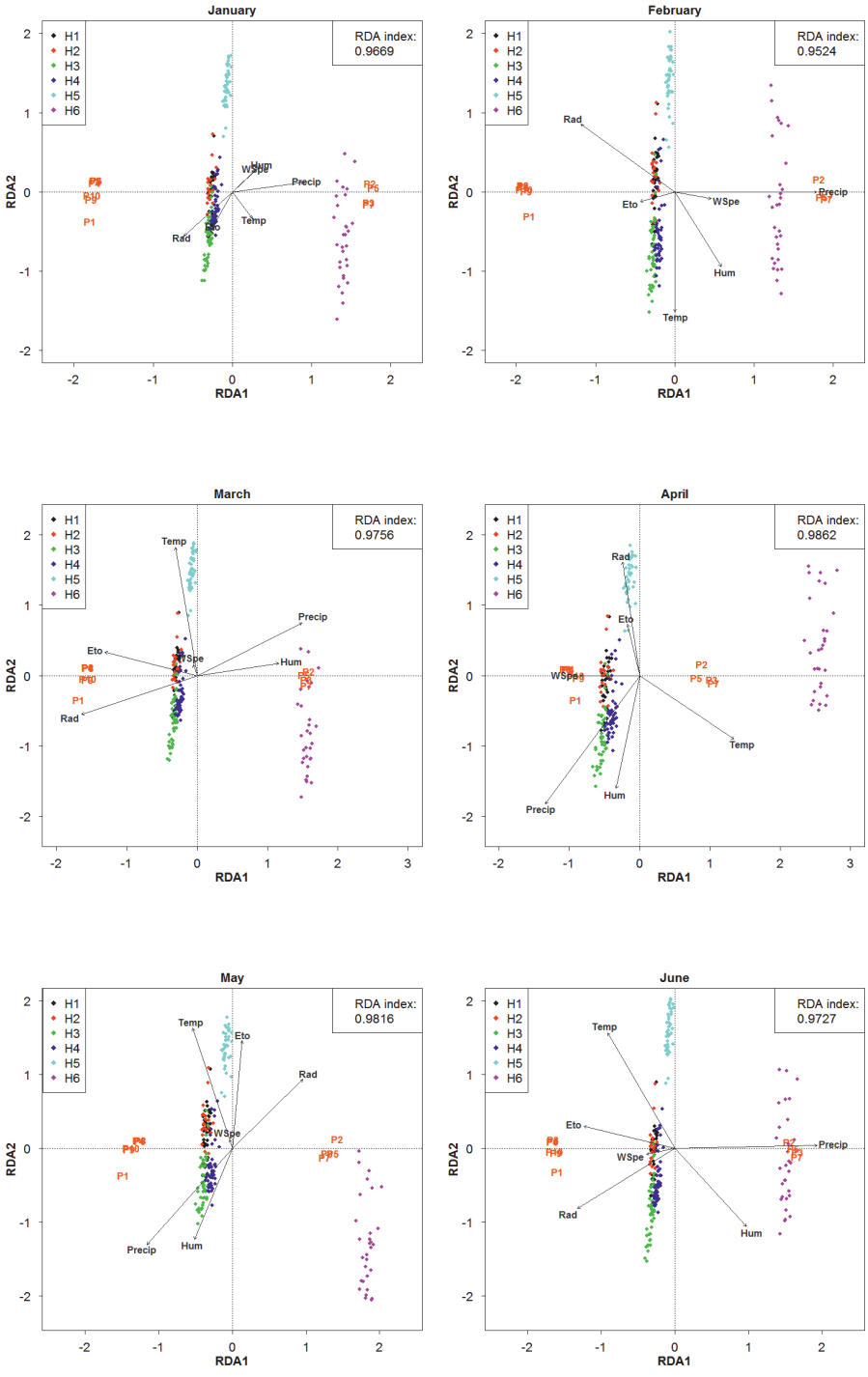
In this section, redundancy analysis (RDA) is applied to generalize the previous results and highlight the cause and effect relationships between two data matrices: one of them, the matrix of explanatory variables, containing in its columns the six accumulated agro-climatic measurements for a specific month ($\mathbf{X}_i$, $i = 1, \ldots, 12$) and the other, the matrix

of response variables ($\mathbf{Y}$), formed by the spectral absorbance associated to the 10 peaks of maximum discrepancy (P_1,P_2,...,P_{12}) represented in Figure 3.

The results of the application of each RDA are shown in the two-dimensional space formed by the two first RDA components (RDA1 and RDA2), where both matrices, $\mathbf{X}_i$ and $\mathbf{Y}$, are represented. The results are drawn, for each month, in Figure 9. Each individual representation shows: *a)* the cases, using different colors for the different harvests (black, red, green, blue, cyan and pink for H1,H2,...,H6, respectively); *b)* the response variables: the absorbance for the spectral peaks ($\mathbf{Y}$, in orange); *c)* the explanatory variables: the agro-climatic measurements for each month ($\mathbf{X}_i$, $i = 1,\ldots,12$, in gray). The redundancy index is greater than 0.95 for all the months (as it is shown at the top right of each graphic), which indicates that the percentage of the total variance of $\mathbf{Y}$ (spectral peaks) explained by the two first RDA components is greater than 95%.

In general lines, the conclusions obtained from the observation of Figure 9 confirm some of the above-mentioned ones, deduced from the bivariate analysis. More in particular, the following results can be enumerated:

- *Cases analysis*: The cases corresponding to the last harvest (H6, in pink) are clearly discriminated or separated from the remaining harvests for all the months: cases of H6 have high scores (in absolute terms) in RDA1 for all the months whereas all the other cases have scores near zero in this axis. RDA2 permits to discriminate the harvest H5 (in cyan) from the others: cases of H5 have high (absolute) scores in RDA2. Cases of H6 have also high scores in RDA2 for months such as October but the groups of cases can be discriminated by the scores in RDA1. The cases associated to H1, H2, H3 and H4 are, in general, overlapped and, so, they are not discriminated by RDA1 and RDA2 (the most important redundancy axes), showing scores near zero in both redundancy axis, in general. RDA statistically modelled the situation previously represented in Figure 3, where spectra corresponding to H6 (and H5, to a lesser degree) are clearly discriminated from the others for some ranges of the original spectra or their first two derivatives.
- *Response variables analysis*: P_2, P_3, P_5, P_7 are clearly discriminated from the other spectral peaks in all the months (being all of them depicted in orange). For all months, the peaks have scores greater than one, in absolute terms, in RDA1, whereas the scores in RDA2 are near zero. In this case, RDA has also a clear correspondence with the representation of Figure 3, as P_2, P_3, P_5, P_7 are relative maxima of the derivative spectra while P_1, P_4, P_6, P_8, P_9 and P_{10} are relative minima.
- *Explanatory variables analysis*: As in bivariate analysis, taking into account the different agro-climatic measurements (represented in gray), the radiation (*Rad*) and the precipitation (*Precip*) are the variables having the highest scores (in absolute terms) in RDA1 (especially, in February, June and December). The sign of the scores (and, so, the correlation) is the opposite for these two variables (in line with the observation of Figure 5 and Bradley et al. (2011)). The humidity (*Hum*)


January
RDA index: 0.9669
February
RDA index: 0.9524
March
RDA index: 0.9756
April
RDA index: 0.9862
May
RDA index: 0.9816
June
RDA index: 0.9727
H1
H2
H3
H4
H5
H6
RDA1
RDA2
Temp
Precip
Rad
Eto
WSpe
Hum

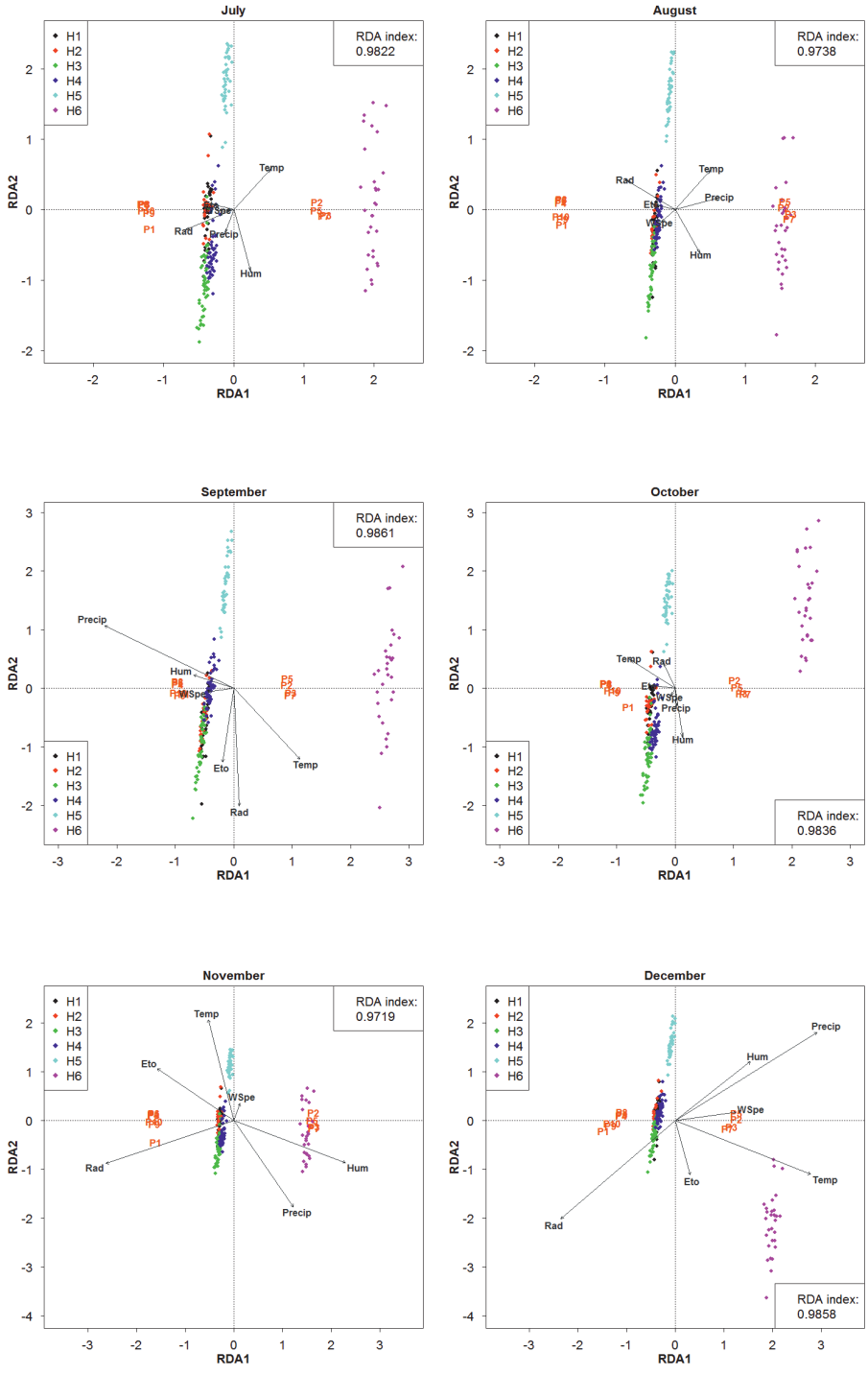


Figure 9: *RDA representations from January to December.*

and the temperature (*Temp*) are clearly discriminated by RDA2 for months such as May or June, having both agro-climatic measurements opposite scores. The evolution of these two variables is also the opposite in Figure 5. Finally, with respect to the cases, response and explanatory variables, in months such as January, March or November, *Precip* shows high scores, in absolute terms, in RDA1 close to the scores of the relative maxima peaks P_2, P_3, P_5, P_7 and the last harvest (H6). Besides, in March, May or June, *Temp* shows high (absolute) scores in RDA2, this variable being near the cases of H5.

4. Conclusions

During recent years NIR spectroscopy has been commonly used because it is a fast, reliable and cost-effective chemical technique. Many studies apply chemometrics analysis to highlight the valuable information contained in NIR spectra of EVOO. Firstly, studies such as Galtier et al. (2007) or Sánchez-Rodríguez et al. (2013) and Sánchez-Rodríguez et al. (2014) show the prediction of the fatty acid profile (quantitative information) from NIR spectra. Other authors (Bertran et al. (2000) or Öztürk et al. (2010)) highlight the potentiality of NIR spectra to analyse the traceability of EVOO in order to their authentication. Casale et al. (2012) characterize PDO olive oil (qualitative information) from NIR spectra.

Moreover, this paper highlights the effect of agro-climatic conditions on spectra of olive oils. In particular, the study show the structure of linear relationships being between two sets of Big Data: NIR spectra of EVOO and agro-climatic data downloaded from the official Andalusian Automatic Weather Stations (AWEs). The graphical analysis of both data sets detects, firstly, an irregular behaviour of (original and derivative) NIR spectra corresponding to the last harvest of extraction of EVOO (H6), in particular, ten peaks of maximum discrepancy, P_1, P_2, ..., P_{10}, are determined. Secondly, the graphical analysis of the series of agro-climatic data shows irregularities in the volume of precipitation (*Precip*) or the wind speed (*WSpe*) accumulated for the previous year. This fact motivates the question about what is the effect of the agro-climatic conditions on NIR spectra or on the chemical compounds of EVOO (as NIR spectra are useful to determine quantitative information of EVOO). The answer is obtained, initially, by using bivariate analysis between the agro-climatic measurements and the spectral absorbance and, then, by extending the previous results by applying RDA. The first RDA component or redundancy axis (obtained when the matrix of spectral absorbance is the response and the matrix of agro-climatic measurements contains the explanatory variables) clearly discriminate the cases of EVOO corresponding to H6 whereas the cases corresponding of H5 are discriminated by the second RDA component. As final conclusions from bivariate and multivariate analysis, the variables monthly accumulating the precipitation (*Precip*) and the radiation (*Rad*) show, in general, the highest (in abso-

lute terms) linear correlation between the spectral absorbance, but having opposite sign. The correlation coefficients associated to wind speed (*WSpe*) are the closest to zero and so, unlike precipitations, the irregularities of the series of *WSpe* at the beginning of the harvest H6 can not be associated with the discrepancy of the EVOO NIR spectra of this harvest.

Therefore, the main contributions of this work are the treatment of the original agro-climatic data, instead of defining a factor with levels associated to the meteorological conditions, and the computational implementation in R to analyse the structure of correlations between this set of Big Data and the EVOO spectral data and efficiently represent the results (see the designed programs in the Supplementary Material). Once the effect of agro-climatic conditions on EVOO NIR spectra has been highlighted by using the Big Data and since NIR spectra contain important qualitative and quantitative information of EVOO, a further study could treat the influence of meteorological aspects in some quality parameters of olive oils, such as the fatty acids content, in order to authenticate the oils and prevent fraudulent practices.

Acknowledgements

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

Appendix A

Table 1: *Automatic weather stations (AWEs).*

Province	Station	Code
Cadiz	Villamartín	1
Cordova	Adamuz	2
	Baena	3
	Belmez	4
	Cabra	5
	Córdoba	6
	El Carpio	7
	Hinojosa del Duque	8
	Hornachuelos	9
	Palma del Río	10
	Santaella	11
Granada	Loja	12
	Pinos Puente	13
Jaen	Alcaudete	14
	Chiclana de Segura	15
	Jaén	16
	Higuera de Arjona	17
	Mancha Real	18
	Marmolejo	19
	Pozo Alcón	20
	San José de los Propios	21
	Santo Tomé	22
Malaga	Antequera	23
	Archidona	24
	Pizarra	25
	Sierra de Yeguas	26
Sevilla	Écija	27
	Osuna	28

Appendix B

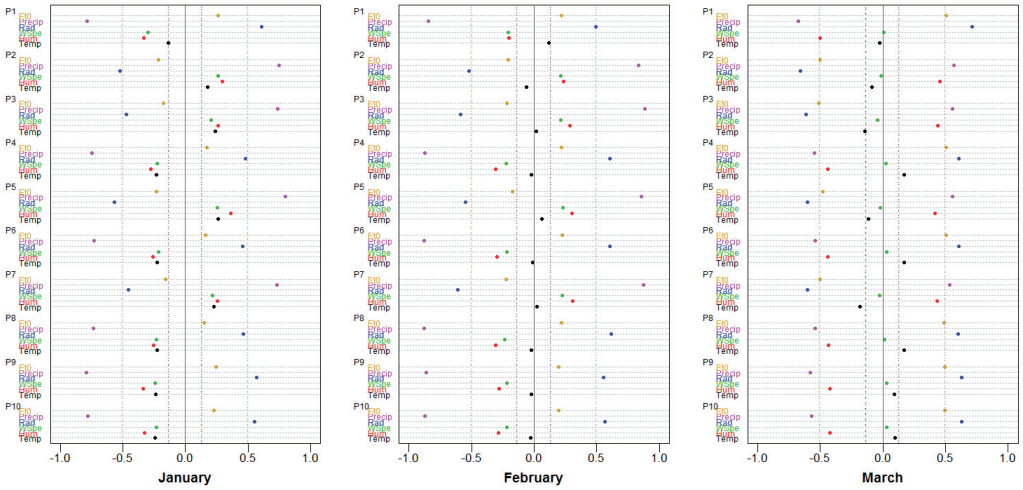


Figure B.1: *Correlations for January, February and March.*

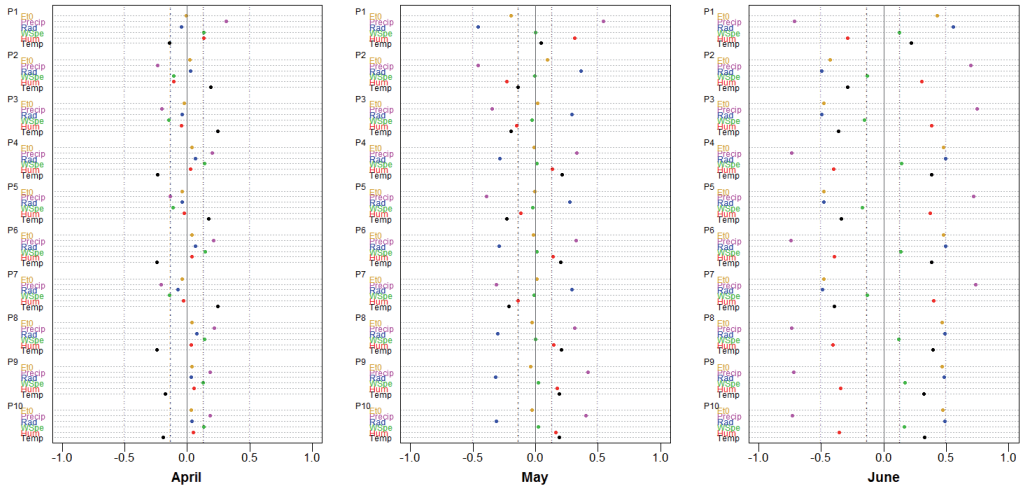


Figure B.2: *Correlations for April, May and June.*

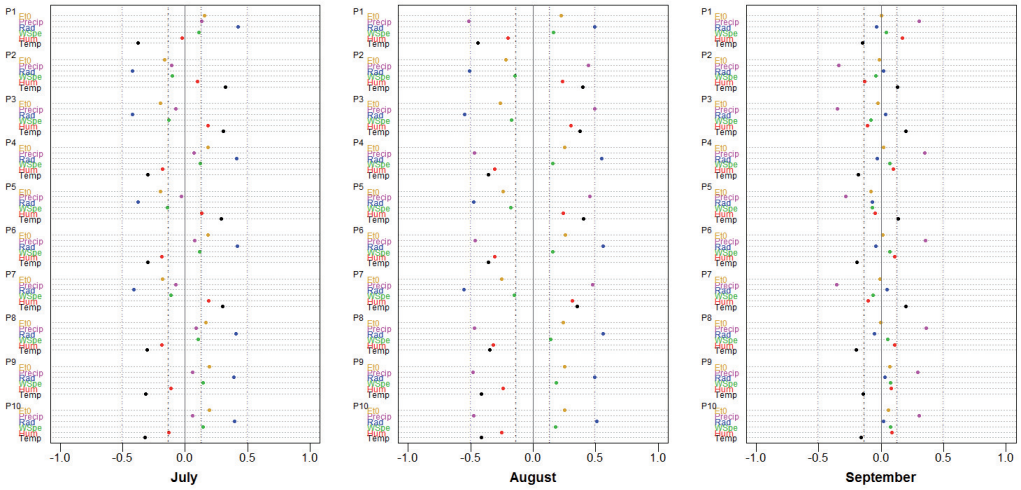


Figure B.3: *Correlations for July, August and September.*

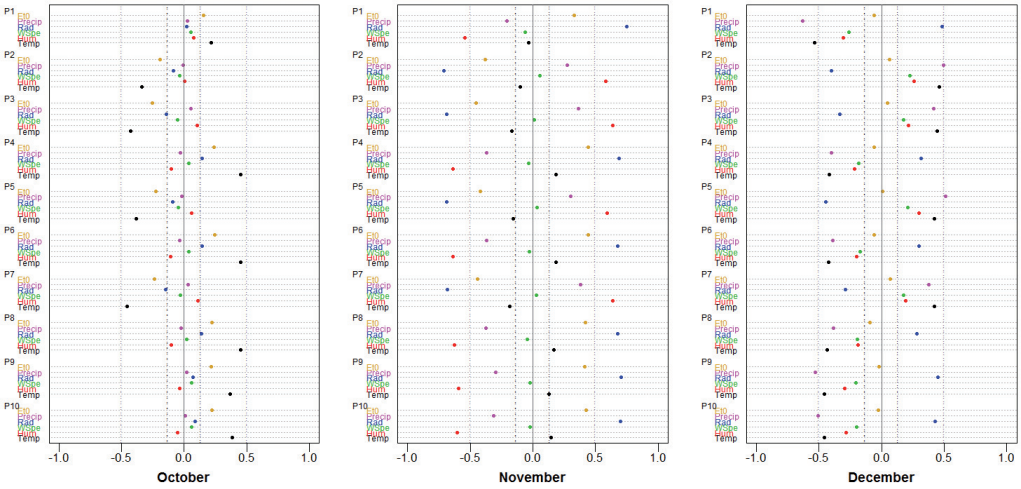


Figure B.4: *Correlations for October, November and December.*

Appendix C

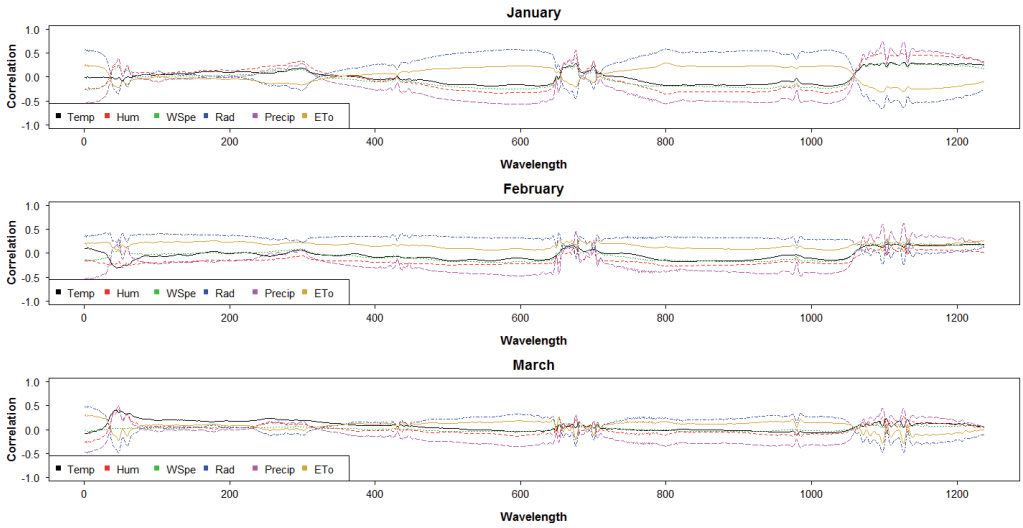


Figure C.1: *Correlations between the NIR spectral absorbance and the accumulated agro-climatic measurement for January, February and March.*

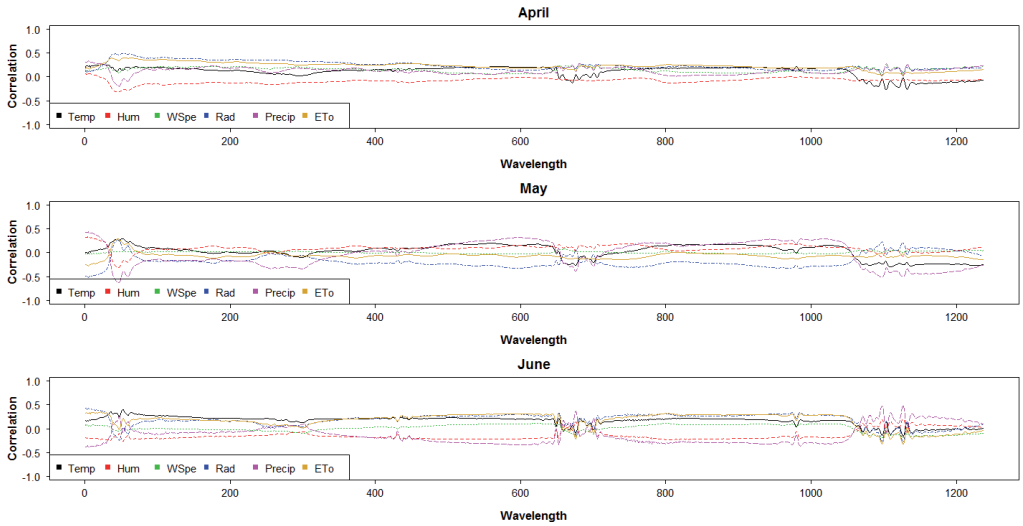


Figure C.2: *Correlations between the NIR spectral absorbance and the accumulated agro-climatic measurement for April, May and June.*

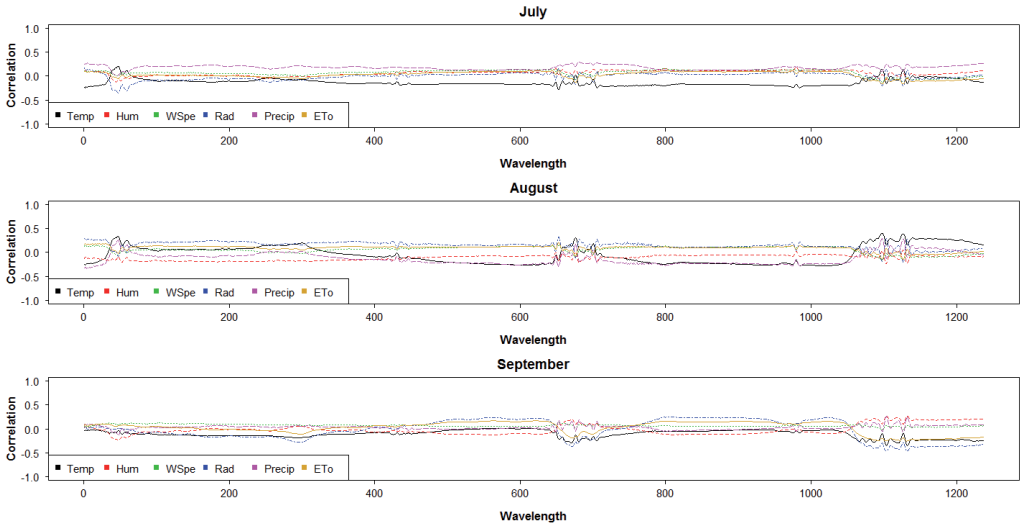


Figure C.3: *Correlations between the NIR spectral absorbance and the accumulated agro-climatic measurement for July, August and September.*

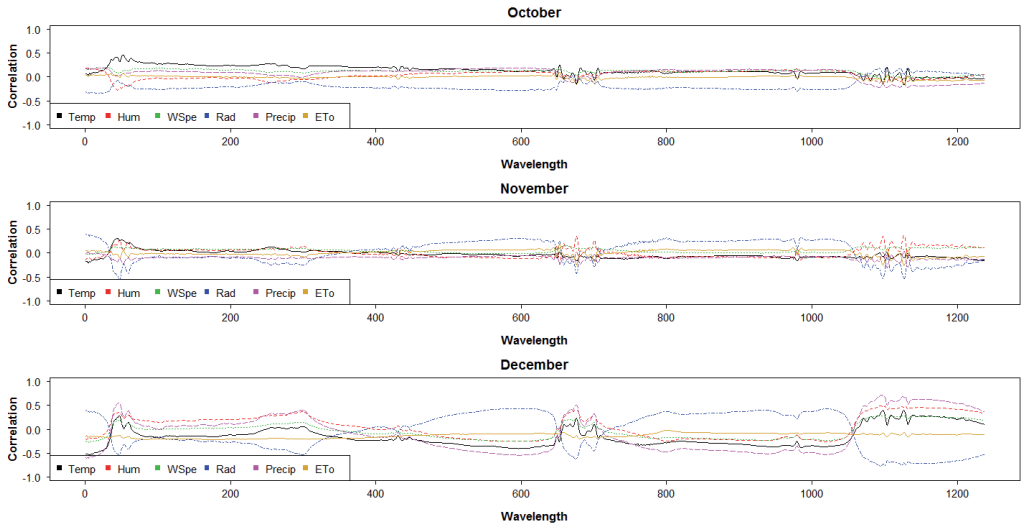


Figure C.4: *Correlations between the NIR spectral absorbance and the accumulated agro-climatic measurement for October, November and December.*

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(2), 289–305.

Alowaiesh, B., Singh, Z. and Kailis, S. G. (2016). Harvesting time influences fruit removal force, moisture, oil content, free fatty acids and peroxide in the oil of Frantoio and Manzanilla olive cultivars. *Australian Journal of Crop Science*, 10(12), 1662–1668. DOI:0.21475/ajcs.2016.10.12.p7737.

Awale, M., Visini, R., Probst, D., Arús-Pous, J. and Reymond, J. L. (2017). Chemical Space: Big Data Challenge for Molecular Diversity. *CHIMIA International Journal for Chemistry*, 71(10), 661–666.

Awan, A. A. (2014). Influence of agro-climatic conditions on fruit yield and oil content of olive cultivars. *Pakistan Journal of Agricultural Sciences*, 51(3).

Back, L. E. and Bretherton (2005). The relationship between wind speed and precipitation in the Pacific ITCZ. *Journal of Climate*, 18(20), 4317–4328.

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.

Bradley, A. V., Gerard, F. F., Barbier, N., Weedon, G. P., Anderson, L. O., Huntingford, C., Aragão, L. E. O. C., Zelazowski, P. and Arai, E. (2011). Relationships between phenology, radiation and precipitation in the Amazon region. *Global Change Biology*, 17(6), 2245–2260.

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.

Chen, J. Y., Zhang, H., Ma, J., Tuchiya, T. and Miao, Y. (2015). Determination of the degree of degradation of frying rapeseed oil using Fourier-transform infrared spectroscopy combined with partial least-squares regression. *International Journal of Analytical Chemistry*, DOI: 10.1155/2015/185367.

Chiang, L., Lu, B. and Castillo, I. (2017). Big Data Analytics in Chemical Engineering. *Annual Review of Chemical and Biomolecular Engineering*, (8), 63–85.

Cornejo, V., Bueno, L. A. and Gines, I. L. (2012). Evaluation of 'Arbequina' olive oils from different growing areas of San Juan, Argentina. *VII International Symposium on Olive Growing*, 1057, 661–667.

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.

Dorey, E., Fournier, P., Léchaudel, M. and Tixier, P. (2016). Modeling sugar content of pineapple under agro-climatic conditions on Reunion Island. *European Journal of Agronomy*, 73, 64–72.

Edmunds, B. A., Clark, C. A., Villordon, A. Q. and Holmes, G. J. (2015). Relationships of preharvest weather conditions and soil factors to susceptibility of sweetpotato to postharvest decay caused by Rhizopus stolonifer and Dickeya dadantii. *Plant Disease*, 99(6), 848–857.

Falasca, S. L., Ulberich, A. C. and Ulberich, E. (2012). Developing an agro-climatic zoning model to determine potential production areas for castor bean (Ricinus communis L.).

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software*, 51(4), 1–28.

Galtier, O., Dupuy, N., Le DrÃ©au, 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.

Hu, Y. and Bajorath, J. (2017). Entering the 'big data' era in medicinal chemistry: molecular promiscuity analysis revisited. *Future Cience OA*, 3(2). DOI: 10.4155/fsoa-2017-0001.

Guo, X. (2016). Application of meteorological big data. *IEEE Communications and Information Technologies (ISCIT), 16th International Symposium*, 273–279.

Jarvis, C. K., Sapirstein, H. D., Bullock, P. R., Naeem, H. A., Angadi, S. V. and Hussain, A. (2008). Models of growing season weather impacts on breadmaking quality of spring wheat from producer fields in western Canada. *Journal of the Science of Food and Agriculture*, 88(13), 2357–2370.

Khokhar, J. S., Sareen, S., Tyagi, B. S., Singh, G., Chowdhury, A. K., Dhar, T., Sign, V., King, I. P., Young, S. D. and Broadley, M. R. (2017). Characterising variation in wheat traits under hostile soil conditions in India. *PLoS One*, 12(6), e0179208.

Legendre, P., Oksanen, J. and ter Braak, C. J. (2011). Testing the significance of canonical axes in redundancy analysis. *Methods in Ecology and Evolution*, 2(3), 269–277.

Leskinen, H. M., Suomela, J. P. and Kallio, H. P. (2009a). Effect of latitude and weather conditions on the regioisomer compositions of α-and γ-linolenoyldilinoleoylglycerol in currant seed oils. *Journal of agricultural and food chemistry*, 57(9), 3920–3926. DOI: 10.1021/jf900068b.

Leskinen, H. M., Suomela, J. P., Yang, B. and Kallio, H. P. (2009b). Regioisomer compositions of vaccenic and oleic acid containing triacylglycerols in sea buckthorn (Hippophae rhamnoides) pulp oils: influence of origin and weather conditions. *Journal of agricultural and food chemistry*, 58(1), 537–545. DOI: 10.1021/jf902679v.

Luciano, R. V., Albuquerque, J. A., Rufato, L., Miquelluti, D. J. and Warmling, M. T. (2013). Weather and soil effects on the composition of 'Cabernet Sauvignon' grape. *Pesquisa Agropecuária Brasileira*, 48(1), 97–104.

Mailer, R. J. (2004). Rapid evaluation of olive oil quality by NIR reflectance spectroscopy. *Journal of the American Oil Chemists' Society*, 81(9), 823–827.

Martínez-Herrera, J., Siddhuraju, P., Francis, G., Davila-Ortiz, G. and Becker, K. (2006). Chemical composition, toxic/antimetabolic constituents, and effects of different treatments on their levels, in four provenances of Jatropha curcas L. from Mexico. *Food Chemistry*, 96(1), 80–89.

Merchak, N., El Bacha, E., Khouzan, R. B., Rizk, T., Akoka, S. and Bejjani, J. (2017). Geoclimatic, morphological and temporal effects on Lebanese olive oils composition and classification: A ^{1}H NMR metabolomic study, *Food Chemistry*, 217, 379–388.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Henry, M., Stevens, H., Szoecs E. and Wagner, H. (2018). *vegan: Community Ecology Package. R package version 2.4-6*. https://CRAN.R-project.org/package=vegan

Orlandi, F., Bonofiglio, T., Romano, B. and Fornaciari, M. (2012). Qualitative and quantitative aspects of olive production in relation to climate in southern Italy. *Scientia Horticulturae*, 138, 151–158.

Ozdemir, Y. (2016). Effects of climate change on olive cultivation and table olive and olive oil quality. *Scientific Papers. Series B, Horticultures, LX*, 65–69.

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.

R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Ramsay, J. O., Wickham, H., Graves, S. and Hooker, G. (2017). *fda: Functional Data Analysis. R package version 2.4.7*. https://CRAN.R-project.org/package=fda.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics*, Series A, 329–358.

Rymbai, H., Laxman, R. H., Dinesh, M. R., Sunoj, V. J., Ravishankar, K. V. and Jha, A. K. (2014). Diversity in leaf morphology and physiological characteristics among mango (Mangifera indica) cultivars popular in different agro-climatic regions of India. *Scientia Horticulturae*, 176, 189–193.

Sacco, A., Brescia, M. A., Liuzzi, V., Reniero, F., Guillou, G., Ghelli, S. and van der Meer, P. (2000). Characterization of Italian olive oils based on analytical and nuclear magnetic resonance determinations. *Journal of the American Oil Chemists' Society*, 77(6), 619–625.

Saeys, W., De Ketelaere, B. and Darius, P. (2008). Potential applications of functional data analysis in chemometrics. *Journal of Chemometrics*, 22(5), 335–344.

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 Journal*, 37(1), 57–78.

Sánchez-Rodríguez, M. I., Sánchez-López, E., Marinas, A., Caridad, J. M., Urbano, F. J. and Marinas, J. M. (2014). New approaches in the chemometrics analysis of infrared spectra of extra-virgin olive oils. *Statistics and Operations Research Transactions Journal*, 38(2), 231–250.

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8), 1627–1639.

Van den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2), 207–219.

Veizi, A., Peçi, E. and Lazaj, L. (2016). Influence of harvesting time in chemical and organoleptic qualities of extra virgin olive oil, *Journal of Multidisciplinary Engineering Science and Technology*, 3(10), 5794–5800.

Wang, X., Song, L., Wang, G., Ren, H., Wu, T., Jia, X., Wu, H.P. and Wu, J. (2016). Operational climate prediction in the era of big data in China: Reviews and prospects. *Journal of Meteorological Research*, 30(3), 444–456.

Woodcock, T., Downey, G. and O'Donnell, C.P. (2008). Confirmation of declared provenance of European extra virgin olive oil samples by NIR spectroscopy. *Journal of Agricultural and Food Chemistry*, 56(23), 11520–11525.

Yang, W., Laaksonen, O., Kallio, H. and Yang, B. (2017). Effects of latitude and weather conditions on proanthocyanidins in berries of Finnish wild and cultivated sea buckthorn (Hippophae rhamnoides L. ssp. rhamnoides). *Food Chemistry*, 216, 87–96.

Zaied, Y. B. and Zouabi, O. (2016). Impacts of climate change on Tunisian olive oil output. *Climatic Change*, 139(3-4), 535–549.

Zheng, J., Yang, B., Ruusunen, V., Laaksonen, O., Tahvonen, R., Hellsten, J. and Kallio, H. (2012). Compositional differences of phenolic compounds between black currant (Ribes nigrum L.) cultivars and their response to latitude and weather conditions. *Journal of Agricultural and Food Chemistry*, 60(26), 6581–6593. DOI: dx.doi.org/10.1021/jf3012739.