Multiblock chemometrics for the discrimination of three extra virgin olive oil varieties

Astrid Maléchaux, Sonda Laroussi-Mezghani, Yveline Le Dréau, Jacques Artaud, Nathalie Dupuy*

Aix Marseille Univ, Univ Avignon, CNRS, IRD, IMBE, Marseille, France

Keywords:
Cultivars
Data fusion
MB-PLS-DA
Gas chromatography
Fatty acids
Mid-infrared

To discriminate samples from three varieties of Tunisian extra virgin olive oils, weighted and non-weighted multiblock partial least squares – discriminant analysis (MB-PLS1-DA) models were compared to PLS1-DA models using data obtained by gas chromatography (GC), or global composition through mid-infrared spectra (MIR). Models performances were determined using percentages of sensitivity, specificity and total correct classification. The choice of threshold level for the interpretation of PLS1-DA results was considered. PLS1-DA models using GC data gave better results than those using MIR data. Even with the most conservative threshold, PLS1-DA on GC data allowed very good predictions for *Chemlali* variety (99% correct classification), but had more difficulty to discriminate *Chetoui* and *Oueslati* samples (95% and 84% correct classification respectively). Non-weighted MB-PLS1-DA models benefiting from the synergy between the two sources of data were more discriminative than simple PLS1-DA, yielding better prediction for *Chetoui* and *Oueslati* varieties (98% and 90% correct classification respectively).

1. Introduction

Olive oil is known for displaying health-promoting effects that depend, among other factors, on its cultivar. Therefore, olive oil authentication has been a growing concern for consumers for many years. As a result, food fraud is a major challenge for both regulatory agencies and producers, as it can negatively impact consumer trust and cause important losses of revenue (Charlebois et al., 2016). High-value products benefiting from quality or origin certifications are an especially attractive target for fraudsters. This is the case of extra virgin olive oils (EVOO) with a Protected Designation of Origin (PDO), which must comply with defined specifications regarding their varietal and geographic origins. Studies aiming to determine the compliance of an EVOO with a reference constituted by the characteristics of a cultivar or the specifications of a PDO can be divided into two main categories. In the first one, samples are treated in order to determine their composition in specific constituents such as triacylglycerols, fatty acids, sterols, volatile compounds, etc (Tena, Wang, Aparicio-Ruiz, García-González, & Aparicio, 2015). The second approach is based on spectroscopic analyses requiring no sample treatment, namely 1H and 13C nuclear magnetic resonance (NMR) (Dais & Hatzakis, 2013), near infrared (NIR), mid infrared (MIR) and Raman spectroscopies (Nenadis & Tsimidou 2017), or fluorescence spectroscopy (Guzmán, Baeten, Pierna, & García-Mesa, 2015).

Furthermore, over the past few decades, the increasing amount of data available from more and more sophisticated analytical techniques, associated with the improvement of computational power allowing to treat this information with multivariate statistical analyses, has spurred the development of methods capable of simultaneously analysing several blocks of data (Brereton et al., 2017). In the field of food chemistry, data can be obtained from different techniques such as electronic sensors, mass spectrometry, gas or liquid chromatography or vibrational spectroscopy. Combining information from complementary analyses can be a way to obtain more reliable classification and prediction results (Callao & Ruisánchez, 2018). In this regard, three types of data fusion strategies are described in a recent review by Borràs et al. (2015): low-level with simple concatenation of data, mid-level using hierarchical models introduced by Wold et al. (1996) or multiblock models developed by Wangen and Kowalski (1989), and high-level combining results from separate models to provide a final prediction using probability estimations.

Most of the articles applying chemometrics to olive oil authenticity consider each analytical technique separately (Gómez-Caravaca, Maggio, & Cerretani, 2016), but to this day few studies have applied data fusion to the discrimination of EVOO origin and even fewer have combined data from spectroscopic and chromatographic analyses

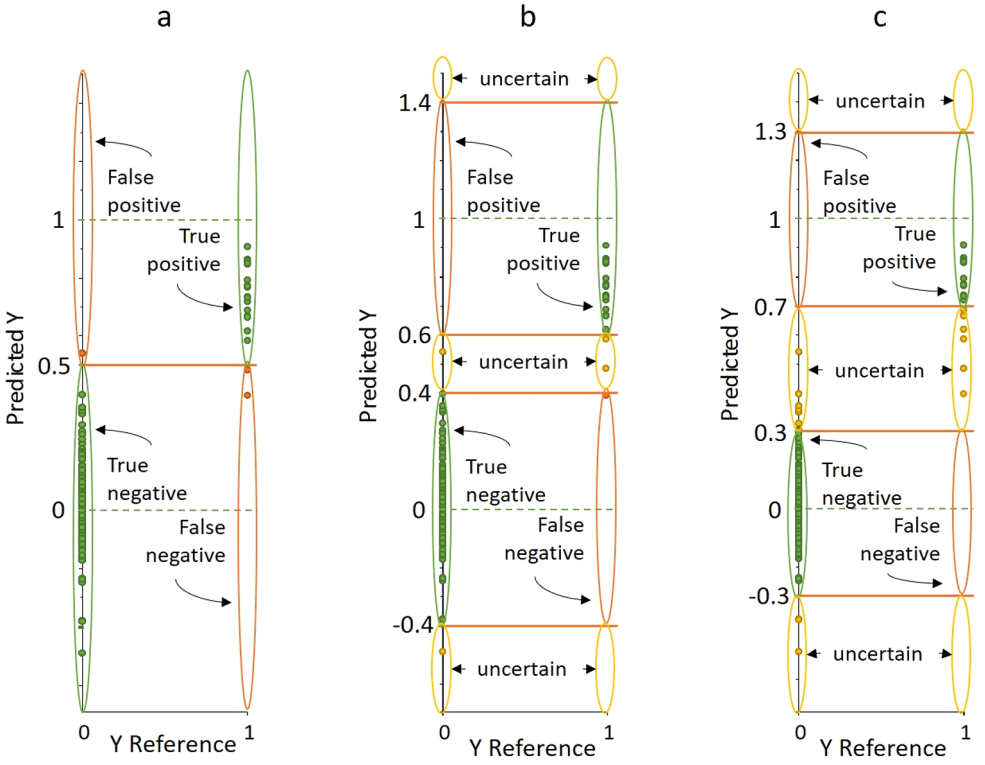

Fig. 1. Definition of the thresholds indicating the True, False or Uncertain attribution of predicted samples to the modelled variety (a: 0.5 threshold, b: 0.4–0.6 threshold, c: 0.3–0.7 threshold).

together. Some articles have studied the simple concatenation of data from two or more sources (de B Harrington, Kister, Artaud & Dupuy, 2009; Casale, Sinelli, Oliveri, Di Egidio, & Lanteri, 2010; Casale, Casolino, Oliveri, & Forina, 2010; Dupuy, Galtier, Ollivier, Vanloot, & Artaud, 2010; Casale et al., 2012; Haddi et al., 2013; Pizarro et al., 2013; Dias, Rodrigues, Veloso, Pereira, & Peres, 2016; Kosma, Badeka, Vatavali, Kontakos, & Kontominas, 2016; Bajoub et al., 2017). Hierarchical models have also been developed based on data from NIR and MIR (Dupuy et al., 2010), spectroscopy and mass spectrometry (Casale et al., 2010), artificial nose, NIR and UV–visible (Forina et al., 2015) or liquid chromatography with two detectors (Bajoub et al., 2017).

To our knowledge, mid-level data fusion approaches using multiblock models have not yet been applied to the discrimination of EVOO varietal origin. Moreover, the combination of GC data giving specific information on the major compounds with MIR data taking into account the global composition of oils, should provide complementary information and is expected to be able to refine the EVOO origin discrimination. This is the purpose of this work: multiblock partial least squares – discriminant analysis models (MB-PLS1-DA) were developed from GC and MIR datasets, with and without weighting the block scores, in order to evaluate their performance against those of the PLS1-DA models applied separately to each dataset. The study was conducted using Tunisian monovarietal EVOO from three cultivars.

2. Materials and methods

2.1. Extra virgin olive oil samples

Sampling was carried out during the 2011–2012 and 2012–2013 harvest years. Three hundred and thirty-four monovarietal EVOO samples from three Tunisian varieties were used for this study: *Chemlali* (n = 187), *Chetoui* (n = 102) and *Oueslati* (n = 45). Tunisian EVOO were obtained in laboratory by oleodoseur extraction system, from handpicked olives and without storage time before the extraction (Laroussi-Mezghani et al., 2015). The quality criteria for all samples were comprised within the ranges established for the "Extra Virgin Olive Oil" category by the trade standard of the International Olive Council (2016).

2.2. Gas chromatography

The transmethylation of the EVOO triacylglycerols and subsequent GC analyses using an Agilent Technology gas chromatograph 7890A equipped with a split/split-less injector, a flame ionization detector and a Supelcowax silica capillary column coated with polyethylene glycol (60 m × 0.25 mm i.d., 0.25 µm film thickness) were conducted following the method described by Laroussi-Mezghani et al. (2015).

2.3. Mid infrared spectroscopy

MIR spectra were recorded between 700 and 4000 cm^{-1} by the accumulation of 64 scans per spectrum with a resolution of 4 cm^{-1} on a Thermo Nicolet Avatar spectrometer equipped with an ATR accessory (Goldengate, Specac), using the same protocol as Galtier et al. (2008).

2.4. Chemometrics

2.4.1. Pre-treatment

Prior to data analysis, the noisy and noninformational regions between 1880 and 2600 cm^{-1} and 3200–4000 cm^{-1} were removed from the MIR spectra. In order to optimize the models, normalisation followed by standard normal variate (SNV) pre-treatments were applied to the spectra to correct the distortions caused by additive and multiplicative effects. The GC data were also normalised.

2.4.2. Data analysis

PLS1-DA models were applied to GC and MIR data separately. Each sample is assigned a binary coding indicating its membership (value equal to 1) or non-membership (value equal to 0) of each class, and a different model is built to predict each class against all the others.

During the calibration process, the PLS1-DA method is trained to compute the "membership values" and a sample is then assigned to the modelled class when its value is above a determined threshold (Granato et al., 2018). However, due to the initial design of PLS for continuous variables, the predicted values are not binary and thus several methods have been proposed to select a threshold that discriminates the results between the expected values 1 or 0: using an arbitrary value of 0.5, determining the optimal threshold with receiver operating characteristic curves, estimating a probability density function to handle unbalanced groups sizes, or defining an interval to take into account the uncertainty of PLS predictions (Lee, Liong, & Jemain, 2018). In this study three thresholds were considered for the calculation of the percentage of correct classification, as presented in Fig. 1(a–c). First, samples with a predicted value over 0.5 were considered positive (belonging to the modelled class) and those with a predicted value under 0.5 negative (outside of the modelled class). However, some samples may have predicted values close to 0.5 or too different from the reference values 0 and 1, indicating that they are not clearly recognised by the model. Thus, in a second approach uncertainty zones were defined to address this issue. Samples were considered positive if their predicted value was between 0.6 and 1.4, negative if predicted between −0.4 and 0.4, and uncertain if predicted between 0.4 and 0.6, under −0.4 or over 1.4. Finally, following the same reasoning, more conservative uncertainty zones were tested with samples considered positive if predicted between 0.7 and 1.3, negative between −0.3 and 0.3, and uncertain otherwise. A sample was considered as true positive, or true negative, if its predicted value was consistent with its expected value of 1, or 0. On the contrary, if the predicted value did not match the expected value the sample was considered as false negative, or false positive (or uncertain, if applicable). The total percentage of correct classification, as well as sensitivity and specificity were calculated according to Eqs. (1)–(3).

$$\%correct\ classification = \frac{true\ positive + true\ negative}{number\ of\ predicted\ samples} \times 100 \quad (1)$$

$$\%sensitivity = \frac{true\ positive}{expected\ positive} \times 100 \quad (2)$$

$$\%specificity = \frac{true\ negative}{expected\ negative} \times 100 \quad (3)$$

MB-PLS1-DA was then applied with one predictor block $\mathbf{X_1}$ consisting of the 15 variables of GC data and a second predictor block $\mathbf{X_2}$ comprising the 948 variables of MIR data, after their respective pre-treatments and mean-centring. Autoscaling was not applied since it could cause the information from the large MIR block to be preponderant over the small GC block (Westerhuis & Coenegracht, 1997). However, two scaling strategies were tested: one with a weighting of the block scores to take into account the number of variables in each block as indicated in Eq. (7), and the other without any weighting. The response matrix $\mathbf{Y}$ contained the 3 varietal origins of the 334 EVOO samples, and three independent models were built to predict each origin against the other two combined. The MB-PLS algorithm used is the one developed by Westerhuis, Kourit & MacGregor (1998), detailed in Eqs. (4)–(13).

$$\mathbf{X_i} = \mathbf{T_i}\mathbf{P_i^T} + \mathbf{E_i} \quad (4)$$

$$\mathbf{Y} = \mathbf{T_s}\mathbf{Q^T} + \mathbf{F} \quad (5)$$

With $\mathbf{X_i}$ the matrix of predictors for block i, $\mathbf{T_i}$ the block scores matrix, $\mathbf{P_i}$ the block loadings matrix, $\mathbf{E_i}$ the residuals, $\mathbf{Y}$ the response matrix, $\mathbf{T_s}$ the super-scores matrix, $\mathbf{Q}$ the response loadings matrix and $\mathbf{F}$ the residuals.

The variable weights ($\mathbf{w_i}$) in Eq. (6) are calculated separately for each block using the response scores ($\mathbf{u}$) and then normalised.

$$\mathbf{w_i} = \mathbf{X_i^T u} \quad (6)$$

The scores ($\mathbf{t_i}$) in Eq. (7) are also computed for each block $\mathbf{X_i}$ so that the covariance between the response $\mathbf{Y}$ and the scores is maximized, and scaled by the square root of the number of variables in the block ($\mathbf{m_i}$).

$$\mathbf{t_i} = \frac{\mathbf{X_i w_i}}{\sqrt{\mathbf{m}_i}} \text{ to maximise } |\mathbf{Y^T} \sum \mathbf{t_i}|^2 \quad (7)$$

The block scores are then combined into a super-matrix ($\mathbf{T}$) and a PLS is performed between $\mathbf{T}$ and $\mathbf{Y}$. The super-weights ($\mathbf{w_s}$) are normalised before calculation of the super-scores ($\mathbf{t_s}$), response loadings ($\mathbf{q}$) and response scores ($\mathbf{u}$), in Eqs. (8)–(11).

$$\mathbf{w_s} = \mathbf{T^T u} \quad (8)$$

$$\mathbf{t_s} = \mathbf{T w_s} \quad (9)$$

$$\mathbf{q} = \frac{\mathbf{Y^T t_s}}{\mathbf{t_s^T t_s}} \quad (10)$$

$$\mathbf{u} = \frac{\mathbf{Y q}}{\mathbf{q^T q}} \quad (11)$$

Eqs. (6)–(11) are repeated until convergence of $\mathbf{u}$. Then the block loadings ($\mathbf{p_i}$) are calculated, and the $\mathbf{X_i}$ and $\mathbf{Y}$ matrices are deflated using the super-scores as indicated in Eqs. (12)–(14). This process, from Eqs. (6)–(14), is repeated for the next latent variable (LV) with the new $\mathbf{X_i}$ and $\mathbf{Y}$.

$$\mathbf{pi} = \frac{\mathbf{X_i^T t_s}}{\mathbf{t_s^T t_s}} \quad (12)$$

$$\mathbf{X_i} \leftarrow \mathbf{X_i} - \mathbf{t_s p_i^T} \quad (13)$$

$$\mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{t_s q^T} \quad (14)$$

For both PLS1-DA and MB-PLS1-DA models, considering the large number of samples these were randomly and equally divided into a calibration set and a prediction set (167 samples each). A first version of the models was constructed with these sets, then a second version was built after permuting the sets. For each version of the PLS1-DA and MB-PLS1-DA models, a "leave one out" cross validation procedure was used on the calibration set to find the optimal number of LV. It had to be large enough to minimise the root mean square error of cross validation (RMSECV), but not too large in order to avoid over-fitting. Thus, the optimal number was chosen as the highest number of LV (n) meeting the criterion from Eq. (15).

$$\frac{\mathbf{RMSECV(n) - RMSECV(n-1)}}{\mathbf{RMSECV(n)}} > 5\% \quad (15)$$

The calibration model was then computed with the selected number of LV. Finally, the prediction set was used to calculate the predicted response according to this calibration model. In addition to the number of LV, determination coefficients of calibration (R^2) and prediction (Q^2), as well as root mean square errors of calibration (RMSEC) and prediction (RMSEP) were calculated.

2.4.3. Software

Chemometrics pre-treatments were applied using The Unscrambler® X (version 10.4, CAMO Software). The pre-treated matrices were then imported to Matlab® (version 7.8 R2009a, MathWorks) for data analysis. MB-PLS1-DA routines, including a calibration with cross-validation step followed by a prediction step, were developed based on the Multi-block Toolbox by van den Berg (2004).

3. Results and discussion

Tunisian EVOO from the three varieties *Chemlali*, *Chetoui* and *Oueslati* were analysed by GC and MIR spectroscopy.

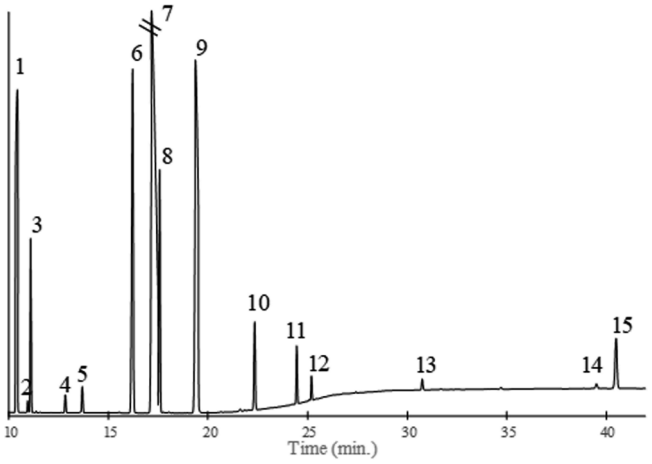

Fig. 2. Example of a chromatogram from VOO with identification of the peaks 1: palmitic acid (16:0), 2: hypogeic acid (16:1 ω9), 3: palmitoleic acid (16:1 ω7), 4: margaric acid (17:0), 5: margaroleic acid (17:1 ω8), 6: stearic acid (18:0), 7: oleic acid (18:1 ω9), 8: z-vaccenic acid (18:1 ω7), 9: linoleic (18:2 ω6), 10: linolenic acid (18:3 ω3), 11: arachidic acid (20:0), 12: gondoic acid (20:1 ω9), 13: behenic acid (22:0), 14: lignoceric acid (24:0) and 15: squalene.

3.1. Gas chromatography profiles

The GC profiles obtained after the transmethylation of triacylglycerols, which represent around 98% of the total content of olive oils, indicate the relative proportions of major compounds (fourteen fatty acids and squalene). An example of chromatogram, with peaks identification from Ollivier, Artaud, Pinatel, Durbec, and Guérère (2003), is shown in Fig. 2. Predominant peaks are due to oleic (18:1ω9), palmitic (16:0) and linoleic (18:2ω6) acids, and two other fatty acids, namely stearic (18:0) and z-vaccenic (18:1ω7) acids, are present in intermediate amounts. Moreover, even other fatty acids present in lesser amounts may play an important part in the discrimination of varietal origin. The mean, maximum and minimum percentages of the fifteen major compounds, as well as the sum of saturated fatty acids (SFA), monounsaturated fatty acids (MUFA) and polyunsaturated fatty acids (PUFA) for each of the three VOO varieties are compiled in Table 1.

Most of the SFA percentages, namely margaric (17:0), stearic (18:0), arachidic (20:0), behenic (22:0) and lignoceric (24:0) acids, do not differ significantly between the three studied varieties. Margaroleic (17:1ω8) acid and the only measured ω3, linolenic acid (18:3ω3), are not discriminant either. Nevertheless, *Chemlali* samples are characterised by a generally higher SFA content, mainly due to their high levels of palmitic (16:0) acid. They contain more palmitoleic (16:1ω7) and z-vaccenic (18:1ω7) acids, but less oleic (18:1ω9) and gondoic (20:1ω9) acids than the other two varieties, resulting in an overall lower MUFA content. The amount of squalene in *Chemlali* samples is also lower. As for *Chetoui* samples, they have average values of total SFA, MUFA and PUFA contents compared to the other two varieties, but they are distinguished by their higher levels of hypogeic (16:1ω9) acid and lower levels of palmitoleic (16:1ω7) acid. Finally, *Oueslati* samples contain less PUFA because of their low levels of linoleic (18:2ω6) acid.

3.2. MIR spectra

MIR spectra contain information on the global composition of the samples, including potential variations in the concentrations of triacylglycerols but also of different families of minor compounds. Some of these constituents, such as squalene, carotenoids, tocopherols, phytosterols and phenolic compounds, have beneficial nutritional properties. An example of MIR spectrum is presented in Fig. 3, with bands attribution according to Aparicio and Harwood (2013). The bands do not result from a single molecule but rather from the vibration of chemical bonds that are present in all the compounds of the sample so that the interpretation is less straightforward than for GC peaks. Well defined bands between 3100 and 1700 cm^{-1} are attributed to C—H, C═O and C═C stretching vibrations, whereas some overlapping bands between 1500 and 700 cm^{-1} have been assigned to C—H, C—O and C—C bending vibrations. Contrary to the noticeable differences in the fatty acid profiles, variations in the global composition are not readily perceptible since the VOO samples have, to the naked eye, similar MIR spectra. The use of chemometrics pre-treatments and modelling is therefore necessary to extract the relevant information from this data.

3.3. Prediction of olive oil variety

The statistical parameters (number of LV, RMSEC, RMSEP, R^2 and Q^2) and results (sensitivity, specificity and total correct classification percentages) from the PLS1-DA prediction models developed for each variety based on GC, MIR and Multiblock data with the first version of the calibration and prediction sets are presented in Table 2. Statistical parameters and results obtained with the second version of the calibration and prediction sets can be found in the Supporting Information.

3.3.1. PLS1-DA on GC data

The GC model for the *Chemlali* variety performs very well, with a RMSEP of 0.11 and Q^2 of 0.97 for 3 LV and can perfectly discriminate *Chemlali* samples from the others using the 0.5 thresholds. Regarding the *Chetoui* variety, the GC models also give good results with a RMSEP of 0.16 and Q^2 of 0.94 for 6 LV, and 99% correct classification with the 0.5 threshold. The model for *Oueslati* samples is slightly less efficient, yielding a RMSEP of 0.24 and Q^2 of 0.74 for 5 LV, but still reaches a correct classification rate of 99% with the 0.5 threshold.

However, when looking at the results obtained with the 0.4–0.6 and 0.3–0.7 threshold, a drop in the percentages of correctly classified

Table 1
Mean, maximum and minimum proportions (%) of fatty acids and squalene for the three varieties (CM: *Chemlali*, CT: *Chetoui*, OU: *Oueslati*).

		16:0	16:1ω9	16:1ω7	17:0	17:1ω8	18:0	18:1ω9	18:1ω7	18:2ω6	18:3ω3	20:0	20:1ω9	22:0	24:0	Squa	SFA	MUFA	PUFA
CM	Mean	17.40	0.06	2.20	0.04	0.07	2.44	57.61	3.11	15.28	0.67	0.44	0.20	0.12	0.07	0.28	20.52	62.38	15.95
	Max	22.75	0.12	3.44	0.07	0.10	2.99	64.32	3.82	21.31	1.17	0.54	0.27	0.16	0.09	0.46	25.77	68.41	22.48
	Min	13.13	0.02	1.57	0.03	0.05	1.99	47.59	2.47	10.75	0.51	0.36	0.14	0.10	0.05	0.11	15.80	53.10	11.47
CT	Mean	11.37	0.13	0.30	0.05	0.05	2.82	66.46	1.31	15.28	0.69	0.48	0.39	0.13	0.05	0.49	14.90	69.77	15.97
	Max	14.47	0.17	0.56	0.09	0.11	3.43	73.53	1.71	21.16	0.90	0.54	0.47	0.16	0.08	0.67	17.76	76.86	21.87
	Min	8.71	0.10	0.17	0.04	0.03	2.34	59.39	0.90	10.14	0.54	0.42	0.31	0.10	0.03	0.37	12.35	62.94	10.92
OU	Mean	11.50	0.09	0.62	0.04	0.05	2.36	70.79	1.79	10.59	0.63	0.44	0.35	0.14	0.07	0.53	14.55	73.92	11.22
	Max	15.75	0.12	0.99	0.04	0.06	3.22	74.47	2.40	15.47	0.75	0.50	0.40	0.15	0.09	0.67	18.63	77.48	16.09
	Min	9.61	0.07	0.43	0.03	0.04	1.88	64.41	1.31	7.97	0.50	0.39	0.28	0.12	0.05	0.36	13.25	67.67	8.72

16:0: palmitic acid, 16:1 ω9: hypogeic acid, 16:1 ω7: palmitoleic acid, 17:0: margaric acid, 17:1 ω8: margaroleic acid, 18:0: stearic acid, 18:1 ω9: oleic acid, 18:1 ω7: z-vaccenic acid, 18:2 ω6: linoleic acid, 18:3 ω3: linolenic acid, 20:0: arachidic acid, 20:1 ω9: gondoic acid, 22:0: behenic acid, 24:0: lignoceric acid, Squa: squalene, SFA: saturated fatty acids, MUFA: monounsaturated fatty acids, PUFA: polyunsaturated fatty acids

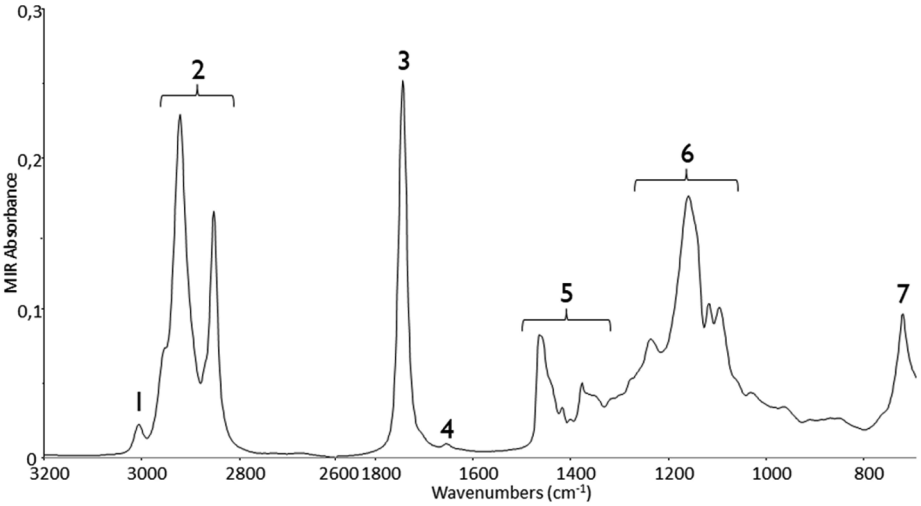

Fig. 3. Example of a MIR spectrum from VOO with identification of the bands 1: =C−H *cis* stretching, 2: C−H stretching, 3: C=O stretching, 4: C=C *cis* stretching, 5: C−H bending, 6: C−O and C−C bending, 7: C−H bending (long chains).

samples indicates that some of them are actually in the uncertainty zones. This can be especially observed for *Oueslati* samples, with a dramatic decrease of sensitivity from 95% to 91% with the 0.4–0.6 threshold, and even to 55% with the 0.3–0.7 threshold, while the specificity is less impacted but goes from 100% to 93% and to 89% for each threshold respectively. The models predicting other two cultivars appear to be more discriminative. Indeed, for *Chetoui* the sensitivity and specificity are still of 98% and 97% with the 0.4–0.6 threshold, and only drop to 94% and 95% respectively for the 0.3–0.7 threshold. The *Chemlali* model is still very good with 100% sensitivity and 99% specificity with both thresholds, showing that most samples are predicted close to their expected values. The poorer performance of the *Oueslati* model may be due to the smaller number of samples from this cultivar, which creates a strong imbalance between the positive and negative classes in this model.

The permutation of calibration and prediction sets brings some modifications to these results, especially for the *Chetoui* and *Oueslati* models, which do not use the same number of LV. The calibration quality parameters are slightly lower, but the prediction parameters are improved. The *Chetoui* model gives slightly better results but with only 3 LV, instead of 6 LV for the previous version. The *Oueslati* model is built with 6 LV, versus 5 LV in the first version, and also has overall better results. Thus, there is some influence of the samples selected in the calibration and prediction sets on the performance of the models.

3.3.2. PLS1-DA on MIR data

PLS1-DA on MIR data is less satisfactory than that using GC data, especially for the *Chemlali* and *Chetoui* models which were very good with GC data. Indeed, the model predicting the *Chemlali* variety has a RMSEP of 0.26 and Q^2 of 0.85 with 4 LV, while for *Chetoui* with 7 LV the RMSEP and Q^2 are respectively of 0.26 and 0.84. The *Oueslati* model also uses 7 LV but its quality parameters are better than that of the model based on GC data, with a RMSEP of 0.21 and Q^2 of 0.79.

Using the 0.5 threshold, the three varieties are still quite well predicted, with total correct classification rates of 93% for *Chemlali*, 94% for *Chetoui* and 99% for *Oueslati*.

The 0.4–0.6 and 0.3–0.7 thresholds identify even more uncertain samples with MIR than with GC data. The sensitivity drops to 82% with the former and to 36% with the latter for *Oueslati* samples, while the specificity is less impacted and only decreases to 98% and to 89%. When using MIR data as opposed to GC data, the model predicting the *Chemlali* origin is more impacted by the change of threshold than the *Chetoui* model. The sensitivity and specificity of the *Chemlali* model decrease to 91% and 89% respectively with the 0.4–0.6 threshold, then

Table 2
Statistical parameters and results (sensitivity, specificity and correct classification rates) of the PLS1-DA models using the first version of the calibration and prediction sets of either GC, MIR, weighted multiblock or non-weighted multiblock data to discriminate the three EVOO varieties (CM: *Chemlali*, CT: *Chetoui*, OU: *Oueslati*).

		CM (Cal: 93, Pred: 94)				CT (Cal: 51, Pred: 51)				OU (Cal: 23, Pred: 22)			
		GC	MIR	MB weight	MB no weight	GC	MIR	MB weight	MB no weight	GC	MIR	MB weight	MB no weight
	LV	3	4	3	4	6	7	6	7	5	7	5	7
	RMSEC	0.10	0.21	0.09	0.09	0.12	0.21	0.12	0.11	0.18	0.19	0.17	0.16
	RMSEP	0.11	0.26	0.11	0.11	0.16	0.26	0.15	0.13	0.24	0.21	0.23	0.20
	R^2	0.98	0.91	0.98	0.98	0.97	0.89	0.97	0.97	0.86	0.84	0.87	0.89
	Q^2	0.97	0.85	0.97	0.97	0.94	0.84	0.95	0.96	0.74	0.79	0.77	0.82
Threshold: 0.5	%Sens	100	94	100	100	100	90	100	100	95	95	95	100
	%Spec	100	92	100	100	99	96	99	100	100	100	100	100
	%CC	100	93	100	100	99	94	99	100	99	99	99	100
Threshold: 0.4–0.6	%Sens	100	91	100	100	98	86	98	100	91	82	91	86
	%Spec	99	89	99	99	97	93	97	100	93	98	94	94
	%CC	99	90	99	99	98	91	98	100	93	96	94	93
Threshold: 0.3–0.7	%Sens	100	78	100	98	94	80	94	98	55	36	64	77
	%Spec	99	84	99	99	95	83	97	97	89	89	89	92
	%CC	99	80	99	98	95	82	96	98	84	82	86	90

LV: number of latent variables, RMSEC: root mean square error of calibration, RMSEP: root mean square error of prediction, R^2: determination coefficient of calibration, Q^2: determination coefficient of prediction, %Sens: sensitivity rate, %Spec: specificity rate, %CC: correct classification rate.

to 78% and 84% with the 0.3–0.7 threshold. For the *Chetoui* model the sensitivity and specificity decrease first to 86% and 93% respectively, and then to 80% and 83%. These results indicate that GC data is more discriminative than MIR, since samples are more clearly predicted as belonging or not to the modelled variety with the former data. However, MIR data still contains valuable information that could be used in case GC-based models do not yield good results, as evidenced by the performance of the MIR-based *Oueslati* model.

The permutation of calibration and prediction sets also leads to variations in the performance of the models. As for the GC-based models, the permutation worsens the quality parameters for the calibration but improves them for the prediction. I this case, the prediction results are better for the *Oueslati* and *Chemlali* models, but not for the *Chetoui* model.

3.3.3. Multiblock PLS1-DA

When applying the MB-PLS1-DA models with the scores weighting, the results are similar to those obtained with GC data for all three varieties with the different thresholds. The models are built with the same number of LV (3 for *Chemlali*, 6 for *Chetoui* and 8 for *Oueslati*) and have the same value of quality parameters. The only noticeable difference is an improvement in the Q^2 for the *Oueslati* model, from 0.74 with GC to 0.77 with MB-PLS. Moreover, the sensitivity with the 0.3–0.7 threshold is improved and goes from 55% with GC alone and 36% with MIR alone to 64% with the multiblock model. Thus, the use of MB-PLS could reduce the influence of the imbalanced number of samples. The *Chemlali* and *Chetoui* models, for which the GC data alone gave better results, are not negatively influenced by the addition of the MIR data and the *Oueslati* model, for which the MIR data alone gave slightly better results, beneficiates from the combination of the two sources of information.

The permutation of calibration and prediction sets slightly changes the results of the individual models, but for the multiblock models with weighting of the scores the results remain close to those of the GC models. These results suggest that the GC data, containing information on the major compounds of olive oil, has a stronger influence than the MIR data, representing all the major and minor compounds, on the weighted multiblock models. This can be highlighted by a study of the contribution of each block to the final model, as presented in Fig. 4(a–c). Indeed, GC data is predominant on all the LVs for each variety and is the main contributor to the first LV. MIR data brings some contribution to the following LV, which indicates a possible synergy between the two sources of information. However, the scaling realized by the MB-PLS algorithm to compensate for the much larger number of variables in the MIR block (948 variables for MIR versus 15 variables for GC) strongly reduces the influence of MIR data on the models. Block weights for the *Chemlali* variety are not affected by the permutation of calibration and prediction sets. However, for the *Chetoui* variety the second version of the model only uses three LV which mostly contain information from the GC block. On the contrary, for the *Oueslati* variety MIR data has more influence on the second version of the model.

In order to take better advantage of the complementary information brought by the MIR data, MB-PLS1-DA models without any scores weighting have been performed to give more importance to this additional source. Indeed, the number of selected LV for these multiblock models are the same as for the MIR-based models for each cultivar. However, the results are improved compared to the previous models, especially for the *Oueslati* and *Chetoui* cultivars. For the *Chemlali* model, the results are close to that of the already effective GC-based model, with only a slightly lower sensitivity of 98% (versus 100%) for the 0.3–0.7 threshold. The prediction quality parameters are improved for the *Chetoui* model, with a RMSEP of 0.13 and Q^2 of 0.96. Moreover, the MB-PLS model without weighting gives perfect predictions with both the 0.5 and the 0.-0.6 thresholds for this cultivar. Using the 0.3–0.7 threshold, the sensitivity and specificity are also better than with GC or MIR data alone, reaching a total correct classification of 98% versus 95% for the GC-based model. The quality parameters are also improved with the MB-PLS model predicting the *Oueslati* cultivar, reaching a RMSEP of 0.20 and Q^2 of 0.82. This model also results in a perfect prediction with the 0.5 threshold. The sensitivity and specificity observed with the 0.4–0.6 threshold are intermediate between those of the GC-based and MIR-based models. Nevertheless, with the 0.3–0.7 threshold the sensitivity and specificity are much better with this MB-PLS model, reaching respectively 77% and 92%. Again, the multiblock

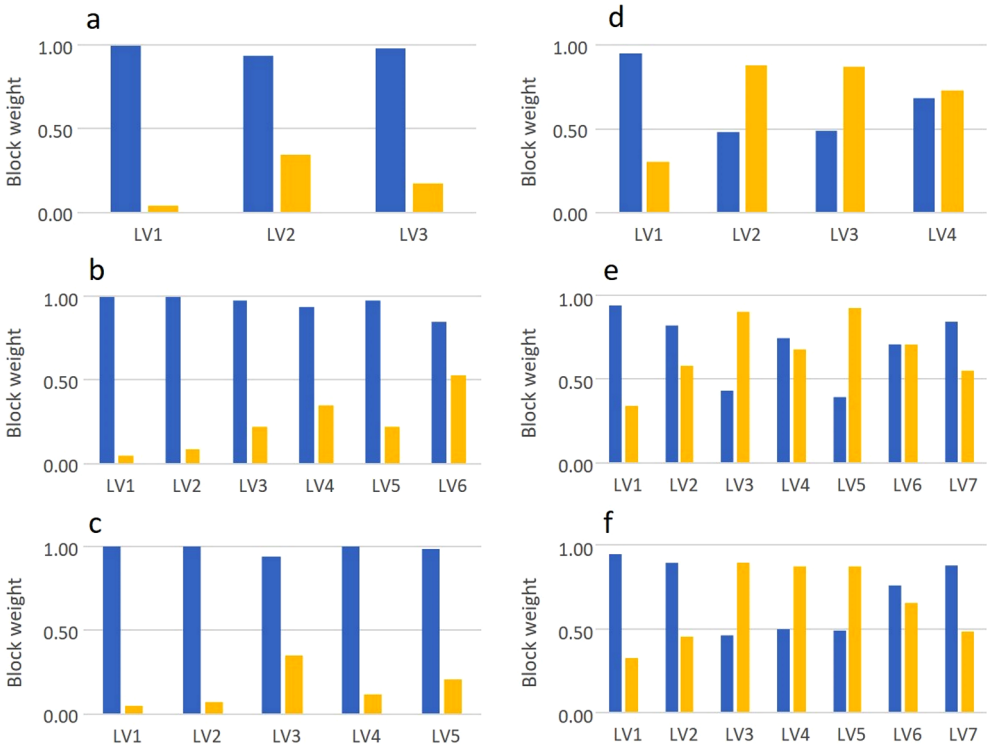

Fig. 4. Weights of the GC (blue) and MIR (yellow) blocks for each latent variable of the MB-PLS models with the first version of the calibration and prediction sets, with weighted block scores (a: *Chemlali*, b: *Chetoui*, c: *Oueslati*) and non-weighted block scores (d: *Chemlali*, e: *Chetoui*, f: *Oueslati*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

approach seems able to correct the issue caused by the imbalanced number of samples in the classes. After the permutation of the calibration and prediction sets the improvement brought by the non-weighted MB-PLS1-DA models is somewhat lost since the results appear to be in-between those obtained with GC and MIR data alone.

The contributions of the blocks without weighting presented in Fig. 4(d–f) show that, despite its much larger number of variables, the MIR block does not overshadow the GC block. On the contrary, there seems to be a good balance between the two sources of information. GC data is still predominant on the first LV for each model, but MIR data has a more important influence on the latter components. A similar pattern can be observed after the permutation of the calibration and prediction sets, but with less influence of the GC data on the latter LVs, which might contribute to the poorer results. Finally, the study of block weights confirms that although variations in the contents of the major compounds of olive oils measured by GC can discriminate most of the samples from the three studied varieties, complementary information about the global composition of the samples detected in their MIR spectra can play a part in the improvement of the prediction models. Thus, the MB-PLS1-DA models without any weighting of the scores can be useful for samples whose origin is difficult to certify from their MIR spectrum or their fatty acid profile only.

4. Conclusion

This study shows that PLS1-DA models using GC data alone give very good results for the discrimination of olive oil origin, especially for the *Chemlali* variety. Using MIR data alone is less efficient, even though it reaches more than 80% of correct classification with the most conservative threshold. Moreover, combining specific information on the major compounds from GC with global information on all major and minor compounds from MIR data can improve the prediction results for the varieties that were not well discriminated with GC data only. In this regard, scaling the block scores to take into account their number of variables strongly reduces the influence of the MIR data. Thus, the results from the weighted MB-PLS1-DA models are close to those of the GC-based models. On the contrary, using non-weighted MB-PLS1-DA models allows for a synergy between the two sources of information and results in better quality parameters and higher sensitivity, specificity and total correct classification percentages for the *Chetoui* and *Oueslati* varieties. These results should nevertheless be considered with caution since the permutation of calibration and prediction sets indicated that the performance of the different models depend on the samples used to develop and test these models. From a food control perspective, MIR analysis is cheaper and faster than GC and could be used as a first screening device. In a second phase, multiblock models combining MIR and GC data can strongly improve the discrimination for the samples that were in the uncertainty zones with the first model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors thank the Olive Tree Institute of Sfax, Tunisia, for providing the olive oil samples.

Funding

This work was financially supported by the French National Agency for Research (ANR) as part of the MedOOmics project, included in the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement number 618127 (ARIMNet2). The authors are grateful to the Tunisian Ministry of Higher Education and Scientific Research and the French Ministry of Foreign Affairs for the project 11G1214 (program PHC-Utique).

References

Aparicio, R., & Harwood, J. (2013). *Handbook of Olive Oil* (2nd ed). Boston, MA: Springer, US.

Bajoub, A., Medina-Rodríguez, S., Gómez-Romero, M., Bagur-González, M. G., Fernández-Gutiérrez, A., & Carrasco-Pancorbo, A. (2017). Assessing the varietal origin of extra-virgin olive oil using liquid chromatography fingerprints of phenolic compound, data fusion and chemometrics. *Food Chemistry, 215*, 245–255.

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment – A review. *Analytica Chimica Acta, 891*, 1–14.

Brereton, R. G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., ... Tauler, R. (2017). Chemometrics in analytical chemistry-part I: History, experimental design and data analysis tools. *Analytical and Bioanalytical Chemistry, 409*, 5891–5899.

Callao, M. P., & Ruisánchez, I. (2018). An overview of multivariate qualitative methods for food fraud detection. *Food Control, 86*, 283–293.

Casale, M., Sinelli, N., Oliveri, P., Di Egidio, V., & Lanteri, S. (2010). Chemometrical strategies for feature selection and data compression applied to NIR and MIR spectra of extra virgin olive oils for cultivar identification. *Talanta, 80*(5), 1832–1837.

Casale, M., Casolino, C., Oliveri, P., & Forina, M. (2010). The potential of coupling information using three analytical techniques for identifying the geographical origin of liguria extra virgin olive oil. *Food Chemistry, 118*(1), 163–170.

Casale, M., Oliveri, P., Casolino, C., Sinelli, N., Zunin, P., Armanino, C., ... Lanteri, S. (2012). Characterisation of PDO olive oil chianti classico by non-selective (UV–Visible, NIR and MIR Spectroscopy) and selective (fatty acid composition) analytical techniques. *Analytica Chimica Acta, 712*, 56–63.

Charlebois, S., Schwab, A., Henn, R., & Huck, C. V. (2016). Food fraud: An exploratory study for measuring consumer perception towards mislabeled food products and influence on self-authentication intentions. *Trends in Food Science & Technology, 50*, 211–218.

Dais, P., & Hatzakis, E. (2013). Quality assessment and authentication of virgin olive oil by NMR spectroscopy: A critical review. *Analytica Chimica Acta, 765*, 1–27.

de B Harrington, P., Kister, J., Artaud, J., & Dupuy, N. (2009). Automated principal component-based orthogonal signal correction applied to fused near infrared – mid-infrared spectra of french olive oils. *Analytical Chemistry, 81*(17), 7160–7169.

Dias, L. G., Rodrigues, N., Veloso, A. C. A., Pereira, J. A., & Peres, A. M. (2016). Monovarietal extra-virgin olive oil classification: A fusion of human sensory attributes and an electronic tongue. *European Food Research and Technology, 242*, 259–270.

Dupuy, N., Galtier, O., Ollivier, D., Vanloot, P., & Artaud, J. (2010). Comparison between NIR, MIR, concatenated NIR and MIR analysis and hierarchical PLS model. Application to virgin olive oil analysis. *Analytica Chimica Acta, 666*(1–2), 23–31.

Forina, M., Oliveri, P., Bagnasco, L., Simonetti, R., Casolino, M. C., Grifi, F. N., & Casale, M. (2015). Artificial nose, NIR and UV–visible spectroscopy for the characterisation of the PDO chianti classico olive oil. *Talanta, 144*, 1070–1078.

Galtier, O., Le Dréau, Y., Ollivier, D., Kister, J., Artaud, J., & Dupuy, N. (2008). Lipid compositions and french registered designations of origins of virgin olive oils predicted by chemometric analysis of mid-infrared spectra. *Applied Spectroscopy, 62*(5), 583–590.

Gómez-Caravaca, A. M., Maggio, R. M., & Cerretani, L. (2016). Chemometric applications to assess quality and critical parameters of virgin and extra-virgin olive oil A review. *Analytica Chimica Acta, 913*, 1–21.

Granato, D., Putnik, P., Kovačević, D. B., Santos, J. S., Calado, V., Rocha, R. S., ... Pomerantsev, A. (2018). Trends in chemometrics: Food authentication, microbiology, and effects of processing. *Comprehensive Reviews in Food Science and Food Safety, 17*(3), 663–677.

Guzmán, E., Baeten, V., Pierna, J. A. F., & García-Mesa, J. A. (2015). Evaluation of the overall quality of olive oil using fluorescence spectroscopy. *Food Chemistry, 173*, 927–934.

Haddi, Z., Alami, H., El Bari, N., Tounsi, M., Barhoumi, H., Maaref, A., ... Bouchikhi, B. (2013). Electronic nose and tongue combination for improved classification of moroccan virgin olive oil profiles. *Food Research International, 54*(2), 1488–1498.

International Olive Council (2016). COI/T.15/NC No 3/Rev. 11 – Trade Standard Applying to Olive Oils and Olive Pomace Oils. http://www.internationaloliveoil.org/estaticos/view/222-standards. Accessed 04.07.18.

Kosma, I., Badeka, A., Vatavali, K., Kontakos, S., & Kontominas, M. (2016). Differentiation of greek extra virgin olive oils according to cultivar based on volatile compound analysis and fatty acid composition: Differentiation of greek extra virgin olive oils. *European Journal of Lipid Science and Technology, 118*(6), 849–861.

Laroussi-Mezghani, S., Vanloot, P., Molinet, J., Dupuy, N., Hammami, M., Grati-Kamoun, N., & Artaud, J. (2015). Authentication of tunisian virgin olive oils by chemometric analysis of fatty acid compositions and NIR spectra. Comparison with maghrebian and french virgin olive oils. *Food Chemistry, 173*, 122–132.

Lee, L. C., Liong, C.-Y., & Jemain, A. A. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. *The Analyst, 143*(15), 3526–3539.

Nenadis, N., & Tsimidou, M. Z. (2017). Perspective of vibrational spectroscopy analytical methods in on-field/official control of olives and virgin olive oil. *European Journal of*

Lipid Science and Technology, 119(1), 1600148.
Ollivier, D., Artaud, J., Pinatel, C., Durbec, J. P., & Guérère, M. (2003). Triacylglycerol and fatty acid compositions of french virgin olive oils. Characterization by chemometrics. *Journal of Agricultural and Food Chemistry, 51*(19), 5723–5731.
Pizarro, C., Rodríguez-Tecedor, S., Pérez-del-Notario, N., Esteban-Díez, I., & González-Sáiz, J. M. (2013). Classification of spanish extra virgin olive oils by data fusion of visible spectroscopic fingerprints and chemical descriptors. *Food Chemistry, 138*(2–3), 915–922.
Tena, N., Wang, S. C., Aparicio-Ruiz, R., García-González, D. L., & Aparicio, R. (2015). In-depth assessment of analytical methods for olive oil purity, safety, and quality characterization. *Journal of Agricultural and Food Chemistry, 63*(18), 4509–4526.
van den Berg, F. (2004). Multi-block Toolbox for MATLAB. http://www.models.life.ku.dk/mbtoolbox. Accessed Jul 18, 2018.
Wangen, L. E., & Kowalski, B. R. (1989). A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics, 3*(1), 3–20.
Westerhuis, J. A., & Coenegracht, P. M. J. (1997). Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics, 11*, 379–392.
Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics, 12*(5), 301–321.
Wold, S., Kettaneh, N., & Tjessem, K. (1996). Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics, 10*(5–6), 463–482.