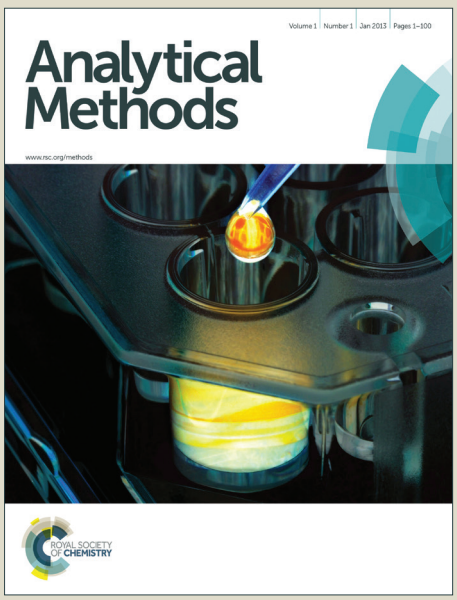
Analytical Methods

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

COMPARISON OF DIFFERENT ANALYTICAL CLASSIFICATION SCENARIOS: APPLICATION FOR GEOGRAPHICAL ORIGIN OF EDIBLE PALM OIL BY STEROLIC (NP)HPLC FINGERPRINTING

Estefanía PÉREZ-CASTAÑO [a,*], Cristina RUIZ-SAMBLÁS [a], Santiago MEDINA-RODRÍGUEZ [a,b], Verónica QUIRÓS-RODRÍGUEZ [a], Ana M. JIMÉNEZ-CARVELO [a], Lucía VALVERDE-SOM [a], Antonio GONZÁLEZ-CASADO [a], Luis CUADROS-RODRÍGUEZ [a]

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

[b] Department of Signal Theory, Networking and Communications, CITIC-UGR, University of Granada, c/ Periodista Rafael Gómez, E-18071 Granada, Spain.

Abstract

This work shows how the best scenario, resulting to apply two chemometric classifiers on different analytical data set from the same sample set, could be chosen according to the classification results. On this way, several classification quality features such as sensitivity (or recall), specificity, positive (or precision) and negative predictive values, Youden index, positive and negative likelihood ratios, F- measure (or F- score), discriminant power, efficiency (or accuracy), AUC (area under the receiver operating curve), Matthews correlation coefficient, Kappa coefficient, overall agreement probability, overall agreement probability from chance and overall Kappa coefficient are described and discussed. As application example, two sterolic chromatographic fingerprints obtained from two different normal-phase HPLC systems are used to discern the geographical origin (South-East Asia, West Africa and South America) of edible palm oil. In each case, two conventional and well-known chemometric classification methods are applied: soft independent modelling by class analogy (SIMCA) and partial least squares-discriminant analysis (PLS-DA).

Keywords

Classification scenario comparison; liquid chromatography sterolic fingerprints; food authenticity; palm oil.

* Corresponding author: phone: +34 958240797; fax: +34 958243328; email: stefani@ugr.es.

Analytical Methods Accepted Manuscript

1. INTRODUCTION

Palm oil is one of the most consumed edible oils in the world for its price and its properties. It is obtained from the fruit of the palm (*Elaeis guineensis*). The plant is native from Western Africa, later extended to South America in the XVI century, and more recently, in the XIX century, it was introduced in Eastern Asia from America. Crude palm oil is a semi-solid fat at room temperature which has a high stability compared with other vegetable oils[1,2,3]. Knowing the geographical origin of the palm oil and the traceability of palm oil supply chain[4] is interesting from an environmental point of view, because the palm oil exploitation is cause of deforestation and loss of biodiversity in some tropical countries[5]. Furthermore, an indication of the geographical origin of food products is an increasingly important consumer demand on food labelling because consumers consider it to be an added-value to the product. Other possible reasons of this growing interest are: specific culinary, organoleptic qualities, or purported health benefits associated with regional products[6].

For the characterization and verification of the geographical origin of palm oils, as well as other food products, it is necessary to find specific qualities derived from its place of production (markers) and which are subject to specific local factors such as climate and terrain[7]. An example of this approach is the use of molecular markers as the DNA fingerprinting. Alternatively chemical markers could be also used. In literature sterol profiles and total sterol contents have been used as tools to value the oil authenticity[8,9,10,11] because each vegetal species has a characteristic compositional profile of sterols[12,13]. This suggests that sterol profiles might be suitable candidates to develop an analytical tool to verify the geographical origin of palm oil.

Phytosterols are a group of bioactive compounds, with a derived cyclopentane-perhydro-phenanthrene four-ring molecular structure. These compounds are present in plants and they are differentiated by the number of carbon atoms of the side chain, and by the nature of the same. There are three classes of sterols: 4-des, 4-mono and 4,4'-dimethylsterols that could be found in free form or esterified with fatty acids and other conjugates[14,15]. Crude palm oil contains about 0.7-0.8 g/kg of total sterols, however it must be taken into mind that the refining processes affect the concentration and the compositional profile of sterols because occur hydrolysis and oxidation processes that destroy sterols.

There are different methods and techniques (chromatographic and non-chromatographic) for the analysis of sterols in vegetable oil. Chromatographic methods are the most commonly used against the non-chromatographic ones because much more information is obtained about the sterols composition present in the sample[15,16,17,18,19].

Depending on the information that is wanted, different sample preparation steps could be applied. Generally, phytosterol analysis includes: an extraction of the lipid fraction, acid

Analytical Methods Accepted Manuscript

hydrolysis or basic hydrolysis (saponification) to release phytosterols, an extraction of the unsaponifiable fraction, the separation or partial purification of sterols and finally a chromatographic analysis. However, in some specific cases, in addition to the above steps, the formation of derivatives of phytosterols is previously required to chromatographic analysis. The routine methods employ a saponification reaction although it could be replaced by a transesterification reaction with similar results[20].

Analytical liquid chromatography has not been much applied for determining the sterol composition of vegetable oils but it is usually performed in reverse phase[16]. Normal-phase HPLC methods can also be used for the absolute quantification of the total amount of phytosterols but these methods show a poor chromatographic resolution and do not provide precise information on the sterol composition[16]. In addition, several conventional detection methods, such as ultraviolet absorption, refractive index, and evaporative light-scattering have been applied. In 2004 the Corona Charged Aerosol detector (CAD) was developed as an alternative, which has ability to accurately measure a wide range of analytes[21]. Nonetheless, only a method has been just reported using CAD for the determination of sterols in vegetable oils[22].

If the chromatographic conditions are properly optimized and a suitable detector is coupled, the yielded chromatogram contains specific and relevant information about the considered product, which could be used for authentication purposes. When the chromatogram is well resolved, the information for each chemical component could be extracted from each peak. Instead, if the chromatogram is shaped on a broad and comprehensive band, the intrinsic information is not evident and the chromatographic fingerprinting methodology should be then applied[23,24]. By reaching that point, the application of multivariate chemometric tools is required to extract the useful information from the chromatographic raw data[25,26,27]. Some examples of the use of chromatographic fingerprints merging with classical chemometric methods have been recently reported by our research group with satisfactory results in the authentication of vegetable edible oils[28,29,30,31]. In addition, data mining classification methods have also been used for edible vegetable oils authenticity applications[32].

In the last few years, some papers have been published using classic supervised pattern recognition methods for vegetable oil classification and authentication based on their sterol composition. In most of them, the data matrices are made up from the sterol contents (concentrations or compositional data)[8,10,33] and only a work set the data matrix from the sterolic chromatographic fingerprint[11]. In the same way, there is not enough background about the authentication of palm oil by using this methodology and, as far as we know, only two studies have been published, but they use fingerprintings from volatile compounds[34] and triacylglycerols[35].

Analytical Methods Accepted Manuscript

In this study, two sterolic chromatographic fingerprints obtained from two different normal-phase HPLC systems are used to discern the geographical origin (South-East Asia, West Africa and South America) of edible palm oil. The aim of this paper is to show how the obtained results from different classification scenarios would be compared in order to select, if possible, the best combination of classification chemometric methods and/or measured analytical data set. The comparison is based on different quality classification metrics which are defined in this work, such as sensitivity (or recall), specificity, positive (or precision) and negative predictive values, Youden index, positive and negative likelihood ratios, F-measure (or F-score), discriminant power, efficiency (or accuracy), AUC (area under the receiver operating curve), Matthews correlation coefficient, Kappa coefficient, overall agreement probability, overall agreement probability from chance and overall Kappa coefficient.

2. MATERIALS AND METHODS

2.1. Instrumentation

The analyses were performed using two different HPLC systems. The first one was a Konik Model 560 (Konik-Tech, Sant Cugat del Valle, Barcelona, Spain) with a quaternary pump, a column oven, an autosampler with a 20 µL loop, and an UV-Vis detector.

The second one was an Agilent 1100 Series (Agilent Technologies, Santa Clara, CA, USA) with a quaternary pump, degasser, autosampler and thermostatted HPLC column compartment Eppendorf CH-30 (Eppendorf, Hamburg, Germany). Detection was carried out with a Corona CAD (ESA Bioscenses Inc., Chemlsford, MA, USA).

2.2. Chemicals

Solid standards of stigmasterol, campesterol and cholestanol (internal standard, IS) were provided from Sigma-Aldrich (Steinheim, Germany) and β-amyrin was from EXTRASYNTHESE (Genay, France).

Sodium methoxide, citric acid monohydrate and anhydre sodium sulphate were provided from Alfa-Aesar (Karlsruhe, Germany), Sigma-Aldrich (Steinheim, Germany), and Panreac Quimica (Barcelona, Spain) respectively. The solvents employed (n-hexane, 2-propanol, methanol and methyl tert-butyl ether (MTBE), BDH Prolabo, HPLC grade), were purchased from VWR International (Madrid, Spain). All the aqueous solutions were prepared with Milli-Q deionized water (Millipore, Bedford, MA).

The nitrogen (99.9999%) used for CAD detector was provided from Air Liquid (Madrid, Spain).

Analytical Methods Accepted Manuscript

2.3. Samples

A total of 102 crude palm oil samples were supplied from RIKILT-Institute of Food Safety Wageningen University, (Wageningen, The Netherland). The samples coming from the main continents of palm oil production: South-East Asia (56 samples from Malaysia, Indonesia, Papua New Guinea and Salomon Islands), West Africa (30 samples, from Ghana, Guinea, Cote d'Ivoire, Nigeria and Cameroon) and South America (16 samples from Brazil). Table 1 shows in detail the origin of the different samples tested.

TABLE 1

2.4. Sample preparation

Prior to chromatographic analysis, a methylation reaction was applied on the palm oil samples. This reaction replaces the usual saponification and isolation processes, and it has the advantage of being less time-consuming and requiring less sample amount. The applied procedure is similar to the one previously described by Biederman[20] and Kamm [36].

The transesterified sample solutions were frozen (-20 ºC) and kept in the dark until analysis. Just before the chromatographic analysis, 500 µL of this solution was added in a 2 mL HPLC vial, and then 120 µL of 0.05% (w/w) cholestanol solution in n-hexane was added as control internal standard. Finally the mixture was diluted with 1000 µL of n-hexane. The vial was sealed and vortexed for 20 s. This solution was prepared just for analysis.

2.5. LC Conditions

HPLC analysis is carried out on a (250 x 4 mm i.d., 5 µm) column Lichrospher® 100 CN maintained at 25 °C. The composition of the eluent was n-hexane/2-propanol (99:1, v/v) at a flow rate of 1.2 mL/min and a run time of 20 min. No gradient was applied. UV detection (Konik equipment) was performed at 202 nm. For CAD monitoring (Agilent equipment), a 100 pA output range was used and nitrogen gas pressure was adjusted to 35 psi.

Chromatographic data handling were performed by a Konikrom Plus software (version 3.0.5) for HPLC Konik, and ChemStation software (version A.10.02) for HPLC Agilent.

2.6. Chemometrics

The raw data files for each chromatogram were exported in a CSV file (*comma-separated values*) from the instrument software to the MATLAB environment (version 7.8, R2009a, The Mathworks Inc. MA, USA). Initially, each chromatographic fingerprinting is coded in a two-data vectors (time/intensity) with 4500 (UV-Vis Konik) and 2400 (CAD ESA) elements (variables), depending of the data acquisition rate of each HPLC detector.

Analytical Methods Accepted Manuscript

All the intensity data vectors, one for each oil sample, from the same chromatographic detector are merged in a single data matrix (X- block matrix) composed of 102 rows (palm oil samples) and a certain number of columns (variables) that varies depending on the measuring time and the rate of acquisition of data of each HPLC detector, as it has been described in the previous paragraph. The elements of X-matrix are the intensities values of the chromatographic signals. In addition, a new column is added to each data matrix specifying the class (geographical continent) of each sample, titled by an alphanumeric code, for example "AF", AM" and "AS" for Africa, America and Africa. This column set up the Y-block of the data matrix.

Next, a preprocessing of each X-matrix was carried out using a home-made MATLAB function, named "MEDINA" (version 07). This function makes use of some of the functions contained in MATLAB Bioinformatics Toolbox™ software to improve the quality of raw chromatographic data, and also the "icoshift" algorithm (version 1.2) for solving signal alignment problems in chromatographic data [37]. Basically, the chromatographic data processing consists of the following stages (see Supplementary Information for details): (1) selection of the interval of interest from chromatograms; (2) decimation of the raw chromatographic data. It makes possible to resample the signal into a more manageable chromatographic data vector, preserving the information contained in the chromatogram (in this case, a decimation factor of 2 was used); (3) de-noising and smoothing of the chromatographic signal using a least-squares digital polynomial filter (i.e., a Savitzky-Golay filter); (4) baseline correction using the "msbackadj" function (available in the above mentioned MATLAB toolbox); (5) alignment of the chromatographic profiles with the "icoshift" algorithm. Finally, a mean centring of the chromatographic data matrix was applied (i.e., the subtraction of the mean from each data vector) prior to the statistical analysis. Once the chromatographic data preprocessing was carried out, it was then possible to use classification and statistical learning tools to create classifiers.

For multivariate chemometric pattern recognition PLS_Toolbox (version 7.5.2, Eigenvector Research, Wenatchee, WA) was used. The performance features of each classifier, described in the next section, were calculated on the validation test from a home-designed MS Excel™ spreadsheet (version 14.0, 2010).

Exploratory analysis and classification methods

Principal components analysis (PCA) is a type of exploratory data non-supervised analysis which can be applied to any X-matrix[38,39,40]. The main aim of PCA is the dimension reduction when the variables are correlated. A few new variables are defined, named principal components (PCs), as linear combination of the original variables in order to explain as much variability as possible with the smallest number of PCs.

Analytical Methods Accepted Manuscript

Two habitually classification methods were then applied. A venetian blinds object out cross-validation procedure was adopted to optimize all the built models.

Soft independent modelling by class analogy (SIMCA) is a well-known class-modeling classification method based on principal component analysis[27, 41]. Each class is independently modelled by a PCA so that each model defines the boundary regions for each class. The number of principal components of each category was determined using the rule of thumb based on the cross-validation, which gives the model optimal prediction properties.

The unknown samples are applied to the model of prediction; they are compared to the defined classes and assigned to a class according to their similarity (analogy). In this study, the recognition is made based on the distance $d_{i,C}$ of each i-sample from each C-class. This is calculated by applying the following equation:

$$d_{i,C} = \sqrt{\left(\frac{Q_{i,C}}{Q_{C(0.95)}}\right)^2 + \left(\frac{T^2_{i,C}}{T^2_{C(0.95)}}\right)^2}$$

where $Q_{i,C}$ and $T^2_{i,C}$ are the computed statistics Q-residuals and the T^2-Hotelling respectively, calculated from the corresponding C-class PCA model, and $Q_{C(0.95)}$ and $T^2_{C(0.95)}$ are the values for a 95% confidence level. The chosen classification threshold[42] was $d_{i,C} = \sqrt{2}$. For the class assignment of a sample, the calculated distance to such class have to be lesser than or equal to $\sqrt{2}$. On the contrary, if the distance is always larger, the sample is unclassified (class no assigned). If a sample is simultaneously assigned to more of two classes (because the distance to both ones was lesser than or equal to $\sqrt{2}$), the sample will be definitively assigned to the class whose distance value is lesser.

Partial least squares-discriminant analysis (PLS-DA) is a linear discrimination method based upon the classical PLS regression method[43] for building predictive models. The goal of PLS regression is to provide dimensionality reduction in an application where the response variable (Y-block) is related to the predictor variables (X-block). The used software is only able to perform binary classifications. So, for n-class classification is necessary build n two-class (binary) models[44].

PLS-DA is applied to develop a model that predicts the representative class value (between 0 and 1) for each sample in each classification. To make a class assignment, the discrimination thresholds and the probability of a sample belonging to a specific class were calculated based on a Bayesian approach. The unknown samples will be correctly classified always than the assigned class value is equal or greater than the threshold value; otherwise the sample will be unclassified. A sample could be classified into two classes if it has a predicted value greater than the threshold value in both classifications. In this case, the sample will be assigned to the class whose predicted value is closer to 1.

Analytical Methods Accepted Manuscript

External validation of classifiers

In order to apply a proper external validation of the classification/prediction models, the original data set was divided into two data sets: (1) a training set, used to establish the chemometric models; and (2) a validation (or testing) set, in order to test the validity of the models. Approximately 30% of the samples from each class were randomly chosen to constitute the validation set. Table 2 shows, more specifically, the composition of the samples that were used in in both sets of calibration and test.

TABLE 2

3. BACKGROUND: QUALITY PERFORMANCE OF CLASSIFICATION SCENARIOS

The empirical evaluation of classification scenarios is a matter of on-going debate between researchers where classification scenario is referred to the combination of classifier and analytical data in a particular case. The assessment of the quality classification performance without focusing on a class is the most general way of comparing the quality of the classification results. In order to quantify this quality, several performance features have been proposed as metrics[45,46,47]. The estimation of such metrics is based on measuring the classifier's ability to distinguish classes and, consequently, to avoid failure in classification. Although most performance features in use today focus on a classifier's ability to identify classes correctly, in certain cases, other properties such as failure avoidance or class discrimination may also be useful[48].

Quality features for classification are built from a contingency table which records correctly and incorrectly assigned examples for each class. The corresponding ternary contingency table is shown in Table 3.

TABLE 3

The quality performance features of the different classifiers are calculated by reducing the ternary contingency table to three binary contingency tables [49] because the quality parameters are described to binary classification. A binary contingency table is a square of 2×2 where the rows represent the number of classifier predictions and the columns are the actual value of class. Table 4 presents a standard contingency table for binary classification.

TABLE 4

Analytical Methods Accepted Manuscript

The final value is obtained from the average of the corresponding features obtained for binary classifiers, weighting with respect to the number of samples in each class.

The different quality metrics used in this paper for evaluating the classification results are shown below:

Sensibility (SENS) (or recall). It indicates the probability of classifying a sample as positive really, *i.e.*, the confidence in a positive result for a sample of the label class is obtained. The range of values for this feature is 0 to 1.

$$\mathrm{SENS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{TL}}$$

Specificity (SPEC). It indicates the probability of classifying a sample as negative really, *i.e.*, the confidence that a negative result for a sample of non-label class is obtained. It is also ranged between 0 and 1.

$$\mathrm{SPEC} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = \frac{\mathrm{TN}}{\mathrm{TnL}}$$

Sensitivity and specificity assess the effectiveness of the classifier on a single class, positive and negative respectively.

Positive predictive value (PPV) (or precision). It estimates the predictive power of the classifier; this metric quantifies the precision of the classifier to identify examples of a given class. PPV measures the proportion of correctly assigned positive examples and its value varies between 0 and 1.

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = \frac{\mathrm{TP}}{\mathrm{AP}}$$

Negative predictive value (NPV). The complement of PPV in this context appears in the form of the negative predictive value (NPV), which measures the proportion of correctly assigned negative examples. The range of values is also between 0 and 1.

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{AN}}$$

Youden's index (YOU). It evaluates the classifier's ability to avoid failure; it is derived from sensitivity and specificity. This parameter varies between 0 and 1.

$$\mathrm{YOU} = \mathrm{SENS} - (1 - \mathrm{SPEC})$$

Analytical Methods Accepted Manuscript

The likelihood ratios (LR). It is possible to distinguish between positive likelihood ratio, LR(+), and negative likelihood ratio, LR(–). The positive likelihood ratio represents the ratio between the probability to predict an example as positive when it is truly positive, and the probability to predict an example as positive when actually it is not positive:

$$\mathrm{LR}(+) = \frac{\mathrm{SENS}}{1 - \mathrm{SPEC}}$$

while the negative likelihood ratio is the ratio between the probabilities to predict an example as negative when it is actually positive, and the probability to predict an example as negative when it is truly negative:

$$\mathrm{LR}(-) = \frac{\text{1-SENS}}{\mathrm{SPEC}}$$

Higher positive likelihood ratio and a lower negative likelihood ratio mean better performance on positive and negative classes respectively.

F-measure (F). It is defined as the harmonic mean of precision and sensibility. It is a composite feature which benefits classifiers with higher sensitivity and challenges classifiers with higher specificity. This metric ranges between 0 and 1.

$$\mathrm{F} = 2 \times \frac{\mathrm{SENS} \times \mathrm{PPV}}{\mathrm{SENS} + \mathrm{PPV}}$$

Discriminant power (DP). It does exactly what its name implies: *i.e.*, it assesses how well a classifier distinguishes between positive and negative examples.

$$\mathrm{DP} = \frac{\sqrt{3}}{\pi}\left(\log\frac{\mathrm{SENS}}{1 - \mathrm{SENS}} + \log\frac{\mathrm{SPEC}}{1 - \mathrm{SPEC}}\right)$$

Efficiency (EFFIC) (or accuracy). The most common metric for classifier evaluation, it assesses the overall effectiveness of the classifier by estimating the probability of the true value of the class label. The EFFIC values are included between 0 and 1.

$$\mathrm{EFFIC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{T}}$$

Area under the ROC curve (AUC) (or correct classification rate). The area under the ROC (Receiver Operating Characteristic) curve is a summary indicator of ROC curve quality that can summarize the performance of a classifier into a single metric. Graphically, ROC is plotted as a curve that gives the true positive rate as a function of false positive rate for the same group. AUC is a measure of the ability of the classifier to avoid errors during

Analytical Methods Accepted Manuscript

classification. The AUC varies between 0 and 1 although, in practice, its values should be larger than 0.5.

$$\mathrm{AUC} = \frac{\mathrm{SENS} + \mathrm{SPEC}}{2}$$

Matthews correlation coefficient (MCC). It measures the overall quality of a method classification since it considers mutually accuracies and error rates, and involve all values of the contingency table.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{FN} + \mathrm{TN}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{FP} + \mathrm{TN})}}$$

MCC ranges from 1 for a perfect prediction to -1 for the worst possible prediction. MCC close to 0 indicates a model that performs randomly.

Kappa coefficient (K). It indicates the proportion of agreement after the chance agreement is removed from consideration[50]. It calculated from a probability rate where the numerator is the percent of units in which beyond-chance agreement occurred, and the denominator is the percent of subjects for which one would not expect any agreement by chance.

$$\mathrm{K} = \frac{\mathrm{P_a} - \mathrm{P_c}}{1 - \mathrm{P_c}}$$

In this equation, P_a is the probability term from agreement and P_c is the probability term from chance.

$$\mathrm{P_a} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{T}} = \mathrm{EFFIC} \quad ; \qquad \mathrm{P_c} = \frac{\mathrm{AP} \times \mathrm{TL} + \mathrm{AN} \times \mathrm{TnL}}{\mathrm{T}^2}$$

The values of K are ranged between 1 (the classifiers are in complete agreement) and 0 (there is no agreement among the classifiers other than what would be expected by chance, as defined by P_c).

An overall value for any of the parameters, which have been previously defined, could also be directly calculated as it is explained below.

Overall agreement probability (overall Pa).

$$\text{overall } \mathrm{P_a} = \frac{\sum \mathrm{a_i}}{\mathrm{T}} = \frac{\sum(\mathrm{TP} + \mathrm{TN})}{\mathrm{T}}$$

Overall chance agreement probability (overall Pc).

Analytical Methods Accepted Manuscript

$$\text{overall } P_c = \frac{\sum(TC_i \times AC_i)}{T^2}$$

Overall kappa coefficient (overall K).

$$\text{overall } K = \frac{\text{overall } P_a - \text{overall } P_c}{1 - \text{overall } P_c}$$

4. RESULTS AND DISCUSSION

An example of the chromatograms obtained from the same palm oil sample from Africa by both HPLC systems, is shown in Fig. 1. As it can be observed, the chromatograms are split in four regions. In order to identify the regions corresponding to the sterolic fraction, three aliquots of this palm oil sample were fortified each one with a representative sterol standard: β-amyrin (a dimethylsterol), stigmasterol and campesterol (two desmethylsterols). Next they were analysed by applying the two analytical chromatographic methods. By inspecting where the height is increased, and in accordance with the assignment carried out by Biedermann[20], it could be concluded that the sterols are divided in regions II and III: the region II is associated to the dimethylsterols while the region III contains the methylsterols and desmethysterols and, possibly, other compounds as the fatty alcohols. The region I was not assigned although the large peak should be probably due to the fatty acids methyl esters, whereas region IV would be due to the terpenic alcohols.

FIGURE 1

As previous exploratory analysis, PCA was performed on the X-matrix in order to perceive similar, dissimilar, typical, or outlier samples. Two PCs were enough to explain 88.8% and 90.6% of the cumulative variance from the HPLC-UV and HPLC-CAD fingerprint data, respectively. Both PC1-PC2 scores and PC1 loading plots for each X-matrix are shown in Figure 2.

FIGURE 2

Both scores plots allow to distinguish two groups separated on the first principal component which are correlated with the AMERICA (left) and ASIA samples (right). On the other hand, the AFRICA samples are not grouped and they do not show any trend but they are dispersed on the plotted space. This fact shows that the sterolic chromatographic profiles from AFRICA

Analytical Methods Accepted Manuscript

samples have not a specific pattern and some of them are similar to the ones from samples from AMERICA or ASIA. This fact could be explained by the African common origin of all palm oils. The PC2 does not provide information about the geographical origin.

The PC1 loadings plot shows a profile that coincides with chromatographic region corresponding to region III of the chromatogram, associated to the Δ^5- and Δ^7-desmethysterols (see Figure 1). Therefore, this region contains the significant information about the geographical origin of the palm oils and it will be selected for building the classification models.

In order to apply a SIMCA classification, three PC models were built, from the corresponding training set samples, for each class (AFRICA, AMERICA and ASIA). The number of chosen PCs for each model was respectively 3, 5 and 5 from HPLC-UV fingerprint data, and 3, 4 and 4 from HPLC-CAD fingerprint data. In all cases, the percentage of explained variance was higher than 96%.

In a similar way, the three-class PLS-DA model was trained. The number of latent variables (LVs) chosen for each model was respectively 5 from HPLC-UV fingerprint data, and 3 from HPLC-CAD fingerprint data, with percentages of explained variance for the X-block and Y-block of the data matrix of 94% and 49% for the first model, and 94% and 43% for the second one, respectively.

Once the classification models are defined, the more probable class is assigned to each sample of the validation set. The contingency tables showing the results of the assignment from each classifier are shown in Table 5.

TABLE 5

In general, the samples from AFRICA and ASIA are better classified than the samples from AMERICA. By comparing the results from the two HPLC fingerprinting, it seems that SIMCA classifies better than PLS-DA. However, only by having a look, it would be difficult to decide the best classifier. In order to have a set of appropriate metrics for making this decision, Table 6 collects the pooled performance features the four classification scenarios; each value has been calculated from the three reduced binary contingency tables by weighing the number of actual samples from each class. The empty cells are the consequence of the zero value obtained for some classification rates in the contingency tables. In addition Table 6 collects the overall values of quality performance features (above mentioned).

TABLE 6

Analytical Methods Accepted Manuscript

The classifiers can now be arranged in decreasing order of performance as:

SIMCA(HPLC-UV) > SIMCA(HPLC-CAD) = PLS-DA(HPLC-CAD) > PLS-DA(HPLC-UV)

This ranking is easily established dealing the main features related to the overall classification reliability: efficiency, AUC and Kappa coefficient. These performance features are the best indicators of the classification ability, although practically the same ranking could be obtained starting from anyone of the tabulated features.

However, in strictly technical terms, none of the tested classifiers show an enough assurance as to be applied in order to discern the geographical origin of any sample of palm oil since, in the best case, one of each six-seven samples (15%) would be erroneously classified.

5. CONCLUSIONS

The chemometric classification methods are widely used for food authentication purposes. As input experimental data set, any unspecific chromatographic signal (formally named chromatographic fingerprint) could be used. To make the most of classification performance, different chemical fractions characteristic of the studied material, different chromatographic conditions and different classification methods could be tried. Later, the best classification scenario has to be select in order to be applied in a real framework. In this work, several classification performance quality features have been presented and discussed.

As application example, two classification methods are applied on two sterolic chromatographic fingerprints obtained from two different normal-phase HPLC systems in order to discern the geographical origin of edible palm oil. For each one of the four classification scenarios, the corresponding quality features have been calculated and used to select the best one. For HPLC-UV fingerprint data, the best classifier is SIMCA classification, whereas for the HPLC-CAD one both classification methods behave on a similar way.

Finally, it is remarkable that all above mentioned parameters are applied jointly to different chromatographic fingerprints for the first time.

Acknowledgment

The authors are very grateful to Rikilt Wageningen University and Research Centre to supply the palm oil samples.

SUPPLEMENTARY INFORMATION

Descriptions of the "MEDINA" function for preprocessing of chromatographic data. Details of various processing options, as well as illustrated examples of the effect of different processing steps on a set of chromatographic data are provided.

Analytical Methods Accepted Manuscript

REFERENCES

1 . Y. Basiron, in *Bailey's Industrial Oil and Fat Products. Edible Oil and Fat Products: Edible Oils*, ed. F. Shaidi, Wiley-Interscience, Hoboken, 6th edn., 2005, vol. 2, ch. 8, pp. 333–429.

2 . L. Siew Wai, in *Vegetable Oils in Food Technology. Composition, Properties and Uses*, ed. Frank, D. G.,Wiley-Blackwell, Chichester, 2nd edn., 2011, ch. 2, pp. 25–58.

3 . O.M. Lai, in *Healthful Lipids*, eds. C. Akoh, O.M. Lai, AOCS Press, Urbana, 2005, pp 731–749.

4 . G. van Duijn, *Lipid Technol.,* 2013. **25**, 15–18.

5 . R.J. Orsato, S.R. Clegg, H. Falcão, *J. Change Manag.*, 2013. **13**, 444–459.

6 . S. Kelly, K. Heaton, J. Hoogewerff, *Trends Food Sci.Tech.*, 2005, **16**, 555-567.

7 . A. Tres, G. van der Veer, M. Alewijn, E. Kok, S. van Ruth, in *Oil Palm: Cultivation, Production and Dietary Components,* ed. S.A. Penna, Nova Science, 2011, ch. 1, pp.1-44.

8. M. R. Alves, S. C. Cunha, J. S. Amaral, J.A. Pereira, M. B. Oliveira, *Anal. Chim. Acta.*, 2005, **549**, 166–178.

9. M.J. Lerma García, G. Ramis Ramos, J.M. Herrero Martínez, E. Simo Alfonso, *Rapid Commun. Mass Sp*. 2008, **22**, 973–978.

10. M.J. Lerma García, E.F. Simó Alfonso, A. Méndez, J.L. Lliberia, J.M. Herrero Martínez, *Food Res. Int.* 2011, **44**, 103–108.

11. D. Gázquez Evangelista, E. Pérez Castaño, M. Sánchez Viñas, M.G. Bagur González, *Food Anal. Method.* 2014, **7**, 912–925.

12. W. Kamm, F. Dionisi, C. Hischenhuber, K.H. Engel, *Food Rev. Int.,* 2001, **17**, 249–290.

13. S. Azadmard-Damirchi, *Food Addit. Contam. A.*, 2010, **27**, 1–10.

14. E. Wasowicz, in *Chemical, Biological, and Functional Aspects of Food Lipids*, eds. Z.E. Sikorski, A. Kolakowska, CRC Press, Boca Raton, 2nd edn, 2011,ch.7, pp. 113–134.

15. H. Saoussem, in *Handbook of Analysis of Active Compounds in Functional Foods, eds.* L. Nollet, F. Toldrá, CRC Press, Boca Raton, 2012, ch. 35, pp. 787–804.

16. L. Nyström, in *Analysis of Antioxidant-Rich Phytochemicals,* eds., Z.R. Xu, L. Howard, John Wiley & Sons, Oxford, 2012, pp. 313–351.

17. S.L. Abidi, *J. Chromatogr. A,* 2001, **935**, 173–201.

18. A. Maija Lampi, V. Piieronen, J. Toivo, in *Phytosterols as Functional Food Components and Nutraceuticals*, ed. P.C. Dutta, Marcel Dekker, New York, 2004, chs. 2,3, pp. 33–73.

19. S. Azadmard-Damirchi, P.C. Dutta, in *Olives and Olive Oil in Health and Disease Prevention*, eds. V. Preedy, R. Watson, Elsevier, London, 2010, ch. 27, pp. 249–257.

20. M. Biedermann, K. Grob, C. Mariani, *Fat Sci Tech.* 1993, **95**, 127–133.

21. G.Tadeusz, L. Frederic, S. Roman, S. Pat, *Anal. Chem.* 2006, **78**, 3186–3192.

22. I. Acworth, B. Bailey, M. Plante, P. Gamache, Simple and direct analysis of phytosterols in red palm oil by reversed-phase HPLC and charged aerosol detection. Thermo Fisher Scientific, 2011.

23. J.M. Bosque Sendra, L. Cuadros Rodriguez, C. Ruiz Samblas, A.P. de la Mata, *Anal. Chim. Acta*., 2012, **724**, 1–11.

24. D.I. Ellis, V.L. Brewster, W.B. Dunn, J.W. Allwood, A.P. Golovanov, R. Goodacre, *Chem. Soc. Rev*., 2012, **41**, 5706–5727.

25. L.A. Berrueta, R.M. Alonso Salces, K. Héberger, *Trends Anal. Chem*., 2007, **35**, 74–86.

26. R. Leardi in *Modern Techniques for Food Authentication*, ed., D-W. Sun, Academic Press / Elsevier, Burlington, 2008, ch. 16, pp. 585–616.

27. P. Oliveri, G. Downey, *J. Chromatogr. A*., 2011, **1158**, 196–214.

Analytical Methods Accepted Manuscript

28. P. de la Mata Espinosa, J.M. Bosque Sendra, R. Bro, L. Cuadros Rodríguez, *Anal. Bioanal. Chem*. 2011, **399**, 2083–2092.
29. P. de la Mata Espinosa, J.M. Bosque Sendra, R. Bro, L. Cuadros Rodríguez, *Talanta,* 2011, **85**, 177–182.
30. C. Ruiz Samblás, L. Cuadros Rodríguez, A. González Casado, F. de Paula Rodríguez García, de P. la Mata Espinosa, J.M. Bosque Sendra, *Anal. Bioanal. Chem*., 2011, **399**, 2093–2103.
31. C. Ruiz Samblás, F. Marini, L. Cuadros Rodríguez, A. González Casado, *J. Chromatogr. B*., 2012, **910**, 71–77.
32. C. Ruiz Samblás, J.M. Cadenas, D.A.Pelta, L.Cuadros Rodríguez, *Anal. Bioanal. Chem.,* 2014, **406**, 2591–2601.
33. T. Galeano Díaz, I. Duran Meras, J. Sánchez Casas, M.F. Alexandre Franco, *Food Control.*,2005, **16**, 339–347.
34. A. Tres, C Ruiz Samblás, G. Van Der Veer, S.M. Van Ruth, *Food Chem.,* 2013, **137**, 142–150.
35. C. Ruiz Samblás, C. Arrebola Pascual, A. Tres, S. Van Ruth, L. Cuadros Rodríguez, *Talanta*. 2013, **116**, 788–793.
36. W. Kamm, F. Dionisi, C. Hischenhuber, H-G. Schmarr, K-H. Engel, *Eur. J. Lipid Sci,. Tech*. 2002, **104**, 756–761.
37. G. Tomasi, F. Savorani, S.B. Engelsen, *J. Chromatogr. A.,* 2011, **1218**, 7832–7840.
38. K.H. Esbensen, P. Geladi, in *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis. Vol. 2: Data Preprocessing. Linear Soft-Modeling. Unsupervised Data Mining*, eds. S.D. Brown, R. Tauler, B. Walczak, Elsevier, Amsterdam, 2009, ch. 13, pp. 211–226.
39. C.B.Y. Cordella, in *Analytical Chemistry*, ed. I.S. Krull, InTech, Rijeka, 2012, ch. 1, pp. 1–46.
40. R. Bro, A.K.Smilde, *Anal. Methods,.* 2014, **6**, 2812–2831.
41. M. Forina, P. Oliveri, S. Lanteri, M. Casale, *Chemom. Intell. Lab. Syst.,* 2008, **93**, 1320–148.
42. F. Marini, *Curr, Anal. Chem.,* 2010, **6,** 72–79.
43. S. Wold, M. Sjostrom, L. Eriksson, *Chemom. Intell. Lab. Syst.,* 2001, **58**, 109–130.
44. D. Ballabio, V. Consonni, *Anal. Methods.,* 2013, **5**, 3790–3798.
45. M .Sokolova, N. Japkowicz, S. Szpakowicz, in *Evaluation Methods for Machine Learning: Papers from the 2006 AAAI Workshop (WS-06-06)*, American Association for Artificial Intelligence, 2006.
46. N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, New York, 2011, ch. 3-4.
47. M. Bekkar, H. Kheliouane Djemaa, T. Akrouf Alitouche, *J. Inform. Eng. Appl.,* 2013, **3**, 27-38.
48. C. Ferri, J. Hernández-Orallo, R. Modroiu, *Pattern Recogn. Lett*., 2009, **30,** 27–38.
49. M. Felking, in *Quality Measures in Data Mining*, ed. F. Guillet, H.J. Hamilton, Springer-Verlag, Berlin Heidelberg, 2007, ch.12, pp. 290–292.
50. K.L. Gwet, Handbook of Inter-Rater Reliability, Advanced Analytics LLC, Gaithersburg, 3rd ed., 2012, ch. 2, pp. 15–46.

Analytical Methods Accepted Manuscript

Table 1. Geographical origin of the 102 palm oil samples.

Asia		Africa		America	
Country	**Samples**	**Country**	**Samples**	**Country**	**Samples**
India	5	Cameroon	2	Brazil	16
Indonesia	24	Ghana	20		
Malaysia	19	Guinea	3		
Papua N. Guinea	7	West Africa	5		
Salomon	1				

Table 2. Continent distribution of the two set of samples.

Set	**Continent**	**Nº samples**
Training set 72 samples (70.6 %)	Africa	20
	America	10
	Asia	42
Validation set 30 samples (29.4 %)	Africa	10
	America	6
	Asia	14

Analytical Methods Accepted Manuscript

Table 3. Multiclass contingency table for ternary classification.

	Actual CLASS 1	Actual CLASS 2	Actual CLASS 3	**TOTAL**
Assigned CLASS 1	$\mathbf{a_1}$	$\mathbf{e_{2,1}}$	$\mathbf{e_{3,1}}$	$\mathbf{AC_1}$
Assigned CLASS 2	$\mathbf{e_{1,2}}$	$\mathbf{a_2}$	$\mathbf{e_{3,2}}$	$\mathbf{AC_2}$
Assigned CLASS 3	$\mathbf{e_{1,3}}$	$\mathbf{e_{2,3}}$	$\mathbf{a_3}$	$\mathbf{AC_3}$
Assigned NO-123	$\mathbf{e_{1,NO}}$	$\mathbf{e_{2,NO}}$	$\mathbf{e_{3,NO}}$	$\mathbf{AC_4}$
TOTAL	$\mathbf{TC_1}$	$\mathbf{TC_2}$	$\mathbf{TC_3}$	**T**

$\mathbf{a_i}$ = the number of assignation agreement of the class "i"; $\mathbf{e_i}$ = the number of assignation error; $\mathbf{TC_i}$ = the total of actual samples from the class "i"; $\mathbf{AC_i}$ = the total of assigned samples to the class "i"; T = the total number of samples. "**NO-123**" represents a fictitious class where the samples that do not assign to any classes, are allocated.

Table 4. Standard contingency table for binary classification.

	LABEL (L) Actual POSITIVE	no-LABEL (nL) Actual NEGATIVE	**TOTAL**
Assigned POSITIVE	**TP**	**FP**	**AP = TP+FP**
Assigned NEGATIVE	**FN**	**TN**	**AN = FN+TN**
TOTAL	**(TL) = TP+FN**	**(TnL) = FP+TN**	**T**

TP = true positive, the number of positive samples that are correctly identified as positive; FN = false negative, the number of positive samples that are misclassified as negative samples; FP = false positive, the number of negative samples that are incorrectly identified as positive samples; TN = true negative, the number of negative samples that are correctly identified as negative samples; AP and AN = the total of assigned positive and negative samples, respectively; TL and TnL = the number of labelled (actual) samples as positive and negative, respectively; T = the total number of samples.

Calculation example for obtaining the contingency table of the binary classification (class 1 / class n1) from the contingency table shown in Table 3:

TP = a_1; **FP** = $e_{2,1} + e_{3,1}$; **FN** = $e_{1,2} + e_{1,3} + e_{1,NO}$; **TN** = $a_2 + a_3 + e_{2,3} + e_{2,NO} + e_{3,2} + e_{3,NO}$

Analytical Methods Accepted Manuscript

Table 5. Contingency tables obtained from both HPLC-UV and HPLC-CAD data for the two classification methods (SIMCA and PLS-DA) when the geographical origin is classified by considering three continents: Africa, America and Asia (a three-class classification). In this table, the sample numbers assigned to each class for the validation set (rows) are shown. Between parentheses, the corresponding rates, in %, in relation to the sample total number of each class (columns).

Fingerprint	Assigned class	SIMCA			PLS-DA		
		Actual class			Actual class		
		Africa	America	Asia	Africa	America	Asia
HPLC-UV	Africa	8 (80.0)	2 (33.3)	3 (21.4)	6 (60.0)	0 (0)	2 (14.3)
	America	0 (0)	2 (33.3)	0 (0)	2 (20.0)	2 (33.3)	0 (0)
	Asia	0 (0)	0 (0)	11 (78.6)	2 (20.0)	3 (50.0)	11 (78.6)
	No assigned	2 (20.0)	2 (33.3)	0 (0)	0 (0)	1 (16.7)	1 (7.1)
HPLC-CAD	Africa	10 (100)	2 (33.3)	3 (21.4)	7 (70.0)	1 (16.7)	2 (14.3)
	America	0 (0)	0 (0)	0 (0)	1 (10.0)	3 (50.0)	1 (7.1)
	Asia	0 (0)	2 (33.3)	8 (57.1)	2 (20.0)	1 (16.7)	11 (78.6)
	No assigned	0 (0)	2 (33.3)	3 (21.4)	0 (0)	1 (16.7)	0 (0)

Analytical Methods Accepted Manuscript

Table 6. Values of quality performance features from two classification methods (SIMCA and PLS-DA) by two fingerprint data (HPLC-UV and HPLC-CAD) for the geographical origin between three continents: Africa, America and Asia (a three-class classification). (i): Pooled performance features of the four 3-class classifiers. (ii): overall values of quality performance features.

	HPLC-UV		HPLC-CAD	
	SIMCA	**PLS-DA**	**SIMCA**	**PLS-DA**
(i) Pooled performance features				
Sensibility (or Recall)	0.70	0.63	0.67	0.70
Specificity	0.92	0.80	0.92	0.85
Positive predictive value (or Precision)	0.87	0.67	0.89	0.72
Negative predictive value	0.86	0.81	0.84	0.84
Youden index	0.62	0.44	0.58	0.55
Positive likelihood rate	–	3.97	–	4.71
Negative likelihood rate	0.32	0.44	0.33	0.35
F-measure	0.76	0.69	0.71	0.76
Discriminant power	–	0.52	–	0.63
Efficiency (or Accuracy)	0.85	0.77	0.82	0.81
AUC (or Correctly classified rate)	0.81	0.72	0.79	0.77
Matthews correlation coefficient	0.66	0.46	0.64	0.55
Agreement probability	0.85	0.77	0.82	0.81
Chance agreement probability	0.56	0.57	0.56	0.56
Kappa coefficient	0.63	0.45	0.59	0.55
(ii) Overall performance features				
Overall agreement probability	0.70	0.63	0.60	0.70
Overall chance agreement probability	0.33	0.36	0.32	0.36
Overall KAPPA coefficient	0.55	0.42	0.41	0.53

Analytical Methods Accepted Manuscript

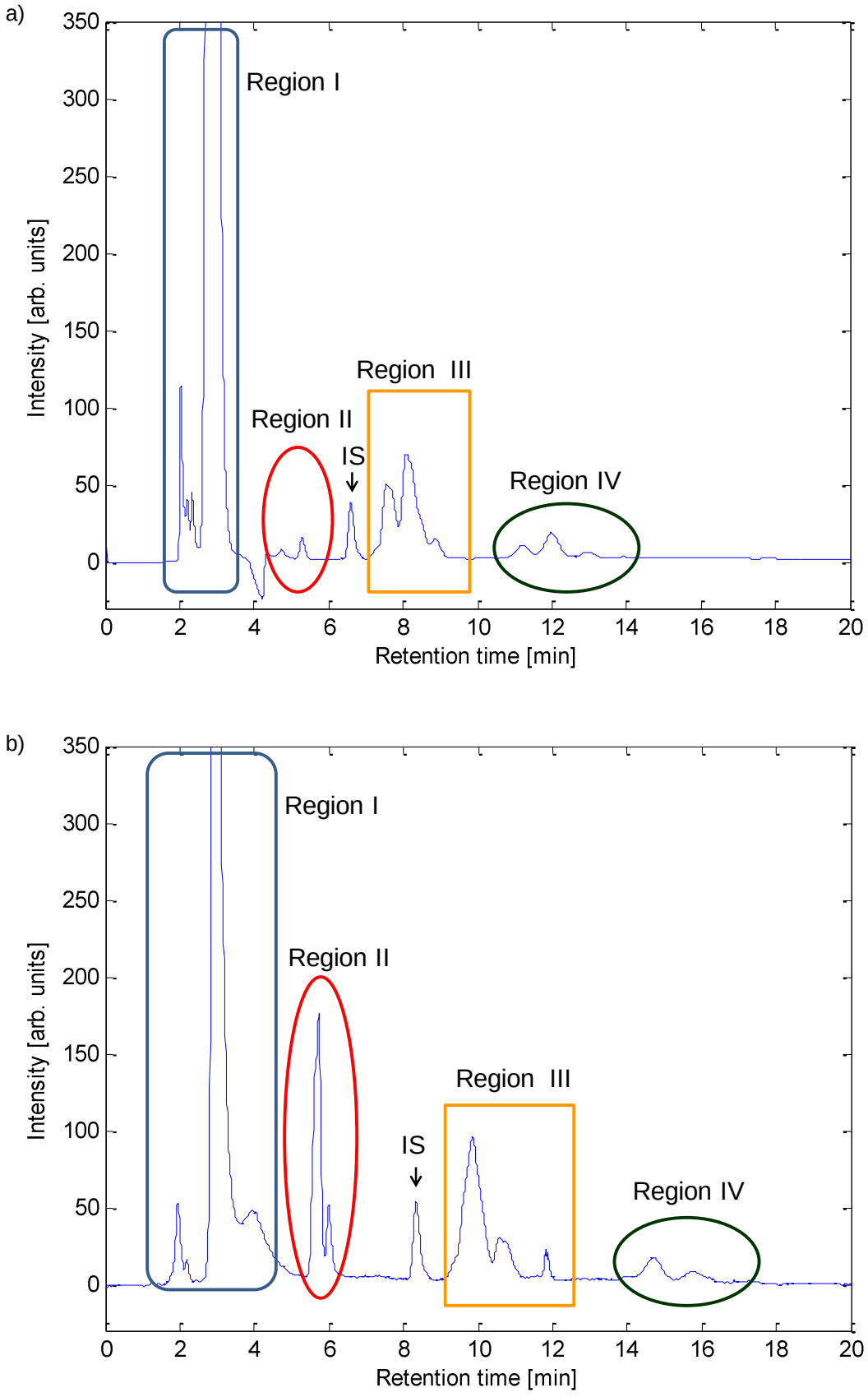
FIGURE CAPTIONS

Figure 1. Chromatograms of the same sample of palm oil from Africa showing the three characteristic regions, obtained from the data of the sterolic fraction by: (a) HPLC-UV, and (b) HPLC-CAD. IS denotes the internal standard. See text for further descriptions.

Figure 2. PC1/PC2 scores and PC1 loadings plots obtained from the data of the sterolic chromatographic data from the palm oil samples of three different continents: America (green squares); Africa (red rhombus); and Asia (blue triangles), by: (a) and (b) HPLC-UV; (c) and (d) HPLC-CAD.

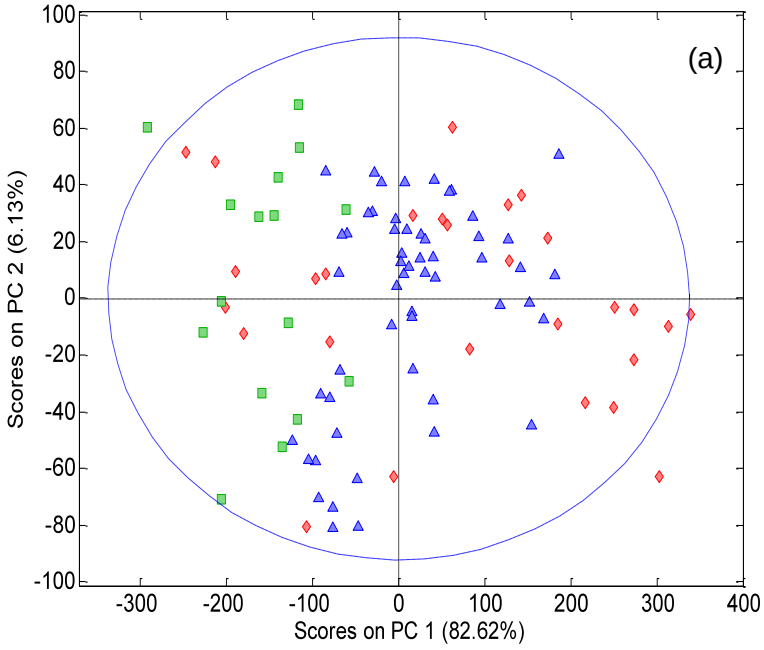
Analytical Methods Accepted Manuscript

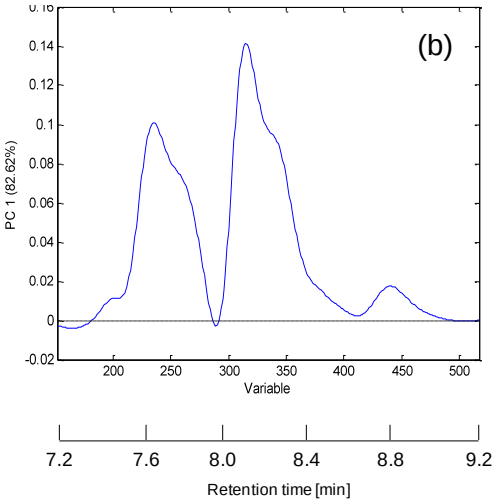
Figure 1

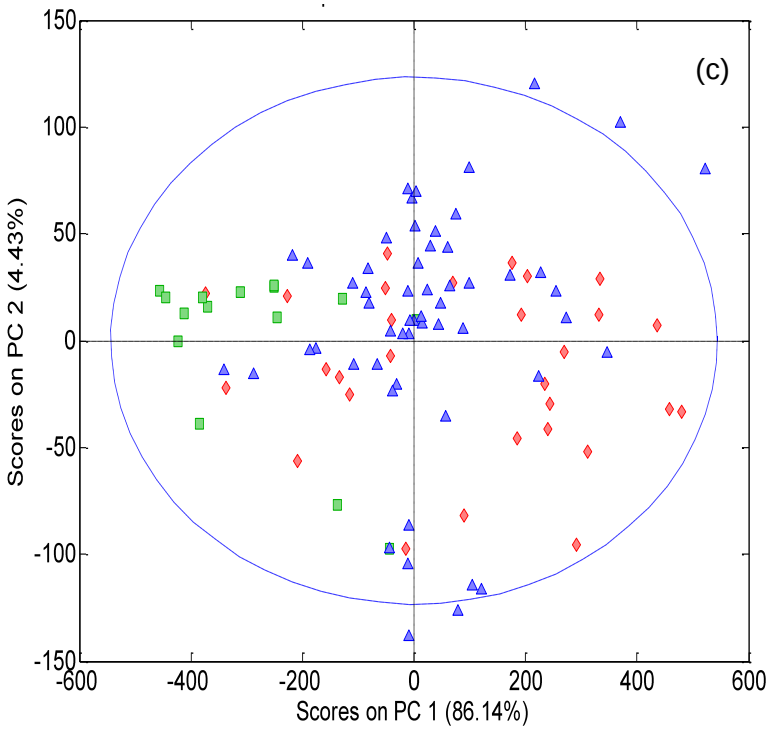


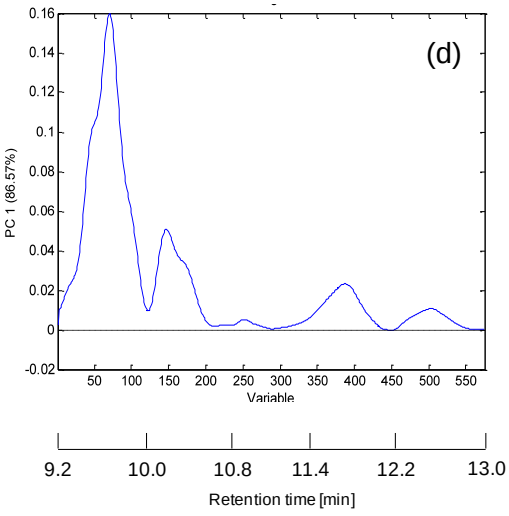
Analytical Methods Accepted Manuscript

Figure 2









Analytical Methods Accepted Manuscript