Origin geographical classification of green coffee beans (*Coffea arabica* L.) produced in different regions of the Minas Gerais state by FT-MIR and chemometric

Geissy de Azevedo Mendes [a], Marcone Augusto Leal de Oliveira [b], Mirian Pereira Rodarte [c], Virgílio de Carvalho dos Anjos [a], Maria Jose Valenzuela Bell [a],*

[a] *Department of Physics, Institute of Exact Sciences, Federal University of Juiz de Fora, Rua José Lourenço Kelmer, s/n São Pedro, 36036-900, Juiz de Fora, MG, Brazil*
[b] *Department of Chemistry, Institute of Exact Sciences, Federal University of Juiz de Fora, Rua José Lourenço Kelmer, s/n São Pedro, 36036-900, Juiz de Fora, MG, Brazil*
[c] *Department of Pharmaceutical Sciences, Faculty of Pharmacy, Federal University of Juiz de Fora, Rua José Lourenço Kelmer, s/n São Pedro, 36036-900, Juiz de Fora, MG, Brazil*

ARTICLE INFO

ABSTRACT

The present work was proposal the potential evaluation of Fourier-Transform Mid-Infrared (FT-MIR) associated with chemometric approach in green beans, in order to discriminate the origin of special Arabica coffees in a single state that has heterogeneous environments. Partial Least Squares Discriminant Analysis (PLS-DA) model presented as result: 3 latent variables, R^2X (cum) = 0.892, R^2Y (cum) = 0.659; Q^2Y (cum) = 0.494, RMSEP = 0.182387, p-value CV-Anova = 0.009, 100% of both sensitivity and specificity and the prediction classification obtained was: 100, 83.33, 100, 83.33% for class 1, class 2, class 3 and class 4, respectively. These results can be considered adequate for the proposed hypothesis. The obtained results that the regions have markers such as trigonelline, chlorogenic and fatty acids, sensitive to absorption in the mid-infrared and that are able to determine the origin of green coffee beans of Arabica. Thus, the FT-MIR associated with chemometrics has the potential to employ speed, modernity and cost reduction in the certification of origin of coffees.

1. Introduction

Coffee is the beverage consumed worldwide and one of the most marketed. There are dozens of known botanical species, but only two are produced on commercial scales: *Coffea arabica* L. (Arabica coffee) and *Coffea canephora* (Robusta coffee). Most of the coffee drinks prepared in the world are produced in from the cultivar *Coffea arabica* L., due to their sensory properties highly appreciated, thus achieving high prices on the markets of export (International Coffee Orga, 2021). In each species, there are different genotypes, which have a prominent factor in the chemical composition of the beans and in the production of specialty coffees (Figueiredo et al., 2019).

Specialty coffees are obtained from the production of Arabica coffee. Its characteristics can be related to the genotype, but mainly to the appropriate geographical environment, in addition to other attributes that add quality to the bean, such as type of harvest, processing, drying (Obeidat et al., 2018; Barrios-Rodríguez et al., 2021).

Green Arabica coffee beans are mainly composed of carbohydrates (59–61%), lipids (11–17%), proteins (10–16%), phenols (6–10%), minerals (4%), fatty acids (2%), caffeine (1–2%), trigonelline (1%) and free amino acids (<1%). During the process of roasting carbohydrates (38–42%), proteins (8–14%), phenols (3–4%) and free amino acids are reduced. On the other hand, there are fewer changes in minerals (5%), fatty acids (3%), caffeine (1–2%), lipids (11–17%) and trigonelline (1%) (Hu et al., 2019).

The amount of these compounds can vary for a number of reasons, including the interaction between the genotype and the geographic environment. Authors (Bessada et al., 2018; Tasew et al., 2020; Worku et al., 2018) point out that environmental factors such as altitude, climate, soil, biome, temperature, insolation and precipitation have a greater impact on the appearance of chemical compounds present in the raw bean and which are fundamental for the production of specialty coffees.

For example, coffees of the same genotype, produced at higher

altitudes, show increased levels of sugars, such as sucrose, lipids, amino acids, trigonelline and chlorogenic acid isomers (Marquetti et al., 2016), consequently, coffees from higher quality are found precisely in these places. Meanwhile, coffees of the same genotype grown at lower altitudes have lower quality in sensory analysis and the amount of these compounds is reduced. Thus, the bean production altitude and region is a decisive factor in the final composition.

Studies indicate the metabolite profiles as a powerful tool to describe the small differences between quality coffees of the same species, cultivated in close origins and that these have recognized potential as sensory quality descriptors of beans and regions. Among the existing metabolites in beans, some are already labelled as potential markers of the region, they are: i) chlorogenic acid isomers, arising from the esterification processes between quinic acid and derivatives trans-cinnamic acid, such as caffeic, p-coumaric and ferulic (Mehari et al., 2016); ii) trigonelline; iii) fatty acids such as linoleic (C18:2cc) and oleic acid (C18:1c). Other markers such as caffeine and sugars are usually associated with bean quality rather than geographic origin.

Furthermore, the interest in local and quality food, with certificates of origin and production environment, is becoming more and more coveted in the coffee sector. However, currently, traceability is paper-based and cup quality verification by a panel of trained tasters is used to identify the region of origin, but papers can be easily counterfeited and it can be complex for tasters to safely differentiate, coffees from different origins (Mehari et al., 2016; Kamiloglu, 2019), in addition to being a highly expensive process.

In line with these trends, the applicability of spectroscopic techniques in coffees has been evaluated. In the literature, the study of pure compounds, isolated and/or in addition to solvents in techniques such as UV–Vis spectroscopy, high performance liquid chromatography, fluorescence, time-resolved fluorescence, mid and near infrared (MIR and NIR) were successful (Núñez et al., 2021a; Talamond et al., 2015; Caporaso et al., 2018). There are also studies aimed at differentiating the addition of adulterants in coffee beans, such as corn, barley, husks or even a different species of coffee, using spectroscopic techniques such as UV–Vis and fluorescence (Dankowska et al., 2017; Núñez et al., 2021b), Raman (Figueiredo et al., 2019; Abreu et al., 2019), NIR (Marquetti et al., 2016; Correia et al., 2018) and MIR (Bona et al., 2017).

However, few studies are intended to differentiate coffees of the same species from neighboring regions (Botelho et al., 2017; Zhu et al., 2021), highlighting the lack of studies in the area. Furthermore, none of them dealt with such differentiation in specialty coffees, which represents the fastest-growing segment in the coffee market. And in this sector, the determination and certification of origin is increasingly desired, becoming essential.

In this sense, spectroscopic techniques have the potential to meet the demands of the coffee market, with fast, biologically adequate and safe techniques, which will minimize the work, making it faster, in addition to adding quality, credibility and value to the final product. Therefore, the present work restricted the geographic area of study to the state of Minas Gerais (Brazil), which has geographic regions with environmental characteristics (Fig. 1), but which also express high sensory quality.

Since this study seeks the development of technological processes applicable in large scale, it was defined, through factorial planning, a sample group in order to optimize the experiments and get reliable results.

Therefore, the objective of this work was to verify the feasibility of geographic segmentation of coffees grown in the State of Minas Gerais by a clean and fast methodology, based on FT-MIR in conjunction with data analysis software multivariate (SIMCA®).

2. Material and methods

2.1. Samples

The raw materials used in this work were green Arabica coffee beans (*Coffea arabica* L.) collected from farms in the four main producing regions of the state of Minas Gerais (Fig. 1), located in Cerrado cities (16° 35′ to 20° 13' S and 45° 20′ to 49° 48′ W), Matas (18° 35′ to 21° 26′ S and 40° 50′ to 43° 36′ W), North (17° 05′ to 18° 09′ S and 40° 50′ to 42° 50′ W) and South (21° 13′ to 22° 10′ S and 44° 20′ to 47° 20′ O).

The choice of the sample group was based on a factorial design (Fig. 2), whose abscissa axis, represents the system variables or regions (C, M, N and S) and on the ordinate axis he levels, authentic replicas for each region.

Random genotypes were chosen for the regions (Catuaí Amarelo,

Fig. 1. Identification of the main coffee producing regions on the map of the State of Minas Gerais: Cerrado in blue, Matas in green, North in yellow and South of Minas in red, their respective altitudes and climates (- Brazilian Institut, 2014). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
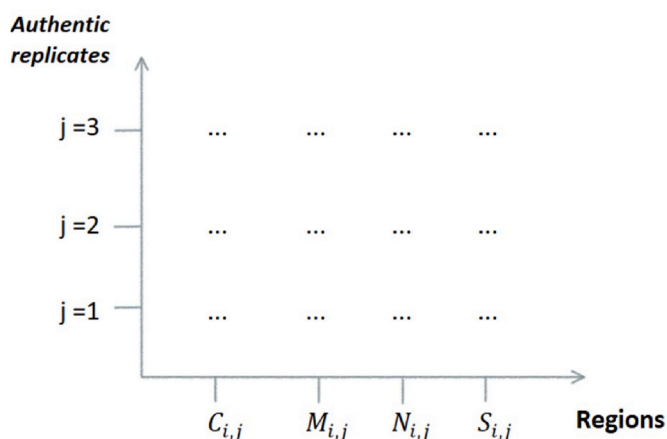
Fig. 2. Schematic representation of the system's factorial planning.

Catiguá MG2, Topázio MG 1190 and New World) and among them, the Yellow Catuaí was common to all regions. The yellow Catuaí cultivars are small (caturra type) and fruit with yellow exocarp (shell). The quality and productivity of the drink is excellent, and adapt well to the regions of the state.

2.2. Sample preparation

The samples come from twelve cities of Minas Gerais, which represent the four regions of the state. In each city, three farms provided coffee samples, totaling 36 samples. Each farm supplied about 2.0 kg of coffee. All beans were harvested mature and submitted natural drying (dry in coconut). All with soft drink rating or higher ($\geq$80 points). After harvesting, the beans were stored equally in a cold chamber at 10 °C until the time of measurements. The bean moisture was around 11% and the defective beans were previously removed for analysis.

The green beans were ground in a mill (Micro Knife Mill - SL-30 - sieve of 20 mesh stainless steel – SOLAB). Then, the samples were pressed with three tons for 3 s with an automated hydraulic press (AtlasTM Power Hydraulic Press T25, SPECAC INC, Fort Washington, PA, USA), transforming them into inserts with a mass of 0.2 g, thickness of 1.4 mm and diameter of 12.7 mm and measurements were taken immediately after its making.

Sensory analyzes of the beans were performed by trained, internationally certified tasters. All procedures were performed according to the protocol described by the Specialty Coffee Association of America (SCAA). In total, 100 g were subjected to the roasting process, about 9 min at 195 °C, which provide a middle roasting point, between the colors determined by discs 65 for ground bean and 55 for whole bean (SCAA/Agtron Roast Color Classification System). The results of these analyzes are shown in Table 1, in addition to the geographic characteristics of each region.

2.3. Instrumentation

The FT-MIR measurements were performed on a VERTEX-70 spectrometer manufactured by Bruker in the Attenuated Total Reflectance (ATR) mode. Measurements were performed on samples prepared as described in item 2.2, in authentic triplicates. In the regions between 4000 cm^{-1} and 400 cm^{-1}, with a spectral resolution of 0.1 cm^{-1}.

2.4. Data treatment and method validation

The spectra were acquired using the OPUS 6.5 Bruker Optik Gmbh 2009 software. The absorbance spectral data obtained from the MIR were organized into matrices, in which each row represented a sample and each column represented a distinct variable (wavenumber). The analyzes of the MIR spectra were carried out in the regions between 3000 and 2800 cm^{-1} and between 1800 and 1400 cm^{-1}, totaling 600 variables. The regions between 2800 and 1800 cm^{-1} and between 1400 and 400 cm^{-1} were excluded from the analysis, as the first is related to the moisture content in the environment and the second is the fingerprint of the molecules, which, because it is the same material, they were similar and did not discriminate between regions.

After checking the assumptions of normality and homoscedasticity, the results were submitted to PLS-DA using the SIMCA-P+ software (version 12, Umetrics, Tvistevägen, SE-907 19 Umeå, Sweden). Calculations were performed for $\alpha = 0.05$. The sampling was carried out by the method of the "blinds", that is, the procedure consisted of ordering the reference values in ascending order and choosing one sample of every three for the prediction set was used. The developed model was validated internally using the leave-one-out (Shao, 1993) cross-validation method (24 samples) and validation (12 samples). It is important to highlight that the variables were scaled and mean-centered. The pre-processing filters, such as multiplicative scatter correction (MSC) and standard normal variate (SNV) were verified for observations, but no significant improvements in the predictive ability of the models were obtained. As there was no preprocessing in X matrix, the spectral data were kept in original dimension. PLS-DA model was

Table 1
Characterization of the evaluated coffee samples.

Region	Sample	Temperature average (°c)	Average annual Precipitation (mm)	Biome	Sensory analysis	Genotype
Cerrado	C1	20.3	1552	100% Cerrado	84.0	Yellow Catuaí
	C2	23.7	1150		80.7	Yellow Catuaí
	C3	21.2	1408		84.4	Topátizo MG 1190
Matas	M1	20.3	1261	100% Atlantic Forest	81.3	Yellow Catuaí
	M2	19.4	1261		81.5	Yellow Catuaí
	M3	20.3	1261		80.9	Yellow Catuaí
North	N1	20.8	1096	46% Cerrado 54% Atlantic Forest	81.3	Yellow Catuaí
	N2	20.8	1096	46% Cerrado 54% Atlantic Forest	81.3	Catiguá MG2
	N3	20.5	1078	43% Cerrado 57% Atlantic Forest	81.0	Yellow Catuaí
South	S1	20.3	1429	100% Atlantic Forest	80.3	Yellow Catuaí
	S2	20.8	1460	16% Cerrado 84% Atlantic Forest	81.2	New World
	S3	20.3	1429	100% Atlantic Forest	81.4	New World

Source: (- Brazilian Institut, 2014; E-G: Weather, 2021; InfoSanbas - Information, 2021).

evaluated considering the root-mean square-error of prediction (RMSEP), R^2X, R^2Y, Q^2Y, p-value from Analysis of Variance testing of Cross-Validated predictive residuals (CV-ANOVA), sensitivity, false negatives, specificity, false positives. CV-ANOVA is a diagnostic tool for assessing the reliability of PLS-DA models introduced in SIMCA-P+ version 12 (Eriksson et al., 2008). CV-ANOVA can be seen as a formal test of the significance of the Q^2Y_{CV} using the F-distribution. CV-ANOVA is fast and has a minimal additional computation time beyond the one of the standard cross-validation. This is in contrast to response permutation testing which is time consuming with large data sets and many components.

3. Results and discussions

3.1. Fourier-Transform Mid-Infrared spectroscopy

In mid-infrared spectra it is possible to identify the compounds responsible for each absorption. In coffee beans, these absorptions are normally bands around 3000 cm^{-1}, which are mainly related to asymmetric CH$_2$ stretches between 2921 and 2908 cm^{-1}, characteristic of lipids (Reis et al., 2013; Craig et al., 2014) and the CH and CH$_3$ symmetric stretch between 2877 and 2850 cm^{-1}, characteristic of caffeine (Craig et al., 2014; Paradkar and Irudayaraj, 2002). There are also bands between 1759 and 1722 cm^{-1}, attributed to carbonyl (CO) vibration in esters (triglycerides) and aldehydes (Reis et al., 2013), mainly lipids and free fatty acids (Vodnar et al., 2010).

Fatty acids are known to be important components of coffee flavour and aroma. By investigating the relationship between fatty acid composition and sensory characteristics of different Bourbon genotypes, grown in different edaphoclimatic conditions, it was possible to describe the relationship of these compounds (fatty acids) with the quality of the drink. Saturated fatty acids, including arachidic, stearic and palmitic acids are potential discriminators of the quality of specialty coffees, indicating better sensory quality. On the other hand, unsaturated fatty acids, including elaidic, oleic, linoleic and linolenic acids may be related to coffees with less intense acidity, fragrance, body and flavour (Figueiredo et al., 2015).

These acids are good markers and can differentiate coffee growing environments (Avelino et al., 2005). Chemical analyses indicate that oleic, linoleic and linolenic acids strongly contribute to environmental discrimination in a city in Minas Gerais (Santo Antônio do Amparo) (Figueiredo et al., 2015; Joët et al., 2010). Furthermore, other fatty acids (palmitic, margaric, arachidonic and eicosanoic and stearic) were also studied and showed high potential to differentiate the environments in which coffees were grown (Bertrand et al., 2008). Of these, linoleic acid was the only potential marker to differentiate coffee samples from the three environments (Figueiredo et al., 2015).

Importantly, this ability of fatty acids to discriminate the origin of the crop has also been demonstrated in other fruits and beans, for example, pistachios (Arena et al., 2007) and olives (Ollivier et al., 2006).

Peaks in the range between 1780 and 1600 cm^{-1} correspond to the carbonyl elongation (C═O) of the vibration of organic compounds, which can be attributed to several compounds, such as carotenoids, chlorogenic acids, alkaloids, polysaccharides, hemicellulose, among others (Sanchez et al., 2018). Trigonelline has several bands in the range of 1650 to 1400 cm^{-1}, which is present in raw and roasted coffee and can be attributed to the axial deformation of the C═C and C═N bonds in the trigonelline aromatic ring (Reis et al., 2013).

Trigonelline and 3-CQA (chlorogenic acid isomer) have higher amounts at altitudes above 1200 m. Raw beans have higher metabolite contents and, consequently, a drink with high scores in sensory analysis. Thus, coffees grown below 1200 m, with a slope facing the sun, have a tendency in raw coffee beans to lower levels of trigonelline and sensory beverage quality with scores below 85 points (Ribeiro et al., 2016). The trigonelline content already allowed the discrimination between the studied environments (Figueiredo et al., 2013).

The 5-CQA does not show a significant interaction between genotype and environment (Figueiredo et al., 2013). However, this isomer is susceptible to different environments, especially altitude, and therefore, the results that show differences regarding this compound are solely due to the environment, not the genotype, making it an excellent marker for regions.

In the region between 1645 and 1610 cm^{-1} there is the presence of NH$_2$, NH and folding of amide II and lactone, originating from the protein (Craig et al., 2014). In the ranges between 1790 and 1310 cm^{-1}, containing C═O, extending vibrations of protein-bound amides (first band of amide 1645 cm^{-1}) and CH$_3$, CH$_2$, asymmetric and symmetric deformations (flexion) of proteins.

In organic materials, the next region of the spectrum is known as "fingerprint", located between 1300 and 900 cm^{-1} and is characterized by vibrational characteristics of proteins, nucleic acids, cell membrane and cell wall components. They can be associated with molecular bonds or with a particular functional group (Vodnar et al., 2010), characterized by vibrations of various types of bonds, such as C–H, C–O and C–N.

In this region, the range between 1165 and 1138 cm^{-1} can be highlighted, characteristic of the C–O stretch, coming from polysaccharides and cellulose. Furthermore, at peaks at 1039 cm^{-1}, there is the C–O stretch for cellulose and 1099 cm^{-1} for carbohydrates (Craig et al., 2014). The region also includes glucose and fructose bands, with specific markers located around 1029 cm^{-1} (Vodnar et al., 2010). The region between 835 and 800 cm^{-1} is responsible for the out-of-plane CH folding of phenolic compounds. However, as the entire study was carried out in coffees, this final part of the spectrum does not contribute to the differentiation of regions.

The FT-MIR spectra of the samples of green Arabica coffee beans evaluated in this work are shown in Fig. 3, which illustrate the main absorption bands responsible for the differentiation of the regions.

In Fig. 3, it can be seen that all samples absorb, in general, in the same regions, but with different intensity. This can be explained by the fact that the samples have the same compounds, only in different amounts, which will exactly be responsible for the differentiation of the samples. Some patterns can be observed in the spectra, such as the fact that the regions with lower absorption intensities are always from the Cerrado . In contrast, those responsible for the highest absorption intensities vary from one peak to another.

Among all these absorption regions, the bands between 3001 and 2800 cm^{-1} and 1800 - 1400 cm^{-1} are responsible for the absorption of compounds that are directly linked to the origin and/or quality of the beans. In these regions, the absorption of lipids is identified, among them, fatty acids such as linoleic acid (C18:2 cc), palmitic acid (C16:0), stearic acid (C18:0), arachidic acid (C20:0) and oleic acid (C18:1c). In addition to the peaks linked to trigonelline and chlorogenic acids, which are equally important in discrimination, because at higher altitudes, there is a greater presence of these compounds (Figueiredo et al., 2015; Avelino et al., 2005; Joët et al., 2010; Bertrand et al., 2008).

Also in this first part, there is a peak at 2852 cm^{-1}, characteristic of the presence of caffeine. However, caffeine is not considered a good marker because it does not depend on the region where it is produced. This fact could make the analysis difficult, as in this work only special coffees were used. On the other hand, higher quality coffees commonly have lower caffeine content (Figueiredo et al., 2018). This fact corroborates the results, as two of the three samples from the Cerrado that had lower absorption intensity are also the samples with the best scores in sensory analyses.

To validate the observations described in this section, a multivariate statistical analysis was necessary due to the similarity and complexity of the spectra. PLS-DA was used to differentiate the green beans into regions, using only the spectra of each sample, see Table 1.

To determine the best results, the PLS-DA was performed between 3001 and 2800 cm^{-1} and 1800-1400 cm^{-1}. The loadings graph (Fig. 4) confirms that in these regions there is absorption of compounds
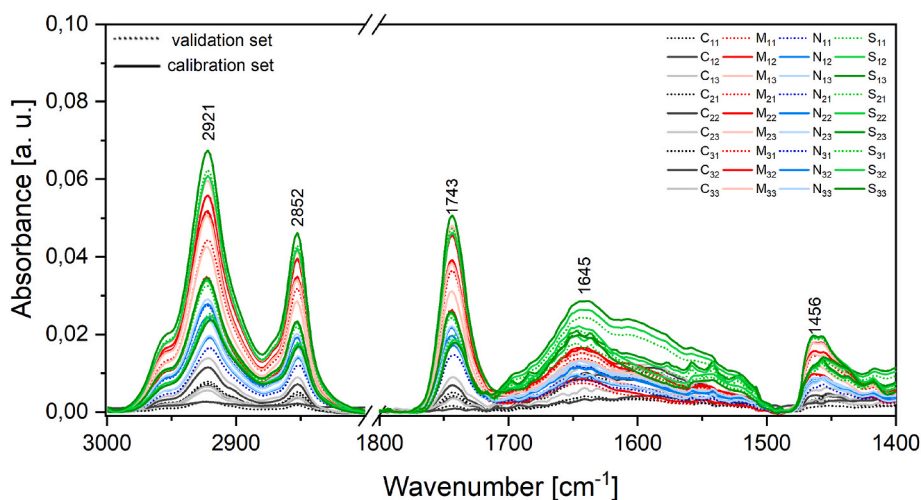
Fig. 3. Spectra obtained in the FT-MIR of samples from the Cerrado (C), South (S), North(N), Matas (M) of green beans. . (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
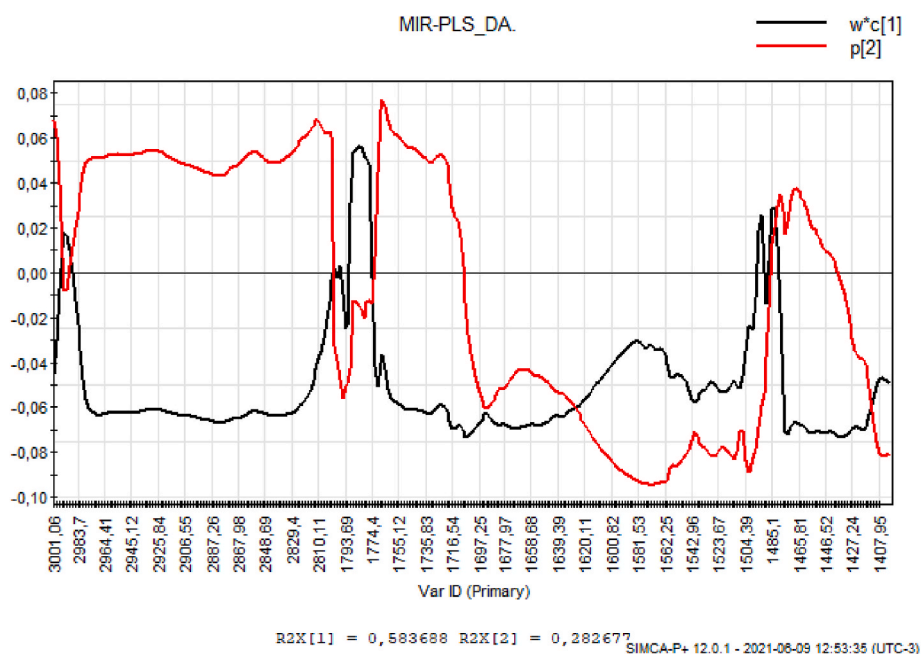

Fig. 4. PLS-DA loadings: Graph of the influence of wavenumbers.

considered to be good region markers, important in discrimination.

The bands between 3000 and 2800 cm^{-1}, 1774 to 1716 cm^{-1} and 1485 to 1427 cm^{-1} can be related to fatty acids, chlorogenic and trigonelline.

In the PLS-DA of green beans, it was possible to group the samples and differentiate each of the regions under study (Fig. 5 a). The first two latent variables explain more than 80% of the results (tPS (International Coffee Orga, 2021) 58% and (tPS (International Coffee Orga, 2021) 28%).

Fig. 5a illustrates the grouping of each of the regions, on the opposite sides of the tPS 1 axis, there is South (green), on the negative part, and Cerrado (black), on the positive side of the axis, the indicating the low correlation between these two regions. Then, on the tPS2 axis, on the positive side, samples from the North (blue) and Matas (red), which show the greatest correlation, but it is still possible to distinguish them.

In order to verify the efficiency of the method, part of the samples were separated to validate the results. From each group, a sample was

taken to validate the results. They are represented in the graph in grey. Note that the method showed good prediction for unknown samples, since PLS-DA model presented as result: 3 latent variables, R^{2}X (cum) = 0.892, R^{2}Y (cum) = 0.659; Q^{2}Y (cum) = 0.494, RMSEP = 0.182387, p-value CV-Anova = 0.0090676 (the common practice is to interpret a p-value lower than 0.05 as denoting a significant model). In Table 2 is showed the CV-Anova results (Eriksson et al., 2008).

The model built have shown 100% of both sensitivity and specificity since all origin geographical regions were correctly classified, according to shown in Fig. 5, item (a), classes highlight in grey colour, such as C1-1, C2-1, C3-1, M1-1, M2-1, M3-1, N1-1, N2-1, N3-1, S1-1, S2-1 and S3-1. The prediction classification obtained was: 100, 83.33, 100, 83.33% for class 1, class 2, class 3 and class 4, respectively. These results can be considered adequate for the analysis performed. Besides, permutation test was performed with 100 permutations and by the results obtained was not observed model overfitting (not shown).

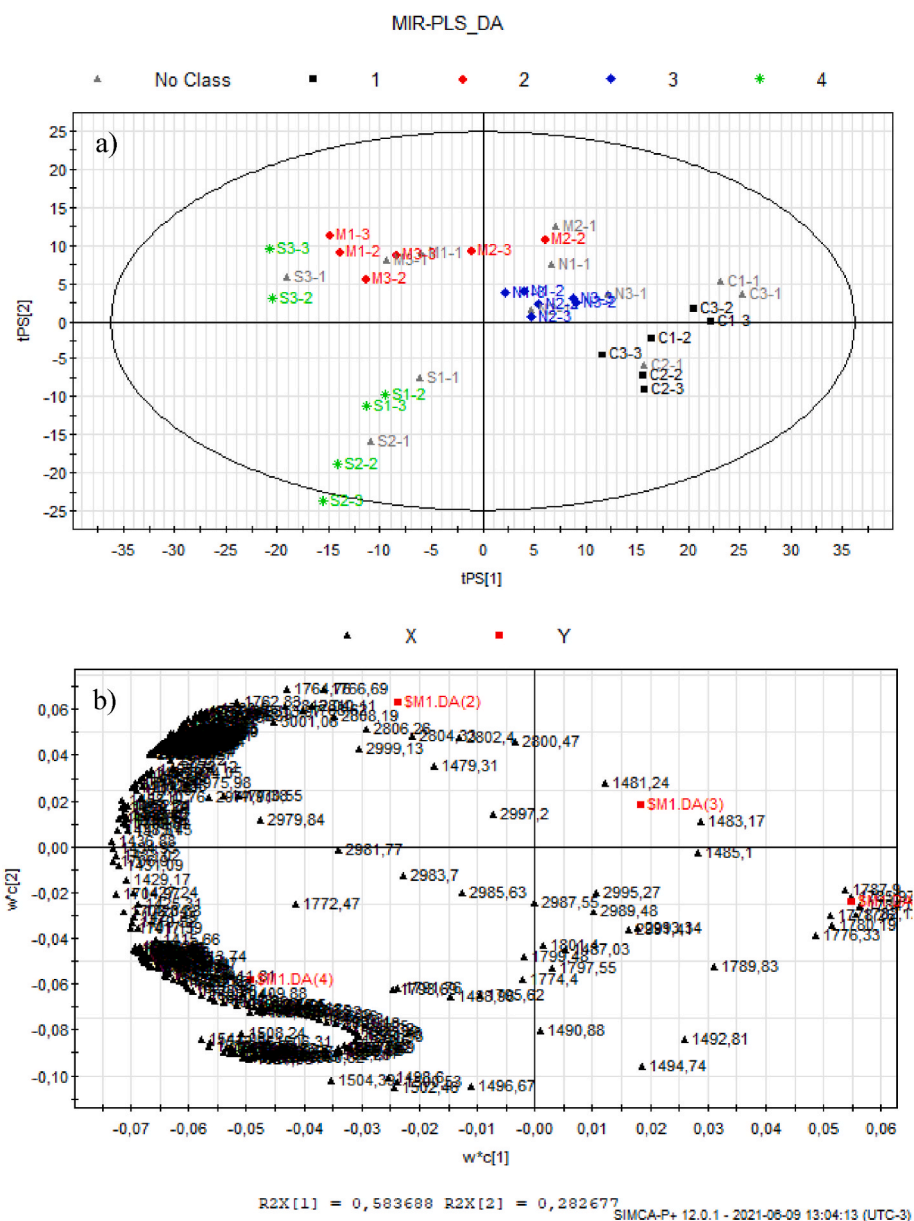In the loadings (Fig. 5b) it is possible to identify the wave numbers

Fig. 5. PLS-DA Model: a) Graph of correlation between samples (scores). Calibration set: Group 1 (black) - Cerrado; Group 2 (red) - Matas; Group 3 (Blue) – North; Group 4 (green) - South. Validation Set: No Class in grey. b) Graph of the influence of wave numbers (loadings). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
CV-ANOVA results.

PLS-DA Model	SS	DF	MS	F	p-value	SD
Total corr.	69	69	1			1
Regression	31.2093	18	1.73385	2.3399	0.0090676	1.31676
Residual	37.7907	51	0.740993			0.86081

that influence each grouping of regions, and that they coincide with the values described in the work.

The evaluation of coffee based on geographic origin has already been successfully carried out using a combination of FT-MIR and unsupervised multivariate analysis (PCA) for different countries, whose differences are very marked (Obeidat et al., 2018). However, in this work, regions within the same state were used, that is, small variability. It was also used a supervised analysis and even so, the results were highly

satisfactory, showing the efficiency of the technique in differentiating regions for green beans.

Furthermore, several studies report that the bean genotype can decisively influence differentiation. However, in this work, each region has at least one sample of the same genotype (catuaí), but this did not interfere with the results, because even if, for example, Cerrado and South had a sample of Catuaí, these samples did not present greater correlation.

The correlation between Matas and North in tPS(International Coffee Orga, 2021) may be associated with the relationship of similarity between environmental characteristics and drink quality in the Matas to the Jequitinhonha region (North), with low moisture deficit and water accumulation in the areas of planting and drying.

Also, the best scores are for altitudes above 850 m. Therefore, the altitude can be divided into three groups (below 850 m; between 851 and 950 m; above 950 m), as shown in Fig. 1. Similarly, the coffee score as well (between 80.9 and 80.9; between 81.0 and 83.9; above 84.0).

Both groups make the samples from Matas and North present similar characteristics in their composition. Another important factor is the biome of each region, in which samples from the Cerrado are exclusively from the Cerrado biome. The samples from the Matas and the South are from the Atlantic Forest biome, only one from the South has less than 20% of the Cerrado, that is, the Atlantic Forest is also predominant in the South. However, samples from the North have distinct biomes, with a percentage between 40% and 50% of the Cerrado, but the highest percentage is from the Atlantic Forest. Adding one more reason why the samples from the North and the Matas have a relationship, as the Matas have a unique Atlantic Forest biome.

In summary, the results lead us to conclude that the FT-MIR associated with PLS-DA, in the regions between 3001 and 2800 cm^{-1} and 1800-1400 cm^{-1}, which represent the absorption of compounds such as trigonelline, chlorogenic and fatty acids, among others, are efficient in discriminating the origin of green Arabica coffee beans and that these compounds were shown to be important markers of the origin of bean production. The technique has the potential to be used efficiently in the sale and export of green beans, as it is able to quickly guarantee the origin of the beans, even within the same state, making inspection and certification of the beans at any time.

4. Conclusions

The results showed that FT-MIR associated with multivariate analysis is efficient in determining the origin of Arabica green beans of the same species from neighboring regions, evidenced by the clear distinction of the four regions: first from Cerrado and South with low correlation, on opposite sides in the first component, followed by North and Matas, more central in the first component, but in different quadrants in the second component. These separations are related to the unique environmental characteristics of each region, which is measured in the absorption of compounds such as chlorogenic acids, fatty acids and trigonelline. Furthermore, by submitting some samples to the developed method, a good prediction was obtained.

Thus, it is possible to certify green Arabica coffee beans by FT-MIR, after harvesting, since the production regions leave unique markers, enabling the possibility of certification of origin and its origin in a quick and safe way.

Credit author statement

Geissy Mendes: Investigation, Methodology, Chemometrics, Formal analysis, Validation. Marcone Augusto Leal de Oliveira:Design of the experiment, Chemometrics; Conceptualization. Mirian Rodarte: Conceptualization, Formal analysis, Methodology. Virgílio de Carvalho dos Anjos: Conceptualization, Founding acquisition, Resources. Maria Jose Valenzuela Bell: Writing, Project administration, Supervision

Declaration of competing interest

Maria Jose Valenzuela Bell reports financial support was provided by National Council for Scientific and Technological Development. Maria Jose Valenzuela Bell reports financial support was provided by Minas Gerais State Foundation of Support to the Research.

References

IBGE - Brazilian Institute of Geography and Statistics: Geographical Location: Mesoregions and Microregions, 2014.

Abreu, G.F., Borém, F.M., Oliveira, L.F.C., Almeida, M.R., Alves, A.P.C., 2019. Raman spectroscopy: a new strategy for monitoring the quality of green coffee beans during storage. Food Chem. 287, 241–248.

Arena, E., Campisi, S., Fallico, B., Maccarone, E., 2007. Distribution of fatty acids and phytosterols as a criterion to discriminate geographic origin of pistachio seeds. Food Chem. 104, 403–408.

Avelino, J., Barboza, B., Araya, J.C., Fonseca, C., Davrieux, F., Guyot, B., Cilas, C., 2005. Effects of slope exposure, altitude and yield on coffee quality in two altitude terroirs of Costa Rica, Orosi and Santa María de Dota. J. Sci. Food Agric. 85, 1869–1876.

Barrios-Rodríguez, Y.F., Rojas Reyes, C.A., Triana Campos, J.S., Girón-Hernández, J., Rodríguez-Gamir, J., 2021. Infrared spectroscopy coupled with chemometrics in coffee post-harvest processes as complement to the sensory analysis. LWT (Lebensm.-Wiss. & Technol.) 145.

Bertrand, B., Villarreal, D., Laffargue, A., Posada, H., Lashermes, P., Dussert, S., 2008. Comparison of the effectiveness of fatty acids, chlorogenic acids, and elements for the chemometric discrimination of coffee (Coffea arabica L.) varieties and growing origins. J. Agric. Food Chem. 56, 2273–2280.

Bessada, S.M.F., Alves, R.C., Costa, A.S.G., Nunes, M.A., Oliveira, M.B.P.P., 2018. Coffea canephora silverskin from different geographical origins: a comparative study. Sci. Total Environ. 645, 1021–1028.

Bona, E., Marquetti, I., Link, J.V., Makimori, G.Y.F., da Costa Arca, V., Guimarães Lemes, A.L., Ferreira, J.M.G., dos Santos Scholz, M.B., Valderrama, P., Poppi, R.J., 2017. Support vector machines in tandem with infrared spectroscopy for geographical classification of green arabica coffee. LWT - Food Sci. Technol. (Lebensmittel-Wissenschaft -Technol.) 76, 330–336.

Botelho, B.G., Oliveira, L.S., Franca, A.S., 2017. Fluorescence spectroscopy as tool for the geographical discrimination of coffees produced in different regions of Minas Gerais State in Brazil. Food Control 77, 25–31.

Caporaso, N., Whitworth, M.B., Grebby, S., Fisk, I.D., 2018. Non-destructive analysis of sucrose, caffeine and trigonelline on single green coffee beans by hyperspectral imaging. Food Res. Int. 106, 193–203.

Correia, R.M., Tosato, F., Domingos, E., Rodrigues, R.R.T., Aquino, L.F.M., Filgueiras, P. R., Lacerda, V., Romão, W., 2018. Portable near infrared spectroscopy applied to quality control of Brazilian coffee. Talanta 176, 59–68.

Craig, A.P., Franca, A.S., Oliveira, L.S., Irudayaraj, J., Ileleji, K., 2014. Application of elastic net and infrared spectroscopy in the discrimination between defective and non-defective roasted coffees. Talanta 128, 393–400.

Dankowska, A., Domagała, A., Kowalewski, W., 2017. Quantification of Coffea arabica and Coffea canephora var. robusta concentration in blends by means of synchronous fluorescence and UV-Vis spectroscopies. Talanta 172, 215–220.

CLIMATE-DATA, 2021. ORG: Weather Data for Cities Worldwide.

Eriksson, L., Trygg, J., Wold, S., 2008. CV-ANOVA for significance testing of PLS and OPLS® models. J. Chemom. 22, 594–600.

Figueiredo, L.P., Borém, F.M., Cirillo, M.Â., Ribeiro, F.C., Giomo, G.S., Salva, T.D.J.G., 2013. The potential for high quality bourbon coffees from different environments. J. Agric. Sci. 5, 87–98.

Figueiredo, L.P., Borem, F.M., Ribeiro, F.C., Giomo, G.S., Taveira Jh da, S., Malta, M.R., 2015. Fatty acid profiles and parameters of quality of specialty coffees produced in different Brazilian regions. Afr. J. Agric. Res. 10, 3484–3493.

Figueiredo, L.P., Borém, F.M., Ribeiro, F.C., Giomo, G.S., Malta, M.R., Taveira Jh da, S., 2018. Sensory analysis and chemical composition of 'bourbon' coffees cultivated in different environments. Coffee Sci 13, 122–131.

Figueiredo, L.P., Borém, F.M., Almeida, M.R., Oliveira, LFC de, Alves, AP. de C., Santos, CM dos, 2019. Raman spectroscopy for the differentiation of Arabic coffee genotypes. Food Chem. 288, 262–267.

Hu, G.L., Wang, X., Zhang, L., Qiu, M.H., 2019. The sources and mechanisms of bioactive ingredients in coffee. Food Funct. 10, 3113–3126.

InfoSanbas - Information on Basic Sanitation: Data Visualization Open on Sanitation, 2021.

International Coffee Organization, 2021. Trade Statistics.

Joët, T., Laffargue, A., Descroix, F., Doulbeau, S., Bertrand, B., kochko, A de, Dussert, S., 2010. Influence of environmental factors, wet processing and their interactions on the biochemical composition of green Arabica coffee beans. Food Chem. 118, 693–701.

Kamiloglu, S., 2019. Authenticity and traceability in beverages. Food Chem. 277, 12–24.

Marquetti, I., Link, J.V., Lemes, A.L.G., Scholz, M.B., dos, S., Valderrama, P., Bona, E., 2016. Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee. Comput. Electron. Agric. 121, 313–319.

Mehari, B., Redi-Abshiro, M., Chandravanshi, B.S., Combrinck, S., Atlabachew, M., McCrindle, R., 2016. Profiling of phenolic compounds using UPLC-MS for determining the geographical origin of green coffee beans from Ethiopia. J. Food Compos. Anal. 45, 16–25.

Núñez, N., Saurina, J., Núñez, O., 2021a. Non-targeted HPLC-FLD fingerprinting for the detection and quantitation of adulterated coffee samples by chemometrics. Food Control 124, 107912.

Núñez, N., Pons, J., Saurina, J., Núñez, O., 2021b. 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 (Lebensm.-Wiss. & Technol.) 147.

Obeidat, S.M., Hammoudeh, A.Y., Alomary, A.A., 2018. Application of FTIR spectroscopy for assessment of green coffee beans according to their origin. J. Appl. Spectrosc. 84, 1051–1055.

Ollivier, D., Artaud, J., Pinatel, C., Durbec, J.P., Guérère, M., 2006. Differentiation of French virgin olive oil RDOs by sensory characteristics, fatty acid and triacylglycerol compositions and chemometrics. Food Chem. 97, 382–393.

Paradkar, M.M., Irudayaraj, J., 2002. Rapid determination of caffeine content in soft drinks using FTIR-ATR spectroscopy. Food Chem. 78, 261–266.

Reis, N., Franca, A.S., Oliveira, L.S., 2013. Quantitative evaluation of multiple adulterants in roasted coffee by diffuse reflectance infrared fourier transform spectroscopy (DRIFTS) and chemometrics. Talanta 115, 563–568.

Ribeiro, D.E., Borem, F.M., Cirillo, M.A., Prado, M.V.B., Ferraz, V.P., Alves, H.M.R., Taveira Jh da, S., 2016. Interaction of genotype, environment and processing in the chemical composition expression and sensorial quality of Arabica coffee. Afr. J. Agric. Res. 11, 2412–2422.

Sanchez, P.M., Pauli, E.D., Scheel, G.L., Rakocevic, M., Bruns, R.E., 2018. Scarminio IS: irrigation and light access effects on Coffea arabica L. Leaves by FTIR-chemometric analysis. J Brazilian Chem Soc 29, 168–176.

Shao, J., 1993. Linear model selection by cross-validation. J. Am. Stat. Assoc. 88, 486–494.

Talamond, P., Verdeil, J.-L., Conéjéro, G., 2015. Secondary metabolite localization by autofluorescence in living plant cells. Molecules 20, 5024–5037.

Tasew, T., Mekonnen, Y., Gelana, T., Redi-Abshiro, M., Chandravanshi, B.S., Ele, E., Mohammed, A.M., Mamo, H., 2020. In vitro antibacterial and antioxidant activities of roasted and green coffee beans originating from different regions of Ethiopia. Int J Food Sci 2020.

Vodnar, D.C., Paucean, A., Dulf, F.V., Socaciu, C., 2010. HPLC characterization of lactic acid formation and FTIR fingerprint of probiotic bacteria during fermentation processes. Not. Bot. Horti Agrobot. Cluj-Napoca 38, 109–113.

Worku, M., de Meulenaer, B., Duchateau, L., Boeckx, P., 2018. Effect of altitude on biochemical composition and quality of green arabica coffee beans can be affected by shade and postharvest processing method. Food Res. Int. 105, 278–285.

Zhu, M., Long, Y., Ma, Y., Chen, Y., Yu, Q., Xie, J., Li, B., Tian, J., 2021. Comparison of chemical and fatty acid composition of green coffee bean (Coffea arabica L.) from different geographical origins. LWT (Lebensm.-Wiss. & Technol.) 140, 110802.