Postprint of Innovative Food Science & Emerging Technologies (2018)
DOI: https://doi.org/10.1016/j.ifset.2018.05.018

Olive oil nutritional labeling by using Vis/NIR spectroscopy and compositional statistical methods

José A. Cayuela-Sánchez*[1], Javier Palarea-Albaladejo[2]

[1]Instituto de la Grasa, CSIC, Campus de la Universidad Pablo de Olavide, Ed. Nº 46, Crtra. De Utrera, Km 1, 41013, Seville, Spain
[2]Biomathematics and Statistics Scotland, JCMB, The King's Buildings, EH9 3FD, Edinburgh, UK
*Correspondence author: jacayuela@ig.csic.es

Abstract

Food nutritional labeling is compulsory in the European Union since 13 December 2016. The olive oil fatty acid composition shows high variation depending mainly on the variety. Thus, olive oil nutritional labeling is problematic for the industry. Besides, the analysis of all batches of olive oil using the official methods is expensive. Therefore, the olive oil industry is seriously concerned about solutions for nutritional labeling. In this study, a new rapid technique to measure the nutrients for the olive oil nutritional labeling, is assessed. A novel partial least squares (PLS) calibration model using log-ratio coordinates has been formulated and successfully tested for predicting the percentages of monounsaturated, saturated, and polyunsaturated fatty acids based on visible and near infrared spectroscopy. The model provided accuracy suitable for labeling, under the rules in force in the European Union. The error was generally much lower than the tolerance.

Industrial relevance: The approach here proposed can be a suitable solution for olive oil nutritional labeling, which is a current challenge for the olive oil industry.

Keywords: compositional data; monounsaturated fat; polyunsaturated fat; saturated fat; nutritional labeling; olive oil.

Abbreviations: EVOO, extra virgin olive oils; FAME, fatty acids methyl esters; MUFA, mono-unsaturated fatty acids; OO, current olive oils; PLS, partial least squares; PUFA, polyunsaturated fatty acids; SFA, saturated fatty acids; TSFA, total saturated fatty acids; TUFA, total unsaturated fatty acids; Vis/NIR, visible and near infrared spectroscopy; VOO, virgin olive oils.

1. Introduction

The regulation of the European Union (CE, 2011) settles the duty of food manufacturers to include nutritional information in the product labels. It has been applicable since 13 December 2016. Olive oil results from the extraction of a substance produced by biosynthesis, in contrast to what happens in foods manufactured according to a composition with several ingredients. The practical challenge of nutritional labeling is different in both cases, since it depends on the diversity of their nutritional features. Compulsory information includes energy value, total fat contents, total saturated fatty acids (TSFA), carbohydrates, sugars, proteins and salt. As voluntary nutritional information, the rule considers other nutrients' values such as mono-unsaturated fatty acids (MUFA) and polyunsaturated fatty acids (PUFA), among others. Regarding olive oil, the most common information included up to date in its nutritional label is total fat, saturated fat, monounsaturated fat and polyunsaturated fat. The producers show voluntarily these two last features. However, the olive oil industry has almost generalized their inclusion in the labeling, since they characterize the product showing its nutritional advantages. It is interesting also that the European Food Safety Agency issued scientific opinion report on the healthy properties provided by olive oil polyphenols (EFSA, 2012). Therefore, the nutritional label information on these bioactive compounds could be well appreciated by the consumers.

In olive oil, the total fat comprises practically 100% of the product, since carbohydrates, sugars, proteins and salt are absent. MUFA are those fatty acids which carbon chain have a single unsaturation. The most common example of this type is oleic acid (C18:1). Its unsaturation locates after the number 9 carbon, and commonly called ω-9. Oleic acid is the olive oil major fatty acid, as detailed later on. Palmitoleic acid (C16:1) is the second MUFA of olive oil, generally lower than 1% (García-González, Infante-Domínguez, & Aparicio, 2013[a]). PUFA are those fatty acids containing more than one double bond in their backbone. Good human health requires diets with small quantities of these compounds, such as the essential fatty acids linoleic (C18:2), ω-6, and linolenic (C18:3), this last called ω-3. Saturated fatty acids (SFA) are those without any unsaturation within their chain. Olive oil includes as major SFA palmitic acid (C16:0), in quantities 8-14%, estearic acid (C18:0), 3-6%, margaric acid (C17:0), araquidic acid (C20:0), and behenic acid (C22:0).

MUFA are the most characteristic fatty acids in olive oil because of their high content of oleic acid. This is helpful, since the positive effect of MUFA on cardiovascular health has been widely demonstrated (Schwingshackl and Hoffmann, 2014; Hernáez et al., 2017). The olive oil fatty acids show high variation depending mainly on the variety. The varieties used to produce olive oil in the world are around 100, although there are more than 2000. The proportions of MUFA in an olive oil depends on many agronomic conditions, the major ones being olive variety and climate. Therefore, olive oils with MUFA proportions relatively small, show PUFA or SFA relatively high. In addition to genetics and climate, agronomic conditions influence the diversity of fatty acids. Oleic acid (18:1), which is the major fatty acid of olive oil, ranges from a minimum 60.94% of Cv. Barnea in Argentina to 84.11% of Cv. Picual in New Zealand (García-González, Infante-Domínguez, & Aparicio, 2013[a]). At the same time, palmitic acid (16:0), the major among those olive oil saturated fatty acids, ranges from 8.13% of Cv. Koroneiki in New Zealand to 19.78% of Cv. Arbequina in Argentina. Diversity also exists within the product manufactured by the major operators in the main producing countries, even when considering some cultivars only. As an example, in the main olive oil producer, which is Spain, palmitic acid ranges from 7.86% in Cv. Gordalilla to 12.55% in Cv. Negral, while oleic acid ranges from 66.49% in Cv. Sevillenca to 81.61% in Cv. Gordalilla (García-González, Infante-Domínguez, & Aparicio, 2013[a]). These facts imply that generic nutritional labeling of olive oil would involve a significant risk of error. Besides, the analysis of all batches of olive oil using the official methods is expensive and complicated. Thus, the olive oil industry is seriously concerned about the best solution for nutritional labeling. Rapid and reliable techniques to achieve this purpose may be an alternative solution. Among the various non-destructive techniques that have offered solutions to these needs so far, near infrared spectroscopy (NIRS) stands out for its important achievements. NIR spectroscopy data analysis is based on multivariate models, in which the spectral data correlate with the analyzed characteristic. Several authors (Armenta, Garrigues, & De la Guardia, 2007; Bendini et al., 2007; Cayuela, Moreda & García, 2013) reported the ability of NIRS to analyze the main features of olive oil quality, such as free acidity or the peroxides value. In fact, a growing number of laboratories use NIRS techniques for these routine analyses, although they are still a minority. The possibility of authenticating the olive oil variety or geographical origin (Galtier et al., 2006, among others), as well as detecting adulteration by NIRS (Azizian et al., 2015) have been also reported, in both cases through NIRS analysis of their acidic composition. NIRS offers several important advantages, as it is a fast, non-destructive and potentially multi-parametric method. In addition, NIRS does not need solvents or reagents, therefore avoiding a significant expense and protecting the environment.

Chemometric methods, using traditional multivariate data analysis, are frequently applied to analyze the fatty acid composition of oils and fats with diverse aims. Thus, NIR data analyses of the olive oil fatty acid composition have been reported (Mailer, 2004; Mossoba et al., 2013, among others). However, standard multivariate analysis techniques are formally designed for ordinary unconstrained data, which take values which are directly meaningful and can be compared across samples. Fatty acid profiles of plant oils are instead generally expressed as relative amounts, using percentages respecting the total weight. Thus, the data information is relative and there are intrinsic co-dependence relationships between components. A higher percentage of one type of fatty acid will necessarily imply lower percentage of, at least, one other fatty acid. Specialized theory and methods for this type of data, so-called compositional data, have been developed in the statistical literature (see e.g. Aitchison, 1986, and Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado, 2015). Issues related to compositional data have been discussed recently regarding volatile fatty acids profile of table olives (Garrido et al., 2017; Garrido et al. 2018) and fatty acid composition of pork meat (Ros-Freixedes and Estany, 2014). For a case in which the composition played the role of explanatory variable, Palarea-Albaladejo et al. (2017) developed a compositional mixed model to explain methane production from ruminal volatile fatty acids in cattle, along with other diet and animal covariates. Partial least squares (PLS) analysis involving compositional data was first discussed in chemometrics by Hinkle and Rayens (1995), although it was not done in terms of orthogonal ILR-coordinates since this was a later development introduced by Egozcue et al. (2003). An application of PLS modelling to discriminant analysis (PLS-DA), which treats the metabolomics profiles as compositions via log-ratios, can be found in Kalivodová et al. (2015). However, to our knowledge, there are no studies using PLS modelling under a compositional approach, to predict the fat composition of vegetable oils from NIR spectroscopy through a convenient log-ratio representation. Neither there are studies on the purpose of using NIRS for olive oil nutritional labeling, which requires a compositional approach.

This study proposes a new rapid technique to measure the nutrients required for olive oil nutritional labeling from Vis/NIR data. For this purpose, a novel compositional PLS calibration model has been formulated, in terms of log-ratio coordinates of the percentage fatty acid composition, to suitably deal with its relative scale. This model has been implemented and successfully tested for estimating the percentage composition of PUFA, MUFA and TSFA. The total unsaturated fatty acids (TUFA) was arithmetically determined from PUFA and MUFA.

2. Material and Methods

2.1. Olive Oils

The robustness of NIRS calibrations depends on the statistical range of the analyzed features. Therefore, several sources provided olive oil samples to assure enough diversity. High quality Extra Virgin Olive Oils (EVOO) from special markets contributed with 70 samples. Olive oils normally found in common markets included 56 EVOO, 5 virgin olive oils (VOO) and 40 nonvirgin olive oils (OO). Moreover, 10 pomace olive oils were included along with 45 EVOO from a collaborative industry and other 45 EVOO samples from a separate research project. These were extracted at the Instituto de la Grasa (CSIC) from olives using a laboratory mill (MC2, Seville, Spain) based on the Abencor system (Martínez, Muñoz, Alba, & Lanzón, 1975). In total, 226 samples were used.

2.2. Spectral Acquisition

The temperature of a body has an important influence on the NIR radiation it reflects and absorbs, thus it is decisive in NIRS (Jiang, Xie, Peng, & Yin, 2008). Therefore, the samples were taken from 4 °C storage and placed in the laboratory 18 h before processing. Before recording spectra, a thermostatic bath (Nahita, London, United Kingdom) fixed at 33 °C held the 20 mL sample containers for 30 min., until temperature stability was reached.

The spectrum of every sample was acquired with the spectrometer Labspec (Analytical Spectral Devices Inc., Boulder). Labspec is equipped with three detectors. The detector for the visible range (350-1000 nm) is a fixed reflective holographic diode array with a sensitivity of 512 pixels. A holographic fast scanner InGaAs detector cooled at -25 ºC covers the wavelength range of 1000-1800 nm. This coupled with a high order blocking filter runs for the 1800-2500 nm interval. The instrument equips internal shutters and automatic offset correction, the scanning speed is 100 ms. The repeatability of the instrument, expressed as standard deviation on the average absorbance of five measures of a white tile between 350 and 2500 nm, is 6.00 10-4 cm-1 mol-1. Using the Labspec, the spectra were registered by transmittance from each sample of VOO directly, without any other treatment. A Hellma quartz spectrophotometric cuvette with 10 mm path length held the samples while their averaged spectra were acquired. The whole spectrum Vis/NIR (350−2500 nm) was registered, each spectral variable matching to a 1 nm interval. Configuration for 50 spectra in continuous acquisition was used, each spectral variable matching to 1 nm interval. Indico Pro software (Analytical Spectral Devices Inc., Boulder, Colorado, USA) was used for this purpose. The registering time was less than a minute for each sample spectrum, all steps included.

2.3. Reference Analysis

The fatty acids compositions were analyzed by gas chromatography (GC) as fatty acid methyl esters (FAME), according to the IUPAC Standard Method (IUPAC, 1987), at the Instituto de la Grasa (CSIC). Briefly, 50 mg of olive oil were dissolved in 2 mL heptane and then transesterified using 300 μL 2 N methanolic potassium hydroxide solution. After decanting, the supernatant was collected. GC analysis was carried out using an Agilent 7697A gas chromatograph (Agilent Technologies, Santa Clara) equipped with a capillary column (poly (90% biscyanopropyl–10% cyanopropylphenyl) siloxane, 60 mÅ, 0.25 mm Φ_i, and 0.20 µm film thickness). Automatic split injection and a flame ionization detector (FID) were used. The carrier gas was hydrogen at a flow rate of 1 mL min^{-1}. The temperatures of the injector and detector were 225 and 250°C, respectively. The oven was programmed at a temperature of 180 °C (10 min), which was then increased 3 °C min^{-1} up to 220 °C (10 min). The injection volume was 1 µL. The fatty acid composition was expressed as percentage of each fatty acid in total fatty acids.

The MUFA, PUFA, TUFA and TSFA percentages were arithmetically calculated from the analyzed fatty acids values. Thus, MUFA was the sum of percentages of the fatty acids palmitoleic (C16:1), heptadecenoic (C17:1), oleic (C18:1) and eicosenoic (C20:1). PUFA was the sum of percentages of the fatty acids linoleic (C18:2) and linolenic (C18:3). TUFA was the sum of percentages of MUFA and PUFA. TSFA was the sum of percentages of the fatty acids palmitic (C16:0), estearic (C18:0), margaric (C17:0), araquidic (C20:0), and behenic (C22:0).

2.4. Principal Component Analysis of the Vis/NIR data

The absorbance data of the whole spectra were pre-treated by mean normalization and Savitzsky-Golay first derivative, with polynomial order 2 and smoothing point 3. The suitability of this treatment has been previously reported (Cayuela et al., 2015). The NIR and Vis/NIR spectral data of the analyzed olive oil samples were reduced by principal component analysis (PCA). This statistical technique projected the data onto low dimensions by computing optimal linear combinations (principal components, PCs) of the measured absorbances across wavelengths. In particular, the two first principal components defined dimensions accounting for the highest percentage of the total variability in the original data and were used to visualize the olive oil samples in an ordinary scatter plot.

2.5. Compositional modelling of fatty acid percentage profiles

Compositional data stand for all kinds of multivariate data representing parts of some whole and, thus, carrying only relative information. This implies that values in each part have meaning only in relation to the other parts. Percentage fatty acid compositions, consisting of mutually exclusive fatty acid categories and expressed as percentages of total fatty acids, correspond to this definition. Percentage compositions are formally defined on a simplex, a constrained subset of the real space formed by vectors of positive values adding up to 100. Compositional data bring some difficulties in relation to the most basic elements of data analysis and modelling like correlations, distances, etc., which are defined according to the geometry of the ordinary real space. It has been shown that the direct use of standard statistical and chemometrics tools on them can introduce artifacts like negative bias in correlation measures, singularity of the covariance matrix, predictions beyond the range of possible values (e.g. the interval [0, 100] in our case) and results which depend on the units of measurement. Obviously, these issues can potentially lead to misleading scientific conclusions. A principled methodology based on using log-ratios between parts of the composition was introduced in the seminal work by Aitchison (1986) and further developed thereof. A key point is that all the relative information in a composition is contained in the ratios between its components. Importantly, working with ratios also guarantees that results do not depend on the scale of measurement of the data. Taking logs of the ratios is mathematically convenient and maps the data onto the real space, where ordinary statistical methods, models and graphs can be used on log-ratio coordinates (Aitchison, 1986; Van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado, 2015).

2.5.1. PLS regression modeling on log-ratio coordinates

According to the above characterization, PLS modelling was based on log-ratio coordinates involving the three fatty acid (FA) categories used as reference, MUFA, PUFA, and TSFA. In particular, we employed an isometric log-ratio (ILR) representation (Egozcue et al., 2003) of the 3-part FA composition, by which its information is projected onto real space by way of two orthogonal coordinates as follows:

$$\mathrm{ILR}_1 = \sqrt{\frac{2}{3}}\ln\frac{\mathrm{MUFA}}{\sqrt{\mathrm{PUFA}\cdot\mathrm{TSFA}}} \qquad \text{and} \qquad \mathrm{ILR}_2 = \sqrt{\frac{1}{2}}\ln\frac{\mathrm{PUFA}}{\mathrm{TSFA}}. \qquad [1]$$

Note that it is possible to define alternative ILR representations, but they all are orthogonal rotations of each other and lead to the same results in terms of the original composition. An ILR-coordinate roughly accounts for the relative importance of some components (in the numerator of the log-ratio) with respect to others (in the denominator). The reduction from three to two dimensions after the ILR transformation is coherent with the actual degrees of freedom of the FA composition, we only need any two components to determine the third. Multivariate PLS regression was conducted using the two ILR-coordinates of the FA composition as response and the Vis/NIR spectra as predictors. Predictions obtained in ILR coordinates were then transformed back into the corresponding predicted FA percentages by inverse ILR transformation. After this, predicted TUFA was obtained by adding predicted percentages of MUFA and PUFA.

A selection of best Vis/NIR spectral variables was conducted prior to multivariate PLS calibration to minimize prediction error using the genetic search algorithm (Hasegawa et al. 1997; Mehmood et al., 2012). The PLS calibration model was fitted by the kernel algorithm to predict the FA ILR-coordinates from the selected (51 out of 237) Vis/NIR spectral variables (scaled by standard deviation). The optimal number of PLS latent components used (10 latent components) was determined by 5-time repeated 10-fold cross validation aiming to minimize the root mean square error of prediction (RMSEP) and maximize the coefficient of determination (R^2) as model performance measures. The prediction performance of the final joint PLS model was evaluated by RMSE and R^2 based a partition of the data into a calibration data set of 75% of the data, used to tune and estimate the model as well as to assess performance using 5-time repeated 10-fold cross-validation, and a test set of 25% of the data.

The prediction performance of the PLS model for the entire FA composition as a whole was assessed by an overall R^2, computed as the following formula:

$$1 - \frac{\text{totvar(ILR residuals)}}{\text{totvar(observed FA)}} \qquad [2]$$

Where totvar, so-called total or metric variance, was obtained as the trace of the covariance matrix of, respectively, the ILR residuals matrix and the observed FA data in ILR-coordinates (ILR FA). Moreover, the metric standard deviation (MSD) of the ILR residuals, obtained as follows:

$$\sqrt{1/(\mathrm{D}-1)\cdot \text{totvar(ILR residuals)}}, \qquad [3]$$

In this case, $D = 3$ was computed. This last statistic provided an overall dispersion measure of the model residuals analogous to RMSE (Van den Boogaart and Tolosana-Delgado, 2013). These statistics were obtained from calibration, cross-validation and test data. For the purpose of comparison with official measurement error tolerance guidelines, analysis of the residuals for each FA category separately was conducted from the cross-validation and test data sets by computing the correlation between predicted and reference values and the mean percent deviation of predictions with respect to the reference data. These differences were also visualized for individual test samples in a scatter plot along with the official error tolerance limits for reference.

All the data analyses and modelling described above were conducted on the R system for statistical computing v3.4 (R Core Team, 2017).

3. Results

3.1. Olive Oil Spectra

The major near-infrared absorption bands of olive oil have been described by Hourant, Baeten, Morales, Meurens, & Aparicio (2000). Near-infrared spectra show various overlapping bands, because their first and second overtones and a combination of fundamental vibrations, mainly carbon–hydrogen (Shenk, Workman, & Westerhaus, 2001). A broad absorbance band exists around 1220 nm, probably due to second overtones of C–H and CH=CH– stretching vibrations from oil. There is other high intensity area related to the C-H first overtone at 1700 nm (García-González, Infante-Domínguez, & Aparicio, 2013[b]), and a combination band at 1880–2100 nm. A high intensity absorbance peak occurs about 2300 nm, caused by a combination of fundamental vibrations from the C-H groups (Hourant, Baeten, Morales, Meurens, & Aparicio, 2000). Besides, the major visible absorption bands of olive oil were made by Moyano, Meléndez, Alba, & Heredia (2008).

Olive oil spectra from the samples analyzed in this work, shown in Fig. 1, agree with the previously indicated reports. A first minor peak occurs next to 415 nm. This area suits to the wavelengths of oil absorption for dark blue colored light. It could be due mainly to carotenoids, as well to pheophytin A, pheophorbide A and pyropheophytin A. A second peak is near 450 nm, matching to blue light absorption, which is characteristic of carotenoids. A third peak appears around at 670 nm, which coincides with chlorophylls absorption (Moyano, Meléndez, Alba, & Heredia, 2008). The high intensity area related to the C-H first overtone at 1700 nm can be seen clearly, as well as the combination band at 1880–2100 nm and the high intensity absorbance peak at 2300 nm, from the combination of fundamental vibrations of the C-H groups.

Fig. 1

3.2. Fatty Acids Characterization

A preliminary exploration of the FA data revealed a very atypical percentage composition of MUFA, PUFA, and TSFA (44.75%, 3.82%, 51.42%) of a commercial sample with registered data, supposedly of olive oil and type ‘acidity lower to 1%’. It was atypical particularly in relation to the relative weight of TSFA (51.42%, whereas for the other samples this was around 16%), thus the possibility of this corresponding to a case of fraud cannot be discarded, and it was left out of the analysis.

Ordinary univariate descriptive statistics of the percentage MUFA, PUFA, TUFA and TSFA in the olive oil samples used in this study are shown in Table 1 for reference. The TUFA ranged from 76.7% to 88.3%, while MUFA ranged from 57.8% to 82.4%, PUFA from 3.1% to 20.2% and TSFA from 11.7% to 23.3%. The most important fatty acid category in olive oil is TUFA, with MUFA in particular being the main contributor in mean (74.60%). The highest variation relative to mean values was shown by PUFA ($C_v = 50.12$). Note that, given the compositional nature of the data, ordinary univariate statistics of central tendency and variability for different FA categories are interrelated and are not considering their particular geometry. Thus, one must interpret them with caution (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado, 2015).

Table 1

3.3. PCA Analysis

A scatter plot based on the two first dimensions obtained from PCA analysis of the olive oil spectral data is shown in Fig. 2. These two first PCs retained 77.5% of the original data variability. Note that a certain 2-group structure can be appreciated along the horizontal axis (first PC) in the graph. It was checked that these two groups corresponded to olive oil samples separated by a MUFA content threshold at 70%. The largest group, with 180 olive oils, corresponded to MUFA greater than 70%. The remaining 52 samples had MUFA less than 70%, 41 of them corresponding to Arbequina olive oils from super-intensive crop system obtained in a research project, 4 to commercial gourmet quality EVOO, 1 to industrial EVOO, 5 to commercial VOO and 1 to commercial OO samples.

A 95% concentration ellipse was estimated to help with the visual identification of outlying spectra. The 9 samples falling beyond the boundaries of the ellipse were identified and not considered for the subsequent analysis. They corresponded to 4 industrial EVOO, 1 commercial EVOO, 1 commercial OO and 3 EVOO from an independent research project. Interestingly, note that 7 out of these 9 outlying spectra corresponded with industrial and research samples. It is frequent with this type of samples to find oils with a higher moisture content, despite having been filtered as the rest ones, which differentiates their spectrum from the other samples with normal moisture content. Although it is not possible to provide moisture content data, since this parameter was not analyzed, we consider that this was the reason why most of these samples were atypical. In the case of the two commercial samples, their spectra may be defective due to methodological factors in their registering process. Hence, we eventually worked with a data set consisting of 223 samples. For each one, we had the basic 3-part FA composition and NIR data along 237 spectral windows. This data set was randomly partitioned into calibration set (75% data, 168 samples) and test set (25%, 55 samples) for subsequent PLS regression analysis.

Fig. 2

3.4. Compositional PLS model on log-ratio coordinates

Figure 3 displays the results from the fitted PLS model for each of the two ILR-coordinates of the FA composition as detailed in Eq. [1]. Figures 3*a* and 3*b* show the respective PLS regression coefficients plots using the pre-selected 51 best Vis/NIR spectral variables. Figures 3*c* and 3*d* show the corresponding observed versus predicted plots. The associated model performance statistics are summarized in Table 2. The most parsimonious model amongst those reaching comparable highest performance following the one-standard error rule (Kuhn and Johnson, 2013) used 10 latent components (see Supplementary File 1). The individual ILR_1 and ILR_2 models provided R^2 equal to 0.95 and 0.90 respectively based on the calibration data (denoted R^2_c). The corresponding cross-validated values R^2_{cv} were 0.92 and 0.83 respectively; with RPDs equal to 3.53 and 2.43 respectively. The coefficients of determination from the test data set, R^2_t, were 0.93 and 0.86 for ILR-coordinates ILR_1 and ILR_2 respectively. Table 2 also includes the calibration, cross-validation and test data based RMSE values of up to 0.10.

Fig. 3

Table 2

3.5. Overall model performance for predicting the FA composition

Predictions from the fitted PLS models on ILR-coordinates were conveniently transformed back to be expressed in terms of the entire 3-part FA percentage composition. We obtained an overall calibration R^2, which accounted for variation in the FA composition as a whole explained by the model, and MSD, which accounted for dispersion in model residuals. They were equal to 0.93 and 0.07 respectively (Table 2). The cross-validated and test data set counterparts were 0.90 and 0.09 respectively in both cases (Table 2). Supplementary File 2 includes the reference and predicted values for the test data set expressed both in ILR-coordinates and in terms of the entire FA percentage composition by ILR back-transformation. Figure 4 illustrates the performance of the model by showing predicted (open triangles) versus reference observed (open circles) FA compositions on a ternary diagram. The axes on the sides of the triangle correspond with MUFA (left), PUFA (right) and TSFA (bottom) percentage contents. The closer a point is to a vertex the higher the relative importance of the corresponding FA in the sample. The region where the data were concentrated was zoomed in for better visualization. The mean FA composition was included for reference (solid square).

Fig. 4

3.6. Assessment of model residuals by FA category

For each individual FA category, Table 3 provides cross-validated and test data based correlation coefficients (*r*) between predicted and observed percentage contents and average percent deviation (% deviation) of predicted with respect to observed percentage content, including results for TUFA as obtained by aggregation of MUFA and PUFA. These measures were useful for the assessment of the results according to current guidance for olive oil nutritional labeling in the European Union, namely in relation to measurement error tolerance which is set at ±20%. The correlation coefficients for MUFA and PUFA were over 0.95 for both cross-validated and test data. For TUFA and TSFA, they were around 0.9. PUFA showed the highest cross-validated average percent deviation (9.61%), whereas for MUFA and TUFA it was close to 1%. A comparable pattern was observed based on test data (Table 3). Figure 5 compares predicted and reference test values for each FA percentage individually, including exact prediction line (in grey) and ±20% tolerance limits (in red) for reference. Predicted values falling beyond the tolerance limits were obtained for TUFA and TSFA in very few isolated samples. They were associated with the lowest percentage contents. Note however that, according to the conceptualization of the FA percentage composition as a whole with values conveying only relative information, these individual statistics and graphical representations are not fully independent from one another and overall measures of performance as provided in Section 3.5 would be preferable.

Table 3

Fig. 5

4. Discussion

The assessment of the performance of the compositional PLS model based on either calibration, cross-validation or test data provided R^2s over 0.9 and RMSEs below 0.1. The obtained differences between predicted and reference FA percentage compositions strongly support the possibility of conducting highly accurate predictions of the FA composition of olive oil samples from Vis/NIR spectroscopy data. Among them, MUFA is the most important category in terms of its relative abundance and also due to its nutritional benefits for human health (García-González, Infante-Domínguez, & Aparicio, 2013; Schwingshackl & Hoffmann, 2014).

The tolerances considered for the olive oil nutritional labeling have been, up to date, detailed in a guidance document only (CE, 2012), which compliance is not compulsory. When the nutritional component is present in less than 4g per 100g, the tolerance is ± 0.8g, whereas when it is present in more than 4g per 100g, the tolerance is ±20%, including measurement uncertainty in both cases. In this study, none of the features analyzed showed mean percentage lower than 4%, thus ±20% tolerance is applicable. Our results show expected percent deviations far within these tolerance limits, with PUFA showing the highest deviation (average deviation of 9.61% from cross-validated data and of 9.59% from test data, Table 3). This agrees with the higher variation coefficient of PUFA shown in Table 1. The predictions for TUFA, as sum of MUFA and PUFA, also satisfied these tolerance limits.

5. Conclusions

The results of this study show that rapid Vis/NIR spectroscopy combined with sensible chemometric modelling can be used for accurate determination of the components required for olive oil nutritional labeling. Measuring the percentages of monounsaturated fatty acids, polyunsaturated fatty acids, and saturated fatty acids, provided accuracy suitable for labeling under the rules in force in the European Union. The data modelling conducted took into account the intrinsic relative and inter-dependent nature of percentage fatty acid compositions. The measured error was generally much lower than the tolerance indicated in European Union guidance documentation, providing then a wide margin of safety. Thus, the approach here proposed can be a suitable solution for olive oil nutritional labeling, which is a current challenge for the olive oil industry.

Acknowledgements

J. A. Cayuela expresses gratitude to the European Regional Development Fund, aswell to the Ministry of Economy and Competitiveness of Spain for funding parts of this study, which was developed within the project Recupera 2020 1.4.4., and to the Spanish Council for Scientific Research, for co-funding the grant 20137R065 corresponding to this project. J. Palarea-Albaladejo was partly supported by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS) and by the Spanish Ministry of Economy and Competitiveness under the project CODA-RETOS MTM2015-65016-C2-1(2)-R.

Conflict of interests

The authors declare no competing interests.

References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Chapman and Hall, London,

UK.

Armenta, S., Garrigues, S., & De la Guardia, M. (2007). Determination of edible oil parameters by near infrared spectrometry. *Analitical Chemical Acta, 596,* 330−337. http://dx.doi.org/10.1016/j.aca.2007.06.028

Azizian, H., Mossoba, M. M., Fardin-Kia, A. R., Delmonte, P., Karunathilaka, S. R., & Kramer, J. K. G. (2015). Novel, rapid identification, and quantification of adulterants in extra virgin olive oil using near-infrared spectroscopy and chemometrics. *Lipids, 50*(7), 705–718. http://dx.doi.org/10.1007/s11745-015-4038-4

Bendini, A., Cerretani, L., Di Virgilio, F., Belloni, P., Lercker, G., & Gallina-Toschi, T. (2007). In-process monitoring in industrial olive mill by means of FT-NIR. *European Journal of Lipid Science and Technology, 109,* 498−504. http://dx.doi.org/10.1002/ejlt.200700001

Cayuela J.A., Moreda W., García J.M. (2013). Rapid Determination of Olive Oil Oxidative Stability and Its Major Quality Parameters Using Vis/NIR Transmittance Spectroscopy. *J. Agric. Food Chem. 61,* 8056−8062. http://dx.doi.org/10.1021/jf4021575

Cayuela, J.A., García, J.F., Moreda, W., Pérez, M.C. 2015. Characterization of some olive oil quality aspects by NIRS analysis of its fatty acids and triglycerides. Poster. 7th Symposium on Recent Advances in Food Analysis. Praga, Czech Republic.

CE (2011). Regulation (EU) No 1169/2011 of the European Parliament and of the Council of 25 October 2011 on the provision of food information to consumers. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32011R1169&from=EN. Last accessed June 2017.

CE (2012). Orientation document for the authorities competent in the control of the compliance with EU legislation on the Regulation (EU) No 1169/2011 on nutrition labeling of foodstuffs. https://ec.europa.eu/food/sites/food/files/safety/docs/labelling_nutrition-vitamins_minerals-guidance_tolerances_1212_en.pdf. Last accessed June 2017.

EFSA (2012). Scientific Opinion on the substantiation of a health claim related to polyphenols in olive and maintenance of normal blood HDL-cholesterol concentrations pursuant to Article 13(1) of Regulation (EC) No 1924/2006. *EFSA Journal* 10(8), 2848.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C. (2003). Isometric Log-ratio Transformations for Compositional Data Analysis. *Mathematical Geology, 35*(3), 279-300.

Galtier, O.; Dupuy, N.; Le Dreau, Y.; Ollivier, D.; Pinatec, C.; Kister, J.; Artaud, J. (2006). Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra. *Analitical Chemical Acta, 595,* 136-144. https://doi.org/10.1016/j.aca.2007.02.033

García-González, D. L., Infante-Domínguez, C., Aparicio, R. (2013^{a}). Tables of Olive Oil Chemical Data. In R. Aparicio, & J. Harwood (Eds.), Handbook of olive oil: Analysis and properties (pp. 739-768). New York: Springer.

Garrido-Fernández, A., Montaño, A., Sánchez-Gómez, A. H., Cortés-Delgado, A., López-López, A. (2017). Volatile profiles of green Spanish-style table olives: Application of compositional data analysis for the segregation of their cultivars and production areas. *Talanta, 169,* 77-84. http://dx.doi.org/10.1016/j.talanta.2017.03.066

Garrido-Fernández, A., Cortés-Delgado, A., López-López, A. (2018). Tentative application of compositional data analysis to the fatty acid profiles of green Spanish-style Gordal table olives. *Food Chemistry, 241,* 14-22. http://dx.doi.org/10.1016/j.foodchem.2017.08.064

Hasegawa, Y. Miyashita, K. Funatsu. (1997). GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *Journal of Chemical Information and Computer Sciences, 37,* 306-310. http://dx.doi.org/10.1021/ci960047x

Hernáez, A., Castañer, O., Goday, A., Ros, E., Pintó, X., Estruch, R., Salas-Salvadó, J., Corella, D., Arós, F., Serra-Majem, L., Martínez-González, M. A., Fiol, M., Lapetra, J., De la Torre, R., López-Sabater, M.C., Fitó, M. (2017). The Mediterranean Diet decreases LDL atherogenicity in high cardiovascular risk individuals: a randomized controlled trial. *Molecular Nutritton & Food Researche, 61*(9), 1601015, 9 pp. http://dx.doi.org/10.1002/mnfr.201601015

Hinkle, J., W. Rayens. (1995). Partial least squares and compositional data: problems and alternatives. *Chemometrics and Intelligent Laboratory Systems, 20,* 159-172. https://doi.org/10.1016/0169-7439(95)00062-3

Hourant, P., Baeten, V., Morales, M. T., Meurens, M., & Aparicio, R. (2000). Oil and fat classification by selected bands of near-infrared spectroscopy. *Applied Spectroscopy, 54,* 1168−1174.

IUPAC (1987). Standard Method 2.302. Standard methods for the analysis of oils, fats and derivatives. Determination of FAMES by capillary GC. Blackwell Scientific: Oxford, Great Britain.

Jiang, H.Y., Xie, L.J., Peng, Y.S., Yin, Y.B. (2008). Study on the influence of temperature on near infrared spectra. *Guang Pu Xue Yu Guang Pu Fen Xi, 28*(7), 1510-1513.

Kalivodová, A., Hron, K., Filzmoser, P., Najdekr, L., Janečková, H., Adam, T. (2015). PLS-DA for compositional data with application to metabolomics. *Journal of Chemometrics, 29*, 21–28. http://dx.doi.org/10.1002/cem.2657

Kuhn, M., Johnson, K. (2013). *Applied Predictive Modelling*. New York, Springer. http://dx.doi.org/10.1007/978-1-4614-6849-3

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

Martínez, J.M., Muñoz, E., Alba, J., Lanzón, A. (1975). Report on the use of the Abencor olive oil yields analyser. *Grasas y Aceites, 26,* 379-385.

Mehmood, K.H. Liland, L. Snipen, S. Sæbø. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems, 118,* 62-69. https://dx.doi.org/10.1016/j.chemolab.2012.07.010

Mossoba, M. M.; Azizian, H.; Tyburczy, C.; Kramer, J. K. G.; Delmonte, P.; Kia, A. R. F.; Rader, J. I. (2013). Rapid FT-NIR Analysis of Edible Oils for Total SFA, MUFA, PUFA, and Trans FA with Comparison to GC. Journal of the American Oils Chemists Society *90*(6), 757-770. https://dx.doi.org/10.1007/s11746-013-2234-z

Moyano, M. J., Meléndez, A. J.; Alba, J., & Heredia, F. J. (2008). A comprehensive study on the colour of virgin olive oils and its relationship with their chlorophylls and carotenoids indexes (I): CIEXYZ non-uniform colour space. *Food Research International, 41,* 505−512. https://doi.org/10.1016/j.foodres.2008.03.007

Palarea-Albaladejo, J., Rooke, J. A., Nevison, I. M., Dewhurst, R. J. (2017). Compositional mixed modeling of methane emissions and ruminal volatile fatty acids from individual cattle and multiple experiments. *Journal of Animal Science, 95,* 2467-2480. https://doi.org/10.2527/jas2016.1339

Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Wiley & Sons, Chichester, UK. https://doi.org/10.1002/9781119003144.ch1

R Core Team. (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Ros-Freixedes, R., and J. Estany. (2014). On the compositional analysis of fatty acids in pork. *The Journal of Agricultural, Biological and Environmental Statistics, 19,* 136–155. https:/doi.org/10.1007/s13253-013-0162-x

Schwingshackl, L., Hoffmann, G. (2014). Monounsaturated fatty acids, olive oil and health status: a systematic review and meta-analysis of cohort studies. *Lipids in Health and Disease, 13*, 154. https://doi.org/10.1186/1476-511X-13-154

Shenk, J. S.; Workman, J. J.; Westerhaus, M. O. (2001). Application of NIR spectroscopy to agricultural products. In: D. A. Burns, and C. W. Ciurcak (Eds.), Handbook of Near Infrared Analysis, 2nd Edition (pp. 419−474). New York: Marcel Dekker.

Van den Boogaart, K. G.; Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*. Springer, Heidelberg, Germany. https:/doi.org/10.1007/978-3-642-36809-7

Figure captions

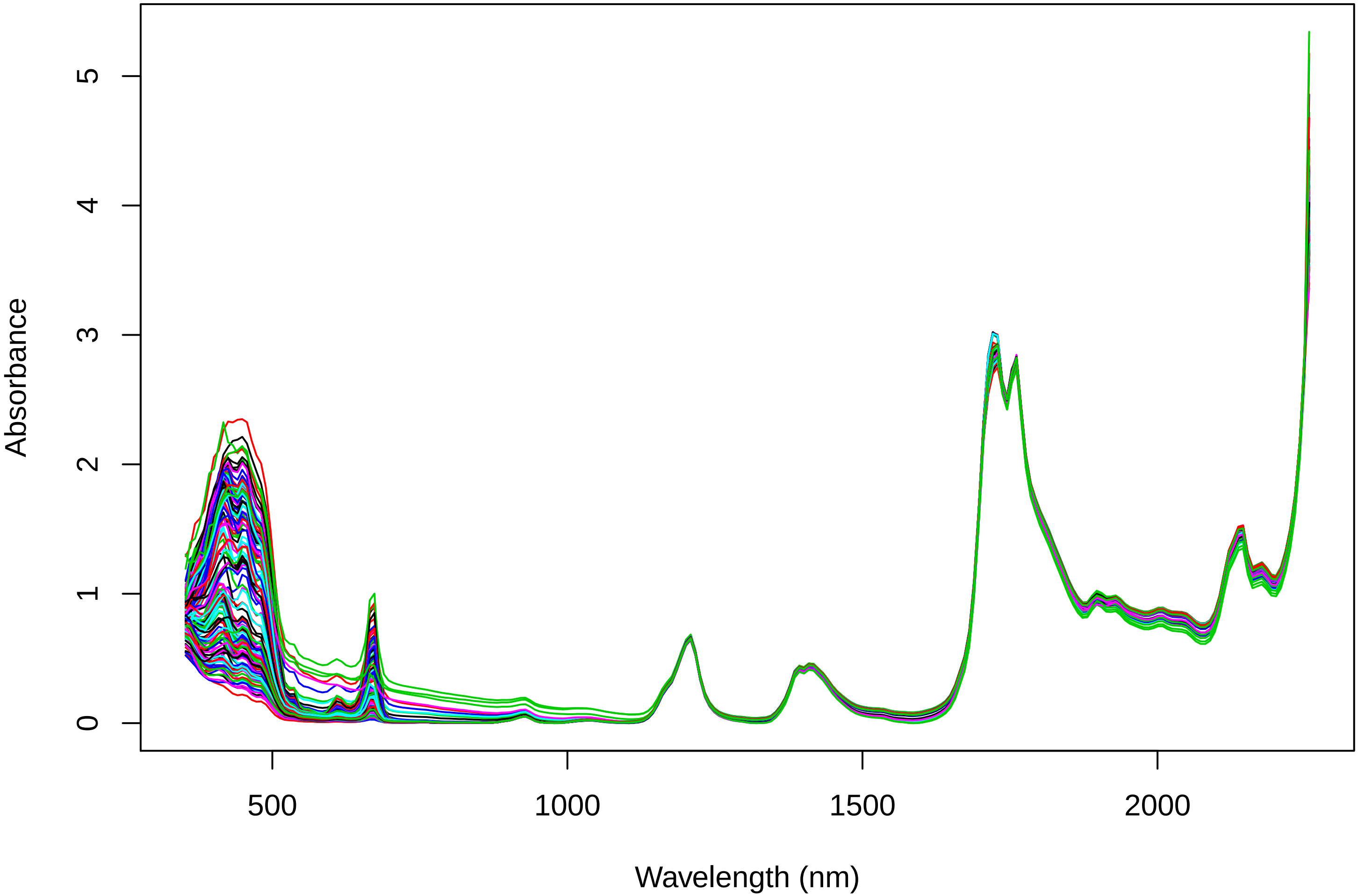
Figure 1. Vis/NIR spectra of the olive oil samples analyzed.

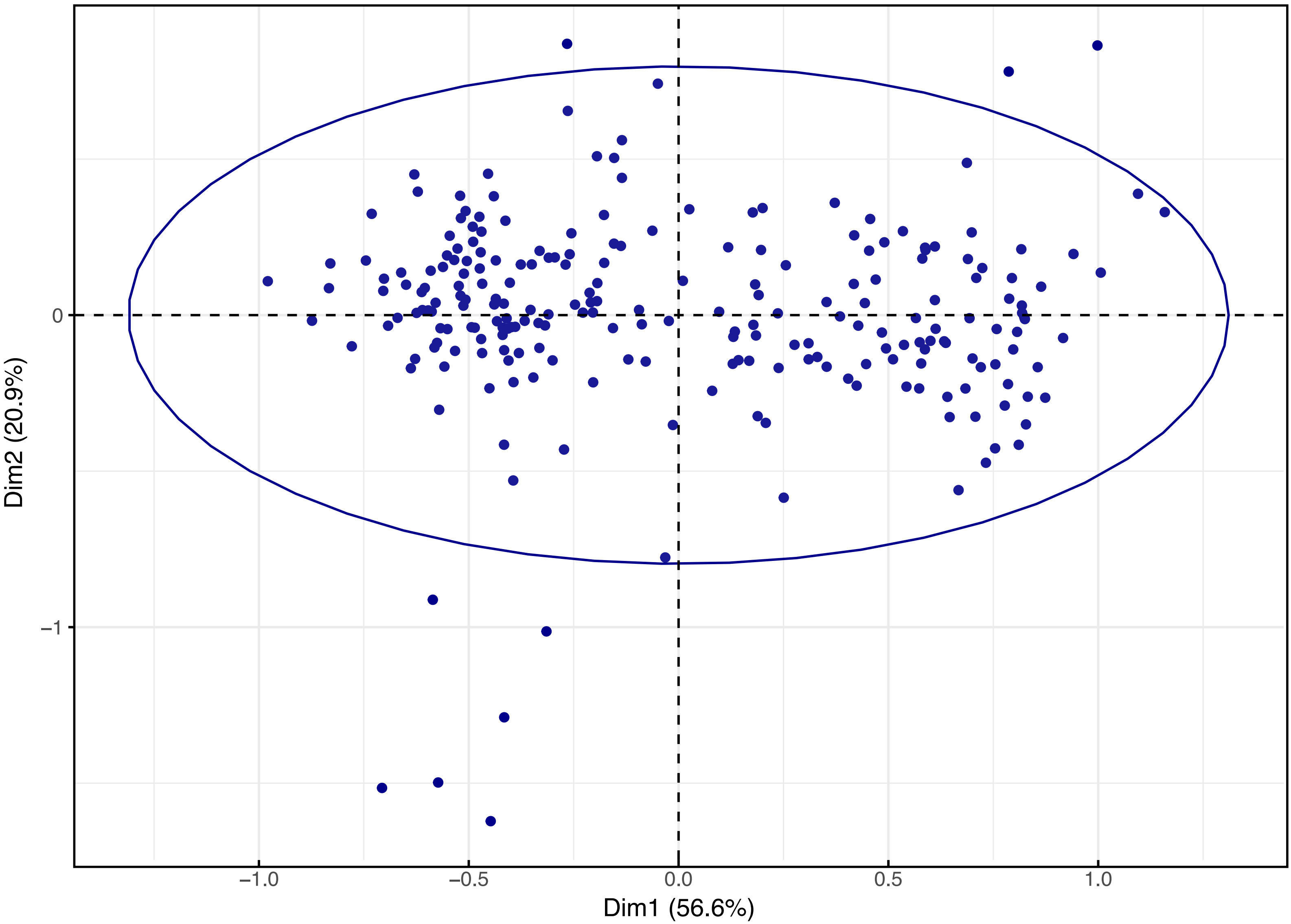
Figure 2. Principal component analysis of olive oil Vis/NIR spectral data (first PC on the horizontal axis and second PC on the vertical axis).

Figure 3. Compositional PLS model results: PLS regression coefficient estimates of individual models for the first (a) and second (b) ILR-coordinates of the FA composition and corresponding predicted versus observed plots (c) and (d) respectively.

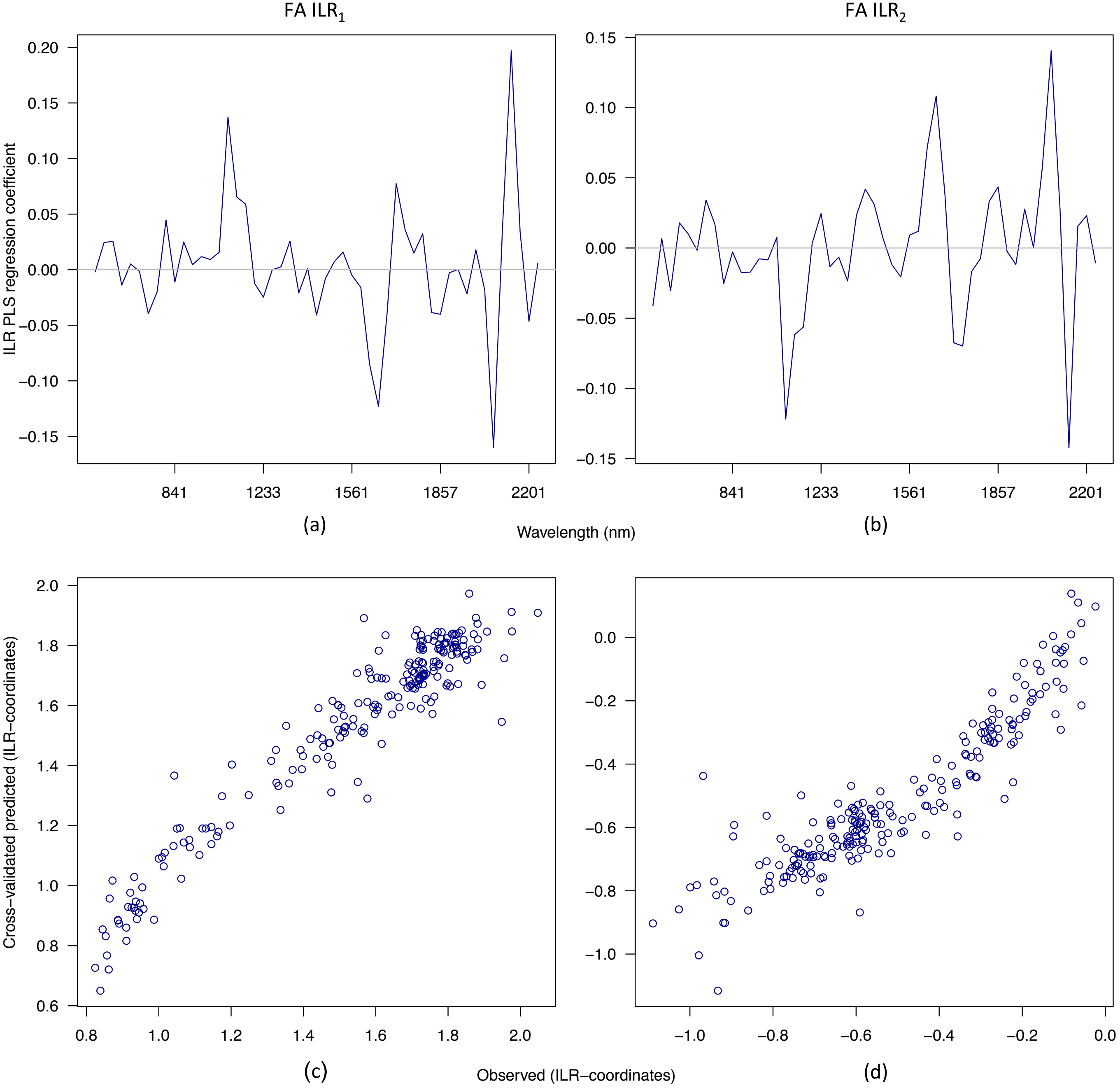
Figure 4. Ternary plot of the predicted and observed FA percentage compositions from the fitted compositional PLS model.

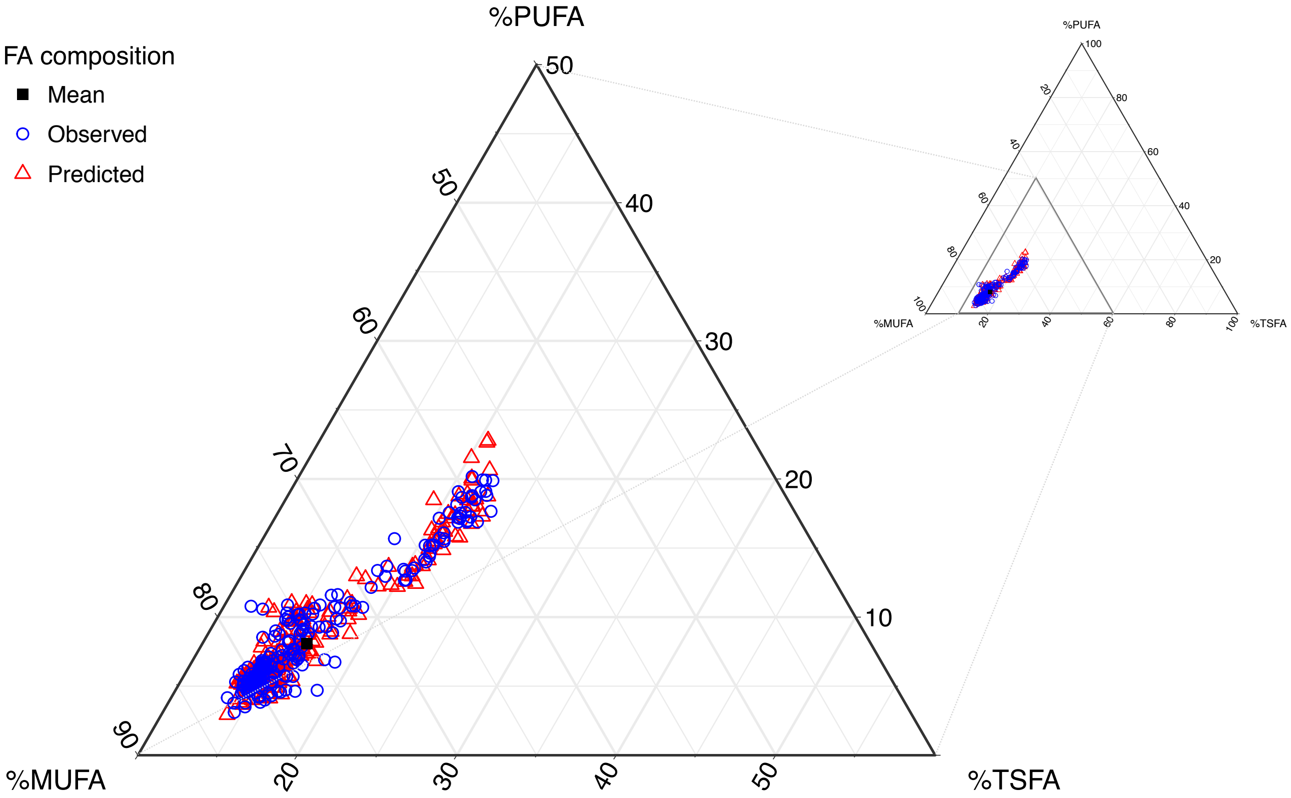
Figure 5. Predicted and observed percentage contents for individual FA categories based on test data (including ±20% tolerance limits according to European Union guidance).


Absorbance
0
1
2
3
4
5
500
1000
1500
2000
Wavelength (nm)



0
−1
−1.0
−0.5
0.0
0.5
1.0
Dim1 (56.6%)
Dim2 (20.9%)


FA ILR$_1$
FA ILR$_2$
ILR PLS regression coefficient
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
841
1233
1561
1857
2201
(a)
(b)
Wavelength (nm)
Cross−validated predicted (ILR−coordinates)
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
(c)
(d)
Observed (ILR−coordinates)


FA composition
■ Mean
○ Observed
△ Predicted
%PUFA
50
40
30
20
10
50
60
70
80
90
%MUFA
20
30
40
50
%TSFA
%PUFA
100
80
60
40
20
20
40
60
80
100
%MUFA
20
40
60
80
100
%TSFA