Microchemical Journal 207 (2024) 111770

Contents lists available at ScienceDirect

Microchemical Journal

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

Polyphenolic profiling of coffee beverages by Liquid Chromatography-High-Resolution mass Spectrometry for classification and characterization

Nerea Núñez [a,b,*], Javier Saurina [a,b], Oscar Núñez [a,b,c]

[a] Department of Chemical Engineering and Analytical Chemistry, Universitat de Barcelona, Martí i Franquès 1-11, E08028 Barcelona, Spain
[b] Research Institute in Food Nutrition and Food Safety, Universitat de Barcelona, Av. Prat de la Riba 171, Edifici Recerca (Gaudí), E08921 Santa Coloma de Gramenet, Spain
[c] Serra Húnter Fellow Programme, Generalitat de Catalunya, Via Laietana 2, E08003, Barcelona, Spain

ARTICLE INFO

Keywords:
Liquid Chromatography-High-Resolution Mass Spectrometry (LC-HRMS)
Polyphenolic Profiling
Coffee Classification and Characterization
Principal Component Analysis (PCA)
Partial least squares (PLS) regression
Partial least squares-discriminant analysis (PLS-DA)

ABSTRACT

The importance of monitoring the presence of bioactive compounds as food attributes for sample classification and characterization is increasing. In this study, targeted Liquid Chromatography coupled with High-Resolution Mass Spectrometry (LC-HRMS) was employed to analyze the chemical profile of polyphenolic compounds as the source of information for the characterization and classification of 306 commercial coffee samples. Coffee holds a distinguished position as one of the most widely popular beverages globally but also one of the most easily adulterated. Regrettably, in recent times, instances of coffee adulteration have been on the rise. Consequently, implementing rigorous quality control measures for coffee becomes imperative to guarantee its quality. The results obtained in this work confirm that the proposed chemical profiles serve as excellent descriptors for sample characterization and classification through the implementation of principal component analysis (PCA) and partial least squares-discriminant analysis (PLS-DA), achieving classification rates higher than 83.3% in PLS-DA validation. Moreover, the proposed LC-HRMS polyphenolic approach was employed to identify and measure adulteration levels in coffee samples using partial least squares (PLS) regression with prediction errors below 7.8%.

1. Introduction

Coffee stands as one of the most widely consumed beverages worldwide. Its unique nature as a beverage arises from its distinctive flavor and the richness of its aroma [1,2]. It is classified within the Rubiaceae family under the Coffea genus. Although a lot of species have been identified, only two are economically significant: *Coffea Arabica* (Arabica coffee) and *Coffea Canephora* (Robusta coffee). Robusta coffee trees are more robust and resistant, requiring less specific climatic conditions for cultivation compared to Arabica counter-parts. In addition, Robusta coffee contains more antioxidant compounds and caffeine, resulting in a much more bitter taste than Arabica coffee, the latter being preferred by consumers. In fact, Robusta seeds are valued at approximately half the price of the Arabica ones [1,2].

The basic chemical composition of green coffee depends on intrinsic factors such as the botanic species, but also on extrinsic factors such as climate, cultivation practices, origin, or roasting degree, among others. The flavor of high-quality coffee can vary considerably between samples of the same species but from different origin regions. In fact, climate or soil composition are relevant because can produce changes in the sensory attributes by the presence of minerals or chemical compounds [1]. Regarding the chemical composition of green coffee, volatile compounds such as alcohols, esters, hydrocarbons, or aldehydes, and non-volatile compounds such as caffeine, carbohydrates, proteins, lipids, trigonelline or polyphenols have been reported [1,3].

Phenolic compounds are secondary plant metabolites that play an important role in the sensory and nutritional quality of fruits, vegetables and other plants. These com-pounds present an aromatic ring with one or more hydroxyl groups, and their structures can vary from simple phenolic molecules to complex polymers. The main polyphenolic classes include phenolic acids, flavonoids, tannins, lignans and stilbenes. Polyphenols are present in many food products of the Mediterranean

* Corresponding autor: Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Martí i Franquès 1-11, E08028 Barcelona, Spain.
E-mail address: nereant7@gmail.com (N. Núñez).

https://doi.org/10.1016/j.microc.2024.111770
Received 18 June 2024; Received in revised form 20 September 2024; Accepted 23 September 2024
Available online 24 September 2024

Diet such as apples, oranges, tomatoes, coffee, tea, wine and olive oil, among others. Natural polyphenols in food products provide important health benefits, especially in terms of antioxidant, anti-inflammatory, antiviral, antihypertensive or anticancer properties [3–5]. For that reason, coffee consumption is associated with health benefits such as a lower risk of type II diabe-tes, or Parkinson's and Alzheimer's diseases due to its high content on some polyphenols [3,6,7]. Hydroxycinnamic acids (e.g., caffeic, ferulic and *p*-coumaric acids) and quinic acid are particularly abundant in green coffee beans and form the so-called chlorogenic acids. In fact, the principal phenolics in coffee are caffeoylquinic acids (representing approximately 80 % of the total chlorogenic acid content), dicaffeoylquinic acids, or feruylquinic acids [1,3,8].

The roasting of green coffee beans involves chemical, physical and sensory transformations. During the process, the beans increase in volume and reduce in weight, appearing new chemical compounds while others are degraded. For example, the chlorogenic acid content in commercial roasted coffees can vary from 0.5 to 6 g/100 g (dry weight), depending on the type of processing, blend, roasting degree, and the roasting method. In addition to chlorogenic acids, other polyphenolic compounds, such as flavonoids or tannins, are also present in coffee beans [1,7].

The process of bringing harvested coffee fruits to consumers as a beverage involves a series of steps, and due to the complexity of the food chain and the involvement of various factors in both production and consumption, unfortunately, coffee adulteration is on the rise, leading to cases of food fraud. Coffee adulteration (as in any other food product) is illegal worldwide, has economic consequences, and can pose serious health problems. For that reason, determining the authenticity of food products by analytical methodologies is crucial to ensure food quality control and food safety [9,10].

The analytical process for assessing polyphenols in food samples consists of three stages: extraction, separation and characterization or quantification [5]. Common extraction techniques are percolation, decoction, heat reflux extraction, Soxhlet extraction, maceration, ultrasound-assisted extraction and supercritical fluid extraction [5,11]. Concerning the separation the most common techniques are capillary electrophoresis (CE) [12] or liquid chromatography (LC) [13]. Some of these methods are coupled to nuclear magnetic resonance (NMR) [14,15], mass spectrometry (MS) [16–19] or ultraviolet (UV) detection [20–22]. For coffee analysis specifically, LC-UV is one of the most popular techniques [20–22]. For example, Craig et al. [20] established an HPLC-UV method for a rapid quantification of seven chlorogenic acid isomers in Arabic green coffee. Moreover, Angeloni et al. [17] developed an HPLC-MS/MS method for the quantification of lignans in 100 % arabica espresso from five different geographical origins. In another work, LC-MS/MS was employed for the quantification of 30 bioactive compounds (including some polyphenols) in Arabica coffees [19].

In this study, a targeted Liquid Chromatography-High-Resolution Mass Spectrometry (LC-HRMS) method was employed for polyphenolic profiling to address the classification and characterization of coffee based on origin, variety, and roasting degree. High-resolution spectrometry was chosen to ensure the unequivocal identification of specific polyphenolsto propose potential markers for coffee authentication. Although this technology is often considered costly and alternative more economical methods have proven effective for coffee classification, the use of high-resolution is crucial at this exploratory stage. The technique allows precise confirmation of compound structures which is essential before validating these markers in future studies using more accessible technologies, such as low-resolution mass spectrometry or liquid chromatography coupled to UV detection.

A total of 306 coffee samples, distributed into three sets, were analyzed with the proposed methodology after brewing and filtration. Polyphenolic profiling was performed and a custom accurate mass database containing 26 polyphenolic compounds previously characterized. For coffee characterization and classification, the obtained LC-HRMS polyphenolic profiles, based on peak area signals, were used as chemical information in Principal Component Analysis (PCA) and Partial Least Squares-Discriminant Analysis (PLS-DA) chemometric methods. The resulting scores and loadings plots provided key chemical information regarding sample patterns and relevant descriptors. Furthermore, the potential use of polyphenols for detecting and quantifying adulteration levels in coffee samples was evaluated through Partial Least Squares (PLS) regression.

2. Materials and methods

2.1. Chemicals and solutions

The mobile phase for the chromatographic method was composed of HPLC grade methanol obtained from PanReac AppliChem (Barcelona, Spain), formic acid (≥98 %) from Sigma-Aldrich (St Louis, MO, USA) and purified water with an Elix 3 Milli-Q purification system from Millipore Corporation (Burlington, MA, USA). Mineral water obtained from Eroski supermarket (Elorrio, Spain) was employed for coffee brewing.

Standard solutions of polyphenols were prepared at 1000 mg·L^{-1} in methanol and diluted to 15 mg·L^{-1} for LC-HRMS analysis. The polyphenolic profiling was based on the following compounds (all of them from Sigma-Aldrich and with a purity > 96 %): quinic acid, pyrogallol, gallic acid, 4-vinylguaiacol, arbutin, 2,5-dihydroxybenzoic, homogentisic acid, pyrocatechol, 4-hydroxybenzoic acid, 4-O-caffeoylquinic acid, chlorogenic acid, caffeic acid, vanillin, 4-methylcatechol, syringaldehyde, ethyl gallate, 3-methylcatechol, *p*-coumaric acid, sinapic acid, ferulic acid, 4,5-di-O-caffeoylquinic acid, 4-ethylcatechol, polydatin, 3,4-di-O-caffeoylquinic acid, 3,4-dihydroxybenzaldehyde and quercetin. The information about chemical structure, molecular formula and CAS number of polyphenols is in Table S1 (supplementary material).

2.2. Instrumentation

A Dionex UHPLC instrument (Thermo Fisher Scientific, San José, CA, USA) equipped with a binary pump and an autosampler coupled to a linear ion-trap (LTQ)-Orbitrap Velos HRMS instrument (Thermo Fisher Scientific) with an electrospray ionization source (ESI) in negative ion mode was employed to analyze the coffee samples. Chromatographic separation was performed in reversed-phase mode with a Kinetex® C18 (100 mm × 4.6 mm, 2.6 μm partially porous particle size) column from Phenomenex (Torrance, CA, USA), kept at room temperature. Mobile phase components were water with 0.1 % formic acid (solvent A) and methanol (solvent B), and the flow rate was 0.4 mL·min^{-1}. The elution program started increasing the methanol percentage in a linear gradient from 3 to 75 % in 30 min; from 30 to 32 min, methanol increased from 75 % to 95 % and was kept at 95 % methanol for 2 min; from 34 to 34.2, the elution program came back to the mobile phase initial conditions (3 % of methanol); finally, the column was equilibrated from 34.2 to 40 min at 3 % methanol. The injection volume used in full-loop mode was 5 μL. For acquisition, ESI source operated in negative ionization mode. Sheath, sweep and auxiliary gases were nitrogen, with a purity higher than 99,98 %, at flow rates of 60, 0 and 10 a.u. (arbitrary units), respectively. The capillary and ESI ionization source temperatures were 350 °C and 25 °C, respectively, and an S-Lens RF level of 50 V was employed. HRMS acquisition was performed in full scan mode from 100 to 1,500 *m/z* at 60,000 full width at half-maximum (FWHM, at *m/z* 200) resolution. An automatic gain control (AGC) of 1 × 106, and a maximum injection time (IT) of 200 ms were also employed. A commercially available calibration solution (Thermo Fisher Scientific) was employed for tune and calibration of the linear ion-trap (LTQ)-Orbitrap Velos HRMS instrument.

2.3. Samples

A total of 306 commercially available coffees (described in Table S2), grouped in three different sets, were analyzed (each set was analyzed and evaluated individually). Sets 1 and 2 comprised a total of 240 commercially available Nespresso® coffee samples purchased from supermarkets in Barcelona (Spain), differing in region of origin, coffee variety (Arabica, Robusta or blends), and roasting degree. In addition, to address the applicability of the proposed methodology for the classification and characterization of coffees produced in nearby countries, set 3 (containing 66 samples) consisted of Vietnam and Cambodia from local supermarkets. Set 3 samples were classified into 5 groups depending on the coffee variety and the region of origin (no information regarding the roasting degree was available). For sets 1 and 2, each sample used is an individual coffee capsule, and these capsules belonged to two different packages. For set 3, there were triplicate samples from the same package as well as triplicates of each sample type from different packages.

A Quality Control (QC) solution was also prepared for every sample set by mixing 50 μL of each sample extract to evaluate the repeatability of the proposed targeted LC-HRMS method and the robustness of the chemometric results, and was injected every ten samples (always behind a Milli-Q water blank) throughout the sequence.

Some adulteration studies were designed using coffee samples belonging to the third set as follows. Three adulteration cases were studied: Vietnam-Arabica vs. Vietnam-Robusta, Vietnam-Arabica vs. Cambodia, and Vietnam-Robusta vs. Cambodia. In any case, the calibration set was composed of mixtures at 20, 40, 60 and 80 % adulteration levels, as well as the corresponding 100 % pure coffee of each class. The validation set included 15, 25, 50, 75 and 85 % adulteration levels. Each blended adulteration level was prepared by quintuplicate, thus resulting in 55 sample extracts for each case under study. Besides, an additional adulterated sample at a 50 % level was employed as the QC solution.

2.4. Sample treatment

Coffees were analyzed without any sample treatment aside from brewing with mineral water. The brewing process for sets 1 and 2 was performed with an espresso machine (Nespresso), always using the same brewing time to reach the same final volume. For set 3, coffees were brewed with an Italian coffee maker, grinding coffee beans when necessary; in this case, ca. 40 g of ground coffee well compressed in the Italian coffee maker and 400 mL of the mineral water were employed. All samples were filtered with 0.45 μm nylon filters (Phenomenex, Alcobendas, Spain) into 2 mL glass vials, which were stored at −4 °C until LC-HRMS analysis.

2.5. Data analysis

Coffee samples were randomly analyzed with the proposed targeted LC-HRMS method. LC-HRMS raw chromatographic data were processed by the TraceFinderTM v3.3 software (Thermo Fisher Scientific) with a user-targeted accurate mass database comprising 26 phenolics. Confirmation criteria, such as chromatographic retention times, accurate mass errors (values below 5 ppm), and isotopic pattern (matches higher than 85 %), were considered to assess the presence of the chemicals. The Polyphenolic Profiles were used to build the different data matrices for PCA, PLS-DA and PLS regression under SOLO 8.6 chemometric software from Eigenvector Research (Manson, WA, USA) [23]. Details of the theoretical foundation of these statistical methodologies are discussed elsewhere [24]. For more information on the concepts and foundations of food classification and authentication and the introduction to the most representative chemometric methods, see the review by Rodionova et al. [25].

X-data matrices to be treated by PCA and PLS-DA consisted of the peak area values of the detected compounds. In each case, a normalization pretreatment, according to the QCs, regarding the overall concentration of the analyte was applied to ensure similar weights to all samples. The Y-data matrix in the PLS-DA models categorized each coffee sample into its respective class. The Y-data matrix for the PLS regression included the blended adulteration levels. The scatter plots of scores and loadings from principal components (PCs) or latent variables were used to study the distribution of samples and compounds. Thus, information regarding correlations and dependences for the targeted polyphenolics with the coffee beverages analyzed was visualized. The optimal number of LVs for PLS-DA and PLS was estimated from the first significant minimum point of the cross-validation (CV) error from a Venetian blind strategy. In addition, the applicability of PLS-DA was proved by validating with an independent prediction set. For this purpose, PLS-DA models were built with 70 % of the sample group as the training set, while the remaining 30 % constituted the prediction set, employing a random selection methodology using Excel algorithm to divide the data into training and validation set. Regarding PLS, models were validated based on the prediction sets described in section 2.3.

3. Results and discussion

3.1. HRMS characterization of targeted polyphenolic compounds

In this work, a total of 26 polyphenolic standards (Table S3) belonging to different families (phenolic acids, flavonoids, stilbenes, and other phenolics), typically reported in coffee beverages [4,5], were characterized by reversed-phase chromatography using a C18 column under gradient elution conditions (see section 2.2) and using acidified water (0.1 % formic acid) and methanol as mobile phase components. These polyphenolic compounds were characterized by HRMS to build a home-made accurate mass database of LC-HRMS Polyphenolic Profiles. For that purpose, the 26 targeted compounds were grouped in several standard mixture solutions (always preventing isobaric compounds), and analyzed with the proposed LC-HRMS method in negative ESI mode. Chromatographic retention time and HRMS spectra at a resolution of 60,000 FWHM were registered, and the obtained data is summarized in Table S3.

3.2. LC-HRMS polyphenolic profiling of coffee samples

The main objective of the present work is to evaluate if polyphenolic profiles, obtained from LC-HRMS raw data using an accurate-mass database of 26 polyphenolic compounds, resulted in good sample chemical descriptors to address the classification and characterization of coffee according to several coffee attributes such as the country of production, variety or roasting degree.

Coffee samples distributed in different sets (see Table S2) were analyzed with the proposed LC-HRMS method. As an example, Fig. 1 shows the LC-HRMS total ion chromatograms (TICs), as well as the extracted ion chromatogram for 4-hydroxybenzoic acid (m/z 137.0241, retention time, RT, 12.31 min) of two selected coffee samples belonging two different varieties (Arabica and Robusta) from set 1. Fig. 1 shows remarkable differences in signal profiles and relative abundances according to the variety. In the supplementary material, as an example, the extracted ion chromatograms for some of the most representative polyphenols in an Arabica coffee sample from Ethiopia (set 2) are shown (Figure S7).

The coffee samples and the corresponding QCs were analyzed to generate the polyphenolic profiles as explained in Section 3.1. To simplify the data, a threshold signal of 1.0×10^5 (peak area) was set in the screening software to consider that a compound may be relevant in a given sample. Besides, accurate mass measurements (with mass errors lower than 5 ppm) and isotopic pattern matches (higher than 85 %) were established as confirmation parameters. A report for each analyzed sample and QC depicting the peak areas of all the targeted compounds

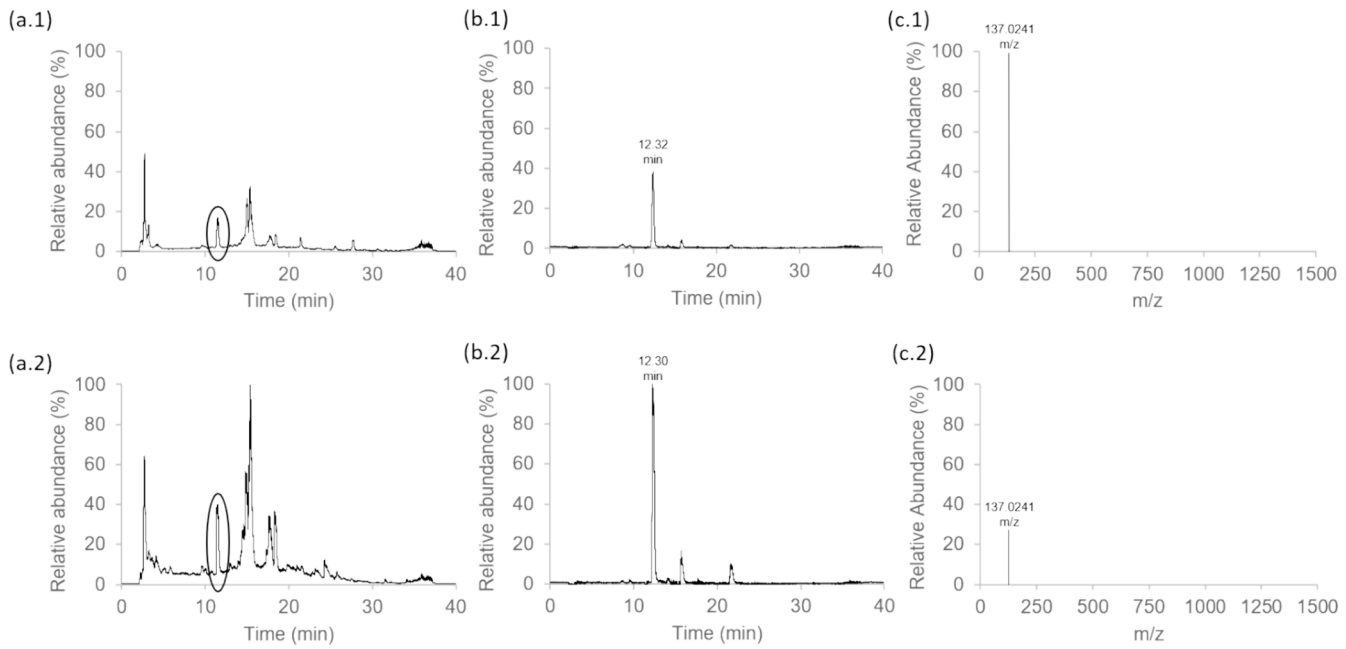

Fig. 1. LC-HRMS (a) total ion chromatograms, (b) extracted ion chromatograms and (c) extracted ion mass spectrum for 4-Hydroxybenzoic for (1) an Arabica coffee from Ethiopia (*m/z* 137.0241, retention time, RT, 12.32 min) and (2) a Robusta coffee from Uganda (*m/z* 137.0241, retention time, RT, 12.30 min).

found with the established confirmation criteria was provided. The list of reported compounds found in the samples includes both targeted polyphenols as well as possible polyphenolic signals coming from polyphenol derivatives. As an example, Table S4 shows the obtained report for a *Master Origin India Coffee* sample from set 2.

As can be seen in Table S4, for this specific Indian coffee, 22 polyphenolic compounds were detected and the peak area signals were provided by the screening software. Several compounds, such as quinic acid or chlorogenic acid, depict high peak area signals compared to other compounds, which may be attributed to their higher reported concentration in coffee samples [22].

As mentioned previously, it is well known that polyphenolic aglycones often form derivatives (glycosylated polyphenols, for example). These derivatives will be chromatographically separated from their aglycones, but fragmentation in the electrospray source will generate the aglycon as a fragment. Hence, as illustrated in Table S4, in some cases the screening software provides matches for a given compound at different retention times than the one corresponding to the standard (data in Table S3). This is the case, for instance, of 2,5-dihydroxybenzoic acid derivative found in the depicted sample data at a retention time of 16.39 min while 2,5-dihydroxybenzoic acid standard elutes at a retention time of 10.94 min under the employed chromatographic separation conditions. In addition, isomeric compounds may also be detected. In any case, as the main objective of the present contribution is the classification and characterization of coffee samples by employing polyphenolic data, but not a deep characterization of the polyphenols found in the analyzed samples, the targeted LC-HRMS profiles that will be used as sample chemical descriptors will be built based on the accurate *m/z* values detected by the screening software at the different retention times (independently if it was the aglycone or one of their derivatives).

3.3. Exploratory principal Component analysis (PCA)

The capability of targeted LC-HRMS Polyphenolic Profiles as chemical markers for categorizing coffee samples according to geographical origin, variety and roasting degree was first evaluated by PCA. In this work, PCA was applied as an exploratory technique to identify patterns in multivariate data, providing an initial insight into the underlying structure of the samples. Consistent with its use in methods like Soft Independent Modeling of Class Analogy (SIMCA), PCA was essential for dimensionality reduction and for highlighting the main trends within the data [25].

First, data matrices (X-data) were built with the polyphenolic peak areas at a specific *m/z* value and retention time for those polyphenols detected in the analyzed samples and QCs. In addition, an autoscaling preprocessing was used to guarantee equal weighting for all variables. The PCA score plots showed that the QCs displayed a linear distribution trend instead of appearing grouped. This QC distribution was related to their injection or-der in the sequence, suggesting a drift in the LC-HRMS polyphenolic signals across the sample sequence. Indeed, a decrease in the QC signal was observed throughout the series, likely attributable to the decay of electrospray ionization performance since the source became dirty during the analysis of the samples. Thus, for the correct interpretation of the results, the X data matrix was corrected with respect to the QCs. Hence, the intensity areas of the compounds in the samples were divided by the corresponding areas of the nearest QC to guarantee a reliable chemometric result interpretation on the classification and characterization studies.

Fig. 2 shows the PCA score plots obtained when using the corrected targeted LC-HRMS Polyphenolic Profiles for set 1 of coffee samples, displayed by labelling the samples according to the coffee variety (Fig. 2. a), to geographical production regions (Fig. 2.b) and to the roasting degree (Fig. 2.c). Similar PCA information is provided for sets 2 and 3 in Figures S1 and S2 (supplementary material), respectively. The classification related to the coffee roasting degree was not examined for sample set 3 due to the lack of information regarding this attribute.

As shown in Fig. 2, samples tend to be clustered based on the coffee attribute under study (geographical production region, variety, and roasting degree). Similar information is shown in the supplementary material for the sets of coffee samples number 2 (Figure S1), and 3 (Figure S2), respectively. These results confirm the potential of the data to ad-dress classification and characterization studies detailed in the following sections.

3.4. Supervised partial least-squares-discriminant analysis

PLS-DA was employed to establish clear boundaries between the different sample classes. Although widely used in food chemistry, this

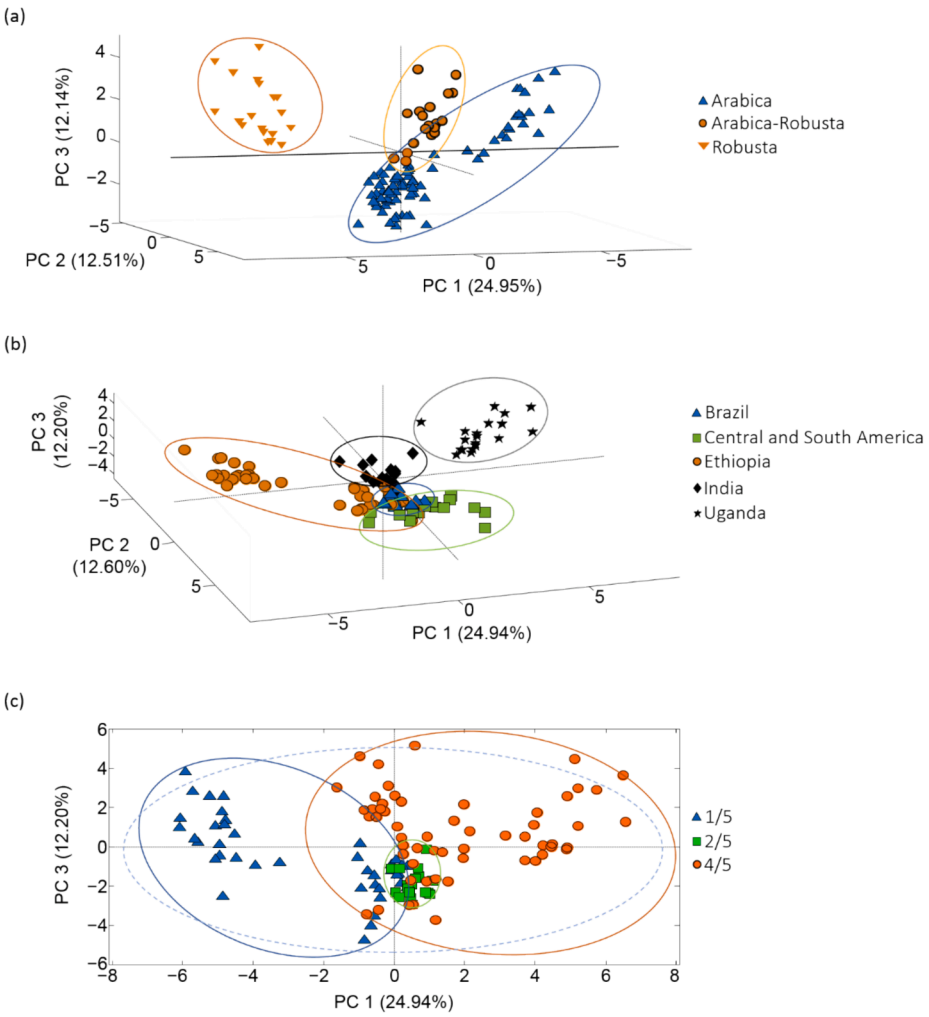

Fig. 2. PCA score plots obtained when corrected targeted LC-HRMS Polyphenolic Profiles were used as sample chemical descriptors to study coffee samples of set 1 according to (a) the variety (score plot of PC1 vs. PC2 vs. PC3), (b) geographical production region (score plot of PC1 vs. PC2 vs. PC3) and (c) the roasting degree (score plot of PC1 vs. PC3).

method should be applied with caution to avoid potential misinterpretations in authentication studies [25]. In our case, PLS-DA efficiently enabled the identification of distinguishing features between the sample groups. The corrected targeted LC-HRMS Polyphenolic Profiles obtained for each set of coffee samples were subjected to a supervised PLS-DA. Coffee samples were categorized based on the three evaluated attributes: the geographical production region, the coffee variety, and the roasting degree; thus, Y-data matrices were designed according to each attribute under study.

Fig. 3 shows the PLS-DA score plots and their respective loading plots obtained for set 1. Figure S3 and Figure S4 (supplementary material) show the equivalent information obtained for the coffee sets 2 and 3, respectively. Furthermore, the values of sensitivities, specificities and class prediction errors by cross-validation are shown in Table S5 for all the obtained multiclass PLS-DA models.

As shown in Fig. 3.a.1, and Figures S3.a.1 and S4.a.1 (supplementary material), the sample classification based on the coffee varieties was excellent for all the coffee sample sets under study. Regardless of the geographical origin of the samples, the differentiation among varieties (Arabica, Robusta, or blended Arabica-Robusta varieties) was perfect, achieving sensitivity and specificity values of 100 % (Table S5) and, consequently, 100 % classification performance.

Furthermore, for the classification of coffee samples according to their geographical production region, good results were also obtained for all analyzed sets of samples (Fig. 3.b.1, S3.b.1 and S4.b.1). As shown in Table S5, for instance, in the case of set 1, satisfactory results were achieved with sensitivity and specificity values higher than 85 % and 95.9 %, respectively, and classification errors lower than 8.8 %. For set 2, sensitivity values of 100 % were obtained for the geographical production regions, and specificity values higher than 88.5 %. In the case of set 3, sensitivity and specificity values of 100 % were achieved, highlighting the capability of the Polyphenolic Profiles for geographical classification and characterization even when coffees are produced under similar climatic conditions.

Finally, the classification of coffee samples based on the roasting degree for sample sets 1 and 2 was also satisfactory, as shown in Fig. 3.c.1 and S3.c.1, with sensitivity and specificity values higher than 90 % and 91 %, respectively, and classification errors values below 7 %.

By studying the obtained PLS-DA loading plots in Fig. 3.a.2, and Figures S3.a.2 and S4.a.2 (supplementary material) it can be seen that Arabica coffee samples are mostly defined by 3,4-di-O-caffeoylquinic acid, and chlorogenic acid, while Robusta coffee samples (or samples with a blended percentage of Robusta coffee), are richer in 4,5-di-O-caffeoylquinic acid, pyrocatechol, 4-ethylcatechol polyphenol derivative or syringaldehyde. These compounds contribute significantly to the coffee classification in agreement with the results described by Król et al. [26] and Bhagat et al. [27].

Regarding the geographical production region, for set 1, as depicted in Fig. 3.b.2, coffees originated from India seem to be more defined by pyrogallol, gallic acid and ethyl gallate. In contrast, 4-vinylguaiacol

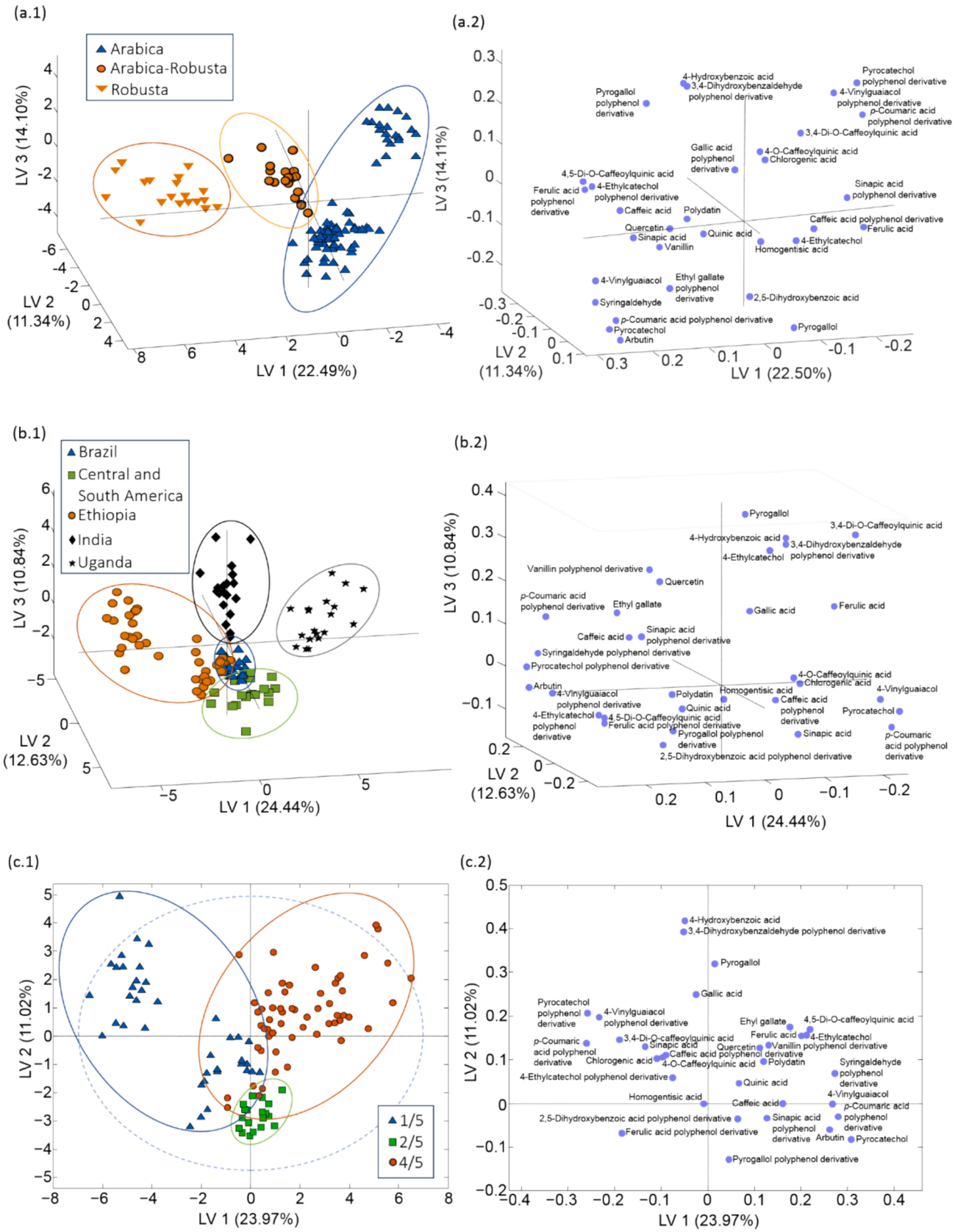

Fig. 3. PLS-DA score and loading plots obtained when corrected targeted LC-HRMS Polyphenolic Profiles were used as sample chemical descriptors to study set 1 according to (a.1) the variety (score plot of LV1 vs. LV2 vs. LV3), (b.1) the geographical production region (score plot of LV1 vs. LV2 vs. LV3) and (c.1) the roasting degree (score plot of LV1 vs. LV2).

seem to be more characteristic of coffees produced in Uganda, while those from Ethiopia are defined by pyrocatechol polyphenol derivative. Finally, samples from Brazil and Central and South America are predominantly characterized by 4-vinylguaiacol. In the case of set 2 (Figure S3.b.2), Ethiopian samples are predominantly defined by sinapic acid or caffeic acid (as with set 1). In contrast, in the case of Indian coffees, also present in set 1, the polyphenols that contribute more to its sample discrimination and classification is 4-vinylguaiacol polyphenol derivative. Homogentisic acid polyphenol derivative seems to play a significant role in the classification of Colombian samples. Finally, 4,5-Di-O-caffeoylquinic acid seem to be a discriminant compounds for Indonesian coffee samples, while 4-O-caffeoylquinic acid for Nicaraguan coffees. The results agree with those described by Angeloni et al. [19] for the quantification of 30 bioactive compounds in Arabica coffee samples from Ethiopia, Brazil, India, Colombia and Costa Rica, and by Craig et al. [20] in Indian coffees. Regarding set 3 (Figure S4.b.2), *p*-coumaric acid stands out as the most characteristic polyphenol for Cambodian coffee samples. Meanwhile, the other employed polyphenols contribute fairly similarly to describing Vietnamese coffee samples, with notable contributions from sinapic acid, ethyl gallate, or 3,4-Di-O-caffeoylquinic acid.

Finally, regarding the roasting degree (data depicted in Fig. 3.c.2 and Figure S3.c.2 of the supplementary material), the samples with 1/5 roasting degree (the less roasted) are mostly defined by pyrocatechol and *p*-coumaric acid polyphenol derivatives. The samples with 2/5 roasting degree are classified thanks to the higher contribution of polyphenols such as sinapic acid polyphenol derivative or caffeic acid. Homogentisic acid and 3,4-dihydroxybenzaldehyde polyphenol derivatives seem to define the roasting degree of 3/5. In contrast, the coffees with a 4/5 roasting degree, are defined, in both sets of samples, by polyphenols such as gallic acid polyphenol derivative, 4,5-di-O-caffeoylquinic acid or 4-ethylcatechol. Finally, ferulic acid polyphenol derivative is a phenolic acid clearly characteristic of the most roasted coffees (5/5 roasting degree). These results agree with Król et al. [26] in their analysis of arabica coffee samples with different roasting levels.

3.5. PLS-DA validation

The feasibility of the proposed methodology for classifying coffees based on the coffee region of origin, variety and roasting degree was also validated. For this purpose, PLS-DA paired models were considered to determine classification rates when comparing a single sample class against all others. Each paired PLS-DA model examined was built using 70 % of samples randomly selected for each group as the training set while the remaining 30 % of the samples were employed as the prediction set.

Table 1 summarizes the optimal number of LVs, as well as the sensitivity, specificity and classification error values achieved for both training and prediction steps for each paired classification model evaluated. In addition, Fig. 4 shows the paired PLS-DA score plots of Y-predicted vs. sample obtained for the three sets of samples when validation based on the coffee varieties was addressed. Similar information is shown in Figures S5 and S6 (supplementary material) when validation based on the geographical production region and coffee roasting degree, respectively, was performed.

Validation results of the classification of coffee samples by paired PLS-DA models (Table 1) are very satisfactory. When addressing the classification of coffees based on their variety, sensitivity and specificity values of 100 % for both training and prediction were obtained. In the case of the geographical production region, sensitivity and specificity values higher than 93.5 % and 94.4 %, respectively, for training, and higher than 83.3 % and 88 %, respectively, for prediction, were accomplished. Finally, for the classification of coffees based on the roasting degree, sensitivity and specificity values were higher than 92.9 % and 92.3 %, respectively, for training, and higher than 83.3 % and

Table 1
LVs, and sensitivity, specificity and classification error values obtained for training and prediction on paired PLS-DA models when studying the classifications of the analyzed coffee samples according to their geographical production region, variety and roasting degree.

			Training			Prediction		
	LVs	Class	Sensitivity (%)	Specificity (%)	Classification Error (%)	Sensitivity (%)	Specificity (%)	Classification Error (%)
Coffee variety								
Set 1	2	Arabica	100	100	0	100	100	0
	2	Arabica-Robusta mixture	100	100	0	100	100	0
	2	Robusta	100	100	0	100	100	0
Set 2	2	Arabica	100	100	0	100	100	0
	2	Arabica-Robusta mixture	100	100	0	100	100	0
Set 3	3	Arabica	100	100	0	100	100	0
	3	Robusta	100	100	0	100	100	0
Coffee geographical production region								
Set 1	3	Brazil	100	100	0	100	90.6	4.7
	2	Central and South America	100	100	0	100	96.7	1.4
	3	Ethiopia	93.5	96.5	5	90	88	11
	4	India	100	100	0	100	100	0
	3	Uganda	100	100	0	100	100	0
Set 2	2	Colombia	100	100	0	100	100	0
	4	Ethiopia	100	100	0	100	100	0
	2	India	100	100	0	100	100	0
	4	Indonesia	100	100	0	100	100	0
	3	Nicaragua	100	94.4	2.8	83.3	92	12.3
Set 3	3	Cambodia	100	100	0	100	100	0
	3	Vietnam	100	100	0	100	100	0
Coffee roasting degree								
Set 1	3	1/5	96.8	94.9	4.2	87.5	95.5	8.5
	5	2/5	100	98.7	0.7	100	96.7	1.7
	4	4/5	92.9	97.7	4.7	90	100	5
Set 2	3	2/5	92.9	96.2	5.5	83.3	100	8.3
	2	3/5	100	100	0	100	100	0
	2	4/5	100	92.3	3.8	90.9	96	6.5
	2	5/5	100	100	0	100	100	0

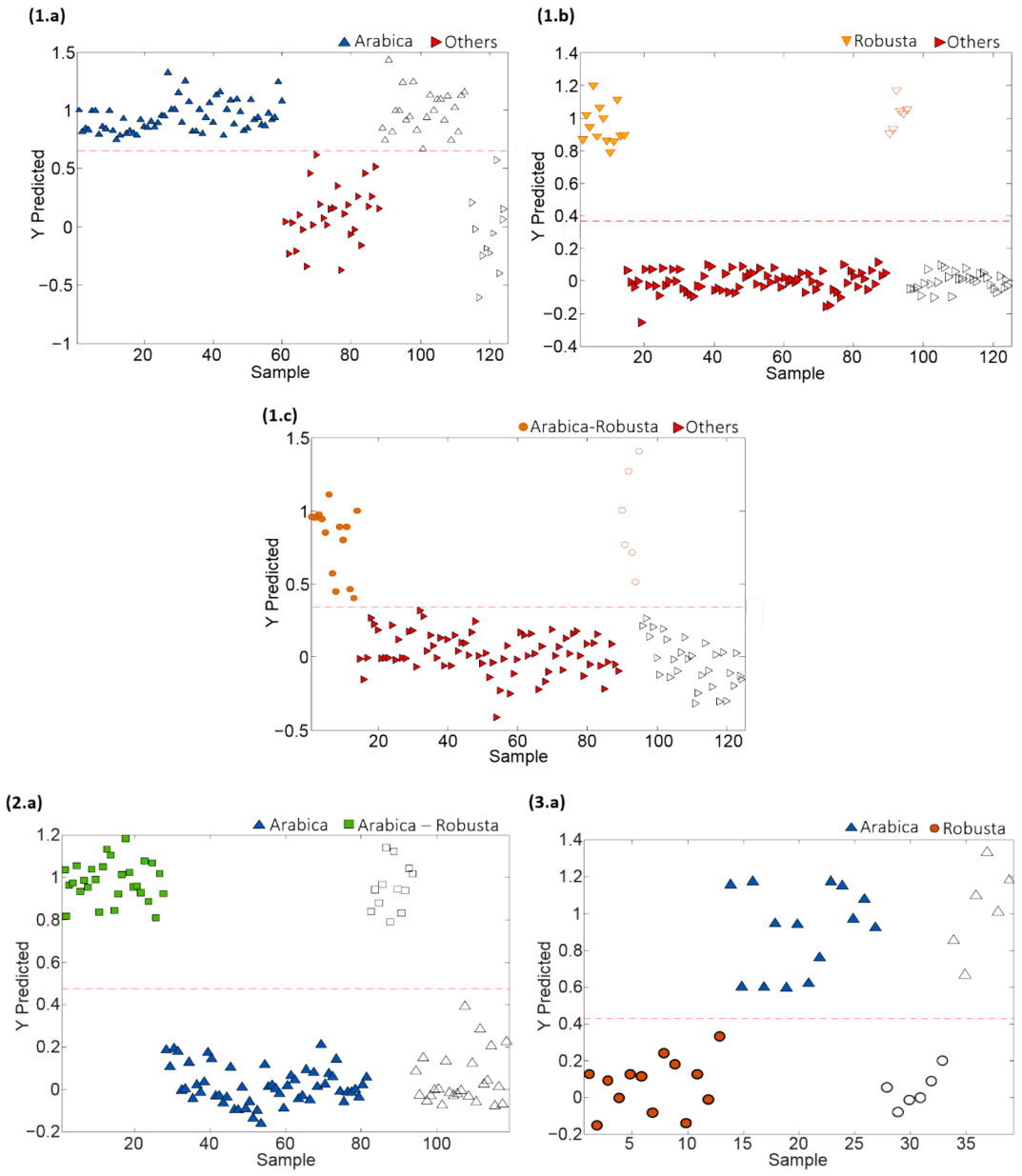

Fig. 4. Paired PLS-DA score plots of Y predicted vs. samples according to the coffee variety for set 1: (1.a) Arabica vs. Others, (1.b) Robusta vs. Others, (1.c) Arabica–Robusta mixture vs. Others; for set 2: (2.a) Arabica vs. Arabica–Robusta mixture; and for set 3: (3.a) Arabica vs. Robusta. Filled and empty symbols correspond to training and prediction sets, respectively. Red lines represent the threshold between classes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

95.5 %, respectively, for prediction.

Classification errors were always lower than 5.5 % for paired PLS-DA training models, and for prediction, values lower than 12.7 % were obtained, which are quite acceptable considering the complexity of the analyzed samples, although it should be high-lighted that 100 % classification rates were accomplished in most of the cases.

The obtained results demonstrate that the proposed LC-HRMS Polyphenolic Profiling of 26 compounds seems to be suitable sample chemical descriptors for the classification and characterization of coffees based on their variety, geographical production region, and roasting degree.

3.6. Detection and quantitation of coffee adulteration by partial least squares regression

The proposed methodology was applied to detect and quantify adulterant percentage levels in blended coffee samples using PLS regression. Three coffee adulteration cases were studied: (i) Vietnamese Robusta coffee adulterated with Cambodian coffee, (ii) Vietnamese Arabica coffee adulterated with Cambodian coffee and (iii) Vietnamese Arabica coffee adulterated with Vietnamese Robusta coffee. The selection of these three cases was based on the proximity of the coffee-growing geographical regions, with similar climatic conditions.

For each adulteration case, independent calibration and validation sets were employed. The calibration sets contained 0 %, 20 %, 40 %, 60 %, 80 % and 100 % adulteration levels, while the validation sets included 15 %, 25 %, 50 %, 75 % and 85 % adulteration levels. Each adulteration level was prepared in quintuplicate, resulting in 55 sample extracts for each case under study. Furthermore, a quality control solution was at a 50 % adulteration level. PLS results for the three adulteration cases are summarized in Table 2. An ex-ample of the PLS predictive performance for Vietnamese Robusta coffee adulterated with

Table 2
Evaluation of the coffee adulteration cases by PLS using corrected targeted LC-HRMS Polyphenolic Profiles as sample chemical descriptors.

	Vietnamese Arabica Coffee adulterated with Cambodian coffee	Vietnamese Robusta Coffee adulterated with Cambodian coffee	Vietnamese Arabica Coffee adulterated with Vietnamese Robusta Coffee
LVs	4	4	3
R^2	0.993	0.983	0.992
Calibration Errors (%)	2.98	4.52	3.08
Prediction Errors (%)	7.78	6.72	6.00

Cambodian coffee is shown in Fig. 5.

As can be seen, PLS was very satisfactory for all the cases studied, with calibration and prediction errors below 4.52 % and 7.78 %, respectively, and correlations higher than 0.983. Compared with the non-targeted LC-HRMS fingerprint approach, prediction errors for the detection of coffee adulterant levels clearly improved with targeted LC-HRMS Polyphenolic Profiles [28]. Thus, the obtained results prove that the developed targeted LC-HRMS Polyphenolic Profile methodology is effective is effective in detecting and quantifying adulteration levels in adulterated coffees produced from nearby geographical production regions.

3.7. Comparison with other scientific publications

This comparative section has been added to analyze and contrast our results with those of previous studies in the field to provide a more comprehensive view of the context of our study. Supplementary Material includes a detailed table (Table S6) summarizing key features and findings of relevant research. This table allows for a direct comparison with other works and facilitates assessing how our results align with or differ from those reported by other researchers. The comparative table includes information on authors, publication year, study objectives, employed methodology, and comparison with the presented work.

By providing this comparison, we aim to offer a clearer understanding of the position of our work within the existing body of literature, as well as insights into how it contributes to advancing the field. The comparative analysis of coffee authentication methodologies, presented in the supplementary material, highlights a broad spectrum of analytical techniques, ranging from Near-Infrared Spectroscopy (NIR) and Gas Chromatography-Mass Spectrometry (GC–MS) to Liquid Chromatography-Mass Spectrometry (LC-MS) and High-Resolution Mass Spectrometry (LC-HRMS). Each of these methods provides valuable insights into coffee classification, variety differentiation, and origin authentication. However, there are clear distinctions in the depth of chemical information they offer, as well as their potential for future application in routine authentication protocols [29–38].

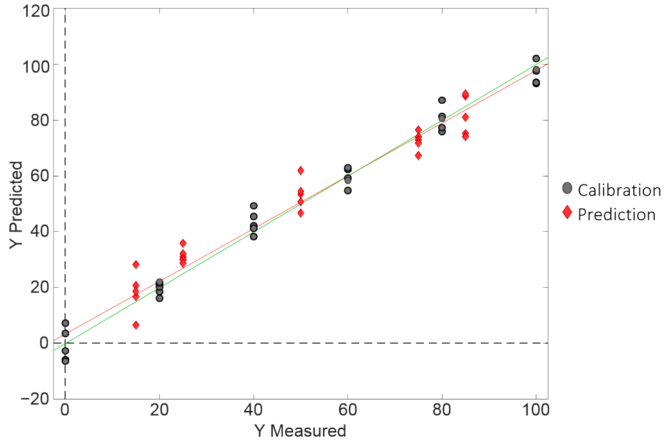

Fig. 5. PLS regression model for the case of coffee Robusta from Vietnam adulterated with coffee from Cambodia.

4. Conclusions

In this work, a LC-HRMS Polyphenolic Profiling methodology based on the screening of 26 characterized polyphenolic compounds through a user accurate mass database was developed. The Polyphenolic Profiles resulted in excellent sample chemical descriptors for the characterization, classification, and future authentication of coffee samples by PLS-DA according to different attributes (e.g., variety, geographical production region and roasting degree). PLS-DA validation models demonstrated the robustness and reliability of the proposed targeted LC-HRMS polyphenolic methodology obtaining, in general, sensitivity and specificity values for training higher than 92.9 % and 92.3 %, respectively, and for prediction higher than 83.3 % and 88 %, respectively. Overall prediction errors in the detection and quantitation levels of adulterations in coffee samples below 7.78 % were accomplished by PLS.

While alternative less costly techniques can achieve comparable classification results, the use of high-resolution mass spectrometry in this exploratory phase is essential for the accurate identification of specific polyphenols. Although this method is not intended for routine authentication, it establishes a solid foundation for the development of simpler and more accessible methods in the future. The precise identification of key polyphenols is crucial for proposing markers that can later be employed with more economical technologies in practical applications.

Finally, regarding the representativeness of the samples, the study presented has to be considered a proof of concept that explores the possibilities of the developed approach. Our work shows some compounds characteristic of different sample typologies (although there could also be others that we have not identified). In this sense, in the particular scenario described in this manuscript or in another more global one, it does seem that the profiling approach with chemometric data treatment is an excellent option for the characterization of coffee or the identification of potential fraud.

Funding

This research was supported by the project PID2020-114401RB-C22 financed by the Agencia Estatal de Investigación (AEI/10.13039/501100011033), and by the Agency for Administration of University and Research Grants (Generalitat de Catalunya, Spain) under the project 2021SGR-00365.

CRediT authorship contribution statement

Nerea Núñez: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation. **Javier Saurina:** Writing – review & editing, Supervision, Software, Investigation, Funding acquisition, Conceptualization. **Oscar Núñez:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.microc.2024.111770.

References

[1] A. Farah, Coffee Constituents, Coffee Emerg. Heal. Eff. Dis. Prev. (2012) 21–58. 10.1002/9781119949893.ch2.

[2] C. Stiefel, B. Lindemann, G.E. Morlock, Non-target bioactive compound profiles of coffee roasts and preparations, Food Chem. 391 (2022) 133263, https://doi.org/10.1016/j.foodchem.2022.133263.

[3] S. Tabrez, M.I. Khan, Polyphenols-Based Nanotherapeutics for Cancer Management (2021), https://doi.org/10.1007/978-981-16-4935-6.

[4] F. Bioactives, CharissM..Galanakiss Editor, n.d.

[5] I. Ignat, I. Volf, V.I. Popa, A critical review of methods for characterisation of polyphenolic compounds in fruits and vegetables, Food Chem. 126 (2011) 1821–1835, https://doi.org/10.1016/j.foodchem.2010.12.026.

[6] A. Ali, H.F. Zahid, J.J. Cottrell, F.R. Dunshea, A Comparative Study for Nutritional and Phytochemical Profiling of Coffea arabica (C. arabica) from Different Origins and Their Antioxidant Potential and Molecular Docking, Molecules 27 (2022), https://doi.org/10.3390/molecules27165126.

[7] T. Stoikidou, A. Koidis, Coffee and tea bioactive compounds, Elsevier Inc., 2022. 10.1016/B978-0-12-823811-0.00006-7.

[8] A. Montis, F. Souard, C. Delporte, P. Stoffelen, C. Stévigny, P. Van Antwerpen, Targeted and Untargeted Mass Spectrometry-Based Metabolomics for Chemical Profiling of Three Coffee Species, Molecules 27 (2022), https://doi.org/10.3390/molecules27103152.

[9] G. Campmajó, N. Núñez, O. Núñez, The Role of Liquid Chromatography-Mass Spectrometry in Food Integrity and Authenticity, in: Kamble, G.S. (ed.) Mass Spectrometry - Future Perceptions and Applications, in: IntechOpen, London, 2019: pp. 3–20. https://doi.org/10.5772/57353.

[10] M. Samad, K. Mohammad, S. Rahman, Techniques to Measure Food Safety and Quality (2021), https://doi.org/10.1007/978-3-030-68636-9.

[11] A. Sridhar, M. Ponnuchamy, P.S. Kumar, A. Kapoor, D.V.N. Vo, S. Prabhakar, Techniques and modeling of polyphenol extraction from food: a review, Springer International Publishing (2021), https://doi.org/10.1007/s10311-021-01217-8.

[12] A. Spisso, F.J.V. Gomez, M. Fernanda Silva, Determination of ellagic acid by capillary electrophoresis in Argentinian wines, Electrophoresis 39 (2018) 1621–1627, https://doi.org/10.1002/elps.201700487.

[13] N. Núñez, O. Vidal-Casanella, S. Sentellas, J. Saurina, O. Núñez, Characterization, classification and authentication of turmeric and curry samples by targeted LC-HRMS polyphenolic and curcuminoid profiling and chemometrics, Molecules 25 (2020) 1–16, https://doi.org/10.3390/molecules25122942.

[14] G.N. Manjunatha Reddy, L. Mannina, A.P. Sobolev, S. Caldarelli, Polyphenols Fingerprinting in Olive Oils Through Maximum-Quantum NMR Spectroscopy, Food Anal. Methods. 11 (2018) 1012–1020, https://doi.org/10.1007/s12161-017-1069-x.

[15] M. Madhava Naidu, G. Sulochanamma, S.R. Sampathu, P. Srinivas, Studies on extraction and antioxidant potential of green coffee, Food Chem. 107 (2008) 377–384, https://doi.org/10.1016/j.foodchem.2007.08.056.

[16] G. Caprioli, F.K. Nzekoue, F. Giusti, S. Vittori, G. Sagratini, Optimization of an extraction method for the simultaneous quantification of sixteen polyphenols in thirty-one pulse samples by using HPLC-MS/MS dynamic-MRM triple quadrupole, Food Chem. 266 (2018) 490–497, https://doi.org/10.1016/j.foodchem.2018.06.049.

[17] S. Angeloni, L. Navarini, G. Sagratini, E. Torregiani, S. Vittori, G. Caprioli, Development of an extraction method for the quantification of lignans in espresso coffee by using HPLC-MS/MS triple quadrupole, J. Mass Spectrom. 53 (2018) 842–848, https://doi.org/10.1002/jms.4251.

[18] Y. Sapozhnikova, Development of liquid chromatography-tandem mass spectrometry method for analysis of polyphenolic compounds in liquid samples of grape juice, green tea and coffee, Food Chem. 150 (2014) 87–93, https://doi.org/10.1016/j.foodchem.2013.10.131.

[19] S. Angeloni, F.K. Nzekoue, L. Navarini, G. Sagratini, E. Torregiani, S. Vittori, G. Caprioli, An analytical method for the simultaneous quantification of 30 bioactive compounds in spent coffee ground by HPLC-MS/MS, J. Mass Spectrom. 55 (2020), https://doi.org/10.1002/jms.4519.

[20] A.P. Craig, C. Fields, N. Liang, D. Kitts, A. Erickson, Performance review of a fast HPLC-UV method for the quantification of chlorogenic acids in green coffee bean extracts, Talanta 154 (2016) 481–485, https://doi.org/10.1016/j.talanta.2016.03.101.

[21] P. Köseoglu Yilmaz, U. Kolak, SPE-HPLC Determination of Chlorogenic and Phenolic Acids in Coffee, J. Chromatogr. Sci. 55 (2017) 712–718, https://doi.org/10.1093/chromsci/bmx025.

[22] K. Belguidoum, H. Amira-Guebailia, Y. Boulmokh, O. Houache, HPLC coupled to UV-vis detection for quantitative determination of phenolic compounds and caffeine in different brands of coffee in the Algerian market, J. Taiwan Inst. Chem. Eng. 45 (2014) 1314–1320, https://doi.org/10.1016/j.jtice.2014.03.014.

[23] Eigenvector Research Incorporated. Powerful Resources for Intelligent Data Analysis. Available online: http://www.eigenvector.com/software/solo.htm (accessed on 15 January 2019), n.d.

[24] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of chemometrics and qualimetrics, Elsevier, Amsterdam, The Netherlands, 1997.

[25] O.Y. Rodionova, P. Oliveri, C. Malegori, A.L. Pomerantsev, Chemometrics as an efficient tool for food authentication: Golden pillars for building reliable models, Trends Food Sci. Technol. 147 (2024) 104429, https://doi.org/10.1016/j.tifs.2024.104429.

[26] K. Król, M. Gantner, A. Tatarak, E. Hallmann, The content of polyphenols in coffee beans as roasting, origin and storage effect, Eur. Food Res. Technol. 246 (2020) 33–39, https://doi.org/10.1007/s00217-019-03388-9.

[27] A.R. Bhagat, A.M. Delgado, M. Issaoui, N. Chammem, M. Fiorino, A. Pellerito, S. Natalello, Review of the role of fluid dairy in delivery of polyphenolic compounds in the diet: Chocolate milk, coffee beverages, matcha green tea, and beyond, J. AOAC Int. 102 (2019) 1365–1372, https://doi.org/10.5740/jaoacint.19-0129.

[28] N. Núñez, J. Saurina, O. Núñez, Liquid Chromatography–High-Resolution Mass Spectrometry (LC-HRMS) Fingerprinting and Chemometrics for Coffee Classification and Authentication, Molecules 29 (2024), https://doi.org/10.3390/molecules29010232.

[29] N. Núñez, X. Collado, C. Martínez, J. Saurina, O. Núñez, Authentication of the Origin, Variety and Roasting Degree of Coffee Samples by Non-Targeted HPLC-UV Fingerprinting and Chemometrics, Application to the Detection and Quantitation of Adulterated Coffee Samples, Foods. 9 (2020) 378, https://doi.org/10.3390/foods9030378.

[30] N. Núñez, C. Martínez, J. Saurina, O. Núñez, High-performance liquid chromatography with fluorescence detection fingerprints as chemical descriptors to authenticate the origin, variety and roasting degree of coffee by multivariate chemometric methods, J. Sci. Food Agric. 101 (2021) 65–73, https://doi.org/10.1002/jsfa.10615.

[31] T.K.L. de Araújo, R.O. Nóbrega, D.D. de S. Fernandes, M.C.U. de Araújo, P.H.G.D. Diniz, E.C. da Silva, Non-destructive authentication of Gourmet ground roasted coffees using NIR spectroscopy and digital images, Food Chem. 364 (2021). 10.1016/j.foodchem.2021.130452.

[32] N. Núñez, J. Pons, J. Saurina, O. Núñez, Non-targeted high-performance liquid chromatography with ultraviolet and fluorescence detection fingerprinting for the classification, authentication, and fraud quantitation of instant coffee and chicory by multivariate chemometric methods, Lwt. 147 (2021), https://doi.org/10.1016/j.lwt.2021.111646.

[33] G. Galarza, J.G. Figueroa, 使用 SPME–GC–MS 表征在不同发酵时间处理的咖啡 (Coffea arabica) 的挥发性化合物, Molecules. 27 (2022).

[34] J. Klikarová, L. Česlová, Targeted and Non-Targeted HPLC Analysis of Coffee-Based, Target. Non-Targeted HPLC Anal. Coffee-Based Prod. as Eff. Tools Eval. Coffee Authent. (2022).

[35] C.H. Lee, I. Te Chen, H.C. Yang, Y.J. Chen, An AI-powered Electronic Nose System with Fingerprint Extraction for Aroma Recognition of Coffee Beans, Micromachines 13 (2022), https://doi.org/10.3390/mi13081313.

[36] A.C.R. Silva, R. Garrett, C.M. Rezende, S.W. Meckelmann, Lipid characterization of arabica and robusta coffee beans by liquid chromatography-ion mobility-mass spectrometry, J. Food Compos. Anal. 111 (2022) 104587, https://doi.org/10.1016/j.jfca.2022.104587.

[37] J.V. Robert, J.S. de Gois, R.B. Rocha, A.S. Luna, Direct solid sample analysis using synchronous fluorescence spectroscopy coupled with chemometric tools for the geographical discrimination of coffee samples, Food Chem. 371 (2022) 131063, https://doi.org/10.1016/j.foodchem.2021.131063.

[38] Y. Zou, M. Gaida, F.A. Franchina, P.H. Stefanuto, J.F. Focant, Distinguishing between Decaffeinated and Regular Coffee by HS-SPME-GC×GC-TOFMS, Chemometrics, and Machine Learning, Molecules 27 (2022), https://doi.org/10.3390/molecules27061806.