HPLC-UV AND HPLC-CAD CHROMATOGRAPHIC DATA FUSION FOR THE AUTHENTICATION OF THE GEOGRAPHICAL ORIGIN OF PALM OIL

Kudirat Abidemi Obisesan[a], Ana M. Jiménez-Carvelo[b], Luis Cuadros-Rodriguez[b],

Itziar Ruisánchez[a]*, M. Pilar Callao[a]

[a] Chemometrics, Qualimetric and Nanosensors Grup, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n, 43007 Tarragona, Spain

[b] Department of Analytical Chemistry, University of Granada, c/Fuentenueva, s.n., E-18071 Granada, Spain.

Abstract

Data fusion combined with a multivariate classification approach (partial least squares-discriminant analysis, PLS-DA) was applied to authenticate the geographical origin of palm oil. Data fusion takes advantage of the synergistic effect of information collected from more than one data source. In this study, data from liquid chromatography coupled to two detectors –ultraviolet (UV) and charged aerosol (CAD)– was fused by high- and mid-level data fusion strategies. Mid-level data fusion combines a few variables from each technique and then applies the classification technique. Principal component analysis and interval partial least squares were applied to obtain the variables selected. High-level data fusion combines the PLS-DA classification results obtained individually from the chromatographic technique with each detector. Fuzzy aggregation connective operators were used to make the combinations. Prediction rates varied between 73% and 98% for the individual techniques and between 87 and 100% and 93 and 100% for the mid- and high-level data fusion strategies, respectively.

Key-words:

Data fusion, multivariate classification, liquid chromatography, authentication, palm oil

*Corresponding author: phone: +34 977558490; fax: +34 977558446; email: itziar.ruisanchez@urv.cat

1. INTRODUCTION

Palm oil is obtained from the palm fruit (Elais Guineensis), originally from West Africa, although at present the main producing countries are Indonesia, Malaysia, Thailand, Colombia and Nigeria. The oil must be refined before it is used for human consumption. The palm fruit produces two different oils: crude palm oil (CPO) and palm kernel oil (PKO) [1,2]. CPO is semi-solid at room temperature and it contains large amounts of saturated fatty acids. The acid occurring in the highest proportion is palmitic acid [3,4]. CPO is known as 'red palm oil' because of the carotenoids in its composition, predominantly α- and β- carotenes [5].

Palm oil is the most commonly consumed oil in the world. It is mainly used in foodstuffs as a food ingredient (e.g. in cakes and pastries) and for frying. Palm oil is cheap, easy to obtain and has low production costs so the number of industrial palm plantations has recently been increasing [6]. This increase in the number of plantations, however, has brought with it the destruction of the rainforest, the expulsion of the indigenous population and serious environmental problems such as deforestation and air pollution [7]. The solution proposed to solve these problems is what is known as 'sustainable palm oil': that is, only palm oil from known forest plantations should be used.

Thus, in 2004, the Roundtable on Sustainable Palm Oil (RSPO) was founded. This organization is defined as “a not-for-profit that unites stakeholders from the 7 sectors of the palm oil industry: oil palm producers, processors or traders, consumer goods manufacturers, retailers, banks/investors, and environmental and social non-governmental organisations (NGOs), to develop and implement global standards for sustainable palm oil.” The RSPO has developed a set of environmental and social criteria which companies must comply with in order to produce Certified Sustainable Palm Oil (CSPO) [8]. The term 'sustainable palm oil' has sometimes been used to make an illicit profit. The labels of some products state that they contain 'certified sustainable palm oil' when in fact the oil used comes from unknown forest plantations. Consequently, customers are increasingly demanding some sort of proof of the geographical origin of products.

To date, most studies on the authentication of oil apply the strategy known as 'chromatographic fingerprinting' [9]. Fingerprinting techniques provide analytical information about the sample in a non-selective way, such as by collecting nonspecific bands, which together with an appropriate multivariate data analysis makes it possible to characterize the oil. Some examples of chromatography fingerprinting-based palm oil characterization used gas chromatography with a mass spectrometer detector [10], gas

chromatography with flame-ionisation detector [11] and liquid chromatography coupled to a charged aerosol detector [10]. The main advantage of this methodology is that the chromatogram is used as a whole analytical signal [12]. Previous studies have shown that liquid chromatography coupled to a charged aerosol detector (HPLC-CAD) [13] and an ultraviolet detector (HPLC-UV) [13-14] combined with multivariate classification techniques can be used to determine the botanical origin of vegetable oil. With these results in mind, the main objective of the present study was to authenticate the geographical origin of palm oil by combining both techniques using data fusion strategies.

Nowadays**,** data fusion is increasingly being used in such fields as metabolomics [15,16], pigments in artworks [17,18], dye degradation processes [19] and food [20-24]. Most fused data comes from spectrometric techniques: near infrared (NIR), mid infrared (MIR), ultraviolet-visible (UV-Vis), Raman, Fluorescence, and mass spectrometry (MS). Very few studies report the fusion of chromatographic data. Those that do deal with gas chromatography (GC) and FT-MIR [25], GC and isotope ratio mass spectrometry (IRMS) [26], and liquid chromatography and GC data [27], all of which are applied to food analysis. To our knowledge, the fusion of liquid chromatographic data has not previously been reported.

The main goal of data fusion is to optimize the information obtained and exploit the synergies of information provided by different techniques [28]. Various strategies can be used to fuse data: low-, mid- and high-level strategies. In low-level data fusion, raw data from more than one source are directly fused (concatenated)~~,~~ after first ensuring that the data are correctly balanced (all the variables must be on the same scale) before they are combined. In mid-level data fusion, a few of the raw variables are fused. In both approaches (low- and mid-level), fused variables are used to build one classification model. In high-level data fusion, the classification results obtained from individual classification models are fused. In this case, a classification model must be developed for each data source.

In this paper, three types of data fusion strategies were studied: one at high- and two at mid-level data fusion. High-level data fusion was chosen because, although it is not used too often, it does have some advantages: it is easy to use and when a new data source becomes available, its classification results can be added to the classification decision rule, thus increasing the versatility of the decision process [22].

Mid-level data fusion, which involves selecting or reducing variables before the classification model is established, was also chosen. The most common approach for

reducing variables is to obtain the scores values for a small number of latent variables independently from the signals of each instrument. The scores values obtained from principal component analysis (PCA) or partial-least squares discriminant analysis (PLS-DA) are fused [28]. Of the various strategies available for selecting variables (iPLS, genetic algorithms, etc.), iPLS has been used here as it identifies which regions of the full chromatograms (variables) are the most influential [29].

The overall performance of the classification process (PLS-DA) was evaluated for each individual chromatographic technique and fusion process.

2. MATERIAL AND METHODS

2.1 Samples

A total of 100 crude palm oil samples were supplied by RIKILT-Institute of Food Safety Wageningen University (Wageningen, The Netherlands) (see table 1 in reference 13). The samples were obtained from the main palm-oil producing continents: South-East Asia (56 samples from Malaysia, Indonesia, Papua New Guinea and Salomon Islands), West Africa (30 samples, from Ghana, Guinea, Ivory Coast, Nigeria and Cameroon) and South America (16 samples from Brazil).

For the chemometric analysis, samples from class 1 were divided into training and test sets, the latter consisting of approximately 20% of randomly selected samples. Because there were few samples representing class 2 and 3, they were all kept in the training set. These two models were evaluated by leave-one-out cross-validation according to Foca et al. [30].

2.2 Sample preparation

Before the chromatographic analysis, a transesterification reaction was applied to the palm oil samples. This reaction is a modification of the original procedure described by Bierderman et al. [31]. More details about the procedure are given in the reference [32]. For the chromatographic analysis, 500 µL of transesterified fraction was added to a 2 mL HPLC vial, and the 120 mL of 0.05% (w/w) cholestanol solution in n-hexane was added as a control internal standard. Finally the mixture was diluted with 1000 mL of n-hexane. All the solvents and other reagents used are shown in the reference [13].

2.3. Instrumentation and software

The analyses were performed using HPLC technique with two different detectors. The first one was a Konik Model 560 (Konik-Tech, Sant Cugat del Vallès, Barcelona, Spain) with a quaternary pump, a column oven, an autosampler with a 20 mL loop, and a UV-Vis molecular absorption detector. The intensity values measured at 202nm and obtained from the retention time gave rise to a data matrix consisting of 100 rows (samples of palm oil) and 3436 variables. For simplicity referred as HPLC-UV in the text.

The second one was an Agilent 1100 Series (Agilent Technologies, Santa Clara, CA, USA) equipped with a quaternary pump, degasser, autosampler and thermostatted column compartment Eppendorf CH-30 (Eppendorf, Hamburg, Germany). Detection was carried out with a corona charged aerosol detector (CAD) (ESA Biosciences Inc., Chelmsford, MA, USA). In this case, the matrix data consisted of 100 samples and 1609 variables. For simplicity referred as HPLC-CAD in the text.

In both cases the chromatography fingerprint was obtained by HPLC using a (250×4 mm i.d, 5 µm) column Lichrospher® 100 CN.

The raw chromatographic data from HPLC-UV and HPLC-CAD were obtained in a CSV file, and then exported to MATLAB format (version R2007). All chemometric treatments were carried out by using the PLS_Toolbox (Eigenvector Research Inc., Wenatchee, WA), for MATLAB software (Mattworks Inc., Natick, MA, USA).

3. DATA ANALYSIS

3.1. Exploratory analysis and partial least squares-discriminant analysis (PLS-DA)

Principal component analysis (PCA) is an unsupervised exploratory analysis that can be used to visualize sample distribution in the multivariate space, check any natural clustering in samples that could influence the subsequent multivariate analysis, and identify possible outliers.

PLS-DA is the PLS regression technique adapted to a supervised classification task. It established a linear regression between a matrix of independent variables (matrix **X**) and an array of dependent variables (matrix **Y**). **Y** is a binary variable that indicates the class to which a sample belongs, where 1 indicates membership and 0 does not. Since this paper aimed to differentiate between three classes and classify them, class 1 samples were encoded as (1,0,0), class 2 as (0,1,0) and class 3 as (0,0,1).

The PLS-DA model uses a value between zero and one to predict the class for each sample. A threshold between 0 and 1 (above which a sample is considered to be a member of the class) is calculated using Bayesian statistics [33]. The Bayesian threshold assumes that the 'y' PLS-DA predicted values are normally distributed and it is selected at the 'y' value at which the number of false positives and false negatives is minimal. More details of the PLS-DA technique can be found in the literature [34,35].

Generalized least squares weighting (GLSW) was used in the application of the PLS-DA model. GLSW reduces the influence of variables (removes variance) from the X-block, which can be associated with interference (background, systematic sampling errors, instrumental drift or differences between samples that should otherwise be similar) [36,37]. A filter matrix is established using the parameter α. Alpha defines how strongly GLSW weights down the variables (interference) and, in our case, it was adjusted to 0.05. Higher values (typically above 0.1) decrease the effect of the filter while lower values (typically 0.01 and below) increase it [35,38].

The optimal number of latent variables (LVs) to be included in each model was chosen using leave-one-out cross-validation to minimize the root mean square-cross validation error (RMSECV) for each class. Finally, this number was selected by reaching a compromise between the optimal values for each class.

3.2. Data fusion

High- and mid-level data fusion strategies were applied. High-level data fusion [21,22], also known as decision data fusion, combines the classification results obtained from each instrumental technique, in this particular case the probability of assignation to one class given by HPLC-UV and HPLC-CAD models. Of the operators available, the minimum, maximum, average and product fuzzy aggregation connective operators were selected because of their conceptual simplicity and ease of implementation. The highest value provided by each operator for the three possible classes was chosen to obtain the sample class assignment. Finally, the class to which a sample is assigned (ensemble decision) was determined by considering the majority of the decisions from all the operators (majority vote). Given the nature of the operators, the concordant assignations for both chromatographic techniques will give the same results after the fusion so, for practical reasons, only non-concordant results were considered.

Mid-level data fusion, also known as variable data fusion, combines variables selected from each instrumental technique and concatenates them into a single vector. So, it requires variables to be reduced before the classification model is established. There are

two ways of reducing the number of variables: a) selecting a specific number of raw variables by implementing variable section methods, or b) generating new variables, which are usually a combination of the raw variables. This study uses both strategies.

Variables were selected by using interval partial least squares (iPLS) to find signal intervals which give better classification results than when all the variables are used [29,39]. The chromatogram signals obtained from each technique were independently divided into a number of intervals of equal size and a PLS-DA model was established for each of these intervals separately. The intervals selected were those with lower RMSECV values than when the full chromatogram model was used. The intervals (variables) selected from each technique were then combined for the classification. With this strategy, the most influential regions of the full chromatograms (variables) can be identified.

Variables were reduced by concatenating the PCA scores obtained individually from each data matrix [24]. The number of PCs was chosen by considering approximately 95% of the cumulative variance. Before concatenation, the scores were normalized (**equation 1)** to avoid the problem of imbalanced data.

$$t_{i,PCj}^{norm} = \frac{t_{i,PCj} - t_{min,PCj}}{t_{max,PCj} - t_{min,PCj}} \quad (1)$$

where $t_{i,PCj}^{norm}$ is the normalized score of sample 'i' at PC_i, and $t_{min,PCj}$ and $t_{max,PCj}$ are respectively the minimum and the maximum score values for PC_j.

4. RESULTS AND DISCUSSION

Figure 1 shows the chromatograms of a randomly selected sample analysed by HPLC using UV-Vis molecular absorption detector and by HPLC coupled to CAD detector. Four regions were identified. Region I contains most of the compounds present in palm oil, the main ones being fatty acids methyl esters. This may be related to the fact that the peak with the highest area appears in region I. Region II is characteristic of phytosterols, mainly dimethyl sterol. Region III is also characteristic of phytosterols, mainly dimethyl and methyl sterols, and other compounds such as fatty alcohols. Region IV is associated with the presence of terpene alcohols. The peak of the internal standard is between region II and III, which was not included in the data analysis.

FIGURE 1

Principal component analysis (PCA) was conducted independently with the auto-scale data from both chromatographic techniques and for all 100 samples in an attempt to determine whether there was a trend among the samples. **Figure 2** shows the scores for the first two PCs of the data analysed with HPLC-CAD independently for the three pre-established classes. The scores of the 56 samples from Asia (Fig. 2a) show that one sample (nº12) is clearly separate from the rest, so it was considered to be an outlier and removed from the data set. Likewise, one sample from America (nº6, Fig. 2c) was also considered to be an outlier and removed from the data set. The 28 samples from Africa show a homogenous score distribution (Fig. 2b). As for the HPLC-CAD data, the first two principal components were evaluated for the HPLC-UV data (graphs not shown) but no samples were identified as outliers.

FIGURE 2

First, to classify palm oil samples according to their geographical origin, independent PLS-DA models were built for each data set source studied (HPLC-CAD and HPLC-UV) for all the 87 samples. The outliers that had been detected by PCA and the 11 test set samples from class 1 were not considered. Of the pre-processing procedures found in the literature on chromatography data (auto-scale, normalized, smoothing with auto-scale and mean centred), auto-scale was chosen. Finally, PLS-DA models were built for both chromatographic data sets using the first six LVs.

Table 1 shows the classification results obtained in optimal conditions for both chromatographic data sets. Overall, classification results were better with the HPLC-CAD model: in the training set 98% of samples were correctly classified for class 1 (Asia), 86% for class 2 (Africa) and 73% for class 3 (America); and in the test set (class 1, Asia), 73% were correctly classified. Of the misclassifications, three samples were wrongly assigned (two from class 1 –one training and one test– and one from class 2), six were assigned to more than one class and three were not assigned at all. When the HPLC-UV model was used, nine samples were wrongly assigned, five samples were multiply assigned and two samples were not assigned. As far as the individual classes are concerned, class 3 had the lowest classification percentages, which may be because there were very few samples.

TABLE 1

Fusion strategies were then used to combine the HPLC-UV and HPLC-CAD data to improve the classification results. All data fusion strategies were used on the same data set (98 samples, without the outliers).

Decision-level fusion combines the probabilities of class assignation (values between 0 and 1) obtained from the individual models (HPLC-UV and HPLC-CAD) using four fuzzy aggregation operators (minimum, maximum, product and average) and the majority vote rule was used for the final sample assignation (ensemble decision), as described in the theory section. Only samples assigned to different classes by the individual classification models were considered in the fusion study (25 out of 98). Of these, only two samples (nº 77 and nº 94) could not be solved by high-level data fusion.

Table 2 shows two examples of samples assigned by the high-level data fusion strategy: one was solved and one was not. Sample nº 92 (true class: class 1, Asia) was properly assigned to class 1 by the HPLC-UV model and multiply assigned to class 1 and class 3 by the HPLC-CAD model. Finally, it was properly assigned to class 1 by the four fuzzy connectors (bold values). Sample nº 77 (true class: class 3, America) was wrongly assigned as class 2 by HPLC-UV and multiply assigned to class 2 and class 3 by HPLC-CAD. And after applying the ensemble decision, it was wrongly assigned to class 2 (bold values). The high-level data fusion strategy provides significantly better classification results than the individual techniques (**table 3**), with percentages higher than 93% for all three classes.

TABLE 2

Mid-level data fusion was used to make a joint analysis of palm oil samples with HPLC-UV and HPLC-CAD data. In the first approach, the PCA scores were independently calculated from the autoscaled raw data. The scores of the first 20 PCs were kept and accounted for 94.5% and 87.7% of the cumulative variance for HPLC-UV and HPLC-CAD data, respectively. These scores were combined (concatenated) to make a single matrix of dimension 98x40, which involved a significant reduction in the number of variables (from 5038 to 40). Scores were then normalized and pretreated with GLSW to obtain the PLS-DA model. As with the previous results, the eleven samples from class 1

were used as the test set and the remaining 87 (from class 1, 2 and 3) as the training set. The classification results obtained using the first six latent variables were significantly better than those obtained using the individual models (**table 3**).

TABLE 3

In the second approach, the variables were selected by iPLS, which shows those variables that most contribute to the correct classification of samples. IPLS was applied independently to each data set. The number of intervals into which each chromatogram was divided was determined by bearing in mind that there were more initial variables in HPLC-UV than in HPLC-CAD, and that the aim of this study was to make just one iPLS iteration.

In the HPLC-UV data, the initial variable set (3430) was divided into 17 intervals of 202 variables. Figure 3a shows the initial 17 intervals. Those intervals whose misclassification rate was lower than or similar to the rate given by the full chromatogram were selected. They were the following: interval 1 (region 1 in the chromatogram, figure 1) which corresponds to variables 1–202; intervals 8 and 9, which correspond to variables 1415–1818 (region 3 in the chromatogram); and intervals 14 and 16, which correspond to variables 2626–2828 and 3030–3232 (region 4 in the chromatogram). It can be observed that no variables from region 2 were selected and that regions 3 and 4 had the greatest influence. Finally, 1012 variables were selected to be concatenated with the ones from the HPLC-CAD data.

FIGURE 3

A similar procedure was applied to the HPLC-CAD data. The raw data (1608 variables) were divided into 12 equally sized subintervals with 134 variables each. Figure 3b shows the 12 intervals. Using the same criteria explained in the paragraph above, the selected intervals were 1, 3, 5, 6, 7, 8, 10 and 11, which corresponded to variables 1–134 (region 1 of the chromatogram), 269–402 (regions 1 and 2 of the chromatogram), 537–1072 (regions 2 and 3) and 1206–1474 (region 4). In this case, regions 2 and 3 were the predominant regions. Finally, 1072 variables were selected to be concatenated with variables from the HPLC-UV data, which led to a matrix of dimensions 98x2085. This

reduced the original number of variables (5042) by more than half to establish the PLS-DA model, which retains the first three latent variables.

Table 3 shows the overall classification results obtained when the two individual PLS_DA models and the three data fusion models were used. A greater percentage of samples were classified with high- and mid-level data fusion than with individual techniques. All three data-fusion strategies improved the classification for the training and test sets (in brackets). This improvement included samples from America (class 3) which were difficult to classify with the independent techniques. A detailed look at the results of high-level and mid-level iPLS data fusion shows that almost 100% of the assignations were correct for the three classes studied (except class 3 with high-level fusion). Although the classifications with mid-level scores data fusion can be considered satisfactory (higher than 90% for classes 1 and 2), the overall results are lower than for the other two fusion strategies with the exception of the test set (class 1) for which 100% of the samples were correctly classified.

5. CONCLUSIONS

This study shows the advantages of applying data fusion strategies that provide complementary information. All three data-fusion methodologies were better at authenticating the geographical origin of palm oil than the individual models (HPLC-UV and HPLC-CAD), which shows that the information obtained from both chromatograms (HPLC technique coupled to two different detectors) has a synergistic effect.

Nowadays, many laboratories have a great deal of analytical equipment so all sorts of data-fusion strategies can be used in multivariate approaches. Mid-level data fusion requires variables to be previously selected by any variable selection technique. In this study, classification results were best when the iPLS method was used.

High-level (decision) data fusion has the extra -advantage that it can be applied to all types of measurement, since it combines individual multivariate results (assignations). Fuzzy aggregation connectives have proven to be a good, simple tool for classification analysis. However, it should be borne in mind that measuring by more than one technique represents an additional cost.

References

1. Mba, O.I.; Dumont, M.J.; Ngadi, M., Palm oil: Processing, characterization and utilization in the food industry – A review, Food Bioscience 10 (2015) 26-41.

2. Ferdous Alam, A.S.A.; Er, A.C; Begum, H., Malaysian oil palm industry: Prospect and problem, J. Food. Agric. Environ. 13 (2015) 143-148.

3. Sambanthamurthi, R.; Sundram, K.; Tan, Y.A., Chemistry and biochemistry of palm oil, Prog. Lipid. Res. 39 (2000) 507-558.

4. Toy Gee, P., Analytical characteristics of crude and refines palm oil and fractions, Eur. J. Lipid. Sci. Tech. 109 (2007) 373-379.

5. Lin, S.W., Palm oil. In Vegetable oils in food technology: composition, properties and uses, Gunstone, F.K., Eds.; Blackwell Publising: Boca Raton (FL) (2002) 59-97.

6. Barcelos, E.; De Almeida Rios, S.; Cunha, R. N.V.; Lopes, R.; Motoike, S.Y.; Babuychuk, E.; Skiycz, A.; Kushnir, S., Oil palm natural diversity and the potential for yield improvent, Front. Plant. Sci. 6 (2015) 1-16.

7. Henson, I.E., A Brief History of the Oil Palm. In Palm Oil: Production, Processing, Characterization, and Uses, O.M. Lai, C.P. Tan, C.C. Akoh Eds.; AOCS Press: Urbana (IL) (2012) 1-31.

8. RSPO, Roundtable on Sustainable Palm Oil. URL (http://www.rspo.org/about) (accessed December, 2016).

9. Tres, A.; Van der Veer, G.; Alewijn, M.; Kok, E.; Van Ruth, S.M., Palm oil authentication: Classical methodology and state-of-the-art techniques. In Oil Palm: Cultivation, Production and Dietary Components, Penna, S.A Ed.; Nova Science Publishers, Inc: Hauppauge (NY) (2011) 1-44.

10. Ruiz Samblás, C.; Arrebola Pascual, C.; Tres A.; Van Ruth, S.; Cuadros Rodríguez, L., Authentication of geographical origin of palm oil by chromatographic fingerprinting of tryacyglycerols and partial least square-discriminant analysis, Talanta 116 (2013) 788-793.

11. Tres, A. Ruiz Samblás, C.; Van der Veer, G.; Van Ruth, S.M., Geographical provenance of palm oil by fatty acid and volatile compound fingerprinting techniques, Food Chem. 137 (2013) 142-150.

12. Cuadros Rodríguez, L., Ruiz Samblás, C., Valverde Som, L., Pérez Castaño, E., & González Casado, A., Chromatographic fingerprinting: an innove approach for food 'identitation' and food authentication- A tutorial, Anal. Chim. Acta 909 (2016) 9-23.

13. Pérez-Castaño, E.; Ruiz-Samblás, C.; Medina-Rodríguez, S.; Quirós-Rodríguez, V.; Jiménez-Carvelo, A.M.; Valverde-Som, L.; González-Casado, A.; Cuadros-Rodríguez, L., Comparison of different analytical classification scenarios: application for the geographical origin of edible palm oil by sterolic (NP)HPLC fingerprinting, Anal. Methods 7 (2015) 4192-4201.

14. Lerma García, M.; Lusardi, R.; Chiavaro, E.; Cerretani, L.; Bendini, A.; Ramis-Ramos, G.; Simó-Alfonso, E., Use of triacylglycerol profiles established by high performance liquid chromatography with ultraviolet-visible detection to predict the botanical origin of vegetable oils, J. Chromatography A. 42 (2011) 7521-7527.

15. Boccard, J.; Rutledge.D.N., A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion, Anal. Chim. Acta 769 (2013) 1-4.

16. Diémé, B.; Mavel, S.; Blasco, H.; Tripi, G.; Bonnet-Brilhault, F.; Malvy, J.; Bocca, C.; Andres, C.R.; Nadal-Desbarats, L.; Emond, P., Metabolomics Study of Urine in Autism Spectrum Disorders Using a Multiplatform Analytical Methodology, J. Proteome Res. 14 (2015) 5273-5282.

17. Ramos, P.M.; Ruisánchez, I., Data fusion dual-domain classification analysis of pigments studied in works of art, Anal. Chim. Acta 558 (2007) 274-282.

18. Ramos, P.M.; Callao, M.P.; Ruisánchez, I., Data fusion in the wavelet domain by means of fuzzy aggregation connectives, Anal. Chim. Acta 584 (2007) 360-369.

19. Fernández, C.; Callao, M.P.; Larrechi, M.S., UV-visible-DAD and 1H-NMR spectroscopy data fusion for studying the photodegradation process of azo-dyes using MCR-ALS, Talanta. 117 (2013) 75-80.

20. Godinho, M.S; Blanco, M.R.; Gambarra Neto, F.F.; Lião.; Sena, M.M.; Tauler, R.; De Oliveira, A.E., Evaluation of transformer insulating oil quality using NIR, fluorescence, and NMR spectroscopic data fusion, Talanta 129 (2014) 143-149.

21. Di Anibal, C.V.; Callao, M.P.; Ruisánchez, I., 1H NMR and UV-visible data fusion for determining Sudan dyes in culinary spices, Talanta 84 (2011) 829-833.

22. Márquez, C.; López, M.I.; Ruisánchez, I.; Callao, M.P., FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud, Talanta 161 (2016) 80-86.

23. Nunes, K.M.; Andrade, M.V.O.; Santos Filho, A.M.P.; Lasmar, M.C.; Sena, M.M., Detection and characterisation of frauds in bovine meat in natura by non-meat ingredient additions using data fusion of chemical parameters and ATR-FTIR spectroscopy, Food Chem. 205 (2016) 14-22.

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

25. Louw, L.; Roux, A.; Treadoux, O.; Tomic, T.; Næs, T.; Nieuwoudt, H.H.; Van Rensburg, P., Characterization of selected South African young cultivar wines using FTMIR Spectroscopy gas chromatography, and multivariate data analysis, J. Agric. Food Chem. 57 (2009) 2623-2632.

26. Longobardi, F.; Casiello, G.; Sacco, D.; Tedone, L.; Sacco, A., Characterisation of the geographical origin of Italian potatoes, based on stable isotope and volatile compound analyses, Food Chem. 124 (2011) 1708-1713.

27. Charve, J.; Chen, C.; Hegeman, A.D.; Reineccius, G.A., Evaluation of instrumental methods for the untargeted analysis of chemical stimuli of orange juice flavor, Flavour Fragr. J. 26 (2011) 429-440.

28. Borrás, E.; Ferré, J.; Boqué, R.; Mestres, M.; Aceña, L.; Busto, O., Data fusion methodologies for food and beverage authentication and quality assessment, A review. Anal. Chim. Acta 891 (2015) 1-14.

29. Di Anibal, C.V.; Callao, M.P.; Ruisánchez, I., 1H NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs, Talanta 86 (2011) 316-323.

30. Foca, G., Cocchi, M.; Li Vigni, M.; Caramanico, R.; Corbellini, M.; Ulrici, A., Different feature selection strategies in the wavelet domain applied to NIR-based quality classification models of bread wheatflours, Chemometrics and Intelligent Laboratory Systems 99 (2009) 91-100.

31. Biedermann, M., Grob, K., and Mariani, Transesterification and on-line LC-GC for determining the sum of free and esterified sterols in edible oils and fats, European Journal of Lipid Science and Technology 95 (1993) 127-133.

32. Jiménez-Carvelo, A., Pérez-Castaño, E., González-Casado, A., & Cuadros-Rodríguez, L., One input-class and two input-class classifications for differentiating olive oil from other edible vegetable oils by use of the normal-phase liquid chromatography fingerprint of the methyl-transesterified fraction, Food Chemistry 221 (2017) 1784-1791.

33. M. Bylesjö, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J. Trygg, OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification, J. Chemometr. 20 (2006) 341–351.

34. Barker, M.; Rayens, W. Partial least squares for discrimination, J. Chemometr. 17 (2003) 166-173.

35. Wise, B.; Gallagher, N.B.; Bro, R.; Shaver, J.M.; Winding, W.; Kich, R.S. PLS_Toolbox 7.0.2 for use with MATLAB. Eigenvector Research Incorporate, Manson (WA).

36. Martens, H.; Høy, M.; Wise, B.M.; Bro, R. and Brockhoff, P.B., Pre-whitening of data by covariance weighted pre-processing, J. Chemometr. 17 (2013) 153-165.

37. Zorzetii, B.M.; Shaver, J.M; Harynuk, J.J., Estimation of the age of a weathered mixture of volatile organic compounds, Anal. Chim. Acta 694 (2011) 31-37.

38. Rozenstein, O.; Paz-Kagan, T.; Salbach, C.; Karnieli, A., Comparing the effect of preprocessing transformations on methods of land-use classification derived from spectral soil measurements, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 99 (2014) 1-12.

39. Andersen, C.M.; Bro, R., Variable selection in regression-a tutorial, J. Chemometr. 24 (2010) 728-737.

Figure captions

Figure 1. Chromatograms of one palm-oil sample analyzed by a) HPLC-UV and b) HPLC-CAD

Figure 2. PC1 vs PC2 score plots of HPLC-CAD data: a) Asia, b) Africa, and c) America

Figure 3. iPLS plot obtained from a) HPLC-UV data, b) and c) HPLC-CAD data. The dashed line is the RMSECV value for the global models. The numbers on the axes along the bottom indicate the optimal latent variables in the iPLS method for each interval.

Table 1

Number of samples assigned to each class using the models established independently with the HPLC-UV and HPLC-CAD data sets. The numbers in brackets refer to the test set data. Class 1: Asia, Class 2: Africa, Class 3: America.

Predicted	HPLC-UV			HPLC-CAD		
	True class			True class		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Class 1	40 (8)	2	0	43 (8)	0	0
Class 2	2 (2)	23	2	0 (1)	24	0
Class 3	0 (0)	1	11	1 (0)	1	11
Multiple	2 (0)	1	2	0 (1)	2	3
No assigned	0 (1)	1	0	0 (1)	1	1
Total	44 (11)	28	15	44 (11)	28	15

Table 2

Class assignation of samples using high-level data fusion. Class 1: Asia; class 2: Africa; class 3: America.

	PLS-DA class assignation values			Ensemble decision
	class 1	**class 2**	**class 3**	
Sample nº92				
HPLC-UV	**0.883**	0.323	0.017	
HPLC-CAD	**0.625**	0.031	**0.627**	
Minimum	**0.625**	0.031	0.017	class 1
Maximum	**0.883**	0.323	0.627	class 1
Product	**0.552**	0.010	0.011	class 1
Average	**0.754**	0.177	0.322	class 1
		Majority vote		**Class 1**
Sample nº77				
HPLC-UV	0.009	**0.651**	0.446	–
HPLC-CAD	0.000	**0.742**	**0.923**	
Minimum	0.000	**0.651**	0.446	class 2
Maximum	0.009	0.742	**0.923**	class 3
Product	0.000	**0.483**	0.412	class 2
Average	0.005	**0.696**	0.685	class 2
		Majority vote		**Class 2**

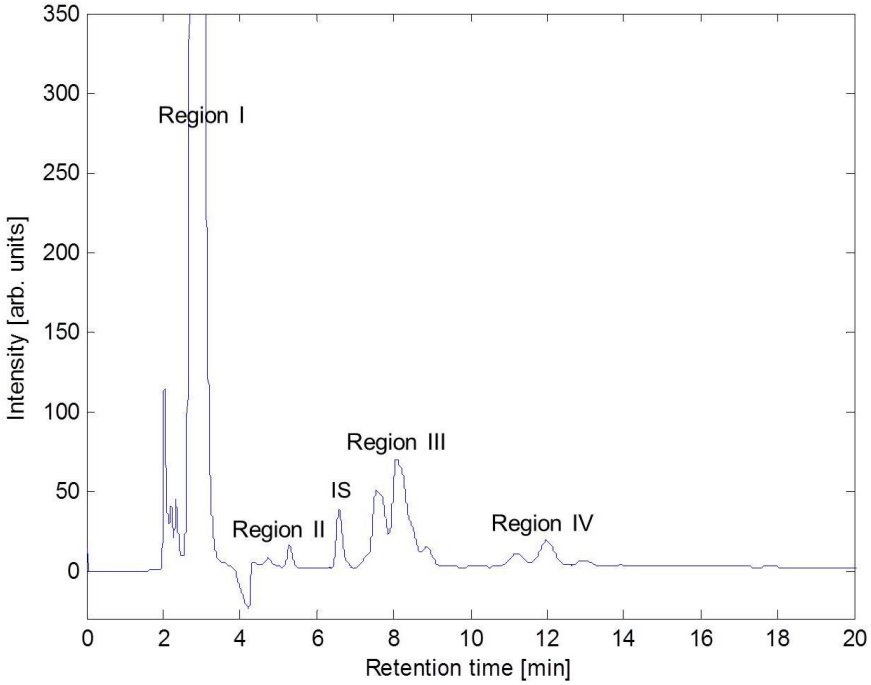
Table 3

Percentages of samples assigned with the PLS-DA models (individual and fusion strategies), Class 1: Asia, Class 2: Africa, Class 3: America.

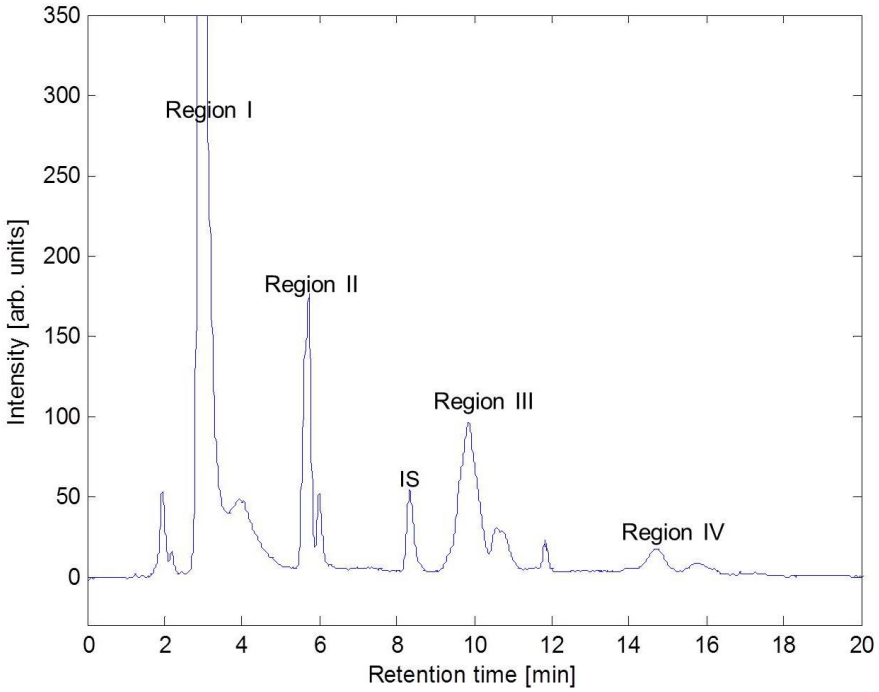
PLS-DA Models	nº LVs	Class 1	Class 2	Class 3
HPLC-UV	6	91 (73)	82	73
HPLC-CAD	6	98 (73)	86	73
High-level_UV-CAD	-	100 (91)	100	93
Mid-level scores_UV-CAD	6	95 (100)	93	87
Mid-level iPLS_UV-CAD	3	100 (91)	100	100

1 *<Figure 1>*

a)



b)

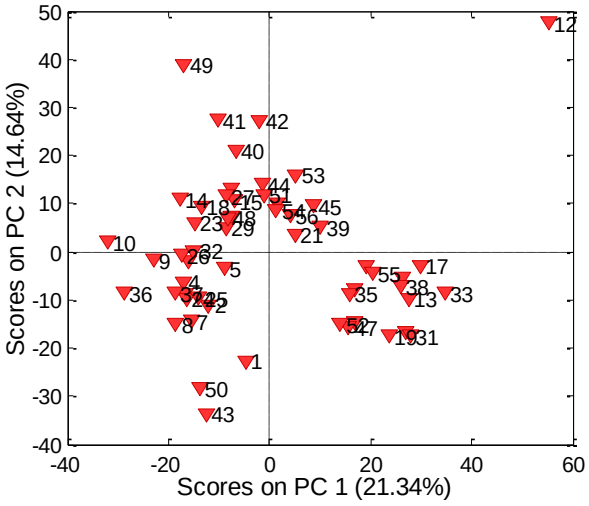


2

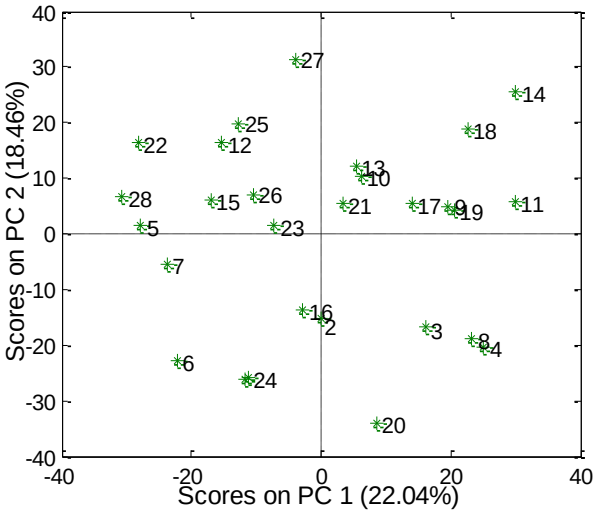
3

<Figure 2>

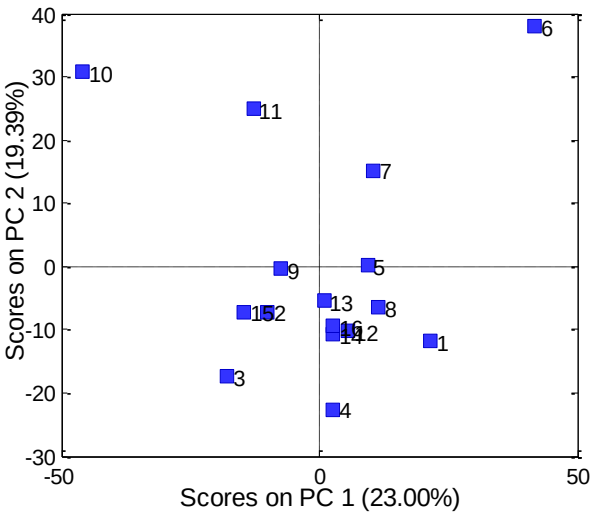
a)


Scores on PC 2 (14.64%)
Scores on PC 1 (21.34%)

b)

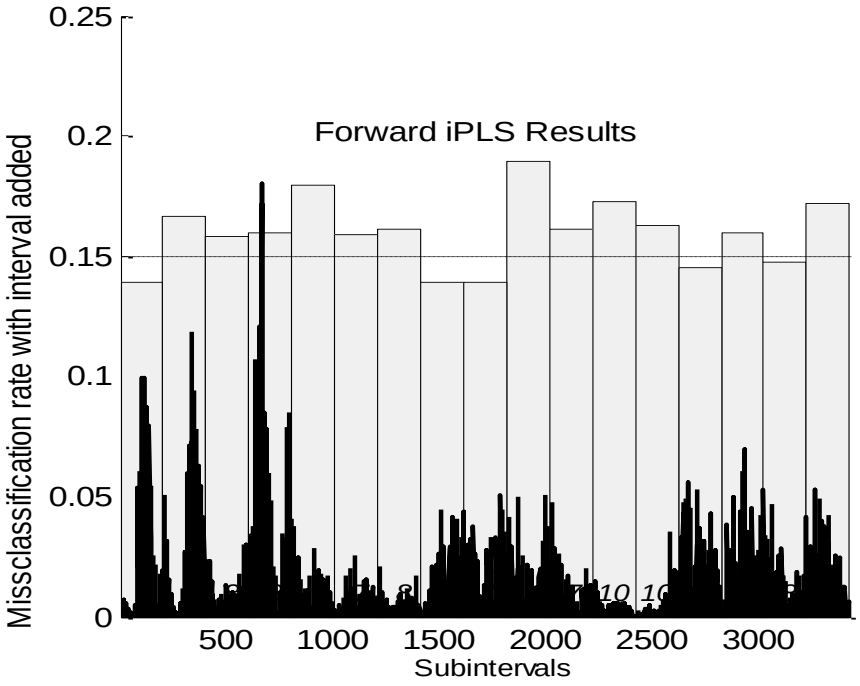

Scores on PC 2 (18.46%)
Scores on PC 1 (22.04%)

c)

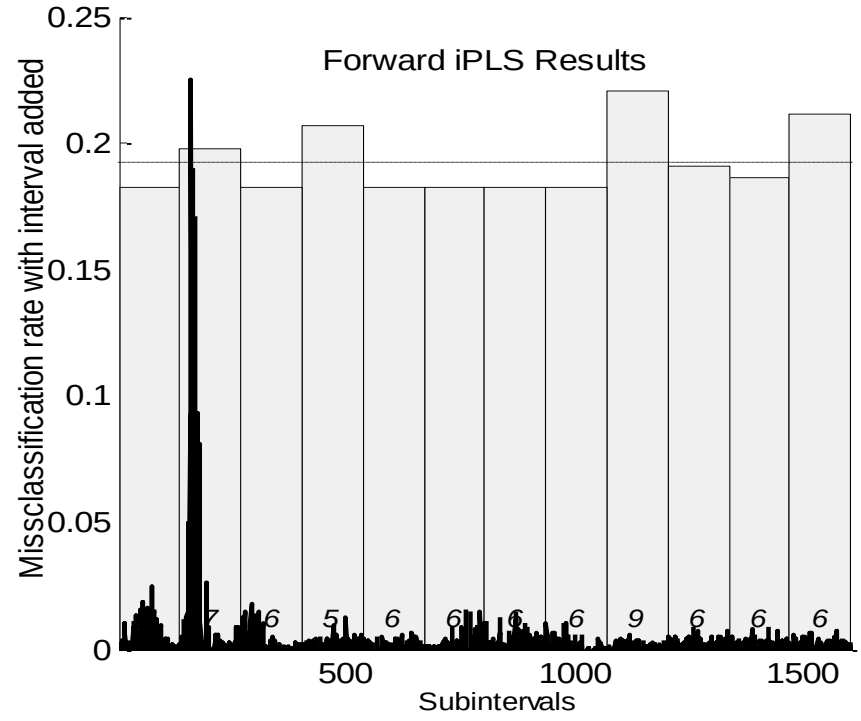

Scores on PC 2 (19.39%)
Scores on PC 1 (23.00%)

<*Figure 3*>

a)


Forward iPLS Results
Missclassification rate with interval added
Subintervals
0.25
0.2
0.15
0.1
0.05
0
500
1000
1500
2000
2500
3000

b)


Forward iPLS Results
Missclassification rate with interval added
Subintervals
0.25
0.2
0.15
0.1
0.05
0
500
1000
1500
7
6
5
6
6
6
6
9
6
6
6