Food Chemistry 292 (2019) 47–57

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Authentication of cocoa bean shells by near- and mid-infrared spectroscopy and inductively coupled plasma-optical emission spectroscopy

Luisa Mandrile[a,*], Letricia Barbosa-Pereira[b], Klavs Martin Sorensen[c], Andrea Mario Giovannozzi[a], Giuseppe Zeppa[b], Søren Balling Engelsen[c], Andrea Mario Rossi[a]

[a] Quality of Life Division, Food Metrology Program, Istituto Nazionale di Ricerca Metrologica, Strada delle Cacce, 91 10135 Torino, Italy
[b] Department of Agricultural, Forestry, and Food Sciences (DISAFA), University of Turin, Largo Paolo Braccini 2, 10095 Grugliasco (TO), Italy
[c] Department of Food Science, University of Copenhagen, Rolighedsvej 26, DK-1958 Frederiksberg, Denmark

ARTICLE INFO

Keywords:
Cocoa bean shell
Food traceability
Data fusion
Near-infrared spectroscopy
Mid-infrared spectroscopy
Inductively coupled plasma

ABSTRACT

The aim of this study was to evaluate the efficacy of a multi-analytical approach for origin authentication of cocoa bean shells (CBS). The overall chemical profiles of CBS from different origins were characterized using diffuse reflectance near-infrared spectroscopy (NIRS) and attenuated total reflectance mid-infrared spectroscopy (ATR-FT-IR) for molecular composition identification, as well as inductively coupled plasma-optical emission spectroscopy (ICP-OES) for elemental composition identification. Exploratory chemometric techniques based on Principal Component Analysis (PCA) were applied to each single technique for the identification of systematic patterns related to the geographical origin of samples. A combination of the three techniques proved to be the most promising approach to establish classification models. Partial Least Squares-Discriminant Analysis modelling of fused PCA scores of three independent models was used and compared with single technique models. Improved classification of CBS samples was obtained using the fused model. Satisfactory classification rates were obtained for Central African samples with an accuracy of 0.84.

1. Introduction

Since the 19th century, cocoa has undergone continuous growth in consumption in a variety of forms, thus having garnered outstanding economic interest from chocolate industries for constant innovation and modernization. As is the case with many other agro-food activities, the cocoa industry produces large amounts of by-products (https://www.icco.org/). Cocoa bean shells (CBS) represent one of the main by-products, almost 12% of the original harvest weight after husking and grinding of dried cocoa seeds, representing a non-negligible disposal problem. Thus, legislation and environmental issues are forcing industries to define process optimization and recovery/recycling strategies. Recently, bioconversion of by-products has attracted the interest of scientific researchers and new strategic visions or dedicated policies are being developed to manage food industry wastes in the most efficient way – abandoning the "take, make and dispose" behaviour and acting out a circular economy paradigm (Sørensen, Aru, Khakimov, Aunskjær, & Engelsen, 2018). The increasing interest surrounding by-products certainly has an environmental basis, but an important role is played by the tendency to reduce the use of synthetic additives in food and replace them with natural substances with high quality/cost ratios (Carocho, Morales, & Ferreira, 2015). Moreover, the demand for new functional foods rich in bio-nutrients, such as polyphenols, fibre, and n-3 fatty acids among others, has driven interest in rich food wastes, such as seed husks (Andrade et al., 2012; Jansman, Verstegen, Huisman, & Van den Berg, 1995), where recycling of vegetal by-products represents one of the valorisation strategies. Therefore, the development of CBS valorisation strategies is aimed at reducing the environmental impacts of cocoa production and promoting conversion of a by-product into a value-added product with applications in the food and healthcare sectors. The definition of the chemical composition of CBS from different countries is meant to evaluate systematic differences due to their origin. Chemical analysis of CBS was carried out, as reported in several research papers, because of its interesting features related to flavour, phenolic compounds and nutritional values (Barbosa-Pereira, Guglielmetti, & Zeppa, 2018; Manzano et al., 2017; Redgwell et al., 2003; Serra Bonvehí & Escolá Jordà, 1998; Martín-Cabrejas, Valiente, Esteban, Mollá, & Waldron, 1994). However, a complete characterization based on different methodologies to highlight similarities and differences in the composition of samples from different countries has not been accomplished yet. In this work, CBS samples from different countries were analysed using three different analytical methods, i.e.

* Corresponding author.
E-mail address: l.mandrile@inrim.it (L. Mandrile).

https://doi.org/10.1016/j.foodchem.2019.04.008
Received 9 October 2018; Received in revised form 1 April 2019; Accepted 1 April 2019
Available online 12 April 2019
0308-8146/ © 2019 Elsevier Ltd. All rights reserved.

near-infrared spectroscopy (NIRS), mid-infrared spectroscopy by attenuated total reflectance (ATR-FT-IR) and inductively coupled plasma-optical emission spectroscopy (ICP-OES), to obtain broad chemical information at both the molecular and elementary levels. The aim of this study was to evaluate the validity of simple and rapid analytical techniques, supported by a chemometric approach, for the identification of differences due to different geographical origins of CBS samples, with the perspective of a future application in traceability and origin authentication of CBS as a food additive.

Nowadays, the exchange of foodstuffs is realised in a complex and interconnected global net, and food products are often involved in fraudulence, false information, contamination risks and counterfeiting. For this reason, it is extremely important to protect and valorise authentic products, including regional specialties. Innovative and reliable strategies to individuate specific markers of origin, as well as characteristic compositional patterns that can be associated with a precise origin, are thus urgently needed (Mandrile, Zeppa, Giovannozzi, & Rossi, 2016). Geographical origin indicators could provide an analytical response to the geographic traceability problem and support the documental certification, which is used today to guarantee food and food-additive provenance. Different techniques such as nuclear magnetic resonance (NMR) and isotope ratio mass spectrometry can play relevant roles in identifying origin indications (Lee et al., 2011). Rapid and non-destructive techniques, such as NIRS, are particularly interesting because of the possibility to obtain an efficient and non-biased overview of the entire sample chemistry (Sørensen, Khakimov, & Engelsen, 2016). The chemical specificity and ease of sampling of NIRS make it an attractive tool for rapid and comprehensive food analysis. The complex patterns of signals revealed by IR analysis, both in the near- and mid-infrared spectral regions are correlated to the contents of the different chemical constituents, such as proteins, fatty acids, carbohydrates, alimentary fibre, and phenolic compounds. Statistics and multivariate data analysis offer powerful tools to identify robust correlations between the chemical constituents and their geographical origins, providing validated models for the recognition of unknown samples with a certain degree of probability (Peres, Barlet, Loiseau, & Montet, 2007; Kelly, Heaton, & Hoogewerff, 2005). In this work, different chemometric approaches were used to calculate both explorative and predictive models. Principal component analysis (PCA) was first applied as a well-established strategy in food science for data exploration and visualisation in order to extract useful information from numerous experimental results (Munck, Nørgaard, Engelsen, Bro, & Andersson, 1998). Moreover, data fusion for multi-block analysis was employed to improve models, gaining information from several different analytical techniques (Biancolillo, Bucci, Magrì, Magrì, & Marini, 2014; Skov, Honoré, Hansen, Næs, & Engelsen, 2014; Silvestri et al., 2014; Zakaria et al., 2010).

2. Materials and methods

2.1. Samples

Fermented and dried cocoa (*Theobroma cacao L.*) samples were selected and collected within the COVALFOOD project funded by European Union's Seventh Framework Programme, involving five Italian chocolate industries. A complete list of 78 samples with associated information concerning supplier, provenance and variety is reported in Table 1S.1 (Supplementary Information). For an easier exploration of the sample pool, charts of geographical and varietal distribution are shown in Fig. 1S.1. All samples were imported as untreated raw materials, and the geographical origin was guaranteed by the supplying industry. All samples were roasted and decorticated in a laboratory in a ventilated oven for 20 min at 130 °C. After roasting, the fragile shells of the beans was separated by mechanical rubbing and removed by vacuum suction. The collected CBS were ground using a Retsch ZM 200 ultracentrifugal mill (RetschGmbh, Haan, Germany) and stored as fine, dry powder (250 μm) in a desiccator in closed containers.

2.2. Near infrared spectroscopy

NIRS spectra of CBS were collected in the spectral range 10,000–4000 cm^{-1} (1000–2500 nm) using an Antaris II FT-NIR spectrometer (Thermo Fisher, Waltham, USA) in diffuse reflectance mode. The integrating sphere accessorise was used to collect diffuse reflected light. CBS samples were analysed without sample pre-treatment. Briefly, 0.1 g of CBS powder was weighed and transferred to a quartz glass vial, which was positioned on the integrating sphere. Each spectrum was collected at a spectral resolution of 8 cm^{-1} and with 32 scans in total. A clean, flat gold surface was used for background collection. Three measurement replicates were collected per sample. All samples were measured in randomized order.

2.3. Mid infrared spectroscopy

ATR-FT-IR spectra in the mid-infrared region (500–4000 cm^{-1}) were collected using a Nicolet FT-IR spectrometer (Thermo Fisher, Waltham, USA) equipped with a Germanium crystal (n = 5.7) for a maximum sample penetration of 1 μm. Each spectrum was collected at a spectral resolution of 4 cm^{-1} and with 64 scans in total. The sample powder was pressed with a conical tip onto the crystal and a pressure of 15 bar was applied. The tip and the crystal were thoroughly washed with ethanol before each measurement to avoid cross contamination. Three spectra were collected for each sample, with resampling for each replicate.

2.4. ICP-OES elemental composition

ICP-OES measurements were performed on an Agilent 5100 Synchronous Vertical Dual View instrument (Agilent, Santa Clara, California, USA), equipped with an EasyFit torch (Agilent P/N G8010-60228). Samples were measured in radial mode, using a plasma flow of 12 ml/min and nebulizer flow of 0.7 ml/min, with a rinse time of 15 s and stabilization time of 15 s, in three replicates. Viewing height was set to 8 mm. Prior to measurement, the samples were digested in an Antor Paar Multiwave (Graz, Austria) GO microwave oven: 5 mg of CBS samples was placed in the oven Teflon tubes, 1 ml of HNO_3 5% v/v was added and the tubes were sealed according to the manufacturer specifications. The temperature ramp was set to reach 180 °C in 5 min, then held constant, and the total treatment lasted 40 min. After digestion, the samples were further diluted with 4 ml HNO_3 5% v/v to obtain a clear solution before being deposited into tubes and placed in the autosampler for ICP analysis. All glassware, tubes and equipment were cleansed in HNO_3 5% v/v as needed.

2.5. Data treatment

Chemometric data analysis was carried out using PLS Toolbox from Eigenvector Research, Inc. (Manson, WA, USA) for Matlab R2015a (Mathworks, Natick, USA). PCA is a linear factorization method uniquely suited for data exploration. As an explorative tool, PCA provides visualization of multivariate data as score points in a model space (Wold, Esbensen, & Geladi, 1987). PCA score plots are useful to explore data and to identify correlations between measured variables and the information of interest, such as geographical provenience of CBS, in this case. Next, PLS-DA (Barker & Rayens, 2003) models were calculated to compare the classification performances of the three different techniques, both separately and contemporarily by joining the three datasets. Ten classes were considered: Central Africa, Ecuador, Gulf of Mexico, Indonesia, Mexico, Peru, São Tomé, Colombia, Venezuela and Brazil. All of the calculated PLS-DA models were validated using leave-one group-out cross validation. The subsets of samples used as tests sets in cross validations correspond to the country of origin. Data pre-

processing details for each technique are reported. Leave-one group-out cross validation was performed using as group vector the country of origin. Sensitivity (True Positive/(True Positive + False Negative)), Specificity (True Negative/(True Negative + False Positive)), Accuracy (correctly classified samples/total samples) and Precision (True Positive/(True Positive + False Positive) were considered as model evaluation parameters for each class in cross validations to compare classification performances of the different techniques.

2.5.1. NIRS data treatment

Pre-processing of NIRS data was applied to extract useful information from the dataset. Absolute absorbance variations and unwanted light scattering were removed using pre-processing of the NIRS data (Martens, Nielsen, & Engelsen, 2003). The most effective pre-processing was chosen based on the minimum differences between replicates on the PCA scores plots relative to the distance between samples. 2nd derivative (Savitzky Golay, filter width 15 and polynomial order 2) coupled with standard normal variate (SNV) normalization was useful to remove random shift of the baseline offset (Barnes, Dhanoa, & Lister, 1989). In addition, the derivatives of spectra were calculated to increase sensitivity to changing data trends. Processed spectra are shown in Fig. 2S.1. Unwanted variability was successfully removed as demonstrated by the narrow grouping of the replicates obtained after processing shown in Fig. 2S.2 in Supplementary Information. PCA was applied to visualize the data and to investigate systematic differences among samples, and variables with particular relevance were identified. A 4 LVs PLS-DA classification model was also calculated to discriminate classes of samples from different geographical areas. The same spectral pre-processing was used.

2.5.2. ATR-FT-IR data treatment

Pre-processing of data was performed to suppress variability associated with unwanted noise. The selection criterion for data pre-processing was the maximized closeness of the PCA scores of technical replicates on PC1, as shown in Fig. 3S.1 in Supplementary Information. Baseline correction (using an asymmetric weighted least squares algorithm, with basis filter of order 2) (Peng et al., 2010) followed by second derivative (Savitzky Golay, filter width 15 and polynomial order 2) and mean centring was selected as optimal pre-processing. PCA models for data visualization and exploration were calculated; a PLS-DA classification model using 4 LVs of the same pre-processed data was also calculated to compare ATR-FT-IR classification capabilities with the other techniques.

2.5.3. ICP-OES data treatment

ICP emission spectra were evaluated for quantification using a calibration curve per element. The calibration curves were estimated using two series of standards prepared by dilution of a certified standard mix (ICP Multi-element standard solution IV, Sigma Aldrich, (Shnelldorf, Germany) containing known concentration of 21 elements (Al, B, Ba, Bi, Ca, Cd, Co, Cr, Cu, Fe, K, Li, Mg, Mn, Mo, Na, Ni, Pb, Sr, Tl and Zn). Standard concentrations were 0, 0.2, 0.4, 0.6, 0.8, 1, 2, 4, 6, 8, 10, 20, 30, 40, 60, 80 and 100 mg/100 g of the certified standard concentration, which was 5 mg/l for all elements, except for potassium, which was 50 mg/l in the standard solution. Three emission wavelengths were monitored for each element, then the intensity revealed for only one λ was selected per element based on the best correlation coefficient of the corresponding calibration curve and trying to avoid interference between different elements: λ_{Al} = 237.3 nm; λ_{B} = 249.7 nm; λ_{Ba} = 455.4 nm; λ_{Bi} = 190.2 nm; λ_{Ca} = 396.8 nm; λ_{Cd} = 228.8 nm; λ_{Co} = 230.8 nm; λ_{Cr} = 206.2 nm; λ_{Cu} = 324.8 nm; λ_{Fe} = 234.4 nm; λ_{K} = 766.5 nm; λ_{Li} = 670.8 nm; λ_{Mg} = 285.2 nm; λ_{Mn} = 259.4 nm; $\lambda_{M\ o}$ = 203.8 nm; λ_{Na} = 589.0 nm; λ_{Ni} = 221.6 nm; λ_{Pb} = 217.0 nm; λ_{Sr} = 421.6 nm; λ_{Tl} = 351.9 nm; λ_{Zn} = 202.5 nm.

The table of results was then imported into Matlab and processed with the PLS Toolbox for PCA model calculation and PLS-DA classification. Autoscaling was performed on the data. Three LVs were considered for PLS-DA classification models. Cross validation was used to evaluate the classification capabilities of the models, leaving one country out at each validation step, as described for the other techniques.

2.5.4. Data fusion

The multi-block tool of the PLS toolbox by Eigenvector was used to fuse the PCA scores from the three single PCA models resulting from the different analytical techniques. A joined model exploiting mid-level data fusion was obtained (Borràs et al., 2015). To make the interpretation clearer, the measurement replicates were averaged, and one matrix line per sample was maintained for the three different original datasets (NIRS, MIR-ATR and ICP). Each block was first decomposed by PCA and the resulting scores were fused into a new dataset. The samples' scores for the most relevant PCs were considered to calculate a new fused model. Seven PCs were considered for MIRS and ICP, and six PCs were considered for NIRS. Thus, twenty initial variables were used to build the new joined PCA model. Default autoscale was applied before data joining. The PLS-DA method was then performed with autoscaled data to obtain a classification model (Ballabio & Consonni, 2013). The class vector was represented by the area of origin. It was composed of 10 classes i.e. Central Africa, Colombia, Ecuador, Gulf of Mexico, Indonesia, Mexico, Peru, São Tomé, Venezuela and Brazil. Unfortunately, the number of samples per class was not balanced due to sample availability. Five latent variables were considered for the PLS-DA model, based on the minimum average classification error in cross validation, using the leave-one-country-out cross validation strategy.

3. Results and discussion

3.1. NIRS characterization of CBS samples

The NIRS profiles show the typical broad bands of overtones and combination bands of vibrational modes associated with the main constituents of vegetal origin materials. Assignments of most bands of the NIR spectrum are reported in Table 2S.1 in the Supplementary Information (Jacobsen, Søndergaard, Møller, Desler, & Munck, 2005). The mean NIR spectra for all CBS samples are shown in Fig. 1a, together with the standard deviation profiles. Similar spectral shapes were obtained for all samples: the same bands were present in all spectra with slight differences in mutual intensities.

Vibrational spectroscopy represents a rapid strategy to gather chemical information of a complex matrix, reducing cost, time and environmental impact of analysis. NIR spectra can be effectively correlated to the main alimentary components, as widely reported in the literature (De Oliveira, Roque, de Maia, Stringheta, & Teófilo, 2018; Dong, Sørensen, He, & Engelsen, 2017; Mandrile et al., 2018).

The sensitivity of NIRS to the botanical variety involved was tested first, since it has been previously demonstrated in the literature that differences in the chemical composition of different varieties of *Theobroma Cacao L.* are present (Elwers, Zambrano, Rohsius, & Lieberei, 2009). The outcome of the PCA on the NIR spectra is shown in Fig. 1b. In contrast to expectations, different botanical varieties did not cause evident systematic clustering of NIR spectra. The scores of NIR spectra for *Forastero* and *Trinitario* samples overlapped in the score plots (Fig. 1b), no separation occurred either in the PC2/PC1 plot, or in the later PCs (plots not shown). This can probably be attributed to the complexity of the sample set, which introduces considerable and confusing variability. However, *Arriba* samples, a specific variety cultivated only in Ecuador (green squares on the scores plot in Fig. 1b), was specifically, even though not selectively, characterized by negative scores on PC1 and by positive scores on PC2, attesting to the capability of NIR spectra to identify common chemical features of *Arriba* samples. The loading profiles (Fig. 2S.3a) and the variance captured (Fig. 2S.4) allow to define what spectral regions are involved in each relevant PC.

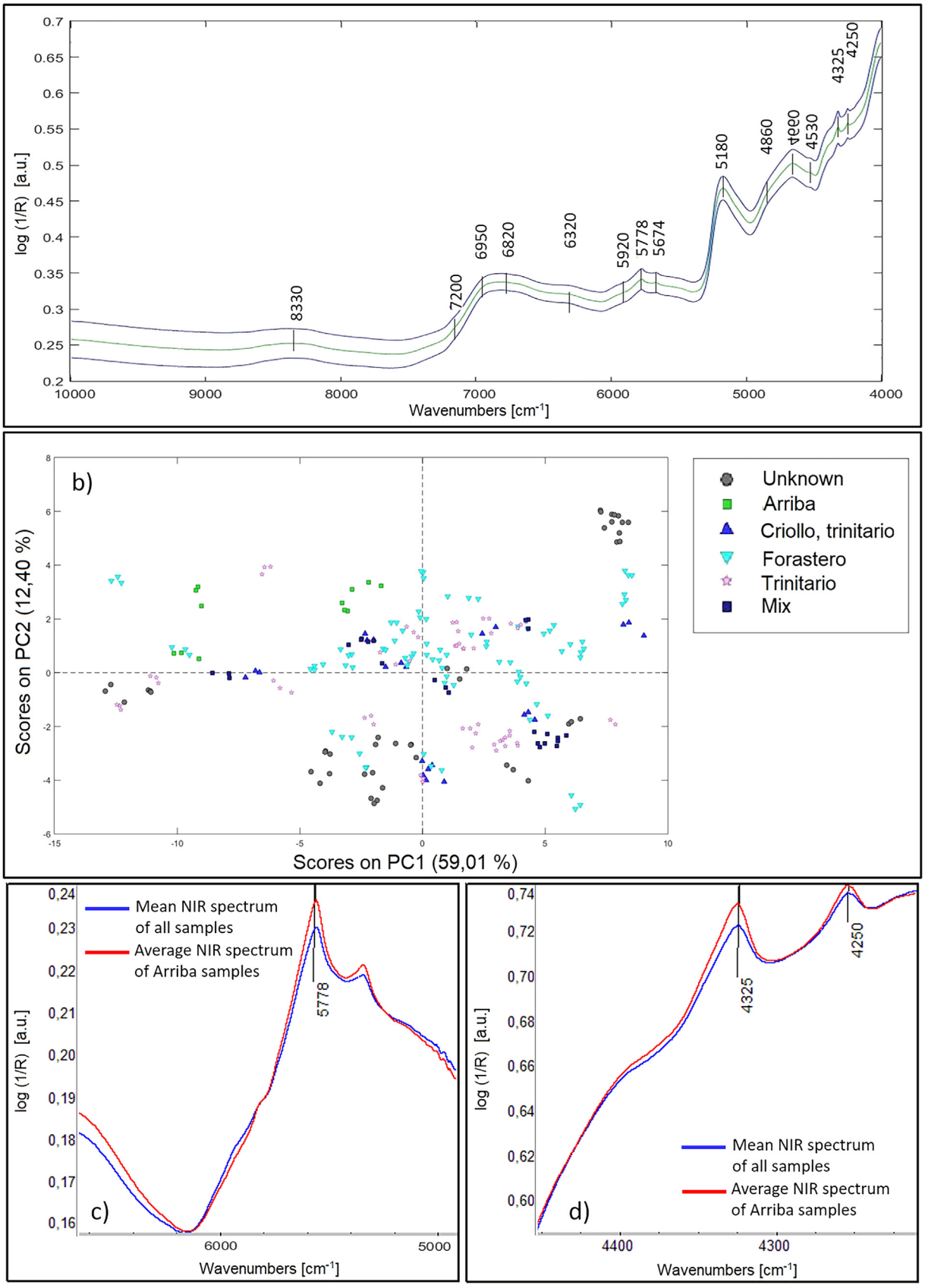


Fig. 1. a) Mean NIR spectrum of all CBS samples (green) and standard deviation limits (blue); b) Scores plot of NIRS data PCA coloured in accordance with variety; c, d) Zoom of average spectrum of *Arriba* samples compared with the mean spectrum calculated considering all other NIR spectra. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

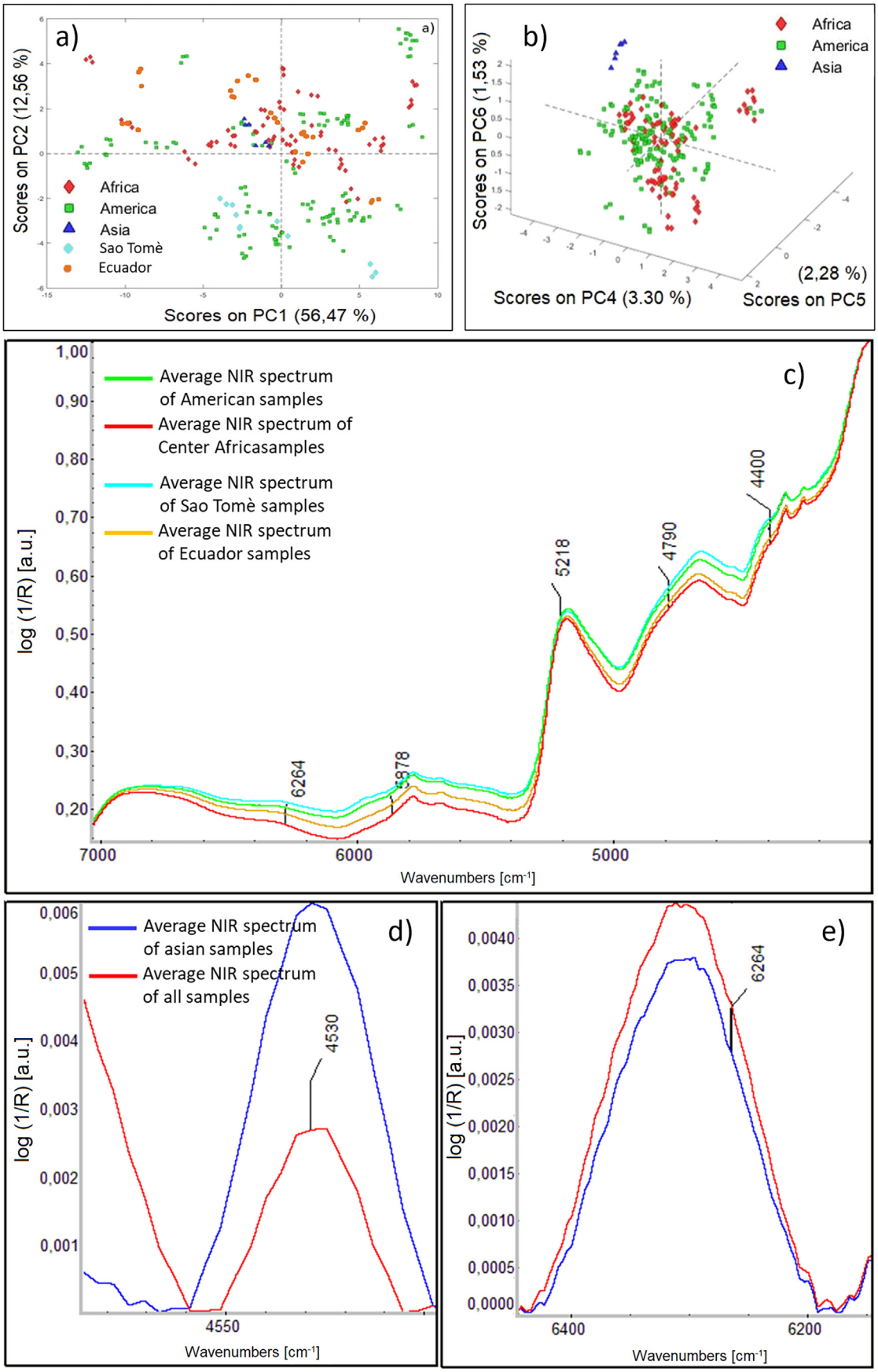
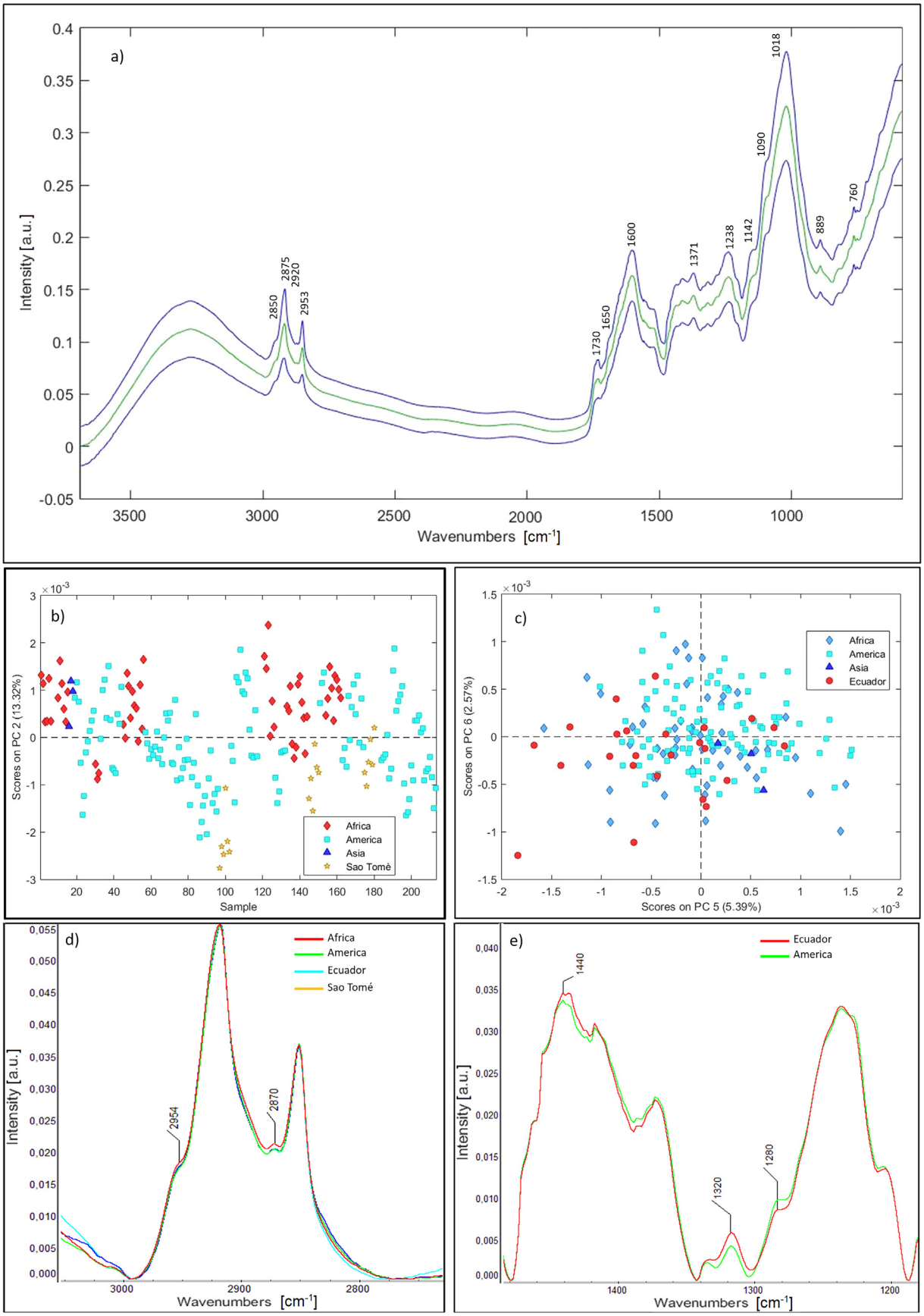


Fig. 2. a) PC2/PC1 scores plot of NIR spectra of CBS sample coloured by geographical origin. b) PC4/PC5/PC6 scores plot of NIR spectra of CBS sample coloured by geographical origin. c) Average NIR spectra of CBS from Africa and America as macro-classes (red and green respectively) and mean spectra of São Tomé and Ecuador groups (light blue and orange respectively); d, e) Zoom on the spectral regions which render Asian samples different from all other CBS samples; (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(caption on next page)

Fig. 3. a) ATR-FT-IR average spectrum of all CBS samples (green) and standard deviation limits (blue); b) PC2 scores plot which highlights common behaviour of African samples; c) PC5/PC6 scores plot that allow to highlight a characteristic trend for Ecuador samples; d) MIR average spectra of CH_x stretching bands of samples from different geographical origins; e) MIR average spectra of Ecuadorian samples compared with American ones in the spectral region where Ecuador samples exhibit distinct characteristics with respect to American samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

PC1, was mainly characterized by fatty acid bands as 5670–5780 cm^{-1} (1st C–H str) and 4325 cm^{-1} (1st C–H str + 1st C–H def CH_2) and 4250 cm^{-1} (1st C–H str + 1st C–H def). In addition, PC1 also captured some regions related to proteins such as 5170–5190 cm^{-1} (2nd C=O of CONH), 5269 cm^{-1} (2nd C=O of COOH), 6320 cm^{-1} (1st N–H str of CONH) and 6535 cm^{-1} (1st N–H str of RNH_2) and 6950 cm^{-1}. PC2, instead, exhibited three maxima at 4400 cm^{-1} (1st O–H str + 1st C–C str, associated with starch), 4763 cm^{-1} (2nd O–H def + 2nd C–O str of starch) and 5000 cm^{-1} (2nd O–H def + 1st C–O def of starch). This indicates that PC2 mostly represents the starch content of the samples. PCA highlighted a major proportion of fatty acids and vegetal proteins in the examined *Arriba* samples, as shown in Fig. 1c, d, whereas lower intensity in the spectral regions was associated with polysaccharide, such as starch (corresponding enlarged spectral regions are not shown for reasons of brevity).

As far as correlations between CBS geographical origins and NIR spectra are concerned, the information provided by the scores plot appears confusing at first sight; however, certain interesting considerations can be identified. Features common to all samples originating from central Africa were noticed in the scores plot (Fig. 2a) when considering PC2. On average, Central African samples (red rhombus in Fig. 2a) showed positive scores on PC2, mainly related to polysaccharide and starch bands (Fig. 2S.3, 2S.4 can be consulted for all attributions of spectral bands to the PCs). Moreover, other common features were noticed in further PCs, such as negative scores on PC3 (Fig. 2S.6b) (where the main contributions are 5218 cm^{-1}, 1st O–H str of phenols, 5878 cm^{-1} 1st C–H str CH_3, 6075 cm^{-1} 1st C–H str of R-CH-CH, 7062 cm^{-1} and 2nd C–H str + 1st C–H def of aromatic compounds) and positive again on PC4 (Fig. 2S.6c) which is related mainly to carbohydrates (4790 cm^{-1} 1st O–H str + 1st O–H def ROH of sucrose and starch, 6264 cm^{-1} and 1st O–H str intramolecular H-bond of starch or glucose). Although the separation of the examined groups was not sufficient for selective discrimination, it was confirmed that the geographical origin information was captured by NIRS. As shown in Fig. 2a, African samples from São Tomé (a small island in the Gulf of Guinea, at latitude 0°) showed features in common with samples coming from America, which on average showed negative scores on PC2. The scores of São Tomé samples (light blue rhombus in Fig. 2a) were mixed with those of Gulf of Mexico samples; this can be attributed to similar environmental and climatic conditions of small islands which influence the chemical composition of cocoa fruits and, therefore, of CBS (see also Fig. 2S.6 a to appreciate the similarities of São Tomé CBS with samples from the islands and coasts of the Gulf of Mexico). Moreover, Ecuadorian samples appeared more similar to African samples than to American ones. Indeed, in Fig. 2a, orange circles corresponding to Ecuadorian samples are mixed with the red rhombus corresponding to samples from Central Africa. In Fig. 2b, the average NIR spectra of the macro classes Africa and America are compared with the spectra of São Tomé and Ecuador, and show peculiar behaviour in contrast with the general trend.

The Asian samples were separated from the others (blue triangles in Fig. 2b) because of high values on PCs 4, 5 and 6. PC4 is characterized by a peak around 4530 cm^{-1}. This spectral region, represented in Fig. 2d was assigned to ROH combination modes, so it can be hypothesized that the sugar content differs for Asian samples with respect to all of the others. The most represented spectral region in PC5 (which is relevant for the clustering of Asian samples) is the side of the peak at 6300 cm^{-1}. This region, represented in Fig. 2e, highlights that the band shapes are relevant, more so than band intensity, in this case. PC6 was also responsible for the following spectral regions: 4466 cm^{-1} (beta-glucan), 5114 cm^{-1} (2nd C=O of esters) and 7147 cm^{-1} typical of R-OH (as already mentioned, Fig. 2S.3, 2S.4 can be consulted for all attributions of spectral bands to the PCs).

The definition of rules to correlate the NIR spectral variability with the geographic area of origin based on the PCA scores plot of NIR spectra is not immediately apparent. However, certain common trends were noticed for samples from the same area, and NIR spectra were demonstrated to contain useful information for geographical provenance analysis.

3.2. ATR-FT-IR spectra

Spectral profiles in the mid-infrared region are shown in Fig. 3a. As for NIRS, ATR-FT-IR spectroscopy was expected to deliver information regarding the chemical composition of CBS samples, including most biochemical species present in the matrix. Although absorption bands in the mid-infrared region were more defined and narrower because primary vibration modes absorb in this spectral region, the visual interpretation of spectra was difficult, especially in the so-called fingerprint region, between 1750 cm^{-1} and 500 cm^{-1}. Main band interpretation is reported in Table 3S.1 in Supplementary Information. (Socrates, 2001; Rubio-Diaz & Rodriguez-Saona, 2010; Li-Chan, Chalmers, & Griffiths, 2011). The region between 2260 and 2440 cm^{-1}, where the CO_2 band is present, was excluded.

MIRS spectra provided information in agreement with NIRS investigations. Signals were more defined and spectral specificity was increased compared to that for NIRS, and PCA score plot investigations revealed an effective strategy to explore spectral similarities. Similarities and differences between samples are ruled by PC1, 2 and 3. The correspondence between PCs and MIR spectral regions was evaluated by analysing Fig. 3S.4, where the MIR spectrum was superimposed over the histogram of the percentage of variance captured by each PC, to understand which bands drive the score distributions on the scores plot. PC1 was mainly dominated by CH_x vibrations in the 3000–2800 cm^{-1} and 1460–1420 cm^{-1} regions (samples with high intensity signals at 2920 cm^{-1} and 1463 cm^{-1} present lower values of PC1). Moreover, the 1730 cm^{-1} peak (C=O stretching) that showed increased intensity in *Arriba* samples was also represented in PC1; PC2 captured variance in the 1700–1650 cm^{-1} region (high values of PC2 mean lower intensity at 1560 cm^{-1} and 1525 cm^{-1} of amide I-II, and lower intensity of the 1690 cm^{-1} shoulder). Several peaks associated with carbohydrates were also relevant, for example the 763 cm^{-1} peak related to pyranose compounds was modelled by PC5. Variety information revealed a certain grouping of *Arriba* samples that showed high PC2 scores and lower intensity of PC5, in agreement with NIRS results. The scores plot coloured by variety information is shown in Fig. 3S.5.

The different geographical provenances drive differentiation between samples, and some general considerations can be extracted from the scores plot (Fig. 3b, c). PC2 certainly explains interesting characteristics of Central African samples that show positive scores on PC2. Samples from São Tomé showed more similarities with samples from the Gulf of Mexico, Venezuela and Colombia, as also attested by NIRS data presented in the previous paragraph. This confirms that similar climatic and environmental conditions are crucial in determining the chemical composition captured by spectroscopic techniques, as previously reported in the literature for cocoa samples (Marseglia et al., 2016). African samples showed higher intensity at 2954 cm^{-1} and

2870 cm^{-1} in the CH_x stretching vibrations (Fig. 3d). Moreover, PC5 and PC6 were relevant to identify features in common between Ecuadorian samples. 87% of Ecuadorian samples were placed to the left of the left diagonal of the PC6/PC5 plot (Fig. 3c). This is due to the ratio between 1280 cm^{-1} (amide III of β-sheet proteins) and 1320 cm^{-1} or 1440 cm^{-1} that allows to separate samples from Ecuador from other American samples, as shown in Fig. 3e. Moreover, low values in PC5 reflected low intensities at 673 cm^{-1} and 1600 cm^{-1} (ring breathing modes of polysaccharides) as already noticed for *Arriba* samples (enlarged spectral regions not shown for reasons of brevity).

The ATR-FT-IR spectrum represents the sum of numerous bands of several functional groups, which are contemporarily present in more than one biochemical compound. Beyond hypothesized interpretations, it should be stressed that an accurate understanding of which peaks and bands drive the score distributions is necessary to avoid misinterpretation. To unequivocally associate the relevant spectral regions to specific classes of compounds remains complicated when a whole and complex matrix such as food is analysed. However, the possibility to identify spectral features that precisely characterize samples from the same origin is an indication that correlations between geographical origin and vibrational spectra can be modelled.

3.3. ICP-OES elemental characterization of CBS samples

Raw ICP-OES results are shown in Table 4S.1 in Supplementary Information. The most abundant elements were Ca, Mg, and K which each had a concentration at least one order of magnitude higher compared to all other elements. Among the secondary elements, particularly relevant were Al, Fe and Li (Barker & Rayens, 2003). Relevant amounts of lead were revealed in all samples (around 0.3 mg/kg), which is a high value compared with the average content of lead in foods reported in 2007 by the Agency for Toxic Substances and Disease Registry (Abadin et al., 2007). All other elements were revealed to be at concentrations lower than 0.2 mg/kg: particularly low concentrations were determined for Ni and Cr. PCA was used to identify major variance directions that can be related to geographical origin. Five samples were identified as very different from the others. These were SB3 and SB4 from Brazil, ICAM10 from Congo, FER8 from Uganda, and FER13 from Côte d'Yvoire. These samples were excluded as outliers because of their very low K content. Boron, potassium, magnesium and calcium were responsible for the most variance captured by PC1, which was not particularly correlated to the provenance of samples. Aluminium, chromium, iron, sodium and nickel were particularly relevant for PC2, whereas cadmium, cobalt and molybdenum, together with calcium and manganese, were mostly represented in PC3, as shown in Fig. 4d.

Examining the PC2/PC3 loadings and scores plot (Fig. 4a, b), high levels of Fe and Al were characteristic for the African continent and for most Central African samples. Moreover, a general deficiency of Ca, K, Mg, and Ni was revealed. Interestingly, some similarities of São Tomé samples to American ones were captured by PC2. A relatively higher content of Fe, Al, Cu and Ni was revealed for these samples; this trend makes São Tomé samples more like American samples than like African samples. Moreover, São Tomé samples were characterized by high content of Ba with respect to other elements. Conversely, Ecuadorian samples did not exhibit any specific elemental profiles.

3.4. Data fusion to merge chemical information provided by the different analytical techniques

The idea of data fusion is to merge information, provided by different analytical determinations, in one single data set, to enhance the quality of the results. The obtained joined PCA model clearly shows that all three datasets provided useful information for the final model. It was noticed that the three most represented variables in PC1 were one from MIR-ATR, one from ICP and one from NIRS (Fig. 5S.1 in Supplementary Information). The scores plot and the loadings projected on the PC2/PC1 space are shown in Fig. 5. The grouping of samples based on geographical origin was improved by the multi-analytical model. Proximity, and hence common features, were appreciated for samples from the same geographical area.

Classification models were calculated to quantify the grouping performances of the joined model compared to the three single models, based on geographical origins. Even though interesting observations were previously discussed for the three techniques separately, and some correlation between geographical origin and composition was defined, single technique outputs were not accurate and precise for the recognition of the geographical origin of samples in predictive classification models. In Table 1, the classification figures of most merit (sensitivity, specificity, error rate, accuracy, precision) relative to PLS-DA classification models for geographical discrimination were reported. The classification performance for sample classes composed of more than 5 samples are shown. Classification results were higher for the joined model compared to each of the three single models for Central Africa, Ecuador and the Gulf of Mexico classes. This experimental evidence was in agreement with literature reports corroborating mid- or high-level data fusion to increase predictive performance of classification models (Doeswijk, Smilde, Hageman, Westerhuis, & Van Eeuwijk, 2011). Single techniques provided nil accuracy and precision for most classes, except for Central Africa. Moreover, after merging information from the three techniques, the accuracy (i.e. correctly classified sample rate) increased.

NIRS, MIRS and ICP profiles together delivered sufficiently accurate information to capture the common features of African samples, and to distinguish them from the others. Unfortunately, this was not the case for the other classes. Low stability emerged during cross validation for Ecuador, Gulf of Mexico and Venezuela classes. Classification results for classes composed of less than 10 samples were not considered statistically valid.

4. Conclusions

Because of the low price and interesting features of CBS, such as the extraordinary similarity to cocoa powder in terms of colour, taste and texture, and the potential beneficial effects on human health, research is needed to assist the valorisation of this food by-product, and to prevent fraud in the cocoa powder market. The present work demonstrated the existence of correlations between geographical origins and composition of CBS samples, even though low specificity for a single country or restricted areas emerged. Some information regarding what samples from the same macro-area have in common was described. The selected techniques provided significant criteria to distinguish sample classes, such as Central African and Ecuadorian samples, with adequate accuracy and precision; however, it is very difficult to precisely determine which chemical species drive this separation using only vibrational spectroscopy for chemical composition analysis. Nevertheless, estimates and trends were determined. The geographical traceability of food based on chemical analysis remains complicated and, invariably, valid rules are rarely identified. The natural variability of most food materials is huge; climatic conditions and process variables represent an intrinsic limit of this field of study. However, the capability to identify leading variables, common trends and general indications using rapid and simple techniques is an encouraging result in this domain. More sensitive and accurate techniques should be employed for an exhaustive investigation. Easy-to-use instrumental analysis still needs the support of more robust analytical strategies for comparison and calibration.

Declaration of interests

None declared.

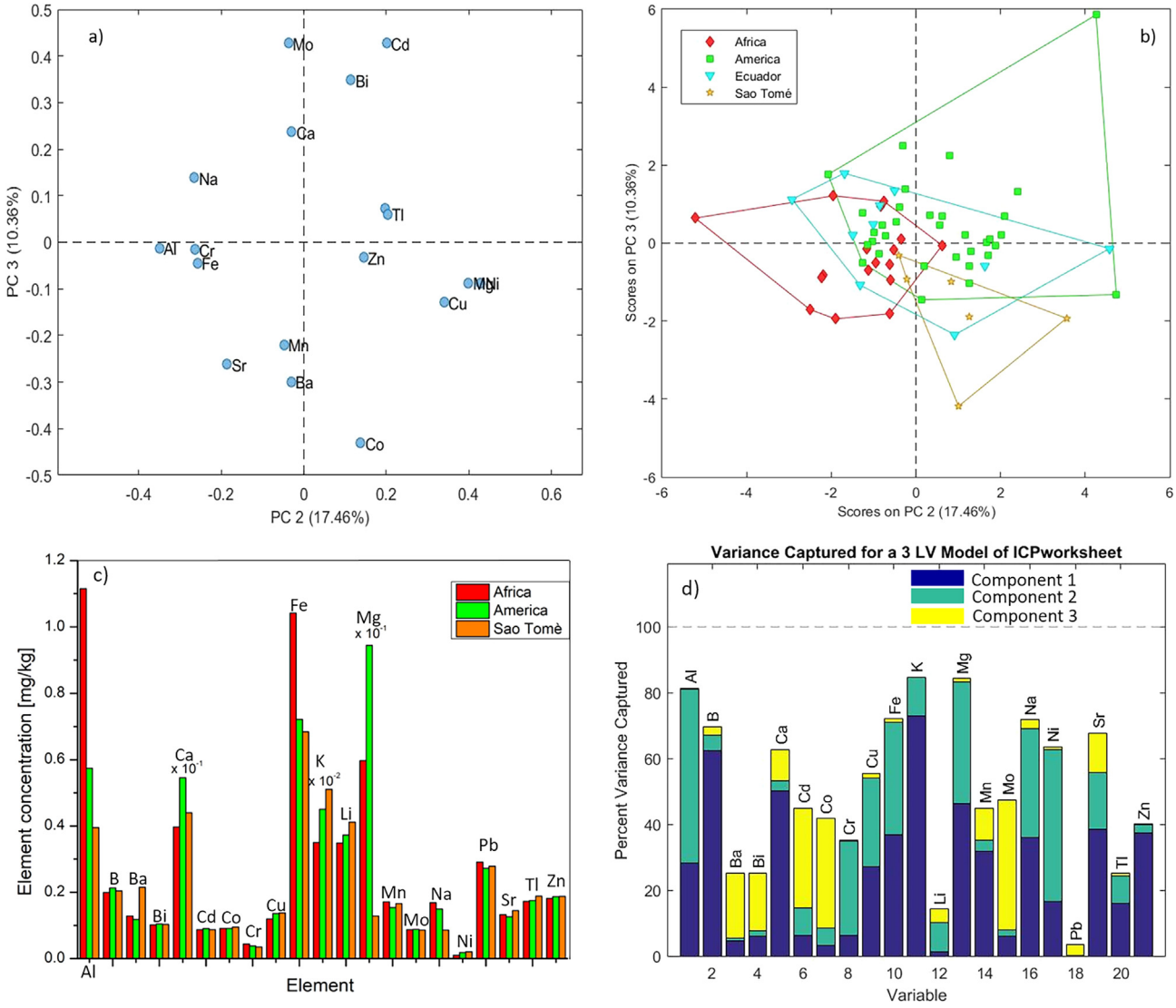


Fig. 4. PCA model of ICP-OES data outputs, 2D a) loading and b) scores plots; c) Histogram of mean data for the macro-classes (Africa and America) investigated, and São Tomé samples that show distinctive features with respect to others; d) Variance captured for each principal component.

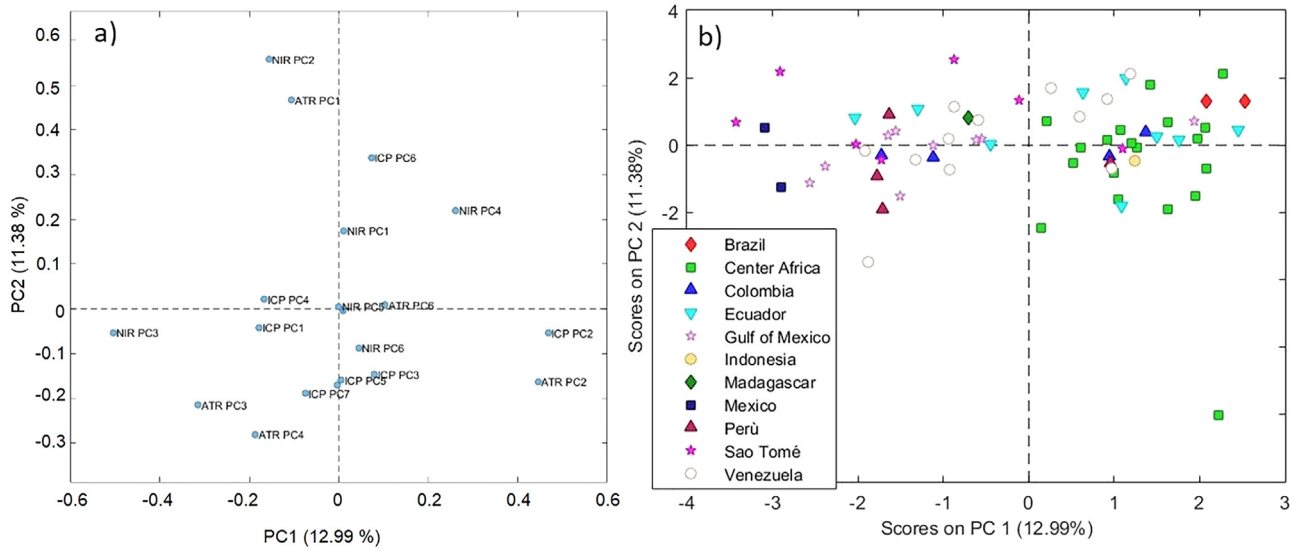


Fig. 5. Joined PCA model of NIRS + ICP + MIRS, a) loadings and b) scores plot on PC1 and PC2.

Table 1
Cross Validation outputs of PLS-Discriminant Analysis classification models for geographical origin discrimination: a) Joined classification model with 5 LVs, classification performances in leave-one origin-out cross validation; b) NIRS PLS-DA model with 4 LVs classification performances in leave-one origin-out cross validation; c) MIRS PLS-DA model with 4 LVs classification performances in leave-one origin-out cross validation; d) ICP-OES PLS-DA model with 3 LVs classification performances in leave-one origin-out cross validation.

Class	Technique	N (number of samples)	Sensitivity	Specificity	Accuracy	Precision
			(true positive ratio)	(true negative ratio)		
Central Africa	a) **Joined**	**22**	**0.68**	**0.92**	**0.84**	**0.79**
	b) NIRS	19	0.68	0.86	0.81	0.00
	c) MIRS	19	0.32	0.70	0.59	0.29
	d) ICP-OES	19	0.50	0.83	0.75	0.50
Gulf of Mexico	a) **Joined**	**9**	**0.33**	**0.82**	**0.76**	**0.21**
	b) NIRS	9	0.00	0.87	0.75	0.00
	c) MIRS	9	0.00	0.82	0.71	0.00
	d) ICP-OES	9	0.00	0.87	0.75	0.00
São Tomé	a) **Joined**	**6**	**0.33**	**0.95**	**0.9**	**0.40**
	b) NIRS	6	0.00	0.91	0.86	0.00
	c) MIRS	6	0.00	0.90	0.83	0.00
	d) ICP-OES	6	0.00	0.92	0.86	0.00
Venezuela	a) **Joined**	10	0.10	0.87	0.76	0.11
	b) NIRS	12	0.00	0.89	0.74	0.00
	c) MIRS	4	0.00	0.89	0.84	0.00
	d) ICP-OES	12	0.00	0.85	0.69	0.00
Ecuador	a) **Joined**	10	0.00	0.87	0.74	0.00
	b) NIRS	10	0.00	0.85	0.72	0.00
	c) MIRS	10	0.00	0.81	0.70	0.00
	d) ICP-OES	10	0.00	0.87	0.73	0.00
Indonesia	a) **Joined**	1	0.00	0.96	0.94	0.00
	b) NIRS	1	0.00	1.00	0.99	0.00
	c) MIRS	1	0.00	1.00	0.99	0.00
	d) ICP-OES	1	0.00	0.98	0.97	0.00
Mexico	a) **Joined**	2	0.00	0.99	0.96	0.00
	b) NIRS	2	0.00	0.96	0.93	0.00
	c) MIRS	2	0.00	0.94	0.91	0.00
	d) ICP-OES	2	0.00	0.97	0.94	0.00
Peru	a) **Joined**	4	0.00	0.89	0.84	0.00
	b) NIRS	4	0.00	0.92	0.87	0.00
	c) MIRS	4	0.00	0.97	0.91	0.00
	d) ICP-OES	4	0.00	0.87	0.81	0.00
Colombia	a) **Joined**	4	0.00	0.95	0.90	0.00
	b) NIRS	4	0.00	0.92	0.87	0.00
	c) MIRS	12	0.00	0.93	0.77	0.00
	d) ICP-OES	4	0.00	0.85	0.80	0.00

Acknowledgements

The present work has been supported by COVALFOOD "Valorisation of high added-value compounds from cocoa industry by-products as food ingredients and additives" project funded by European Union's Seventh Framework Programme for research and innovation under the Marie Skłodowska-Curie grant agreement No 609402 - 2020.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2019.04.008.

References

Andrade, K. S., Gonçalvez, R. T., Maraschin, M., Ribeiro-do-Valle, R. M., Martínez, J., & Ferreira, S. R. (2012). Supercritical fluid extraction from spent coffee grounds and coffee husks: Antioxidant activity and effect of operational variables on extract composition. *Talanta, 88*, 544–552.

Abadin, H., Ashizawa, A., Stevens, Y. W., Llados, F., Diamond, G., Sage, G., ... Swarts, S. G. (2007). Toxicological profile for lead. *US Department of Health and Human Services, 1*, 582.

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods, 5*(16), 3790–3798.

Biancolillo, A., Bucci, R., Magrì, A. L., Magrì, A. D., & Marini, F. (2014). Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication. *Analytica chimica acta, 820*, 23–31.

Barbosa-Pereira, L., Guglielmetti, A., & Zeppa, G. (2018). Pulsed electric field assisted extraction of bioactive compounds from cocoa bean shell and coffee silverskin. *Food and Bioprocess Technology, 11*(4), 818–835.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17*(3), 166–173.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy, 43*(5), 772–777.

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.

Carocho, M., Morales, P., & Ferreira, I. C. (2015). Natural food additives: Quo vadis? *Trends in Food Science & Technology, 45*(2), 284–295.

De Oliveira, I. R., Roque, J. V., de Maia, M. P., Stringheta, P. C., & Teófilo, R. F. (2018). New strategy for determination of anthocyanins, polyphenols and antioxidant capacity of *Brassica oleracea* liquid extract using infrared spectroscopies and multivariate regression. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 194*, 172–180.

Doeswijk, T. G., Smilde, A. K., Hageman, J. A., Westerhuis, J. A., & Van Eeuwijk, F. A. (2011). On the increase of predictive performance with high-level data fusion. *Analytica chimica acta, 705*(1–2), 41–47.

Dong, Y., Sørensen, K. M., He, S., & Engelsen, S. B. (2017). Gum Arabic authentication and mixture quantification by near infrared spectroscopy. *Food Control, 78*, 144–149.

Elwers, S., Zambrano, A., Rohsius, C., & Lieberei, R. (2009). Differences between the content of phenolic compounds in Criollo, Forastero and Trinitario cocoa seed (Theobroma cacao L.). *European Food Research and Technology, 229*(6), 937–948.

Jacobsen, S., Søndergaard, I., Møller, B., Desler, T., & Munck, L. (2005). A chemometric

evaluation of the underlying physical and chemical patterns that support near infrared spectroscopy of barley seeds as a tool for explorative classification of endosperm genes and gene combinations. *Journal of Cereal Science, 42*(3), 281–299.
Jansman, A. J., Verstegen, M. W., Huisman, J., & Van den Berg, J. W. (1995). Effects of hulls of fava beans (Vicia faba L.) with a low or high content of condensed tannins on the apparent ileal and fecal digestibility of nutrients and the excretion of endogenous protein in ileal digesta and feces of pigs. *Journal of Animal Science, 73*(1), 118–127.
Kelly, S., Heaton, K., & Hoogewerff, J. (2005). Tracing the geographical origin of food: The application of multi-element and multi-isotope analysis. *Trends in Food Science & Technology, 16*(12), 555–567.
Lee, A. R., Gautam, M., Kim, J., Shin, W. J., Choi, M. S., Bong, Y. S., ... Lee, K. S. (2011). A multianalytical approach for determining the geographical origin of ginseng using strontium isotopes, multielements, and 1H NMR analysis. *Journal of Agricultural and Food Chemistry, 59*(16), 8560–8567.
Li-Chan, E., Chalmers, J., & Griffiths, P. (Eds.). (2011). *Applications of Vibrational Spectroscopy in Food Science*. John Wiley & Sons.
Mandrile, L., Fusaro, I., Amato, G., Marchis, D., Martra, G., & Rossi, A. M. (2018). Detection of insect's meal in compound feed by Near Infrared spectral imaging. *Food Chemistry*.
Mandrile, L., Zeppa, G., Giovannozzi, A. M., & Rossi, A. M. (2016). Controlling protected designation of origin of wine by Raman spectroscopy. *Food Chemistry, 211*, 260–267.
Manzano, P., Hernández, J., Quijano-Avilés, M., Barragán, A., Chóez-Guaranda, I., Viteri, R., & Valle, O. (2017). Polyphenols extracted from Theobroma cacao waste and its utility as antioxidant. *Emirates Journal of Food and Agriculture, 29*(1), 45.
Marseglia, A., Acquotti, D., Consonni, R., Cagliani, L. R., Palla, G., & Caligiani, A. (2016). HR MAS 1H NMR and chemometrics as useful tool to assess the geographical origin of cocoa beans – Comparison with HR 1H NMR. *Food Research International, 85*, 273–281.
Martens, H., Nielsen, J. P., & Engelsen, S. B. (2003). Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Analytical Chemistry, 75*(3), 394–404.
Martín-Cabrejas, M. A., Valiente, C., Esteban, R. M., Mollá, E., & Waldron, K. (1994). Cocoa hull: A potential source of dietary fibre. *Journal of the Science of Food and Agriculture, 66*(3), 307–311.
Munck, L., Nørgaard, L., Engelsen, S. B., Bro, R., & Andersson, C. A. (1998). Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemometrics and Intelligent Laboratory Systems, 44*(1), 31–60.
Peng, J., Peng, S., Jiang, A., Wei, J., Li, C., & Tan, J. (2010). Asymmetric least squares for multiple spectra baseline correction. *Analytica chimica acta, 683*(1), 63–68.
Peres, B., Barlet, N., Loiseau, G., & Montet, D. (2007). Review of the current methods of analytical traceability allowing determination of the origin of foodstuffs. *Food Control, 18*(3), 228–235.
Redgwell, R., Trovato, V., Merinat, S., Curti, D., Hediger, S., & Manez, A. (2003). Dietary fibre in cocoa shell: Characterisation of component polysaccharides. *Food Chemistry, 81*(1), 103–112.
Rubio-Diaz, D. E., & Rodriguez-Saona, L. E. (2010). *Application of vibrational spectroscopy for the study of heat-induced changes in food components. Handbook of Vibrational Spectroscopy*.
Serra Bonvehí, J., & Escolá Jordà, R. (1998). Constituents of cocoa husks. *Zeitschrift für Naturforschung, 53c*, 785–792.
Silvestri, M., Elia, A., Bertelli, D., Salvatore, E., Durante, C., Vigni, M. L., ... Cocchi, M. (2014). A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines. *Chemometrics and Intelligent Laboratory Systems, 137*, 181–189.
Skov, T., Honoré, A. H., Hansen, H. M., Næs, T., & Engelsen, S. B. (2014). Chemometrics in foodomics: Handling data structures from multiple analytical platforms. *TRAC-Trends Analytical Chemistry, 60*, 71–79.
Socrates, G. (2001). *Infrared and Raman Characteristic Group Frequencies: Tables and Charts*. John Wiley & Sons.
Sørensen, K. M., Khakimov, B., & Engelsen, S. B. (2016). The use of rapid spectroscopic screening methods to detect adulteration of food raw materials and ingredients. *Current Opinion in Food Science, 10*, 45–51.
Sørensen, K. M., Aru, V., Khakimov, B., Aunskjær, U., & Engelsen, S. B. (2018). Biogenic amines: A key freshness parameter of animal protein products in the coming circular economy. *Current Opinion in Food Science*.
Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*, 37–52.
Zakaria, A., Shakaff, A. Y. M., Adom, A. H., Ahmad, M., Masnan, M. J., Aziz, A. H. A., ... Kamarudin, L. M. (2010). Improved classification of *Orthosiphon stamineus* by data fusion of electronic nose and tongue sensors. *Sensors, 10*(10), 8782–8796.