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

Authentication of cocoa bean shells by near-infrared and mid-infrared spectroscopy and inductive 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] and 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*

Corresponding author Luisa Mandrile, tel +39 011 3919329; e-mail l.mandrile@inrim.it

Keywords: cocoa bean shell, food traceability, data fusion, near infrared spectroscopy, mid infrared spectroscopy, inductive coupled plasma.

Abstract

The aim of this study was to evaluate the efficacy of a multi-analytical approach for origin authentication of cocoa beans shells (CBS). The overall chemical profiles of cocoa bean shells from different origins were collected and measured using diffuse reflectance near-infrared spectroscopy (NIRS) and attenuated total reflectance mid-infrared spectroscopy (ATR-FT-IR) for molecular composition, as well as inductive coupled plasma-optical emission spectroscopy (ICP-OES) for elemental composition. Exploratory chemometric

techniques were employed to identify systematic patterns related to the geographical origin of samples based on each technique using Principal Components Analysis (PCA). A combination of the three techniques proved to be the most promising approach to establish classification models. Partial Least Squares-Discriminant Analysis model of the fused PCA scores of three independent models was used and compared with single technique models. CBS samples were better classified by the fused model. Satisfactory classification rates were obtained for Central Africa samples with accuracy of 0.84.

1. Introduction

Since the 19th century cocoa has seen a continuous growth of consumption in a variety of forms, leading to an outstanding economic interest of chocolate industries for constant innovation and modernization. As many other agro-food activities, cocoa industry produces large amounts of by-products (https://www.icco.org/). Cocoa bean shells (CBS) is one of the main by-products, which represents the 12 % of weight after husking and grinding of dried cocoa seeds. CBS represents a non-negligible disposal problem and thus legislation and environmental issues are forcing industries to define process optimization and recovery/recycling strategies. Recently, bioconversion of by-products has raised the interest of scientific research and in several countries strategic vision or dedicated policies are being prepared to manage food industry wastes in the most efficient way – abandoning the "take, make and dispose" behavior and instead acting out a circular economy paradigm (Sørensen, Aru, Khakimov, Aunskjær, Engelsen, 2018). The increasing interest for byproducts has certainly an environmental basis, but an important role is played by the tendency to reduce the use of synthetic additives and replace them with natural substances in food. Research concerning new natural additives with high quality/costs ratio is increasing nowadays (Carocho, Morales, Ferreira, 2015). Moreover, the demand of new functional foods, rich in bio compounds such as polyphenols, fiber, n-3 fatty acids etc., drives interest for rich food wastes, such as seeds husks (Andrade, Gonçalvez, Maraschin, Ribeiro-do-Valle, Martínez, Ferreira, 2012; Jansman, Verstegen, Huisman, Van den Berg, 1995). Vegetal by-products are rich of nutrients, such as fiber, polyphenols, minerals and their recycling represent one of the valorization strategies. The development of CBS valorization strategies is aimed at reducing the environmental impact of the cocoa production and provides information to promote conversion of a by-product into added-value products with application in food and healthcare sectors. The definition of the chemical composition of CBS from different countries is meant to evaluate the systematic

49 differences due to their origin. Chemical analysis of CBS has been carried out in several research papers because
50 of its interesting features related to flavor, phenolic compounds and nutritional values (Barbosa-Pereira,
51 Guglielmetti, Zeppa, , 2018; Manzano, Hernández, Quijano-Avilés, Barragán, Chóez-Guaranda, Viteri, Valle,
52 2017; Redgwell, Trovato, Merinat, Curti, Hediger, Manez, 2003; Serra Bonvehí, and Escolá Jordà, 1998;
53 Martín- Cabrejas, Valiente, Esteban, Mollá, Waldron, 1994;), however a complete characterization, using
54 different methodologies to highlight similarities and differences in composition of samples from different
55 countries has not been accomplished yet. In this work, CBS samples from different countries were analyzed with
56 three different analytical methods. Near infrared spectroscopy (NIRS), Mid infrared spectroscopy by attenuated
57 total reflectance (ATR-FT-IR) and inductively coupled plasma-optical emission spectroscopy (ICP-OES) were
58 used to collect a wide chemical information, both molecular and elementary. The aim of this study was to
59 evaluate the validity of simple and rapid analytical techniques, supported by a chemometric approach, for the
60 identification of differences due to different geographical origin of samples of CBS, with the perspective of a
61 future application for traceability and origin authentication of CBS as food additive.
62 Nowadays, the exchange of food is realized in a complex and interconnected global net, and food products are
63 often exposed to frauds, false information, contamination risk and counterfeiting (Regattieri, Gamberi, Manzini,
64 2007). For this reason, it is extremely important to protect and valorize authentic products, including regionals
65 specialties. Innovative, reliable strategies to individuate specific markers of origin, as well as characteristic
66 compositional patterns that can be associated to a precise origin are needed (Mandrile, Giovannozzi, Zeppa,
67 Rossi, 2016). Geographical origin indicators should provide an analytical response to the geographical
68 traceability problem and support the documental certification, which is used today to guarantee food and food-
69 additives provenience. Different techniques such as NMR and isotope ratio mass spectrometry can play a
70 relevant role to provide origin indicators (Lee, et al., 2011). Rapid and non-destructive techniques, such as near
71 infrared spectroscopy, are particularly interesting because of the possibility to obtain an efficient and non-biased
72 overview of the sample chemistry (Sørensen, Khakimov, Engelsen, 2016; Sørensen, He, Engelsen, 2017). The
73 chemical specificity and ease of sampling of NIR spectroscopy make it an attractive tool for rapid and
74 comprehensive food analysis. The complex pattern of signals revealed by IR analysis, both in the near and mid
75 infrared spectral region, is correlated to the content of the different chemical constituents, such as proteins, fatty
76 acids, carbohydrates, alimentary fibers and phenolic compounds. Statistics and multivariate data analysis offer

77 powerful tools to identify robust correlations between measured data and geographical origin, and validated
78 models can provide useful methods for the recognition of unknown samples, with a certain probability (Peres,
79 Barlet, Loiseau, Montet, 2007; Kelly, Heaton, Hoogewerff, 2005). In this work, chemometrics was used for data
80 analysis to calculate at first explorative, and subsequently predictive, models. Principal Component Analysis for
81 data exploration and visualization is a well-established strategy to allow the extraction of useful information
82 from numerous experimental results in food science (Munck, Nørgaard, Engelsen, Bro, Andersson, 1998).
83 Moreover, data fusion for multi-block analysis was used to improve models, gaining information from several
84 different analytical techniques (Skov, Honoré, Hansen, Næs, Engelsen, 2014; Zakaria, et al, 2010, Silvestri et. al,
85 2014).
86
87 **2. Material and Methods**

88 *2.1 Samples*

89 Fermented and dried cocoa (Theobroma cacao L.) samples were selected and collected within COVALFOOD
90 project funded by European Union's Seventh Framework, involving five Italian chocolate industries. A complete
91 list of 78 samples with the associated information about supplier, provenience and variety is reported in table
92 1S.1 in supplementary information. For an easier exploration of the sample pool, charts of geographical and
93 varietal_distribution are shown in figure 1S.1. All samples were imported as untreated raw materials, and the
94 geographical origin was guaranteed by the supplying industry. All samples were roasted and decorticated in
95 laboratory in a ventilated oven for 20 min at 130°C. After roasting, the fragile shell of the beans was separated
96 by mechanical rubbing and removed by hoover suction. The collected cocoa bean shells (CBS) were ground
97 using an ultra-centrifugal mill Retsch ZM 200 (RetschGmbh, Haan, Germany) and stored as dry fine powders
98 (250 μm) in a desiccator in closed containers.

99 *2.2 Near infrared spectroscopy*

100 NIR spectra of CBS were collected in the spectral range 10000 - 4000 cm^{-1} (1000 - 2500 nm) using an Antaris II
101 FT-NIR spectrometer (Thermo Fisher, Waltham, USA) in diffuse reflectance mode. The integrating sphere
102 accessorize was used to collect diffuse reflected light. CBS was analyzed without sample pretreatment; 0.1 g of

103 powder in a quartz glass vial located over the integrating sphere. 32 scans were collected per each sample with
104 spectral resolution of 8 cm^{-1}. A clean flat golden surface was used for background collection. Three
105 measurement replicates were collected per sample. All samples were measured in randomized order.

106 *2.3 Mid infrared spectroscopy*

107 ATR-FT-IR spectra in the mid infrared region between 500 - 4000 cm^{-1} were collected using Nicolet FT-IR
108 spectrometer (Thermo Fisher, Waltham, USA), Germanium crystal (n = 5.7) for total reflection was used which
109 allows a maximum sample penetration of 1 µm. 64 scans were needed for a good signal to noise with 4 cm^{-1}
110 resolution. The sample powder was pressed with a conical tip on the crystal, the pressure applied was 15 Bar.
111 The tip and the crystal were washed with ethanol between one sample analysis and the following. Three spectra
112 were collected for each sample, resampling at each replicate.

113 *2.4 ICP-OES elemental composition*

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

124 *2.5 Data treatment*

125 Chemometric data analysis was carried out using PLS Toolbox from Eigenvector Research, Inc. (Manson, WA)
126 for Matlab R2015a (Mathworks, Natick, USA). Principal Components Analysis (PCA) method is a linear
127 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 scores plot are useful to explore data and to find correlation between measured variables and the information of interest, such as geographical provenience of CBS, in this case. Then PLS-DA (Barker and Rayens 2003) models were calculated to compare the classification performances of the three techniques separately with the results obtained by joining the three datasets and considering all information contemporarily. Ten classes were considered: Central Africa, Ecuador, Gulf of Mexico, Indonesia, Mexico, Peru, São Tomé, Colombia, Venezuela and Brazil. All the calculated PLS-DA models were validated using leave-one group-out cross validation. The subsets of samples used as tests sets in cross validation corresponds to the country of origin. For each technique data preprocessing details 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 validation to compare classification performances of different techniques.

2.5.1 NIRS data treatment

Preprocessing of NIRS data was applied to extract useful information from the dataset. Absolute absorbance variations and unwanted light scattering were removed using preprocessing of the NIRS data (Martens et al, 2003). The most effective preprocessing was chosen based on the minimum differences between replicates on the PCA scores plots relative to the distance between samples. 2^{nd} 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 data trends changings. Processed spectra were shown in figure 2S.1. Unwanted variability was successfully removed as demonstrated by the narrow grouping of the replicates obtained after processing shown in figure 2S.2 in supplementary information. PCA was applied to visualize data and to investigate systematic differences among samples, and variables with peculiar relevance were identified. 4LVs PLS-DA classification model was also calculated to discriminate classes of samples from different geographical areas. Same spectra preprocessing was used.

154 *2.5.2 MIRS data treatment*

155 Preprocessing of data was performed to suppress useless variability associated to unwanted noise. The selection
156 criterion for data preprocessing was the maximized closeness of the scores of technical replicates on PC1, as
157 shown in figure 3S.1 in supplementary information. Baseline correction (using asymmetric weighted least
158 squares algorithm, with basis filter of order 2) (Peng, Peng, Jiang, Wei, Li, Tan, 2010) followed by second
159 derivative (Savitzky Golay, filter width 15 and polynomial order 2) and mean centering was selected as optimal
160 preprocessing. PCA model for data visualization and exploration was calculated; PLS-DA classification model
161 using 4 LVs of the same preprocessed data was also calculated to compare MIRS classification capabilities with
162 the other techniques.

163 *2.5.3 ICP-OES data treatment*

164 ICP emission spectra were evaluated for quantification using a calibration curve per element. The calibration
165 curves were estimated using two series of standards prepared by dilution of a certified standard mix (ICP Multi-
166 element standard solution IV, Sigma Aldrich, Germany) containing known concentration of 21 elements (Al, B,
167 Ba, Bi, Ca, Cd, Co, Cr, Cu, Fe, K, Li, Mg, Mn, Mo, Na, Ni, Pb, Sr, Tl, Zn). Standard concentrations were 0, 0.2,
168 0.4, 0.6, 0.8, 1, 2, 4, 6, 8, 10, 20, 30, 40, 60, 80 100 mg/100g of the certified standard concentration, which was 5
169 mg/l for all elements, out of Potassium that was 50 mg/l in the standard solution. Three emission wavelengths
170 were monitored per each element, then the intensity revealed for only one λ was selected per each element based
171 on the best correlation coefficient of the corresponding calibration curve and trying to avoid interferences
172 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
173 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;
174 λ_{Mn}=259.4 nm; λ_{Mo}=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;
175 λ_{Zn}=202.5 nm.

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

180 *2.5.4 Data fusion*

181 The multi-block tool of PLS toolbox by Eigenvector was used to fuse the PCA scores from the three single PCA
182 models of the different analytical techniques. A joined model exploiting mid-level data fusion was obtained (;
183 Borràs, et al, 2014). To make the interpretation clearer, the measurement replicates were averaged, and one
184 matrix line per each sample was maintained for the three different original datasets (NIRS, MIR-ATR and ICP).
185 Each block was first decomposed by PCA, and the resulting scores were fused into a new dataset. The samples'
186 scores for the most relevant PCs were considered to calculate a new fused model. Seven PCs were considered for
187 MIRS and ICP, and six PCs were considered for NIRS. Thus, twenty initial variables were used to build the new
188 joined PCA model. Default autoscale was applied before joining data. PLS-DA method was then performed with
189 autoscaled data to obtain a classification model (Ballabio, Consonni, 2013). The class vector was represented by
190 the area of origin. It was composed of 10 classes i.e. Central Africa, Colombia, Ecuador, Gulf of Mexico,
191 Indonesia, Mexico, Peru, São Tomé, Venezuela, Brazil. Unfortunately the number of samples per each class was
192 not balanced, due to sample availability. Five latent variables were considered for the PLS-DA model, based on
193 the minimum average classification error in cross validation, using leave-one country-out cross validation
194 strategy.

195 **3. Results and Discussion**

196 *3.1 NIRS spectroscopy characterization of CBS samples*

197 The NIRS profiles show the typical broad bands of overtones and combination bands of vibrational modes
198 associated to the main constituents of vegetal origin materials. The assignment of the most bands of the NIR
199 spectrum are reported in table 2S.1 in the supplementary information (Jacobsen, et. al. 2011). The mean NIR
200 spectra of all CBS samples is shown in figure 1 a, together with the standard deviation profiles. Similar spectral
201 shape was obtained for all samples, the same bands are present in all spectra with slight differences in mutual
202 intensities.

203 **Figure 1**

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

The sensitivity of NIRS to the botanical variety was tested at first, since it has been previously demonstrated in literature that differences in the chemical composition of different varieties of *Theobroma Cacao L.* were present (Elwers, Zambrano, Rohsius, Lieberei, 2009). The outcome of the PCA on the NIR spectra is shown in figure 1 b. In contrast with expectations, different botanical varieties did not cause evident systematic clustering of NIR spectra. The scores of NIR spectra of *Forastero* and *Trinitario* samples were overlapped in the scores plot (figure 1 b), no separation occurred neither in the PC2/PC1 plot, nor in the later PCs (plots not shown). This can be probably attributed to the complexity of the samples' set, that introduces a lot of confusing variability. However, Arriba samples, a specific variety cultivated in Ecuador only (green squares on the scores plot in figure 1b), was specifically, even though not selectively, characterized by negative scores on PC1 and positive scores on PC2 attesting the capability of NIR spectra to catch common chemical features of Arriba samples. The loadings profiles (figure 2S.3 a) and the variance captured (figure 2S.4) hallows to define what spectral regions are involved in each relevant PC. PC1, is mainly characterized by fatty acids bands as 5670-5780 cm^{-1} (1^{st} C-H str) and 4325 cm^{-1} (1^{st} C-H str + 1^{st} C-H def CH_2), 4250 cm^{-1} (1^{st} C-H str + 1^{st} C-H def). In addition PC1 captures also some regions related to proteins such as 5170-5190 cm^{-1} (2^{nd} C=O of CONH), 5269 cm^{-1} (2^{nd} C=O of COOH), 6320 cm^{-1} (1^{st} N-H str of CONH) and 6535 cm^{-1} (1^{st} N-H str of RNH_2) and 6950 cm^{-1}. PC2, instead, shows three maxima at 4400 cm^{-1} (1^{st} O-H str + 1st C-C str, associated to starch), 4763 cm^{-1} (2^{nd} O-H def + 2^{nd} C-O str of starch) and 5000 cm^{-1} (2^{nd} O-H def + 1st C-O def of starch), this means that PC2 mostly represents the starch content into the samples. PCA highlighted a major content of fatty acids and vegetal proteins in the examined Arriba samples as shown in figure 1 c, d, whereas lower intensity in the spectral regions associable to polysaccharides, such as starch, was measured (corresponding enlarged spectral region not shown for brevity reasons).

As far as correlations between the geographical origin and NIR spectra are concerned, the information provided by the scores plot seems confused at a first look, however some interesting considerations can be underlined.

231 Common features of all samples coming from central Africa were noticed in the scores plot (figure 2 a) when
232 considering PC2. On average, central Africa samples (red rhombus in figure 2 a) show positive scores on PC2
233 (related to polysaccharides and starch bands mainly). Moreover other common features were noticed in further
234 PCs, such as negative scores on PC3 (figure 2S.6 b) (where the main contributions are 5218 cm^{-1}, 1st O-H str of
235 phenols, 5878 cm^{-1} 1st C-H str CH_3, 6075 cm^{-1} 1st C-H str of R-CH-CH, 7062 cm^{-1}, 2nd C-H str + 1st C-H def of
236 aromatic compounds) and positive again on PC4 (Figure 2S.6 c) which is related mainly to carbohydrates (4790
237 cm^{-1} 1st O-H str + 1st O-H def ROH o sucrose and starch, 6264 cm^{-1}, 1st O-H str intramolecular H-bond of starch
238 or glucose). The mentioned frequency is marked on the enlarged spectrum in figure 2 c, since it resulted to be
239 particularly relevant for the differentiation of the spectra from different areas, even though it does not correspond
240 to main peak in the spectrum. Although the separation of the examined groups is not sufficient for selective
241 discrimination, it was confirmed that the geographical origin information is captured by NIRS. As shown in
242 figure 2 a, African samples from São Tomé (a little island in Guinea Gulf, at latitude 0°) show features in
243 common with samples coming from America, which on average showed negative scores on PC2. The scores of
244 São Tomé samples (light blue rhombus in figure 2 a) are mixed with Gulf of Mexico Samples, this can be
245 attributed to similar environmental and climatic conditions of the little islands, that influences the chemical
246 composition of Cocoa fruits, and therefore of CBS (see also figures 2S.6 a to appreciate similitudes of São Tomé
247 with samples from the islands and coasts of Gulf of Mexico). Moreover, Ecuador samples seemed more similar
248 to the African samples than to the American, indeed, in figure 2 a, orange circles corresponding to Ecuador
249 samples are mixed with red rhombus corresponding to samples from Central Africa.
250 The Asian samples are separated from the others (blue triangles in figure 2 b), because of high values on PCs 4,
251 5 and 6. PC4 characterized by a peak around 4530 cm^{-1}. This spectral region, represented in figure 2 c is
252 assigned to ROH combination modes, so it can be hypothesized that sugars' content differs for Asian samples
253 with respect to all the others. The most represented spectral region in PC5 (which is relevant for the clustering of
254 Asian samples) is the side of the peak at 6300 cm^{-1}. This region, represented in figure 2 d, highlights that the
255 bands' shape is relevant, more than its intensity in this case. PC6 is also responsible for the following spectral
256 regions: 4466 cm^{-1} (beta-glucan), 5114 cm^{-1} (2nd C=O of esters) and 7147 cm^{-1} typical of R-OH (figures 2S.3,
257 2S.4 can be consulted for all attribution of spectral bands to the PCs).
258 **Figure 2**

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

3.2 ATR-FT-IR spectra

Spectral profiles in the mid infrared region are shown in figure 3 a. As well as for NIRS, ATR-FT-IR spectroscopy is expected to deliver information about the chemical composition of CBS samples including most of biochemical species present in the matrix. Although absorption bands in the mid infrared region are more defined and narrower because primary vibration modes absorb in this spectral region, the visual interpretation of spectra is difficult, especially in the so-called fingerprint region, between 1750 cm^{-1} and 500 cm^{-1}. Main bands 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 2260-2440 cm^{-1}, where CO_2 band is present, was excluded.

Figure 3

MIRS spectra provided information in agreement with NIRS investigation. Signals are more defined and spectral specificity is increased compared to NIRS, and PCA scores plots investigation resulted an effective strategy to explore spectra similarities. Similarities and differences between samples are ruled by PC1, 2 and 3. The correspondence between PCs and MIR spectral regions was evaluated analyzing figure 3S.4, where the MIR spectrum was superimposed over the histogram of the percentage of variance captured by each PC, to understand what bands drive the scores distribution on the scores plot. PC1 is mainly dominated by CH_x vibrations in the 3000-2800 cm^{-1} and 1460-1420 cm^{-1} region (samples with high intensity of signals at 2920 cm^{-1} and 1463 cm^{-1} present lower values of PC1), moreover 1730 cm^{-1} peak (C=O stretching) that showed increased intensity in Arriba samples is also represented in PC1; PC2 captures variance in 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 to carbohydrates are also relevant, for example 763 cm^{-1} related to pyranose compounds is modeled by PC5. Variety information reveals a certain grouping of Arriba sample that show high PC2 scores

285 and lower intensity of PC5 in Arriba samples, in agreement with NIRS results. The scores plot colored by variety
286 information is shown in figure 3S.5.

287 The different geographical provenience drives a differentiation between samples and some general
288 considerations can be extracted from the scores plot (figure 3 b,c). PC2 certainly explains interesting
289 characteristics of Central Africa samples, that show positive scores on PC2. Samples from São Tomé showed
290 more similarities with samples from Gulf of Mexico, Venezuela and Colombia, as attested also by NIRS data
291 shown in the previous paragraph. This confirms that similar climatic and environmental conditions are crucial in
292 determining the chemical composition captured by spectroscopic techniques, as previously reported in literature
293 for cocoa samples (Marseglia, et al, 2017). African samples show higher intensity at 2954 cm^{-1} and 2870 cm^{-1} in
294 the CH_x stretching vibrations (Figure 3 d). Moreover, PC5 and PC6 were relevant to identify features in common
295 between Ecuadorian samples. 87% of Ecuador samples were placed to the left of the left diagonal of the
296 PC6/PC5 plot (figure 3 c). This is due to the ratio between 1280 cm^{-1} (Amide III of β-sheet proteins) and 1320
297 cm^{-1} or 1440 cm^{-1} allows to separate samples from Ecuador from other American samples, as shown in figure 3 e.
298 Moreover low values in PC5 reflect low intensities at 673 cm^{-1} and 1600 cm^{-1} (ring breathing modes of
299 polysaccharides) as already noticed for Arriba samples (enlarged spectral regions not shown for brevity reasons).
300 The ATR-FT-IR spectrum represents the sum of numerous bands of several functional groups, which are
301 contemporarily present in more than one biochemical compound. Beyond the hypothesized interpretation, it
302 should be stressed that an accurate understanding of what peaks and bands drive the scores distribution should
303 by managed carefully to avoid misinterpretation. To univocally associate the relevant spectral regions to specific
304 classes of compounds remains complicated when a whole complex matrix such as food is analyzed. However,
305 the possibility to identify spectral features that precisely, characterize samples from the same origin is an
306 indication that a correlation between geographical origin and vibrational spectra can be modeled.

307 *3.3. ICP-OES elemental characterization of CBS samples*

308 The raw ICP-OES results are shown in Table 4S.1 in supplementary information. The most abundant elements
309 are by far Ca, Mg, K which have a concentration at least one order of magnitude higher compared to all other
310 elements. Among the secondary elements, particularly relevant were Al, Fe and Li (Barker and Rayens 2003).

311 Relevant amounts of lead were revealed in all samples (around 0.3 mg/kg), which is a high value compared with
312 the average content of lead in foods reported in 2007 by the Agency for Toxic Substances and Disease Registry
313 (Abadin H., et al. 2007). All other elements were revealed in concentration lower than 0.2 mg/kg, particularly
314 low concentrations were determined for Ni and Cr. PCA was used to identify major variance directions that can
315 be related to geographical origin. Five samples were identified as very different from the others. They were SB3,
316 SB4 from Brazil, ICAM10 from Congo, FER8 from Uganda and FER13 from Côte d'Yvoire. These samples
317 were excluded as outliers because of their very low K content. Boron, Potassium, Magnesium and Calcium are
318 responsible of the most variance captured by PC1, which resulted not to be particularly correlated to provenience
319 of samples. Aluminum, Chromium, Iron, Sodium and Nickel are particularly relevant for PC2, whereas
320 Cadmium, Cobalt and Molybdenum together with Calcium and Manganese are mostly represented in PC3, as
321 shown in figure 4 d.

322 Examining the PC2/PC3 loadings and scores plot (figure 4 a, b), high levels of Fe and Al resulted to be
323 characteristic for African continent for most of Central Africa Samples, moreover a general deficiency of Ca, K,
324 Mg, Ni, was revealed. Interestingly some similitudes of São Tomé samples with American samples were
325 captured by PC2. Precisely a relatively higher content of Fe, Al, Cu and Ni was revealed for this samples, this
326 trend makes São Tomé samples more like American than to African samples. Moreover, São Tomé samples are
327 characterized by high content of Ba with respect to others. Conversely, Ecuador samples did not show any
328 specific elemental profile.

329 **Figure 4**

330 *3.4 Data fusion to merge chemical information provided by the different analytical techniques*

331 The idea of data fusion is to merge information, provided by different analytical determinations, in one single
332 data set, to enhance the quality of the results. The obtained joined PCA model clearly shows that all the three
333 datasets provide useful information for the final model. It was noticed that the three most represented variables
334 in PC1 were one from MIR-ATR, one from ICP and one from NIRS (figure 5S.1 in supplementary information).
335 The scores plot and the loadings projected on the PC2/PC1 space are shown in figure 5. The grouping of samples

336 based on the geographical origin was improved by the multi analytical model. Proximity, and hence common
337 features, were appreciated for samples from the same geographical area.
338 Classification models were calculated to quantify the grouping performances of the joined model compared to
339 the three single models, based on the geographical origin. Even though interesting observations were previously
340 discussed for the three techniques separately, and some correlation between geographical origin and the
341 composition was defined, single technique outputs were not accurate and precise for the recognition of the
342 geographical origin of samples in predictive classification models. In table 1 the most classification figure of
343 merit (sensitivity, specificity, error rate, accuracy, precision) relative to PLS-DA classification models for the
344 geographical discrimination were reported. The classification performances for the samples' classes composed of
345 more than 5 samples were shown. Classification results were higher for the joined model compared to each of
346 the three single models for Central Africa, Ecuador and Gulf of Mexico classes. This experimental evidence was
347 in agreement with literature findings corroborating mid-level or high-level data fusion to increase predictive
348 performance of classification models (Doeswijk, T. G., Smilde, A. K., Hageman, J. A., Westerhuis, J. A., & Van
349 Eeuwijk, F. A. 2011). Single techniques provide null accuracy and precision for most classes, out of Central
350 Africa. Moreover, merging information from the three techniques, the accuracy (correctly classified samples
351 rate) increased.

352 **Table 1**

353 NIRS, MIRS and ICP profiles together deliver sufficiently accurate information to capture the common features
354 of African samples, and to distinguish them from all the others. Unfortunately, the same is not confirmed for the
355 other classes. Low stability emerged during cross validation for Ecuador, Gulf of Mexico and Venezuela classes.
356 Classification results for classes composed of less than 10 samples were not considered statistically valid.

357 **4. Conclusions**

358 Because of the low price and interesting features of CBS, such as the extraordinary similarity to cocoa powder in
359 terms of color, taste and texture, and the potential beneficial effects on human health, research is needed to assist
360 the valorization of this food by-product, and to prevent fraud in cocoa powder market. The present work
361 demonstrats the existence of correlations between the geographical origin and the composition of CBS samples,
362 even though low specificity for the single country or restricted areas emerged. Some information about what

363 samples from the same macro-area have in common was described. The selected techniques provided significant
364 criteria to distinguish sample classes, such as Central Africa and Ecuador samples with adequate accuracy and
365 precision, however it is very difficult to precisely determine what chemical species drive this separation only
366 using vibrational spectroscopy for chemical composition analysis. Nevertheless, estimates and trends were
367 determined. The geographical traceability of food based on chemical analysis remains complicated and always
368 valid rules are rarely identified. The natural variability of most food materials is huge, climatic conditions and
369 process variables represent an intrinsic limit of this field of study. However, the capability to identify leading
370 variables, common trends and general indications using rapid and simple techniques is an encouraging result in
371 this domain. More sensitive and accurate techniques should be used for an exhaustive investigation. Easy-to-use
372 instrumental analysis still needs the support of heavier analytical strategies for comparison and calibration.

373 **Acknowledgements**

374 The present work has been supported by COVALFOOD "Valorisation of high added-value compounds from
375 cocoa industry by-products as food ingredients and additives" project funded by European Union's Seventh
376 Framework programme for research and innovation under the Marie Skłodowska-Curie grant agreement No
377 609402 - 2020 researchers: Train to Move (T2M).

378 **References**

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

382 Abadin, H., Ashizawa, A., Stevens Y.W., Llados, F., Diamond, G., Sage, G., Citra, M., Quinones, A., Bosch S.
383 J., and Swarts, S. G., ATSDR, U. (2007). Toxicological profile for lead. US Department of Health and Human
384 Services, 1, 582.

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

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

389 Barbosa-Pereira, L., Guglielmetti, A., & Zeppa, G., 2018. Pulsed Electric Field Assisted Extraction of Bioactive
390 Compounds from Cocoa Bean Shell and Coffee Silverskin. *Food and Bioprocess Technology*, 11(4), 818-835.

391 Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):
392 166–173.

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

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

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

399 Delwiche, S. R., Pitt, R. E., Norris, K. H. (1991). Examination of Starch- Water and Cellulose- Water
400 Interactions With Near Infrared (NIR) Diffuse Reflectance Spectroscopy. *Starch- Stärke*, 43(11), 415-422.

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

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

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

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

412 Jacobsen, S., Søndergaard, I., Møller, B., Desler, T., Munck, L. (2005). A chemometric evaluation of the
413 underlying physical and chemical patterns that support near infrared spectroscopy of barley seeds as a tool for
414 explorative classification of endosperm genes and gene combinations. *Journal of Cereal Science*, 42(3), 281-
415 299.

416 Jansman, A. J., Verstegen, M. W., Huisman, J., Van den Berg, J. W. (1995). Effects of hulls of fava beans (Vicia
417 faba L.) with a low or high content of condensed tannins on the apparent ileal and fecal digestibility of nutrients
418 and the excretion of endogenous protein in ileal digesta and feces of pigs. *Journal of Animal Science*, 73(1), 118-
419 127.

420 Kelly, S., Heaton, K., Hoogewerff, J. (2005). Tracing the geographical origin of food: The application of multi-
421 element and multi-isotope analysis. *Trends in Food Science & Technology*, 16(12), 555-567.

422 Lee, A. R., Gautam, M., Kim, J., Shin, W. J., Choi, M. S., Bong, Y. S., Hwang G.S., Lee, K. S. (2011). A
423 multianalytical approach for determining the geographical origin of ginseng using strontium isotopes,
424 multielements, and 1H NMR analysis. *Journal of agricultural and food chemistry*, 59(16), 8560-8567.

425 Li-Chan, E., Chalmers, J., Griffiths, P. (Eds.). (2011). Applications of vibrational spectroscopy in Food Science.
426 John Wiley & Sons.

427 Luykx, D. M., Van Ruth, S. M. (2008). An overview of analytical methods for determining the geographical
428 origin of food products. *Food Chemistry*, 107(2), 897-911.

429 Magagna, F., Guglielmetti, A., Liberto, E., Reichenbach, S. E., Allegrucci, E., Gobino, G., ... & Cordero, C.
430 (2017). Comprehensive Chemical Fingerprinting of High-Quality Cocoa at Early Stages of Processing:
431 Effectiveness of Combined Untargeted and Targeted Approaches for Classification and Discrimination. *Journal*
432 *of agricultural and food chemistry*, 65(30), 6329-6341.

433 Mandrile, L., Fusaro, I., Amato, G., Marchis, D., Martra, G., & Rossi, A. M. (2018). Detection of insect's meal
434 in compound feed by Near Infrared spectral imaging. *Food Chemistry*.

435 Mandrile, L., Zeppa, G., Giovannozzi, A. M., & Rossi, A. M. (2016). Controlling protected designation of origin
436 of wine by Raman spectroscopy. *Food chemistry*, 211, 260-267.

437 Manzano, P., Hernández, J., Quijano-Avilés, M., Barragán, A., Chóez-Guaranda, I., Viteri, R., Valle, O. (2017).
438 Polyphenols extracted from Theobroma cacao waste and its utility as antioxidant. *Emirates Journal of Food and*
439 *Agriculture*, 29(1), 45.

440 Marseglia, A., Acquotti, D., Consonni, R., Cagliani, L. R., Palla, G., & Caligiani, A. (2016). HR MAS 1H NMR
441 and chemometrics as useful tool to assess the geographical origin of cocoa beans–Comparison with HR 1H
442 NMR. *Food Research International*, 85, 273-281.

443 Martens, H., Nielsen, J. P., & Engelsen, S. B. (2003). Light scattering and light absorbance separated by
444 extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures.
445 *Analytical Chemistry*, 75(3), 394-404.

446 Martín- Cabrejas, M. A., Valiente, C., Esteban, R. M., Mollá, E., Waldron, K. (1994). Cocoa hull: a potential
447 source of dietary fibre. *Journal of the Science of Food and Agriculture*, 66(3), 307-311.

448 Munck, L., Nørgaard, L., Engelsen, S. B., Bro, R., Andersson, C. A. (1998). Chemometrics in food science—a
449 demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific
450 significance. *Chemometrics and Intelligent Laboratory Systems*, 44(1), 31-60.

451 Peng, J., Peng, S., Jiang, A., Wei, J., Li, C., & Tan, J. (2010). Asymmetric least squares for multiple spectra
452 baseline correction. *Analytica chimica acta*, 683(1), 63-68.

453 Peres, B., Barlet, N., Loiseau, G., Montet, D. (2007). Review of the current methods of analytical traceability
454 allowing determination of the origin of foodstuffs. *Food Control*, 18(3), 228-235.

455 Redgwell, R., Trovato, V., Merinat, S., Curti, D., Hediger, S., Manez, A. (2003). Dietary fibre in cocoa shell:
456 characterisation of component polysaccharides. *Food Chemistry*, 81(1), 103-112.

457 Regattieri, A., Gamberi, M., Manzini, R. (2007). Traceability of food products: General framework and
458 experimental evidence. *Journal of food engineering*, 81(2), 347-356.

459 Rubio- Diaz, D. E., Rodriguez- Saona, L. E. (2010). Application of Vibrational Spectroscopy for the Study of
460 Heat- Induced Changes in Food Components. *Handbook of Vibrational Spectroscopy.*

461 Schwanninger, M., Rodrigues, J. C., Fackler, K. (2011). A review of band assignments in near infrared spectra of
462 wood and wood components. *Journal of Near Infrared Spectroscopy*, 19(5), 287-308.

463 Serra Bonvehí, J., and Escolá Jordà, R. (1998). Constituents of Cocoa Husks, *Z. Naturforsch.* 53c, 785-792.

464 Silvestri, M., Elia, A., Bertelli, D., Salvatore, E., Durante, C., Vigni, M. L., ... & Cocchi, M. (2014). A mid level
465 data fusion strategy for the Varietal Classification of Lambrusco PDO wines. Chemometrics and Intelligent
466 Laboratory Systems, 137, 181-189.

467 Skov T., Honoré A.H., Hansen H.M., Næs T., S.B. Engelsen,, (2014). Chemometrics in Foodomics: Handling
468 data structures from multiple analytical platforms, TRAC-Trends. *Analytical Chemistry*, 60, 71-79.

469 Socrates, G. (2001). Infrared and Raman characteristic group frequencies: tables and charts. John Wiley & Sons.

470 Sørensen, K. M., Khakimov, B., & Engelsen, S. B. (2016). The use of rapid spectroscopic screening methods to
471 detect adulteration of food raw materials and ingredients. *Current Opinion in Food Science*, 10, 45-51.;

472 Sørensen, K. M., Aru, V., Khakimov, B., Aunskjær, U., & Engelsen, S. B. (2018). Biogenic Amines: a key
473 freshness parameter of animal protein products in the coming circular economy. *Current Opinion in Food*
474 *Science.*

475 Wold S., Esbensen K., Geladi P. (1987). Principal component analysis. *Chemometrics and Intelligent*
476 *Laboratory Systems*, 2, 37-52.

477 Zakaria, A., Shakaff, A.Y.M., Adom, A.H., Ahmad, M., Masnan, M.J., Aziz, A.H.A., Fikri, N.A., Abdullah,
478 A.H. and Kamarudin, L.M. (2010). Improved classification of Orthosiphon stamineus by data fusion of
479 electronic nose and tongue sensors. *Sensors*, 10(10), 8782-8796.

480 **FIGURE CAPTIONS**

481 **Figure 1**–a) Mean NIR spectrum of all CBS samples (blue) and standard deviation limits (green); b) Scores plot
482 of NIRS data PCA colored in accordance with variety; c, d) Zoom of average spectrum of Arriba samples
483 compared with the mean spectrum calculated considering all other NIR spectra.

484 **Figure 2**– a, b) PCA scores plot of NIR spectra of CBS sample colored by geographical origin. c, d) Zoom on
485 the spectral regions which make Asian samples different from all other CBS samples.

486 **Figure 3**– a) ATR-FT-IR average spectra; b) PC2 scores plot which highlight common behavior of African
487 samples; c) PC5/PC6 scores plot that allow to highlight characteristic trend for Ecuador samples; d) MIR
488 average spectra of CHx stretching bands of samples different geographical origin; e) MIR average spectra of
489 Ecuador sample compared with Americans in the spectral region where Ecuador samples show distinct
490 characteristics with respect to American samples.

491 **Figure 4**– PCA model of ICP-OES data outputs, 2D a) loading and b) scores plots; c) Histogram of mean data
492 for the considered macro-classes (Africa and America) and São Tomé samples that show peculiar feature with
493 respect to others; d) Variance captured per each principal component.

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

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

501

502

503

504

505

506

Declaration of interests

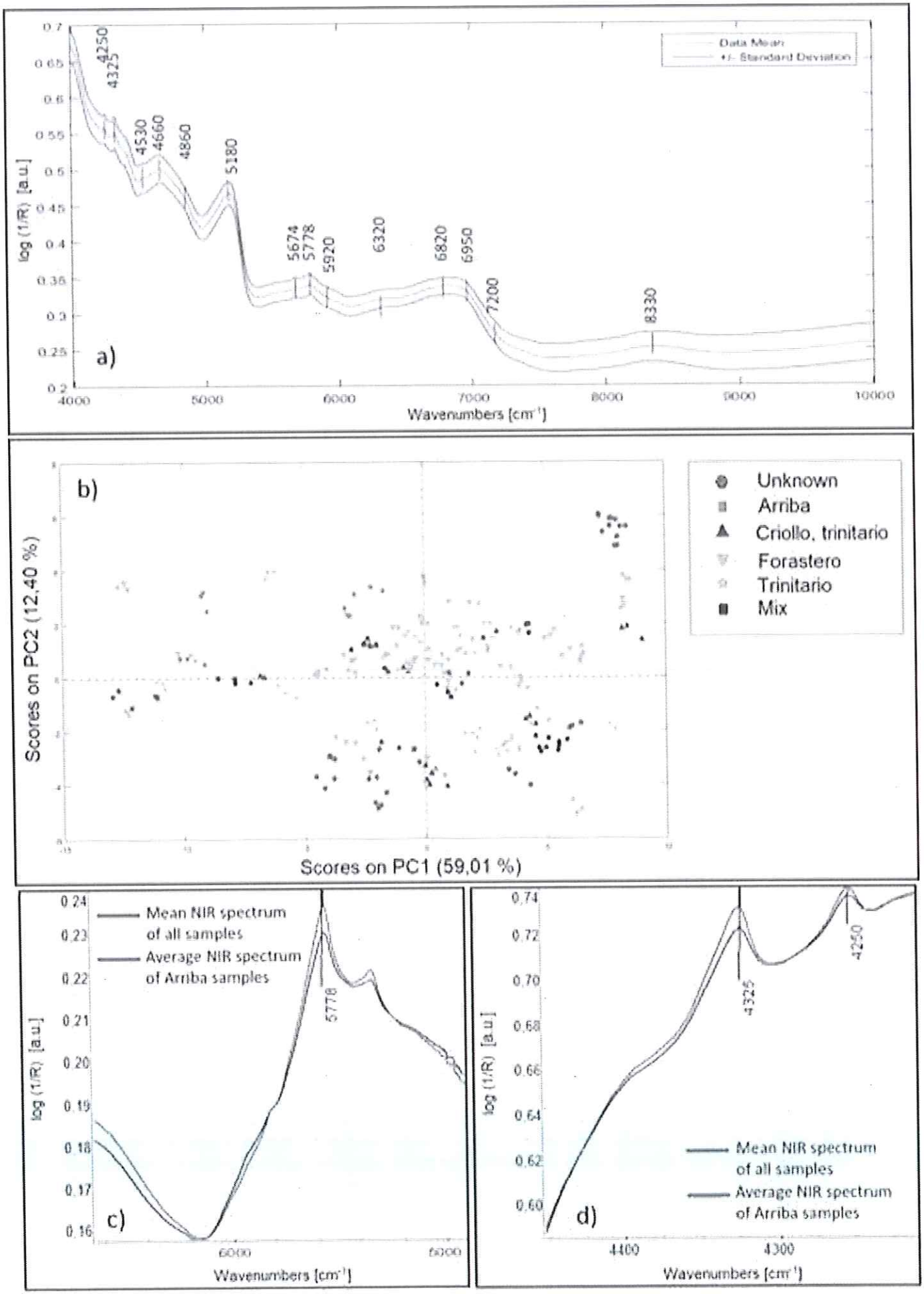
☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

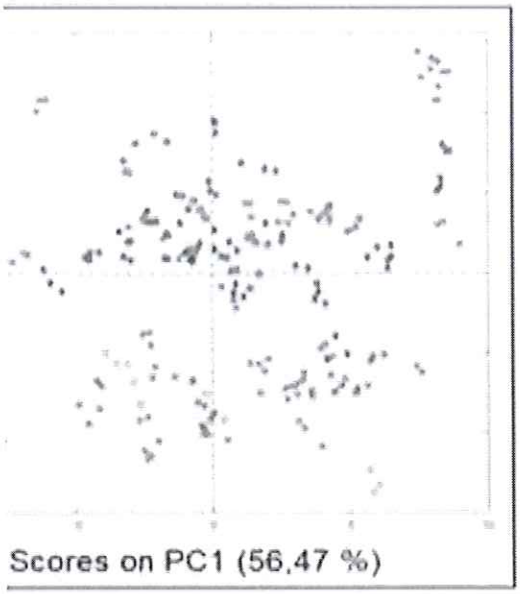
Class	Technique	N	Sensitivity (true positive ratio)	Specificity (true negative ratio)	Accuracy	Precision
Central Africa	a) **Joined**	22	0.68	0.92	0.84	0.79
	b) NIRS	19	0.68	0.86	0.81	0.00
	c) MIRS	19	0.32	0.70	0.59	0.29
	d) ICP-OES	19	0.50	0.83	0.75	0.50
Gulf of Mexico	a) **Joined**	9	0.33	0.82	0.76	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

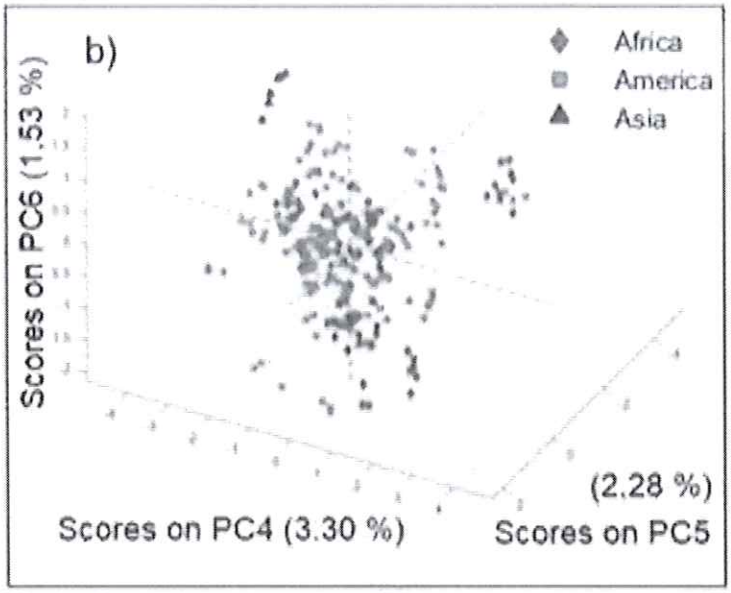
Table 1: Cross Validation outputs of PLS-Discriminant Analysis classification models for geographical origin discrimination: a) Joined classification model with 5 LVs, classification performances in leave-one origin-out cross validation; b) NIRS PLS-DA model with 4 LVs classification performances in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

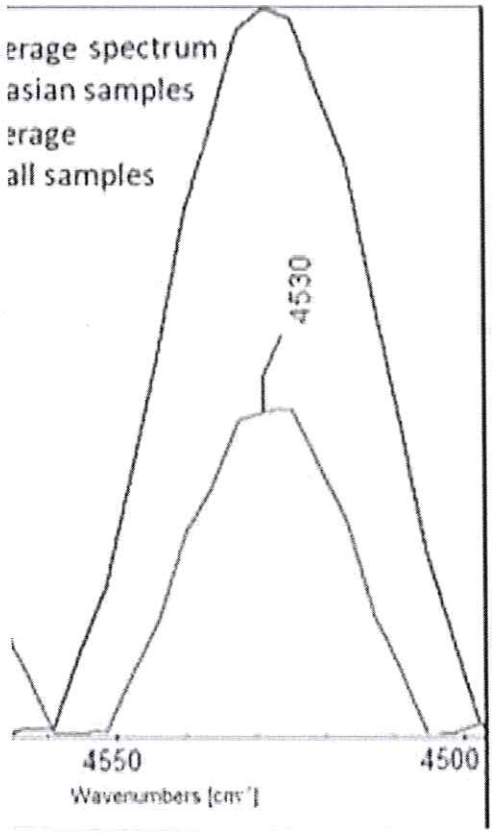
lick here to download high resolution image

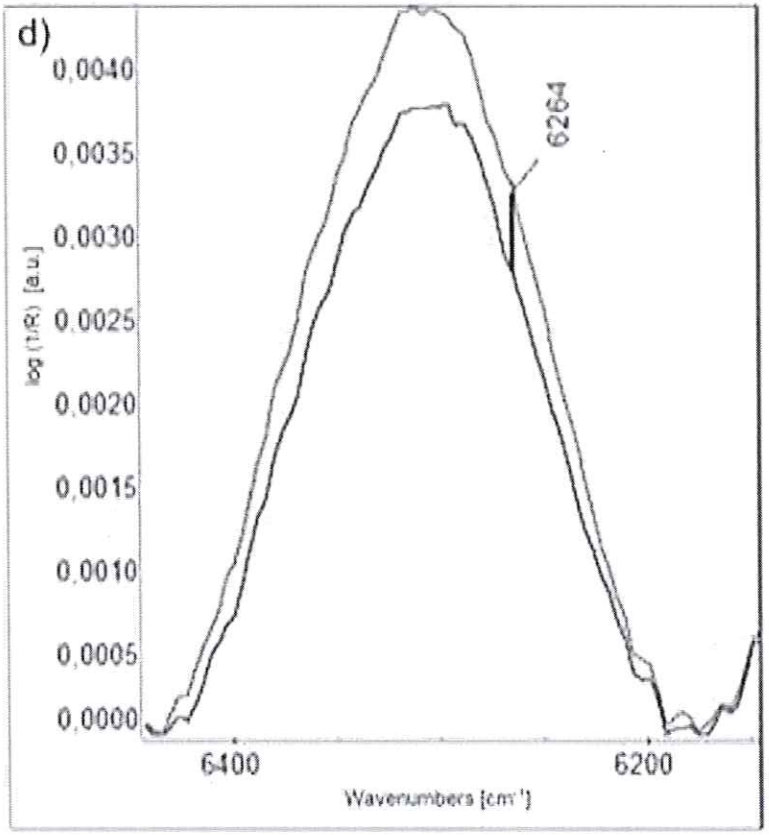


ge

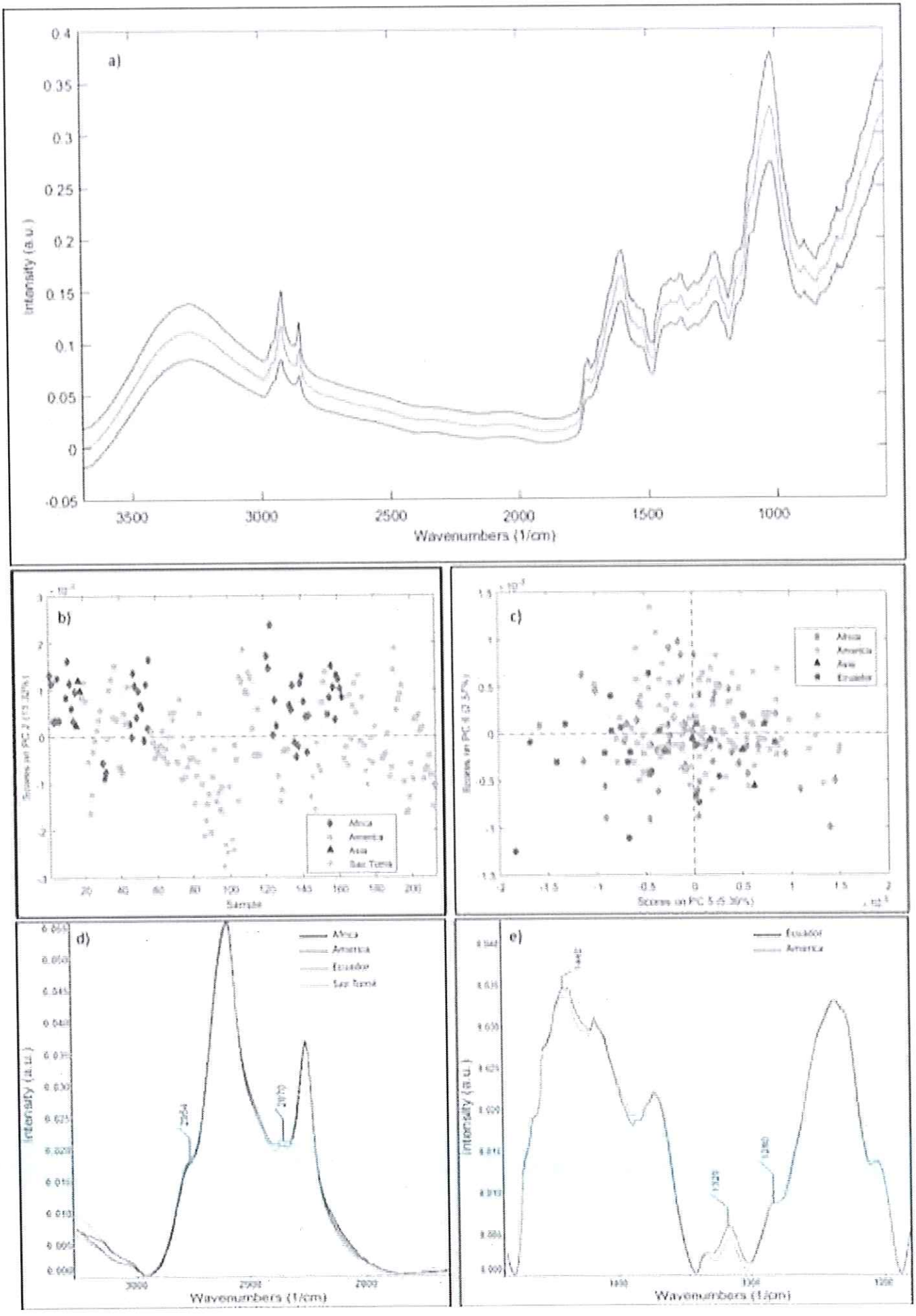

Scores on PC1 (56,47 %)


b)
Africa
America
Asia
Scores on PC6 (1,53 %)
(2,28 %)
Scores on PC4 (3.30 %)
Scores on PC5

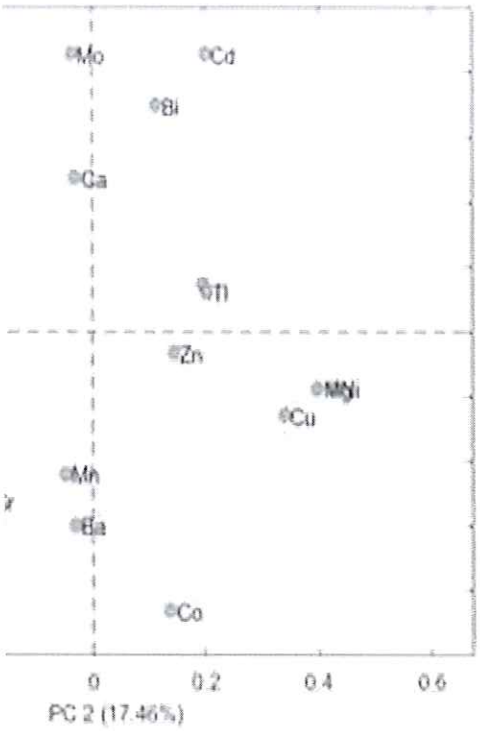

erage spectrum
asian samples
erage
all samples
4530
4550
4500
Wavenumbers [cm⁻¹]

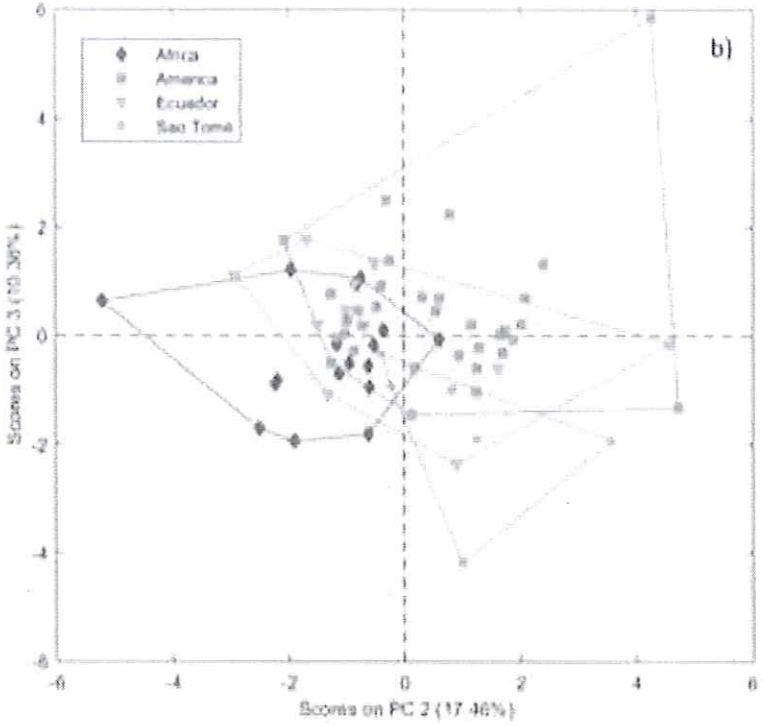

d)
0,0040
0,0035
0,0030
0,0025
0,0020
0,0015
0,0010
0,0005
0,0000
log (1/R) [a.u.]
6264
6400
6200
Wavenumbers [cm⁻¹]

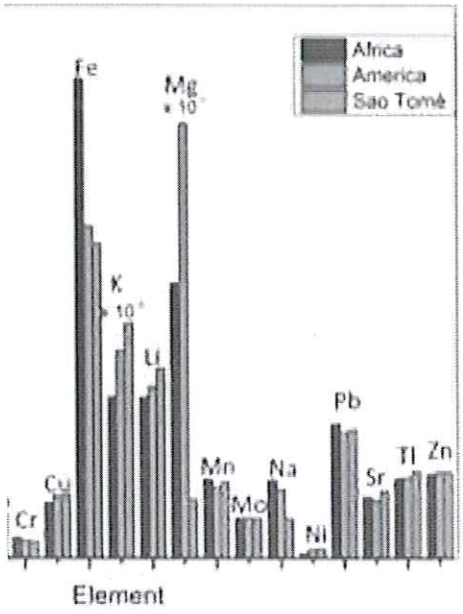
lick here to download high resolution image

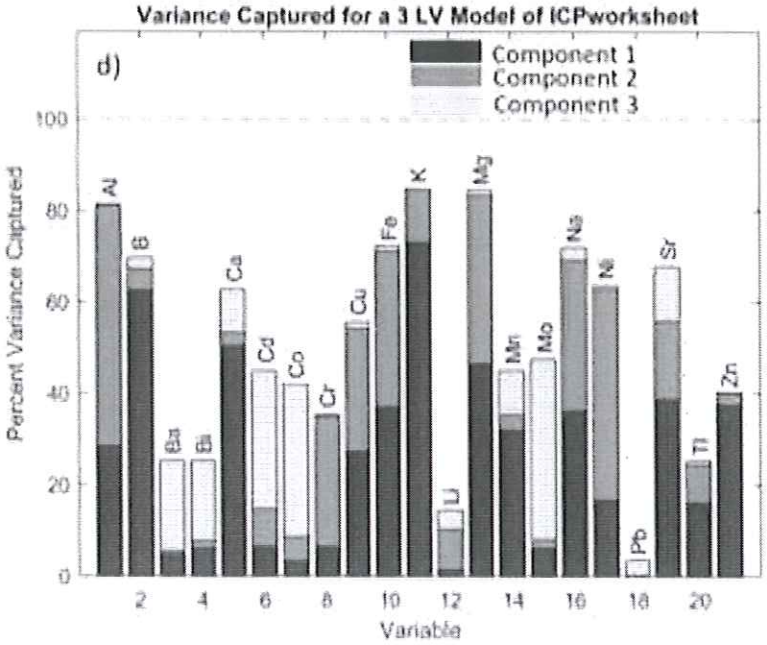


ge

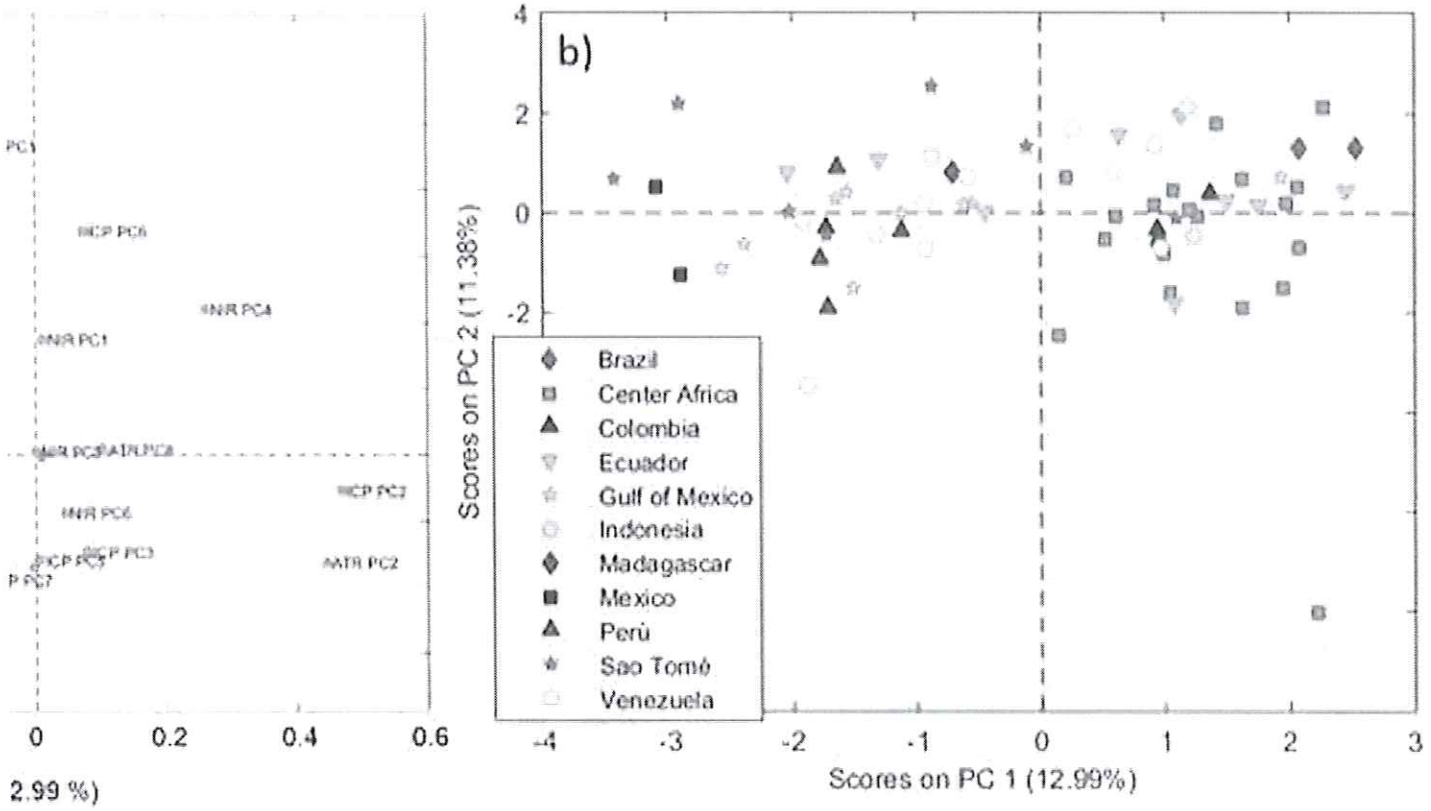

Mo
Cd
Bi
Ga
Tl
Zn
Mg
Cu
Mn
Ba
Co
0
0.2
0.4
0.6
PC 2 (17.46%)


b)
Africa
America
Ecuador
Sao Tome
Scores on PC 3 (10.36%)
Scores on PC 2 (17.46%)


Africa
America
Sao Tomé
Fe
Mg
K
Li
Pb
Cr
Cu
Mn
Mo
Na
Ni
Sr
Tl
Zn
Element


Variance Captured for a 3 LV Model of ICPworksheet
d)
Component 1
Component 2
Component 3
Percent Variance Captured
Al
B
Ba
Bi
Ca
Cd
Co
Cr
Cu
Fe
K
Li
Mg
Mn
Mo
Na
Ni
Pb
Sr
Tl
Zn
Variable

ge


b)
Brazil
Center Africa
Colombia
Ecuador
Gulf of Mexico
Indonesia
Madagascar
Mexico
Perú
Sao Tomé
Venezuela
Scores on PC 2 (11.38%)
Scores on PC 1 (12.99%)
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
0
0.2
0.4
0.6
2.99 %)