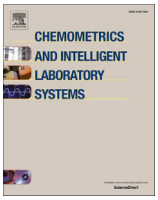
Chemometrics and Intelligent Laboratory Systems 240 (2023) 104907

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



^{1}H NMR, FAAS, portable NIR, benchtop NIR, and ATR-FTIR-MIR spectroscopies for characterizing and discriminating new Brazilian Canephora coffees in a multi-block analysis perspective

Michel Rocha Baqueta [a,f,*], Patrícia Valderrama [b,**], Manuela Mandrone [c], Ferruccio Poli [c], Aline Coqueiro [d], Augusto Cesar Costa-Santos [a], Ana Paula Rebellato [a], Gisele Marcondes Luz [a], Rodrigo Barros Rocha [e], Juliana Azevedo Lima Pallone [a,***], Federico Marini [f,****]

[a] Department of Food Science and Nutrition, School of Food Engineering, State University of Campinas – UNICAMP, Campinas, São Paulo, Brazil
[b] Universidade Tecnológica Federal do Paraná – UTFPR, Campo Mourão, Paraná, Brazil
[c] University of Bologna, Department of Pharmacy and Biotechnology (FaBiT), Bologna, Italy
[d] Department of Chemistry, Federal University of Technology – Paraná (UTFPR), Ponta Grossa, PR, 84017-220, Brazil
[e] Empresa Brasileira de Pesquisa Agropecuária, EMBRAPA Rondônia, Porto Velho, Rondônia, Brazil
[f] Department of Chemistry, University of Rome "La Sapienza", Piazzale Aldo Moro 5, 00185, Rome, Italy

ARTICLE INFO

Keywords:
Coffee
Conilon
Data fusion
Multi-block
Fine Amazonian Robusta coffee

ABSTRACT

Different analytical techniques, mixing single and multi-block chemometric analyses in supervised and unsupervised approaches, and the selection of variables in the coffee discrimination domain have been reported. Molecular and atomic spectroscopic techniques (^{1}H NMR, portable NIR, benchtop NIR, ATR-FTIR-MIR, and FAAS) were used to characterize and discriminate Brazilian Canephora coffees of specific producers, including two with geographical indication, and also to differentiate them from the Arabica. The sample set comprised 100 Canephora samples of different geographical origins in Brazil (Conilon from Espírito Santo, Amazonian Robusta from indigenous and non-indigenous producers of Rondônia, and Conilon from Bahia) and Arabica coffee (25 samples). ComDim exploratory multi-block analysis was first used to evaluate the contributions of all the different data blocks that characterized the samples and determine the calibration and validation sets. Multi-block discrimination by the SO-PLS-LDA was then used to validate the feasibility of discrimination, and to identify the relevant analytical techniques. The discrimination based on all the blocks presented 100% correct discrimination on training, cross-validation, and test sets, and suggested that only a single block (benchtop NIR) is needed to achieve a perfect discrimination. PLS-DA was then applied to the data from portable NIR to evaluate whether comparable performances could be achieved; all test samples were correctly discriminated with the exception of two Robusta Amazônico from non-indigenous producers and two Conilon from Espírito Santo, indicating that very accurate results can be obtained also using a portable instrument. Finally, the use of CovSel-DA allowed the discrimination to be optimized on the most cost-effective analytical technique (portable NIR). CovSel-LDA models selected the best variables for benchtop and portable NIR, impacting positively portable NIR performance and suggesting that portable NIR could bring comparable or at least only slightly worse performance than benchtop NIR for discrimination.

* Corresponding author. Department of Food Science and Nutrition, School of Food Engineering, State University of Campinas – UNICAMP, Campinas, São Paulo, Brazil.
** Corresponding author.
*** Corresponding author.
**** Corresponding author.
E-mail addresses: michelbaqueta@gmail.com (M.R. Baqueta), pativalderrama@gmail.com (P. Valderrama), jpallone@unicamp.br (J.A.L. Pallone), federico.marini@uniroma1.it (F. Marini).

https://doi.org/10.1016/j.chemolab.2023.104907
Received 29 April 2023; Received in revised form 24 June 2023; Accepted 26 June 2023
Available online 28 June 2023
0169-7439/© 2023 Elsevier B.V. All rights reserved.

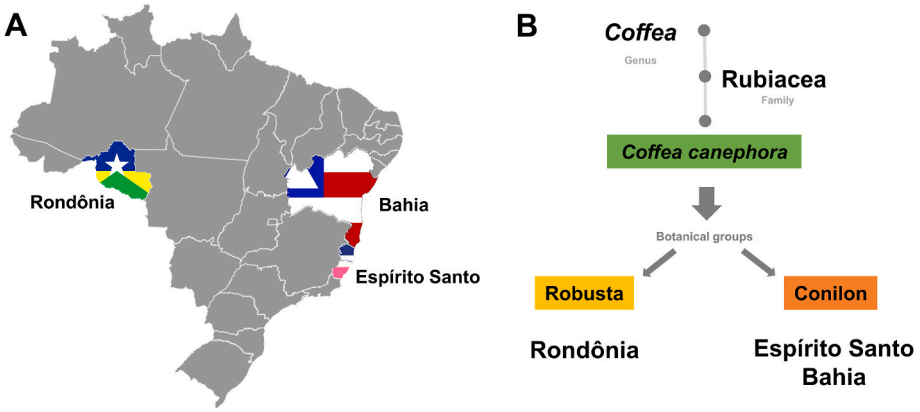


Fig. 1. Graphical representation showing the Brazilian map with the states of Rondônia, Bahia and Espírito Santo highlighted with their national flags (A) and Canephora division between botanical varieties Conilon and Robusta (B).

1. Introduction

Spectroscopic techniques have played an important role in the omics era to solve analytical problems in food evaluation. Although the use of many of them is well established, one challenge of these techniques is data analysis, because they provide large or very large data sets. Another current challenge is to try to combine them to obtain valuable chemical information in different perspectives. However, only the use of intelligent data analysis—more appropriately of chemometrics— can have the advantage of analyzing this information and interpreting it individually or in combination in an efficient way.

Roasted coffee is a challenging matrix for analysis because it is affected by many factors and has many compounds resulting from the roasting process that translate into its sensory and chemical characteristics [1]. After a long time, the industry is changing its perception about the Canephora coffee species, which has always been considered inferior. Although this is still a challenge, it is gaining interest for its adaptability to varying climates [2] and Brazil has helped to change this perception. Unique and distinctive Canephora coffees of specific Brazilian geographical origins have been emerging. Canephora production ranks Brazil as the second-largest producer of this species [3]. Rondônia, Espírito Santo and Bahia are the main producing states [4]. There are two botanical varieties of Canephora commonly cultivated in Brazil: Conilon and Robusta. Rondônia typically produces Robusta coffee [5], while Espírito Santo and Bahia produce Conilon [6]. Fig. 1 illustrates this information.

An evolution in the quality standards of Brazilian Canephora coffees has been observed and, as a result, the Robusta from Rondônia and Conilon from Espírito Santo have been registered with a geographical indication (GI) [7]. Conilon from Bahia has not yet a GI register and there is no prevalence of superior quality to date. There is a special interest in the Robusta from Rondônia, which is produced in the Brazilian Amazon region. This Amazonian coffee is named Robusta Amazônico and has a peculiar production divided into indigenous and non-indigenous coffee producers [8]. There is little information available in the literature about new Brazilian Canephora coffees, their different producers, and varieties. There were no reports of GI verification for them as well. In coffee analytical chemistry, most investigations are related to Arabica, and consequently studies with Canephora, particularly those that have achieved a higher level of quality, have only rarely been considered.

The study of the Brazilian Canephora coffees is quite new, therefore, to perform a characterization based on different analytical techniques trying to get the most complementary information possible from the different instruments is of relevance. The possibility to acquire multiple data/information from different sources for the same samples and then integrate them can lead to better interpretation of the results, because it is possible to identify the contribution of each method in distinguishing the samples [9]. The combination of molecular and atomic spectroscopic techniques in coffee analytical chemistry, as well as in analytical chemistry in general is still not comprehensive.

This study addresses the integration of molecular and atomic spectroscopy information to obtain an evaluation of Canephora coffees in a multi-block analysis perspective. Five analytical techniques, including (1) portable near-infrared – portable NIR, (2) benchtop NIR, (3) attenuated total reflectance Fourier transform mid-infrared – ATR-FTIR-MIR, (4) nuclear magnetic resonance – ^{1}H NMR, and (5) flame atomic absorption spectrometry – FAAS were chosen to provide information that is distinct but at the same time complementary. NIR spectroscopy has an appeal of fast and direct determination requiring small amounts of sample and no sample pre-treatment. Its portable version has the advantage of being cheaper and, in perspective, to be suitable for on-site application, e.g., directly in the coffee plantation. ATR-FTIR-MIR shares many of the analytical features of NIR characteristics. ^{1}H NMR provides metabolite fingerprints and complements the molecular information obtained with NIR and ATR-FTIR-MIR spectroscopies. Not least, elemental composition obtained with FAAS can be used to obtain other perspective of characterization not achievable through the other analytical techniques which mostly analyze the organic fraction of coffee.

The multiple analytical techniques have been used to characterize Brazilian Canephora coffees of specific regions/producers (including two with geographical indication register) and that have been reaching the specialty quality level, and to differentiate them from the specialty Arabica. Multi-block data analysis methods such as ComDim (Common Dimensions) for exploratory analysis [10], and a multi-block discrimination method, namely SO-PLS-LDA (sequential and orthogonalized partial least squares-linear discriminant analysis) [11], considered the information obtained by the different analytical blocks for the analysis. Once these results were obtained and verified that NIR spectroscopy was dominant in multi-block characterization and discrimination, PLS-DA (partial least squares discriminant analysis) and CovSel-LDA (covariance selection-linear discriminant analysis) [12] were used to compare the performance of the portable and benchtop NIR spectrometers for discrimination.

The main contribution of this study is to comparatively evaluate the feasibility of discrimination using different analytical techniques to highlight the role of each spectroscopic technique in the framework of a real application. Furthermore, the literature is limited with respect to the discrimination of Canephora coffee, either from an instrumental or chemometric perspective, using different analytical techniques or mixing single and multi-block chemometric analyses in supervised and unsupervised approaches, as well as the selection of variables. Although the discrimination of coffee in the domain of chemometrics has been a

topic of interest for many years [13], differently from what happens for Arabica, the number of studies exclusively on the discrimination of Canephora is still limited, especially for Canephora that are reaching a special high-quality level. It was shown that it is not an easy task to discriminate the high-quality Conilon from Espírito Santo Brazil based on different fermentation times of the beans using NMR spectroscopy, with accuracies lower than 50% in some cases [1]. In opposite, benchtop NIR spectroscopy has demonstrated be more efficient discriminate Brazilian Canephora coffees, their origins, cultivars from Western Brazilian Amazon, and specialty Canephora and Arabica [14]. More recently, benchtop NIR provided an accuracy of 96% for the discrimination of Amazonian Robusta coffee of different origins, while for the portable NIR the correct classification rate was 92% [15]. Specific local Robusta producers of Rondônia Brazil have been discriminated recently using synchronous fluorescence spectroscopy with sensitivity and specificity varying from 90 to 100% depending on the city of origin [16]. The geographical origin has been the main discriminating factor for Canephora and until now no article has taken into account that there are two distinct genetic groups to be discriminated: the group represented by the Conilon variety, and the group known as the Robusta variety.

2. Materials and methods

2.1. Coffee samples

A total of 128 coffee samples from different producers were collected, where 25 were Robusta Amazônico samples from indigenous producers, 25 were Robusta Amazônico samples from non-indigenous producers, 25 were Conilon samples from Espírito Santo, and 25 were Conilon samples from Bahia. The 25 other samples were specialty Arabica of different origins in Brazil and sensory profiles. Three low-quality Canephora samples were included to comparison as a regular Canephora. Robusta Amazônico and Conilon from Espírito Santo, which were registered with geographical indication, were provided by the EMBRAPA Rondônia, Porto Velho, Brazil, which guaranteed their authenticity.

Green coffee samples (100 g per sample) were roasted to a medium degree in a Probat sample roaster according to the Uganda Coffee Development Authority protocol [17] with the initial temperature of 160 °C and 190 °C at the end, with a time ranging from 7:30 min–9 min to achieve the desired profile. The samples were cooled, milled, and sieved through a 20-mesh sieve for particle size standardization.

2.2. Instrumental analyses

2.2.1. Benchtop near-infrared spectroscopy (benchtop NIR)

Powder samples were directly analyzed using a PerkinElmer Fourier Transform NIR spectrophotometer Spectrum 100 N equipped with a glass cuvette. Reflectance mode was used. Each spectrum was digitized with 32 scans from 1000 to 2500 nm with a resolution of 4 nm. Roasted and ground coffee samples were analyzed in a random sequence at room temperature (22 °C) by placing them directly on the instrument. Three different aliquots of the sample were used, recording the corresponding spectra which were, then, averaged. Before analysis, the blank was evaluated using a NIR reflectance standard.

2.2.2. Portable near-infrared spectroscopy (portable NIR)

A MicroNIR spectrometer (microNIR™ 1700) from JDSU Uniphase Corporation with a glass cuvette was used to obtain spectra with portable NIR. This spectrometer covers the 906–1676 nm range. Powder samples were directly analyzed in a random sequence at room temperature (22 °C) by placing them directly on the portable NIR. Three different sample aliquots were used, and the spectrum of each aliquot was recorded in the automatic reflectance mode, with 50 scans and a resolution of 6.25 nm, resulting in a measurement time of 0.50 s; spectra of the three aliquots were then averaged. The blank was evaluated using a standard NIR reflectance (Spectralon™) with a diffuse reflection coefficient of 99%, while a dark reference (zero–to simulate non-reflection) was obtained with the lamp off. The dimensions of this spectrometer were 45 mm in diameter and 42 mm in height, weighting about 60 g.

2.2.3. Mid-infrared spectroscopy (ATR-FTIR-MIR)

Mid-infrared spectroscopic signals were recorded in an IRAffinity-1S spectrometer (Shimadzu, Kyoto, Japan) utilizing a horizontal Attenuated total reflectance (ATR) accessory, which contains a zinc selenide (ZnSe) crystal at a 45° angle (PIKE Technologies, Madison, USA). Powder samples were directly analyzed in random order and spectra were obtained in the wavenumber range between 4000 and 600 cm^{-1}, at a nominal resolution of 4 cm^{-1} and 32 scans, with the aid of IR Solution software (Shimadzu, Kyoto, Japan). For each sample, spectra were collected in triplicate and then averaged. Background correction was performed recording ambient air.

2.2.4. Nuclear magnetic resonance (^1H NMR)

The metabolite extraction and NMR procedure were applied according to a previous coffee study [18]. The reagents were purchased from Sigma Aldrich, except deuterated solvents as H_2O-d_2 and 3-(trimethylsilyl)-propionic-2,2,3,3-d4 acid sodium salt–TMSP were purchased from Eurisotop. Extraction was performed with 0.1 g of sample to 1.5 mL of phosphate buffer (90 mM, pH 6.0) in H_2O-d_2 containing 0.01% of TMSP as standard for 1 h in a bath maintained to 90 °C. After extraction, samples were centrifuged for 10 min (17,000×*g*), then the supernatant were analyzed.

NMR spectra were recorded at 298 K using a Varian 14.4 T NMR instrument (600.13 MHz operating at ^1H frequency) equipped with a high-field triple resonance probe, using H_2O-d_2 for the internal lock. Relaxation delay of 2.0 s observed pulse of 5.80 μs, and the sum of 256 scans were acquired for each sample. The acquisition time was 16 min and the spectral width of 16.00 ppm. A presaturation sequence was used to suppress the residual water signal at 4.83 ppm (power = 22 Hz, presaturation delay = 2 s). The free induction decays (FIDs) were Fourier transformed, and the resulting spectra were phased, baseline-corrected, and calibrated for TMSP at 0.00 ppm. The spectral intensities were reduced to integrated regions of equal width (0.04 ppm) corresponding to the interval of 0.00–10.00 ppm after normalization with respect to the signal of the standard at 0.00 ppm using the NMR MestReNova software (Mestrelab Research, Spain). The regions from 5.00 to 4.50 ppm were excluded from the analysis due to residual water signals. The identification of metabolites was based on the chemical shifts, coupling constants, and comparison with data [19] from the available literature on coffee.

2.2.5. Flame atomic absorption spectrometry (FAAS)

Standard solutions of calcium – Ca, magnesium – Mg, zinc – Zn, iron – Fe, manganese – Mn, copper – Cu (LabSynth, Diadema, SP, Brazil), and potassium – K (SpecSol, Quimlab, Jacareí, SP, Brazil) at 1000 mg L^{-1} were used. Other reagents/materials were hydrogen peroxide 30% (v/v) (LabSynth, Diadema, SP, Brazil); lanthanum oxide (Sigma Chemical Co., St. Louis, USA), commercial diluted nitric acid 50% (v/v) (Merck, Darmstadt, Germany), ultrapure water (Sartorious, Germany), acetylene gas (Messer Gases, Brazil), filter paper of 9 cm diameter (Nalgon REF 3551, Germany), ultrasonic bath (model 1400, Unique, Brazil), and digester block (model M242, Quimis, Brazil).

The procedure of flame atomic absorption spectrometry (FAAS) was carried out according to Baqueta et al. [20] slightly modified. The samples were prepared through wet mineralization in an open system (digester block) using dilute nitric acid. Blanks and 0.6 g of each sample (roasted ground coffee powder) were mineralized with 6 mL of diluted nitric acid and 2 mL of hydrogen peroxide for 4 h, filtered and diluted with ultrapure water before minerals quantification. A PerkinElmer AAnalyst 200 equipped with a deuterium lamp for correction of the

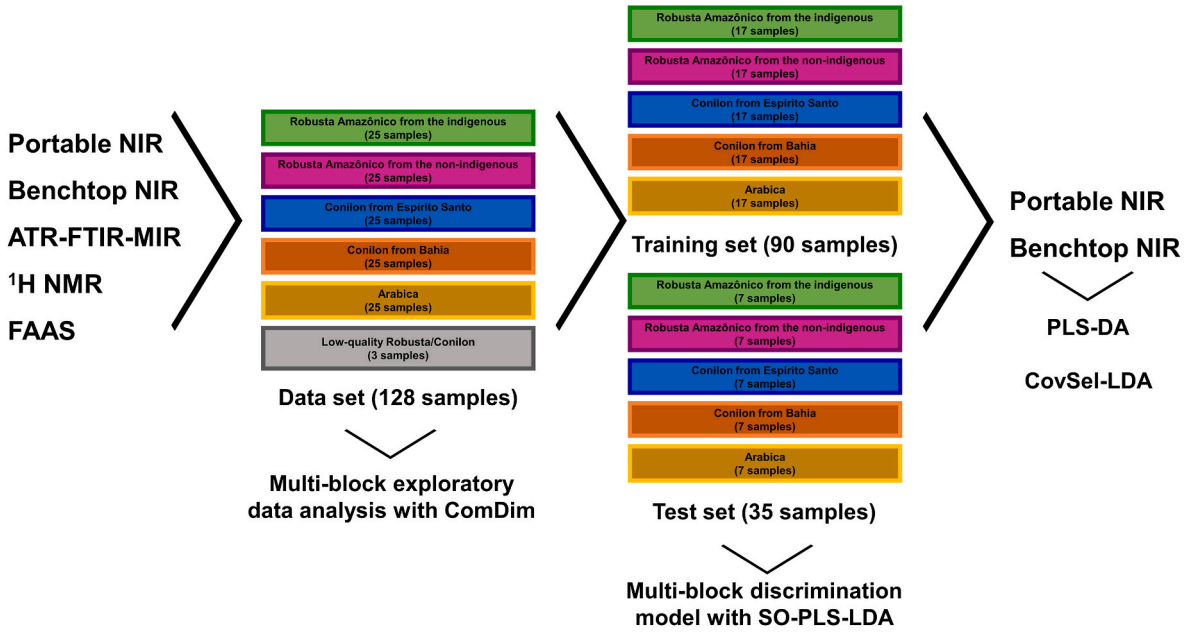


Fig. 2. Graphical representation of the data set acquired by the different analytical techniques and the division into a training and a test set.

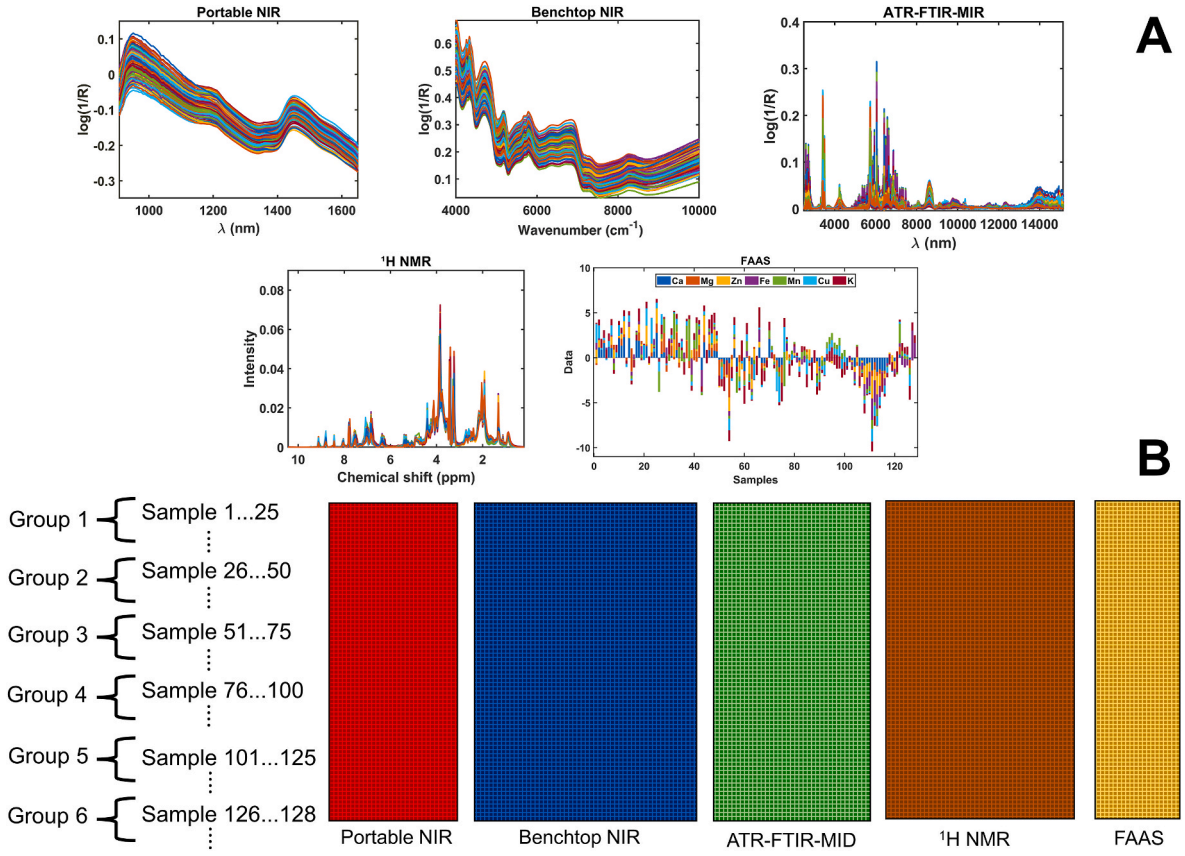


Fig. 3. Multi-platform information acquired for the 128 samples, showing (A) spectra obtained using portable and benchtop NIR, ATR-FTIR, and NMR, and mineral composition (mg/100 g) obtained with FAAS and (B) data organization in matrices for multi-block analysis.

background radiation was used to determine the essential minerals. The samples were nebulized and mixed with air-acetylene flame (2.5/10 l h^{-1}) at about 2000 °C. Hollow cathode lamps (PerkinElmer, Norwalk, USA) for Fe – 248.3 nm, Ca – 422.67 nm, Cu – 324.75 nm, Mg – 279.48 nm, Mn – 285.21 nm, and Zn – 213,86 nm were used. For K determination, the equipment was configured for atomic emission. Each sample was mineralized in triplicate and then read.

The method was checked and considered adequate to determine the essential minerals [21]. The relative standard deviation (RSD) (%) values were 3.9 for Ca, 7.2 for Mg, 7.4 for Zn, 6.7 for Fe, 4.8 for Mn, 8.4

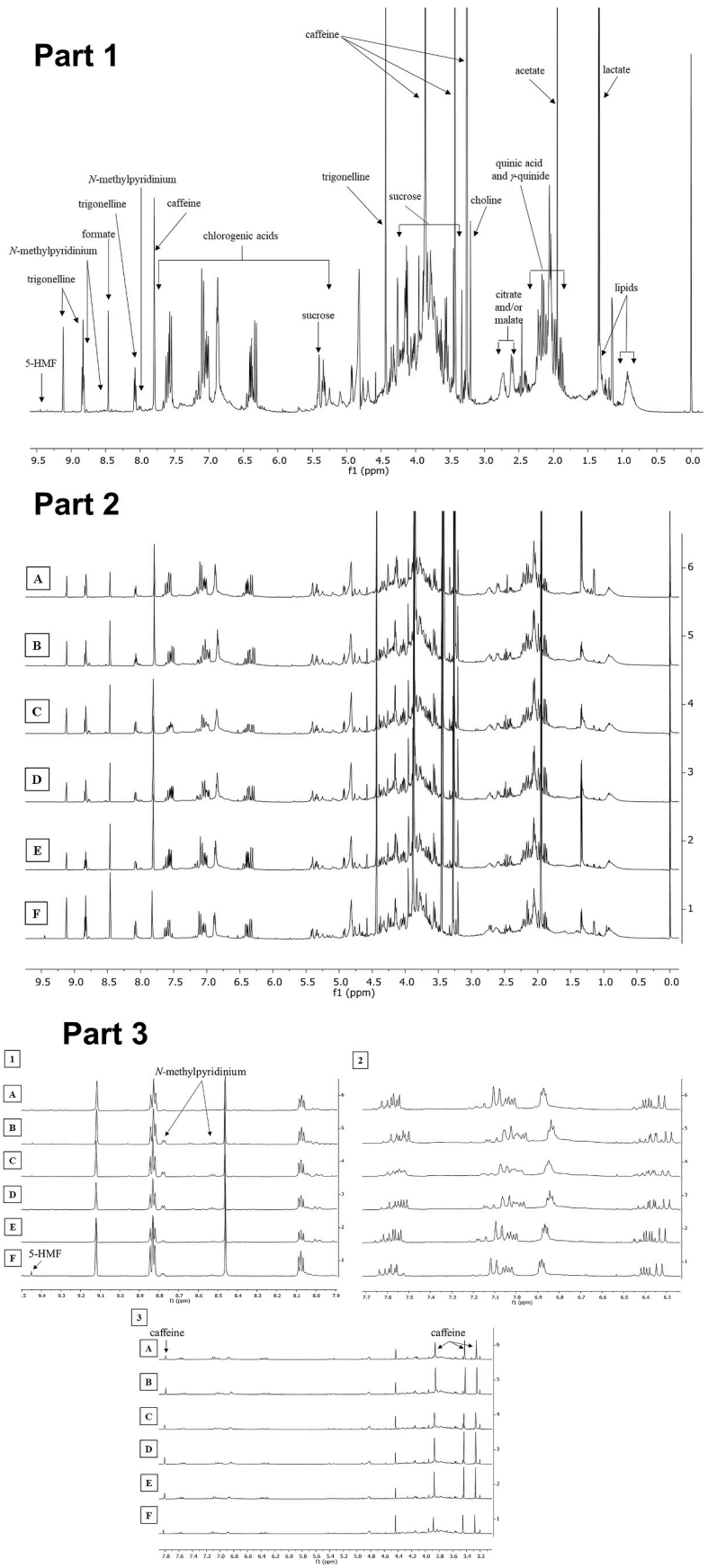


Fig. 4. Part 1 – One representative ^{1}H NMR spectra of Robusta Amazônico cultivated by the indigenous coffee with resonance signals assigned; **Part 2 –** ^{1}H NMR fingerprints of different coffee groups. (A) Robusta Amazônico from indigenous producers; (B) Robusta Amazônico from non-indigenous producers; (C) low-quality Canephora; (D) Conilon from Espírito Santo; (E) Conilon from Bahia; (F) Arabica; **Part 3** – ^{1}H NMR comparison for different coffee groups. (1) Expansion of the range 7.9–9.5 ppm. (2) Expansion of the range 6.2–7.7 ppm, and (3) Expansion of the range 3.1–7.9.

Table 1
Preprocessing approaches applied on the different data blocks.

Technique	Preprocessing
Portable NIR	2nd derivative + mean centering
Benchtop NIR	1st derivative + mean centering
ATR-FTIR-MID	Baseline correction +1st derivative + mean centering
^{1}H NMR	Pareto scaling
FAAS	Autoscaling

for Cu, and 8.6 for K. The correlation coefficients were >0.9997 for Ca; >0.9996 for Mg; >0.9991 for Zn; >0.9996 for Fe; >0.9995 for Mn; >0.9999 for Cu; and >0.9901 for K. The LOD (limit of detection) and LOQ (limit of quantification) (mg/100 g) for Ca were (LOD: 0.11, LOQ: 0.18); Mg (LOD: 0.0001, LOQ:0.0001); Zn (LOD: 0.15, LOQ: 0.24); Fe (LOD: 0.60, LOQ: 1.01); Mn (LOD: 0.002, LOQ: 0.004); Cu (LOD: 0.09, LOQ: 0.15); K (LOD: 0.0004, LOQ: 0.006).

2.3. Chemometric methods

After data collection, the data were imported in Matlab R2019a (The Mathworks, Natick, MA). The matrices collecting the data acquired by the different techniques had the following dimensions: 128 × 125 in the case of portable NIR, 128 × 6001 for benchtop NIR, 128 × 1733 for ATR-FTIR-MIR, 128 × 248 in the case of ^{1}H NMR, and 128 × 7 for FAAS. Samples in the five data matrices representing the different data blocks (one per technique) were always organized in the same sequence: from sample 1 to 25 the Robusta Amazônico samples from indigenous producers (class 1), from sample 26 to 50 the Robusta Amazônico samples from non-indigenous producers (class 2), from 51 to 75 the Conilon samples from Espírito Santo (class 3), from 76 to 100 the Conilon samples from Bahia (class 4), from 101 to 125 the Arabica samples (class 5), and from 126 to 128 the low-quality Canephora. Choice of the most suitable preprocessing for the different blocks was made based on the previous experience with similar data.

2.3.1. ComDim multi-block exploratory data analysis

ComDim multi-block analysis [10] is an exploratory data analysis technique which focused on the extraction of common components that jointly explain as much as possible of the variance of the different blocks. This allows to understand relations between the variables in the different blocks and to relate the corresponding underlying latent structure to the differences observed among the samples. Models with different differentiation purposes/sample classes were built to extensively explore the potential of the different techniques together for characterization. The concatenated matrix always had 8114 variables considering all data blocks and the number of samples in each model was varied. Ten common components (CCs) were calculated to evaluate and discuss the contributions or saliences from all the data blocks on sample distribution through scores. CCs with great contributions of the different data blocks and explaining high total variance were selected.

2.3.2. SO-PLS-LDA multi-block discrimination

In order to externally validate the models, samples were divided into a training and a test set for multi-block discrimination by the Duplex algorithm [22] using a strategy based on taking into account the variability of all five data blocks differently pretreated described in Ref. [15], i.e., applying the Duplex algorithm on the ComDim scores, individually for each category. A total of 18 samples from each class were selected for model building and selection, resulting in 90 training samples. The remaining 35 samples (7 from each class) constituted the test set. The 3 low-quality Canephora samples were not considered for the discrimination study. A graphical representation of the composition of the data set is shown in Fig. 2.

SO-PLS [23] is a multi-block regression method, which, by introducing the dummy matrix coding for class belonging and applying linear discriminant analysis to the predicted responses or the scores, can be extended to discrimination purposes [11,24]. It sequentially extracts information from each data block and use it for the prediction of the desired response(s), in this case, the dummy binary **Y**, which is made of 5 columns, one for each of the 5 category: Robusta Amazônico from indigenous producers (class 1), Robusta Amazônico from non-indigenous producers (class 2), Conilon from Espírito Santo (class 3), Conilon from Bahia (class 4), and Arabica (class 5). Firstly, the algorithm calculated a SO-PLS model and then LDA was applied on the predicted **Y** [25]. The discriminant ability of the model was determined by sensitivity (equation (1)) and specificity (equation (2)) rates. These parameters can assume values from 0 to 100% and are calculated for training, cross-validation, and prediction sets considering true positive (TP) and negative (TN), and false-positive (FP) and negative (FN) rates.

$$Sensitivity = \frac{N(true\ positives)}{N(true\ positives) + N(false\ negatives)} \quad (1)$$

$$Specificity = \frac{N(true\ negatives)}{N(true\ negatives) + N(false\ negatives)} \quad (2)$$

where the true positives and the false negatives are the samples of the particular class which are correctly predicted as belonging to that class or erroneously misclassified as belonging to other classes, respectively. On the other hand, the true negatives and the false positives are the samples from other categories which are correctly predicted as not belonging to the specific class or wrongly classified as belonging to the particular class, respectively.

2.3.3. PLS-DA and CovSel-LDA for benchtop and portable NIR discriminations

PLS-DA was then applied to discriminate the coffee samples in the five classes based on the data of specific individual blocks (the NIR ones) [15]. In order to try to obtain better results, another discrimination method employing variable selection was applied for these data sets. Covariance Selection (CovSel) [12] is a variable selection method whose algorithm shares similar traits with PLS. Indeed it can be thought as a PLS algorithm with binary weights, rather than real-valued ones, this characteristics being exploited for feature reduction [26]. CovSel algorithm can also be extended to discrimination problems by introducing the usual dummy binary **Y**, and applying LDA either on the selected variables or on the predicted **Y**.

3. Results and discussion

Each sample was analyzed by the five instrumental techniques. The different spectral data obtained are shown in Fig. 3A. The mineral composition obtained with FAAS is also graphically represented. Drawing conclusions by visual analysis of the spectroscopic results is difficult, because there was an overlap of signals among the samples. The same occurred with the results of the essential mineral composition. However, a special interest was given to the NMR spectra interpretation expanded in Fig. 4.

Fig. 4 shows the identification of compounds by the ^{1}H NMR spectroscopy. One sample of Robusta Amazônico cultivated by the indigenous was chosen to present the chemical characterization. Assignments of the main metabolites present in the roasted coffee extracts were depicted in Fig. 4 – Part 1. The ^{1}H NMR fingerprints for the different coffee groups obtained for the coffee samples were shown in Fig. 4 – Part 2.

Fig. 4 – Part 3 brings a comparison between the spectra. Comparing the ^{1}H NMR chemical profiles for the different coffee groups was possible to observe some main differences (Fig. 4 – Part 3 in 3.1, 3.2 and 3.3). A semiquantitative analysis based on the internal standard used in the ^{1}H NMR analysis allowed differences in the amount of some compounds in the different coffee groups. Arabica coffee showed the highest

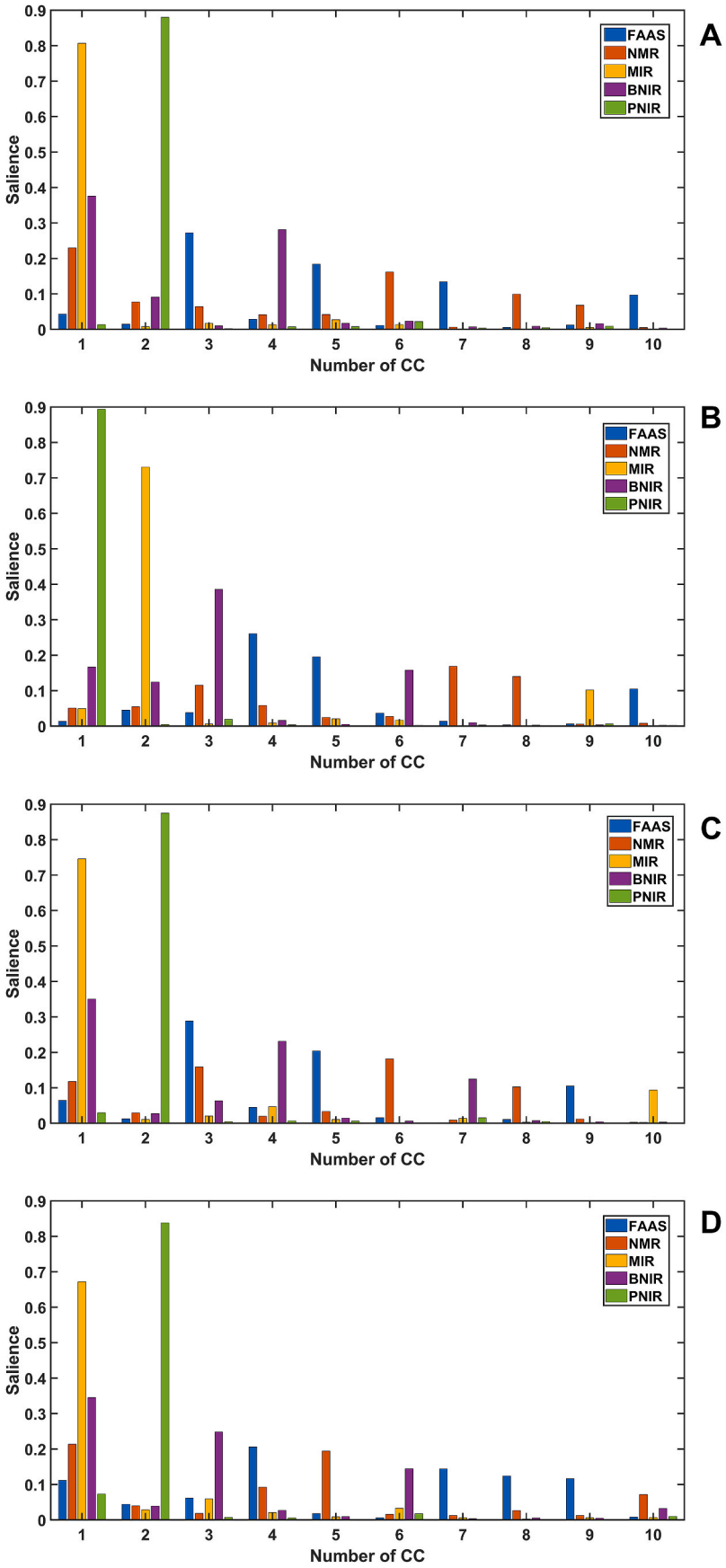


Fig. 5. Contributions of the different data blocks along the 10 CCs for differentiating all samples (A), only Canephora (B), exclusively GI Canephora (C), and indigenous and non-indigenous Robusta Amazônico producers (D).

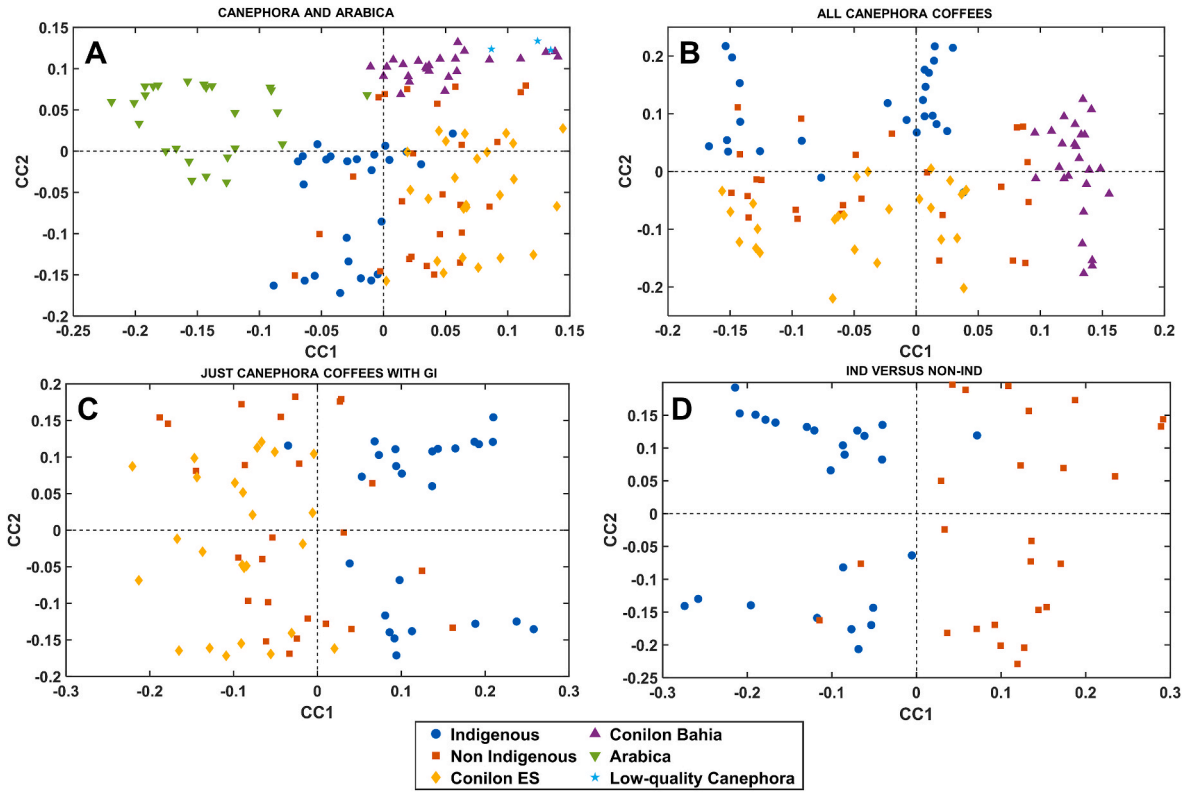


Fig. 6. Scores of ComDim analysis applied for the sample set comprising 103 Canephora samples of different geographical origins in Brazil (Conilon from Espírito Santo – 25 samples, Amazonian Robusta from indigenous – 25 samples and non-indigenous producers of Rondônia – 25 samples, Conilon from Bahia – 25 samples, and low-quality Canephora – 3 samples) and 25 Arabica coffee samples. The first analysis was performed with all 128 Canephora and Arabica samples (A); The second analysis considered only the Robusta Amazônico samples originated from the indigenous and non-indigenous producers of Rondônia, Conilon from Espírito Santo, and Conilon from Bahia (B), the third analysis was applied exclusively for Robusta Amazônico samples originated from the indigenous and non-indigenous producers of Rondônia, and Conilon from Espírito Santo, which were registered with geographical indication (C), and the last analysis considered exclusively Rondônia samples, which were the Robusta Amazônico samples from the indigenous and non-indigenous producers (D).

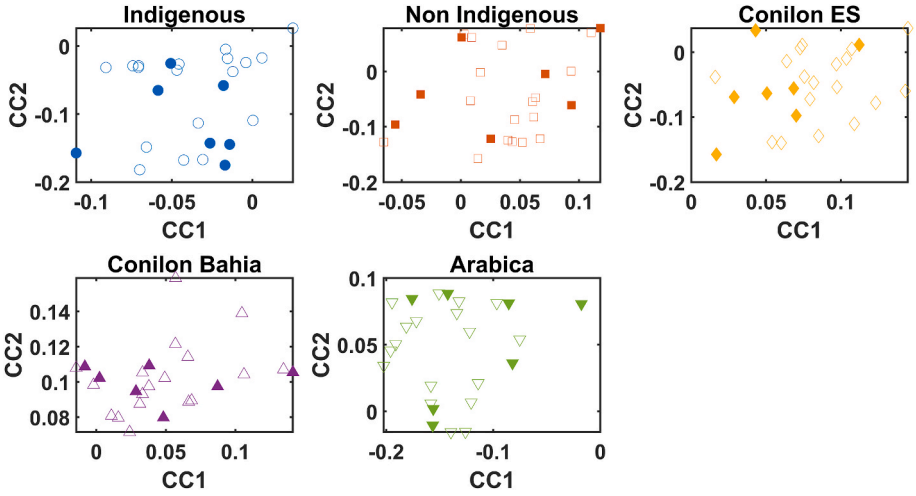


Fig. 7. Division of the samples of each category into training (hollow markers) and test set (filled markers).

amount of 5-HMF (5-hydroxymethylfurfural), with resonance at 9.45 ppm, while indigenous Robusta Amazônico presented the lowest amounts of 5-HMF and *N*-methylpyridinium with resonances at 8.5 and 8.8 ppm (Fig. 4 – Part 3 in 3.1). Conilon from Espírito Santo and low-quality Canephora coffees showed the lowest amount of *N*-methylpyridinium with resonances at 8.53 and 8.78 ppm. 5-HMF and *N*-methylpyridinium are formed through the roasting process [27]. The expansion of the region of 6.2–7.7 ppm (Fig. 4 – Part 3 in 3.2) showed the resonances of the chlorogenic acids (quinic esters of hydroxycinnamic acids, CGAs). CGAs are important active compounds and the most important class of coffee polyphenols; they can be divided into the major groups: caffeoylquinic acids (CQAs), being 5-*O*-caffeoylquinic acid (5-CQA), the most common CQA; di-caffeoylquinic acids (di-CQAs), feruloylquinic acids (FQAs), p-coumaroylquinic acids (p-CoQA), and caffeoylferuloylquinic acids (CFQA) [28,29]. The flavor of roasted coffee is highly influenced by CGAs, that directly interfere in the cup

Table 2
Results of modeling with the multi-block strategy based on all the data blocks (SO-PLS-LDA) and individual analysis and comparison between benchtop and portable NIR for discriminating the samples based on PLS-DA and CovSel-LDA methods.

Sets	SO-PLS-LDA based on all the blocks					
	Parameters	1	2	3	4	5
Training	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
Cross-validation	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
Test	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
	PLS-DA for benchtop NIR					
Sets	Parameters	1	2	3	4	5
Training	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
Cross-validation	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
Test	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
	PLS-DA for portable NIR					
Sets	Parameters	1	2	3	4	5
Training	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
Cross-validation	Sensitivity	94.4	88.9	77.8	100.0	100.0
	Specificity	97.2	95.8	97.2	100.0	100.0
Test	Sensitivity	100.0	71.4	71.4	100.0	100.0
	Specificity	89.3	100.0	96.4	100.0	100.0
	CovSel-LDA for benchtop NIR					
Sets	Parameters	1	2	3	4	5
Training	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
Cross-validation	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
Test	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
	CovSel-LDA for portable NIR					
Sets	Parameters	1	2	3	4	5
Training	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
Cross-validation	Sensitivity	100.0	100.0	100.0	100.0	100.0
	Specificity	100.0	100.0	100.0	100.0	100.0
Test	Sensitivity	100.0	85.7	71.4	100.0	100.0
	Specificity	89.3	100.0	100.0	100.0	100.0

*Robusta Amazônico samples from indigenous producers (class 1), Robusta Amazônico samples from non-indigenous producers (class 2), Conilon samples from Espírito Santo (class 3), Conilon samples from Bahia (class 4), Arabica samples (class 5).

quality [30]. Although it was possible to observe differences in the composition of chlorogenic acids in the different analyzed coffees, there are already dozens of CGAs identified and for these compounds to be properly identified, future analyzes such as 2D NMR spectroscopy and mass spectrometry are necessary. It was also notorious that Arabica coffee presented the lowest amount of caffeine, with resonances at 7.80; 3.86; 3,43, and 3.26 ppm, while non-indigenous Robusta Amazônico showed the highest quantities; this is in agreement of the literature when comparing Arabica and Robusta contents of caffeine [27]. For trigonelline and choline, non-indigenous Robusta Amazônico and Arabica coffees showed the highest amount while Conilon from Bahia and from Espírito Santo presented the lowest amounts.

3.1. ComDim multi-block exploration

Each data block had a preprocessing applied to remove unwanted sources of variability and enhance the results (Table 1) to perform the ComDim analysis.

The data were organized in matrices as shown in Fig. 3B. To investigate the correlation among the investigated blocks, four ComDim analyses were performed, each time considering a different number of samples to be included in the analysis. A maximum of ten common components (CCs) was calculated in each case, where the saliences/contributions of the different data blocks along the 10 CCs can be analyzed in Fig. 5. The contributions/saliences from the different pretreated data blocks changed when calculating models with different sample classes. However, a prominent contribution of the MIR and benchtop NIR spectroscopies was always observed in the four analyses, mainly on CC1 (Fig. 5). The first five CCs always explained more than 98% of the total variance in the models. However, the first two CCs were always more informative and provided the greatest contribution from the different data blocks analyzed jointly. Therefore, CC1 and CC2 scores were jointly analyzed to obtain interpretations regarding the samples.

The scores of the different ComDim brought information of relevance for the study of the Canephora coffees (Fig. 6). It was possible to observe that the sample diversity varied considerably between groups and even within groups. The first analysis with all samples immediately showed a partial distinction between Canephora and Arabica (in green) along CC1 (Fig. 6). It was also showed that the three low-quality Canephora samples (in cyan) were closely grouped with those Conilon from Bahia (in purple), that currently are not predominantly of specialty quality. CC2 showed a partial distinction of Arabica (in green) and Conilon from Bahia (in purple) in relation to the others. This shows how diverse they were compared to the other Canephora coffees with GI. Although there was contribution from almost all data blocks in distinguishing CC1 (see saliences in Fig. 5A), the MIR block dominated. In CC2, the result suggested that the portable NIR spectra brought the most important information (Fig. 5A).

Other ComDim was developed only with Canephora samples (Fig. 6B) in order to ignore the species factor in the differentiation observed in the previous model (Fig. 6A). In the model considering exclusively Canephora coffees (Fig. 6B), there was a well-defined distinction of the Robusta from indigenous (in blue) and Conilon from Bahia (in purple). Robusta from the non-indigenous (in orange) and Conilon from Espírito Santo (in yellow) overlapped. In this model, the portable NIR was found to have the most significant contribution in CC1, while the MIR was the most significant in CC2 (Fig. 5B).

A third ComDim where the samples from Bahia were removed to consider only the Canephora coffees with GI was built (Fig. 6C). It contained the Robusta from the indigenous and non-indigenous, and Conilon from Espírito Santo. Almost the same sample distribution discussed in the previous model (Fig. 6B) was observed, with no considerable differences. The same sample distinction trend was observed. Although this has occurred, the contribution of the data blocks has varied. MIR data block contributed more in CC1, while portable NIR more in CC2 (Fig. 5C).

po It was possible to see that most of them were dispersed differently in CC1. However, it was observed also a distinction within the Robusta of non-indigenous producers, which could be associated with the diversity of cultivars available within this class but of unknown characteristics. Two subgroups appeared within the indigenous group in this and other ComDim analyses and may be differences between local producers. The data blocks that contributed the most were again the MIR and portable NIR for CC1 and CC2, respectively (Fig. 5D). In general, multi-block exploratory data analysis with different ComDim suggested that MIR and NIR (both benchtop and portable) spectroscopies were dominant techniques for bringing the necessary information for a preliminary distinction of samples, either by species, geographical origin, or manufacturer.

3.2. Multi-block discrimination with SO-PLS-LDA analysis

As anticipated, the split of the samples into training and test set was performed by applying the duplex algorithm on the scores of the ComDim multi-block analysis. In particular, to obtain a representative and reliable selection of training and test samples, the duplex selection was carried out individually on each category. The sample selection for the

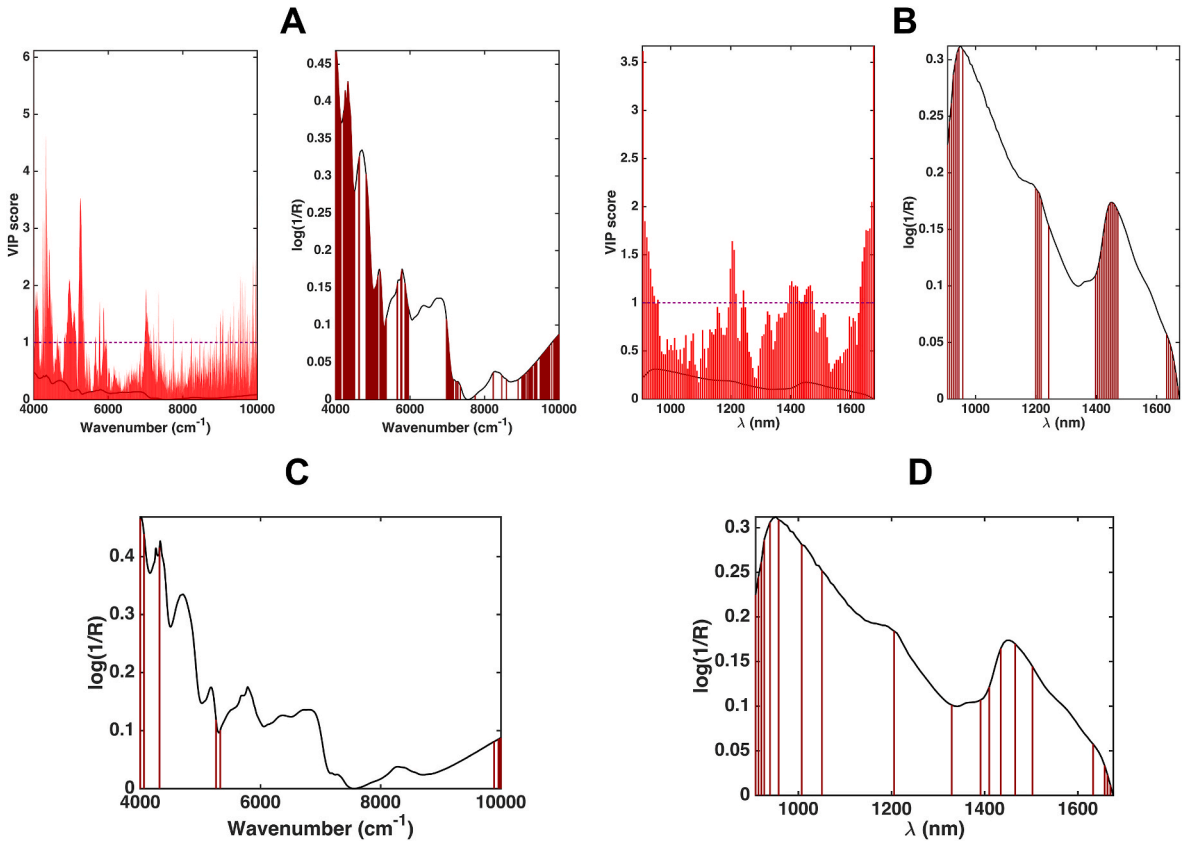


Fig. 8. Graphical representation of the VIP scores extracted from the SO-PLS-LDA model for benchtop NIR (A) and portable NIR (B), and variables selected by CovSel-LDA for benchtop NIR (C) and portable NIR (D).

individual classes is shown in Fig. 7.

The discrimination results on training, cross-validation and test sets are presented in Table 2. With reliably selected sample sets, SO-PLS-LDA based on all the five blocks was applied and resulted in 100% correct discrimination on training, cross-validation, and test sets, using 5 latent variables (LVs) on benchtop NIR and 0 LVs for all the others. Although SO-PLS-LDA considered the information contained in all blocks, it suggested that only a single block (benchtop NIR) is needed to achieve a perfect discrimination of the five coffee classes. Indeed, the LVs from SO-PLS-LDA model were selected from benchtop NIR block only. This finding encouraged a further comparison of the use of NIR spectroscopy in coffee discrimination with the benchtop and portable versions available.

3.3. Comparison between benchtop and portable NIR performance

PLS-DA models were built for portable NIR and benchtop NIR to compare their performance (Table 2). They confirmed the effectiveness of NIR spectroscopy. The PLS-DA on benchtop NIR, which, as already discussed is completely equivalent to the best SO-PLS-LDA model (since LVs were extracted only from that block) had a 100% correct discrimination on the training, cross-validation, and test sets. For portable NIR, 100% discrimination on all classes was achieved on the training sets. In cross-validation and prediction, there were incorrect discriminations for classes 1, 2 and 3 using the portable NIR. For classes 4 and 5, the discriminations were perfect reaching 100%. It encouraged to check if with portable NIR alone could possibly have comparable or at least only slightly worse performance than benchtop NIR using other discrimination strategy based on CovSel-LDA to select the best variables for benchtop NIR and portable NIR. CovSel-LDA improved portable NIR performance as can be seen in Table 2.

Fig. 8 shows the values of the variable importance in projection (VIP) index extracted from the SO-PLS-LDA model for benchtop NIR (A) and portable NIR (B), and variables selected by CovSel-LDA for benchtop NIR (C) and portable NIR (D). It was found that a significantly smaller number of variables were considered to obtain the discrimination results.

4. Conclusions

The multi-block characterization and discrimination of Brazilian Canephora coffees has brought new contributions to the study of this coffee species that is on the rise and comes from different regions of Brazil. The multi-block exploratory analysis showed that although there was a contribution from all the different techniques for characterization, the multi-block discrimination showed that NIR spectroscopy dominated for this purpose. Due to this finding, comparisons between benchtop NIR and portable NIR were proposed. Portable NIR provided slightly inferior results to benchtop NIR through variable selection.

Besides the relevant aspects of chemometrics, this study brought results that are of importance for coffee science in particular. The ability to discriminate preliminarily or definitively between Brazilian Canephora coffees opens the possibility of establishing a more reliable certification system than the current ones based on sensory or physical analyses of the beans.

Author statement

Michel Rocha Baqueta: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft. Patrícia Valderrama: Supervision, Writing - Review & Editing. Manuela Mandrone: Methodology, Formal analysis, Ferruccio Poli: Methodology, Formal analysis. Aline Coqueiro: Methodology, Formal analysis. Augusto Cesar Costa-Santos: Methodology, Formal analysis. Ana Paula Rebellato:

Methodology, Formal analysis. Gisele Marcondes Luz: Methodology, Formal analysis. Rodrigo Barros Rocha: Resources. Juliana Azevedo Lima Pallone: Writing - Review & Editing, Funding acquisition, Project administration, Supervision. Federico Marini: Conceptualization, Software, Validation, Supervision, Writing - Review & Editing.

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

The authors do not have permission to share data.

Acknowledgments

This work was supported by the São Paulo Research Foundation (FAPESP) (grant #2019/21062-0 and grant #2022/04068-8, and process #2022/03268-3), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (process 306606/2020-8, process 402441/2022-2 and process 310982/2020-0), Agência Brasileira de Desenvolvimento Industrial (process 23200.19/0070-2-01 and process 30.20.90.027.00.00). The authors would like to thank the EMBRAPA Rondônia, cooperatives and coffee producers for their invaluable help with the acquisition of the samples for the study.

References

[1] B. Zani Agnoletti, W. dos Santos Gomes, G. Falquetto de Oliveira, P. Henrique da Cunha, M. Helena Cassago Nascimento, Á. Cunha Neto, L. Louzada Pereira, E. Vinicius Ribeiro de Castro, E. Catarina da Silva Oliveira, P. Roberto Filgueiras, Effect of fermentation on the quality of conilon coffee (Coffea canephora): chemical and sensory aspects, Microchem. J. 182 (2022), 107966, https://doi.org/10.1016/J.MICROC.2022.107966.

[2] ICO, International Coffee Organization, Botanic Aspects, 2022. https://www.ico.org/pt/botanical_p.asp.

[3] ICO, International Coffee Organization, Coffee Production Report, 2021. http://www.ico.org/prices/po-production.pdf.

[4] CONAB, Acompanhamento da Safra Brasileira de café - safra 2023 - Primeiro levantamento, Obs. Agrícola. 10 (2023) 10–11. https://www.conab.gov.br/info-agro/safras/cafe.

[5] A.L. Teixeira, R.B. Rocha, M.C. Espindula, A.R. Ramalho, J.R.V. Júnior, E.A. Alves, A.M.P. Lunz, F. de F. Souza, J.N.M. Costa, C. de F. Fernandes, Amazonian robustas-new coffea canephora coffee cultivars for the western brazilian amazon, Crop Breed. Appl. Biotechnol. 20 (2020) 1–5, https://doi.org/10.1590/1984-70332020v20n3c53.

[6] M.F. Lemos, C. Perez, P.H.P. da Cunha, P.R. Filgueiras, L.L. Pereira, A.F. Almeida da Fonseca, D.R. Ifa, R. Scherer, Chemical and sensory profile of new genotypes of Brazilian Coffea canephora, Food Chem. 310 (2020), https://doi.org/10.1016/j.foodchem.2019.125850.

[7] Brazil, Brazilian Coffees with Geographical Indication, 2021. https://www.gov.br/agricultura/pt-br/assuntos/sustentabilidade/indicacao-geografica/arquivos-publicacoes-ig/brazilian-coffees-with-geographical-indication.

[8] A.O. Zacharias, C. Rosa Neto, E.A. Alves, R.K. da Silva, Modelo de negócio: cafés especiais robustas amazônicos, 2021. https://www.embrapa.br/en/busca-de-publicacoes/-/publicacao/1135107/modelo-de-negocio-cafes-especiais-robustas-amazonicos.

[9] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication, Anal. Chim. Acta 820 (2014) 23–31, https://doi.org/10.1016/J.ACA.2014.02.024.

[10] V. Cariou, D. Jouan-Rimbaud Bouveresse, E.M. Qannari, D.N. Rutledge, ComDim Methods for the Analysis of Multiblock Data in a Data Fusion Perspective, Elsevier, 2019, https://doi.org/10.1016/B978-0-444-63984-4.00007-7.

[11] A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multi-block classification, Chemometr. Intell. Lab. Syst. 141 (2015) 58–67, https://doi.org/10.1016/j.chemolab.2014.12.001.

[12] A. Biancolillo, F. Marini, J.M. Roger, SO-CovSel, A novel method for variable selection in a multiblock framework, J. Chemom. 34 (2020) 1–21, https://doi.org/10.1002/cem.3120.

[13] M.R. Baqueta, A. Coqueiro, P.H. Março, P. Valderrama, Multivariate classification for the direct determination of cup profile in coffee blends via handheld near-infrared spectroscopy, Talanta 222 (2021), 121526, https://doi.org/10.1016/J.TALANTA.2020.121526.

[14] M.R. Baqueta, E.A. Alves, P. Valderrama, J.A.L. Pallone, Brazilian Canephora coffee evaluation using NIR spectroscopy and discriminant chemometric techniques, J. Food Compos. Anal. 116 (2023), https://doi.org/10.1016/j.jfca.2022.105065, 0–2.

[15] M.R. Baqueta, P. Valderrama, A. Alves, Discrimination of Robusta Amazônico Co Ff Ee Farmed by Indigenous and Non-indigenous People in Amazon : Comparing Benchtop and Portable NIR Using ComDim and Duplex, 2023, https://doi.org/10.1039/d3an00104k.

[16] 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.

[17] UCDA, UCDA - Uganda Coffee Development Authority and CQI - Coffee Quality Institute, Fine Robusta Standards and Protocols: Technical Standards, Evaluation Procedures and Reference Materials for Quality-Differentiated Robusta Coffee, 2010. https://www.coffeestrategies.com/wp-content/uploads/2015/04/compiled-standards-distribute1.1.pdf. (Accessed 1 February 2020).

[18] M.R. Baqueta, A. Coqueiro, P.H. Março, M. Mandrone, F. Poli, P. Valderrama, Integrated 1H NMR fingerprint with NIR spectroscopy, sensory properties, and quality parameters in a multi-block data analysis using ComDim to evaluate coffee blends, Food Chem. 355 (2021), 129618, https://doi.org/10.1016/j.foodchem.2021.129618.

[19] A. Coqueiro, M.R. Baqueta, M. Mandrone, F. Poli, P. Valderrama, Coffee assessment using 1H NMR spectroscopy and multivariate data analysis: a review, in: F. Atta-ur-Rahman, M.I. Choudhary (Eds.), Appl. NMR Spectrosc, Bentham Science Publishers, 2021, pp. 35–58. https://books.google.com.br/books/about/Applications_of_NMR_Spectroscopy_Volume.html?id=u2lZEAAAQBAJ&redir_esc=y.

[20] M.R. Baqueta, F.P. Pizano, J.D. Villani, S.J.H. Toro, A.P.A. Bragotto, P. Valderrama, J.A.L. Pallone, Kurtosis-based projection pursuit analysis to evaluate South American rapadura, Food Chem. 368 (2022), 130731, https://doi.org/10.1016/j.foodchem.2021.130731.

[21] Inmetro Brazil, The national institute of metrology, stardardization and industrial quality, DOQ-CGCRE-008, ReVision (2020), 09 (p. 28.

[22] F. Westad, F. Marini, Validation of chemometric models - a tutorial, Anal. Chim. Acta 893 (2015) 14–24, https://doi.org/10.1016/j.aca.2015.06.056.

[23] T. Næs, O. Tomic, B.H. Mevik, H. Martens, Path modelling by sequential PLS regression, J. Chemom. 25 (2011) 28–40, https://doi.org/10.1002/cem.1357.

[24] V. Giannetti, M.B. Mariani, F. Marini, P. Torrelli, A. Biancolillo, Grappa and Italian spirits: multi-platform investigation based on GC–MS, MIR and NIR spectroscopies for the authentication of the Geographical Indication, Microchem. J. 157 (2020), 104896, https://doi.org/10.1016/J.MICROC.2020.104896.

[25] A. Biancolillo, T. Næs, The sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions, Data Handling Sci. Technol. 31 (2019) 157–177, https://doi.org/10.1016/B978-0-444-63984-4.00006-5.

[26] J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, CovSel: variable selection for highly multivariate and multi-response calibration. Application to IR spectroscopy, Chemometr. Intell. Lab. Syst. 106 (2011) 216–223, https://doi.org/10.1016/j.chemolab.2010.10.003.

[27] C. Ciaramelli, A. Palmioli, C. Airoldi, Coffee variety, origin and extraction procedure: implications for coffee beneficial effects on human health, Food Chem. 278 (2019) 47–55, https://doi.org/10.1016/J.FOODCHEM.2018.11.063.

[28] M.N. Clifford, S. Marks, S. Knight, N. Kuhnert, Characterization by LC-MS n of four new classes of p-coumaric acid-containing diacyl chlorogenic acids in green coffee beans, J. Agric. Food Chem. 54 (2006) 4095–4101, https://doi.org/10.1021/jf060536p.

[29] D. Perrone, A. Farah, C.M. Donangelo, T. de Paulis, P.R. Martin, Comprehensive analysis of major and minor chlorogenic acids and lactones in economically relevant Brazilian coffee cultivars, Food Chem. 106 (2008) 859–867, https://doi.org/10.1016/J.FOODCHEM.2007.06.053.

[30] A. Farah, M.C. Monteiro, V. Calado, A.S. Franca, L.C. Trugo, Correlation between cup quality and chemical attributes of Brazilian coffee, Food Chem. 98 (2006) 373–380, https://doi.org/10.1016/J.FOODCHEM.2005.07.032.