







DOI: <http://dx.doi.org/10.1590/1807-1929/agriambi.v27n10p803-810>

Uses of mid-infrared spectroscopy and chemometric models for differentiating between dried cocoa bean varieties¹

Utilização de espectroscopia de infravermelho médio e modelos quimiométricos para discriminação de variedades de cacau seco

Gentil A. Collazos-Escobar^{2,3} , Yeison F. Barrios-Rodríguez² ,
Andrés F. Bahamón-Monje^{2*}  & Nelson Gutiérrez-Guzmán² 

¹ Research developed at Universidad Surcolombiana/South Colombian Coffee Research Center/Department of Agricultural Engineering, Neiva, Colombia

² Universidad Surcolombiana/South Colombian Coffee Research Center/Department of Agricultural Engineering, Neiva, Colombia

³ Universitat Politècnica de València/Grupo de Análisis y Simulación de Procesos Agroalimentarios/Departamento de Tecnología de Alimentos, Valencia, Spain

HIGHLIGHTS:

The ATR-FTIR technique discriminates between cocoa genotypes by chemometric methods.

The LDA and PLS-DA chemometric models achieve high reliability in the discrimination of food matrices.

It will be possible to improve the processes of dry cocoa bean classification in the industry.

ABSTRACT: Generally, the taxonomic classification of cocoa beans is based on the theobromine/caffeine ratio determined using high-performance liquid chromatography (HPLC). However, this technique involves laborious and time-consuming calculations. Attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy is a valuable, effective, and rapid tool for analyzing the chemical composition of food products. The objective of this study was to examine the potential of ATR-FTIR combined with chemometric tools such as principal component analysis (PCA), linear discriminant analysis (LDA), and partial least squares regression-discriminant analysis (PLS-DA) to discriminate between the Trinitario and Forastero dry bean cocoa varieties defined by theobromine and caffeine measurements via HPLC. The cocoa varieties were evaluated using HPLC analysis of 36 dry cocoa bean samples to determine the theobromine/caffeine ratio. Moreover, ATR-FTIR spectra were analyzed in the mid-infrared (MIR) region, and signals associated with theobromine and caffeine were identified and analyzed using the LDA and PLS-DA models. The LDA and PLS-DA models allowed the satisfactory differentiation between cocoa varieties, providing overall prediction capacity values of $98.2 \pm 1.8\%$ and $96.1 \pm 2.4\%$, respectively. The results show the potential of ATR-FTIR spectroscopy for the reliable, fast, and easy differentiation of dried cocoa beans.

Key words: *Theobroma cacao*, ATR-FTIR, multivariate statistical tools, theobromine, caffeine

RESUMO: Geralmente, a classificação taxonômica das sementes de cacau foi feita pela relação teobromina/caféina por cromatografia líquida de alto desempenho (HPLC). No entanto, esta técnica envolve determinações laboriosas e demoradas; assim, a espectroscopia atenuada de infravermelho de reflexão total (ATR-FTIR) é uma ferramenta valiosa, eficaz e rápida para analisar a composição química dos produtos alimentares. Neste estudo, o objetivo era examinar o potencial da ATR-FTIR combinado com ferramentas quimiométricas como a análise de componentes principais (PCA), a análise discriminante linear (LDA), e a análise menos regressiva quadrada parcial (PLS-DA) examinada para a discriminação entre as variedades de cacau Trinitario e Forastero definidas pelas medições de teobromina e caféina via HPLC. As variedades de cacau foram avaliadas pela análise por HPLC de 36 amostras de grãos de cacau secos para a relação teobromina/caféina. Além disso, os espectros ATR-FTIR foram analisados em regiões de infravermelhos médios (MIR), e os sinais associados à teobromina e à caféina foram identificados e analisados através dos modelos LDA e PLS-DA. Os modelos LDA e PLS-DA proporcionaram uma discriminação satisfatória do cacau, fornecendo valores de capacidade de previsão global de $98.2 \pm 1.8\%$ e $96.1 \pm 2.4\%$, respectivamente. Os resultados mostram o potencial da técnica de espectroscopia ATR-FTIR para uma diferenciação fiável, rápida e fácil das amêndoas de cacau secas.

Palavras-chave: *Theobroma cacao*, ATR-FTIR, ferramentas estatísticas multivariadas, teobromina, caféina

• Ref. 269563 – Received 15 Nov, 2022

* Corresponding author - E-mail: gentilcollazosescobar09@gmail.com

• Accepted 10 May, 2023 • Published 16 Jun, 2023

Editors: Ítalo Herbet Lucena Cavalcante & Walter Esfrain Pereira

This is an open-access article distributed under the Creative Commons Attribution 4.0 International License.



INTRODUCTION

Cocoa is a globally important agricultural commodity and its quality depends on many factors, including post-harvest processing, genotype, geographical origin, agronomic management, climate, and soil conditions (Barrientos et al., 2019; Kongor et al., 2016).

Criollo, Forastero, and Trinitario are three commonly available varieties (Żyzelewicz et al., 2018). The cultivars show genetic variability in their theobromine/caffeine ratio, which can therefore be used for their differentiation (Carrillo et al., 2014). High-performance chromatography (HPLC) has facilitated the identification of these compounds; however, it is expensive. Therefore, a combination of chemometrics and Fourier transform infrared spectroscopy technique (FTIR) has been used to determine the composition or purity of products (Vasquez-Vuelvas et al., 2020) with satisfactory results in various fields of research, as was described by Juybar et al. (2020).

The powerful integration of FTIR and chemometric tools has been successfully applied to food discrimination, differentiation, defect detection, quality prediction, and adulteration (Christou et al., 2018). Other studies have included the discrimination of espresso coffees (Belchior et al., 2019), quantification of defects in coffee (Craig et al., 2015), detection of food fraud (El Darra et al., 2017), antioxidant capacity of cocoa (Batista et al., 2016), classification of cocoa varieties (Barbin et al., 2018), and differentiation of coffee processed using different post-harvest techniques (Barrios et al., 2020).

As was previously reported, the cocoa variety was ascertained by measuring the ratio of theobromine and caffeine by HPLC; however, multivariate models based on FTIR information have not been used to predict the variety. This study aimed to examine the potential of attenuated total reflection (ATR)-FTIR combined with chemometric tools such as principal component analysis (PCA), linear discriminant analysis (LDA), and partial least squares regression-discriminant analysis (PLS-DA) for differentiating between Trinitario and Forastero dry bean cocoa varieties defined by theobromine and caffeine measurements via HPLC.

MATERIAL AND METHODS

Thirty-six cocoa samples of *Theobroma cacao* L. were obtained directly from different farmers in the growing areas of Huila, Colombia. The origins and geographic locations of the cocoa samples are listed in Table 1. Raw cocoa samples (60 kg) were obtained from the fruits, fermented for 8 d in a wooden box (30 × 30 × 30 cm), and turned every 24 hours to guarantee uniform fermentation. The samples were then spread on a meshed wooden tray with an area of approximately 120 × 120 cm and raised 130 cm above ground level.

The sun-drying process was conducted daily between 9 am and 4 pm until a moisture content of 6-7% on a wet basis (% w.b.) was achieved. This was monitored using a grain moisture tester (Gehaka G600, Gehaka AGRA, São Paulo, Brazil), and the dried cocoa samples (5 g) were dehydrated in an oven (UF55, Memmert GmbH + Co.KG, Schwabach, Germany) at 105°C for 24 h to determine the moisture content. The results were expressed as dry matter percentage (% d.b.).

The water activity (a_w) of the samples was evaluated using a vapor sorption analyzer (VSA). To measure a_w , 2-3 g of dried and ground cocoa was placed inside a VSA (Aqualab Decagon Devices, Inc., Pullman, WA, USA), with prior calibration of the dew point sensor with four saturated aqueous salt solutions purchased from the instrument's manufacturer: 13.41 m LiCl ($0.25 \pm 0.003 a_w$), 8.57 m LiCl ($0.50 \pm 0.003 a_w$), 6.0 m NaCl ($0.76 \pm 0.003 a_w$), and 2.33 m NaCl ($0.92 \pm 0.003 a_w$). All measurements were performed in triplicates.

Dried cocoa beans without the skin were ground independently using a blender (Oster®, Colombia). The aqueous extractions were made in triplicate with 100 mg of dried cocoa powder in 25 mL Milli-Q water at 85°C for 25 min in a water bath (WNE 45, Memmert, Schwabach, Germany) and stirred in a magnetic plate at 800 rpm for 10 min. The extracts were centrifuged at 9,000 rpm for 10 min with an EBA 200 (Hettich, Kirchlegern, Germany) centrifuge and filtered with 0.22 µm nylon filters.

Analysis was performed using an Agilent 1260 Infinity II series liquid chromatography instrument (Agilent Technologies, Santa Clara, CA, USA) with a Poroshell 120-

Table 1. Cocoa samples origin and geographic location

Municipality	Latitude	Longitude	Altitude (masl)	Municipality	Latitude	Longitude	Altitude (masl)
Colombia	3° 24' 35.73" N	74° 48' 38.94" W	736	Algeciras	2° 43' 14.1" N	75° 15' 51.6" W	863
Colombia	3° 23' 9.74" N	74° 49' 18.31" W	750	Algeciras	2° 32' 48.8" N	75° 18' 17.1" W	1,020
Tello	2° 59' 49" N	75° 4' 20.5" W	1,222	Algeciras	2° 32' 50.8" N	75° 17' 19.3" W	1,101
Tello	2° 59' 56" N	75° 3' 52.2" W	1,211	Rivera	2° 49' 19" N	75° 14' 2.2" W	660
Tello	3° 1' 7.9" N	75° 3' 39.6" W	1,172	Rivera	2° 44' 50.6" N	75° 14' 58.9" W	874
Palermo	2° 50' 41.1" N	75° 33' 35.9" W	1,106	Colombia	3° 21' 32.54" N	74° 50' 17.8" W	700
Palermo	2° 49' 29.7" N	75° 33' 58.7" W	1,185	Tello	3° 0' 18.2" N	75° 3' 49.6" W	1,297
Santa María	2° 57' 23" N	75° 30' 21.2" W	843	Tello	2° 59' 55.4" N	75° 4' 34.1" W	1,019
Santa María	2° 56' 57" N	75° 33' 34.3" W	1,197	Tello	2° 58' 55" N	75° 4' 9.2" W	984
Teruel	2° 48' 10.1" N	75° 33' 17.1" W	1,060	Tello	2° 59' 8.6" N	75° 4' 16" W	1,025
Teruel	2° 46' 53.9" N	75° 31' 32.4" W	815	Palermo	2° 51' 32" N	75° 30' 56.9" W	769
Íquira	2° 39' 20.2" N	75° 36' 32.8" W	834	Palermo	2° 49' 19.46" N	75° 33' 52.62" W	1,200
Íquira	2° 39' 11.2" N	75° 38' 27.4" W	1,179	Santa María	2° 51' 42.2" N	75° 33' 36.2" W	834
Nátaga	2° 32' 16.8" N	75° 49' 30.6" W	1,300	Teruel	2° 46' 57.2" N	75° 32' 34.1" W	845
Nátaga	2° 32' 32.58" N	75° 49' 50.5" W	1,190	Teruel	2° 42' 22" N	75° 35' 55.7" W	885
Nátaga	2° 32' 19.1" N	75° 49' 9.2" W	1,345	Íquira	2° 38' 17.4" N	75° 39' 14.5" W	960
Gigante	2° 25' 17.13" N	75° 25' 17.13" W	943	Gigante	2° 23' 30.95" N	75° 31' 47.37" W	916
Gigante	2° 21' 40.94" N	75° 31' 46.23" W	1,040	Campoalegre	2° 36' 7.3" N	75° 20' 7.9" W	424

C18 (2.7 μm , 4 μm – 4.6 \times 150 mm) column. The injection volume was 20 μL with a flow rate of 1 mL min^{-1} . Separation was performed using isocratic elution with methanol (Merck, Darmstadt, Germany) and water containing 0.2% acetic acid (20:80 v/v) for 10 min. The detection was performed using a diode array detector (DAD) at 280 nm. Theobromine and caffeine were identified by comparing their retention times and UV-spectra of the standards (Borja Fajardo et al., 2022).

An Agilent Cary 630 FTIR spectrophotometer (Agilent Technologies, Santa Clara, CA, USA) with an ATR sampling accessory was used for the ATR-FTIR measurements, which were performed in a dry atmosphere at room temperature (20 \pm 0.5 $^{\circ}\text{C}$) (Craig et al., 2018). Approximately 0.5 g of dried cocoa powder (0.5 g) was placed in a sampling accessory and pressed. All spectra were obtained in triplicate and recorded within the mid-infrared (MIR) range of 4000–650 cm^{-1} with 4 cm^{-1} resolution and 16 scans. They were subjected to background subtraction. The ATR-FTIR standard spectra of theobromine (CAS 83-67-0 purity \geq 98.0%) and caffeine (CAS 58-08-2 purity \geq 99.0%) were determined in triplicate under the same spectral conditions and were considered to correlate with the signals of these compounds in the dried cocoa samples.

The theobromine and caffeine concentrations were expressed as means \pm standard deviations (SD) in triplicate for every other sample. One-way analysis of variance (\leq 0.05) was performed using the STATGRAPHICS Centurion XVIII (Manufacturers Inc., Rockville, MD, USA). To compensate for and remove the bias linked to the experimental assessment of the spectrum, the infrared spectral data were preprocessed using baseline correction. Subsequently, multiplicative dispersion correction (MSC) was applied. Data processing was performed using R statistical software (version 3.6.3, R statistics, St. Louis, MO, USA) and the ChemoSpec (Hanson, 2022) R package.

Multivariate statistical analysis of the spectroscopic data was performed using principal component analysis (PCA) to explain the data variability. To detect and remove outlier observations from the experimental data, multivariate control statistics, such as the residual sum of squares (RSS) and Hotelling T^2 were used. Subsequently, LDA and PLS-DA were used to develop classification models to differentiate between the dried cocoa bean varieties established using HPLC. The analyses were performed using R statistical software with the DiscrMiner and Factoextra packages.

For LDA model construction, it was used fewer uncorrelated variables than the subjects. Therefore, the orthogonal eigenspace from the PCA (all principal components that summarized 100% variability of the original dataset) was used as the input for the LDA model. Additionally, the PCA scores for the LDA model training were screened using the mean decrease accuracy criterion of the Random Forest algorithm, which was computed using the randomForest R-package. To achieve the “leave one out” cross-validation of the LDA and PLS-DA models, the samples were randomly divided 100 times into calibration (75%) and validation (25%) data sets. The chemometric models were trained with the calibration dataset (using cross-validation) and external validation (with 25% of the remaining data) was subsequently performed for each iteration.

The predictability of the classification models was evaluated based on the overall accuracy (%) and sensitivity (Eq. 1), specificity (Eq. 2), precision (Eq. 3), recall (Eq. 4), and F-score (Eq. 5). Additionally, to select the best classifier, a multifactor analysis of variance (ANOVA) was performed considering the models and iterations as factors and the different classification goodness metrics of the validation dataset as responses, employing a comparative mean with least significant difference (LSD) test intervals ($p \leq 0.05$). Residual validation of all ANOVA models was performed using Shapiro-Wilk's test to verify residual normality, the Ljung-Box test to check residual independence, and a multifactor ANOVA was performed on square residuals to verify the homoscedasticity hypothesis. Statistical assumptions were verified at $p \leq 0.05$ using STATGRAPHICS Centurion XVIII (Manugistics, Inc., Rockville MD, USA).

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

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

$$\text{Precision} = \frac{N(\text{true positives})}{N(\text{true positives}) + N(\text{false positives})} \quad (3)$$

$$\text{Recall} = \frac{N(\text{true positives})}{N(\text{true positives}) + N(\text{false negatives})} \quad (4)$$

$$\text{F Score} = 2 \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \quad (5)$$

RESULTS AND DISCUSSION

Firstly, the water activity and moisture content of the dried cocoa bean samples were between 0.32 and 0.42 a_w and 6.38 and 7.52% d.b., respectively. The cocoa varieties identified using HPLC are listed in Table 2. A theobromine/caffeine relationship was observed, and dried cocoa beans were classified into the Trinitario and Forastero varieties. According to Carrillo et al. (2014) and Samaniego et al. (2020), theobromine/caffeine relationship values between 3 and 9 are indicative of the Trinitario genotype, and values higher than 9 can be classified as the Forastero variety.

As can be observed, the theobromine concentrations in both varieties were higher than caffeine concentrations, thus indicating that theobromine was the predominant compound in the extracts, which is consistent with previous findings (Carrillo et al., 2014; Hernández-Hernández et al., 2022). Furthermore, significant differences in methylxanthine levels were observed. The Trinitario variety had higher theobromine and caffeine contents than the Forastero variety, as is shown in Table 2. These results are similar to those reported by Carrillo et al. (2014) for Colombian cocoa beans from different growing areas (Samaniego et al., 2020) in Ecuador. The differences in methylxanthine observed between our results and those

Table 2. Cocoa varieties determined using HPLC

Theobromine/Caffeine	Variety	Theobromine/Caffeine	Variety
8.630 ± 0.241	Trinitario	8.061 ± 0.717	Trinitario
8.263 ± 0.372	Trinitario	6.335 ± 0.482	Trinitario
8.900 ± 0.392	Trinitario	4.750 ± 0.257	Trinitario
8.012 ± 0.102	Trinitario	8.702 ± 1.121	Trinitario
7.178 ± 0.523	Trinitario	8.066 ± 2.115	Trinitario
8.362 ± 0.343	Trinitario	9.107 ± 0.648	Forastero
8.733 ± 0.808	Trinitario	10.473 ± 0.894	Forastero
6.893 ± 0.189	Trinitario	11.128 ± 0.297	Forastero
7.396 ± 0.515	Trinitario	14.429 ± 0.776	Forastero
5.988 ± 0.164	Trinitario	9.284 ± 0.465	Forastero
5.409 ± 0.908	Trinitario	10.026 ± 0.564	Forastero
6.624 ± 0.327	Trinitario	9.965 ± 0.634	Forastero
4.769 ± 0.118	Trinitario	12.126 ± 0.138	Forastero
6.860 ± 0.276	Trinitario	10.239 ± 0.397	Forastero
5.783 ± 0.247	Trinitario	9.301 ± 1.192	Forastero
7.688 ± 0.817	Trinitario	9.601 ± 0.310	Forastero
7.601 ± 0.503	Trinitario	12.910 ± 2.262	Forastero
8.452 ± 0.872	Trinitario	14.035 ± 0.514	Forastero

in the literature can be attributed to different post-harvest treatments, geographical origins, agronomic management, climate, and soil conditions (Kongor et al., 2016). Theobromine and caffeine constitute approximately the total concentration of alkaloids in *Theobroma cacao* and its derivative products, with the former being the most abundant methylxanthine and the latter found only in small amounts (Bartella et al., 2019). The methylxanthine concentrations in the Trinitario and Forastero cocoa varieties are shown in Table 3.

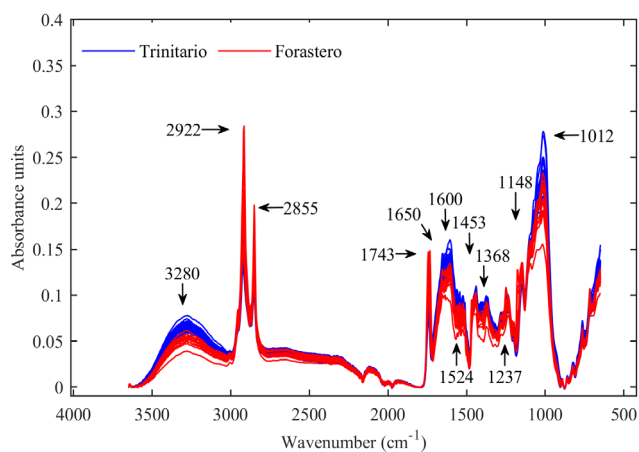
A variety of dried cocoa bean samples were prepared (Table 2), and the theobromine/caffeine ratio could be related to bean quality (Álvarez et al., 2012). The results could be valuable for determining the genotype of dried cocoa beans and their application in the cocoa industry, because the Forastero variety has been regarded as a precursor of ordinary and basic quality notes and is the primary raw material used in 80% of the global chocolate production. The Trinitario variety is a hybrid resulting from crossing Criollo (more aromatic and floral notes, less bitter, and smoother than Forastero), which is used in close to 10-15% of chocolate production, and is known to produce some wine flavor notes (Quiroz-Reyes & Fogliano, 2018).

Figure 1 shows the ATR-FTIR spectra of 36 dried cocoa bean samples comprising two varieties previously established using HPLC (23 Trinitario and 13 Forastero samples). According to Riswahyuli et al. (2020), all MIR peaks within the wavenumber range of 4000-650 cm^{-1} can be classified into two different regions: the functional (4000-1500 cm^{-1}) and fingerprint (1500-650 cm^{-1}) regions. To describe the results obtained in Figure 1, Durak & Depciuch (2020) reported chemical compounds and functional groups in plants that were associated with the infrared spectrum range: 3290-3270 cm^{-1} to water and polysaccharides; 2950-2910, 2870-2840, 1800-1765, 1750-1700, and 1380-1300 cm^{-1} to lipids; 1650-1620 cm^{-1} to

Table 3. Theobromine and caffeine concentration of Trinitario and Forastero dried cocoa samples

Variety	Theobromine (mg g^{-1})	Caffeine (mg g^{-1})	Theobromine/Caffeine
Trinitario	20.3 ± 2.8a	2.9 ± 0.8a	7.3 ± 1.4a
Forastero	18.7 ± 2.1b	1.7 ± 0.3b	10.97 ± 1.92b

Means with different letters indicate statistically significant differences ($p \leq 0.05$) according to Fisher's test. Results are expressed as mean ± standard deviation

**Figure 1.** Attenuated total reflectance-Fourier transform infrared spectra of 'Trinitario' and 'Forastero' dried cocoa beans in the mid-infrared 4000-650 cm^{-1} range

water and proteins; 1600-1545 and 1460-1430 cm^{-1} to proteins and lipids; 1260-1235 cm^{-1} to polysaccharides; 1160-1145 and 1120-1100 cm^{-1} to ester, 1080-1015 cm^{-1} with glycosidic bonding; and 840-810 cm^{-1} to the aromatic ring.

The observed ATR-FTIR spectra showed typical vibration patterns of biological material constituents, such as proteins, lipids, and carbohydrates, reflecting the composition of dried cocoa beans and the influence of their genotype on absorbance unit variation. According to Baker et al. (2014), the most important spectral regions are the fingerprint region (1450-600 cm^{-1}), amide I and II regions (1700-1500 cm^{-1}), and the higher wavenumber region of 3500-2550 cm^{-1} , which is associated with stretching vibrations (S-H, C-H, N-H, and O-H).

Several ATR-FTIR studies have been conducted on cocoa beans and chocolate. The bands at 963 cm^{-1} , 1018 cm^{-1} , 1076 cm^{-1} , and 1112 cm^{-1} were associated with the C-O and C-C stretching modes (Hu et al., 2016; Batista et al., 2016). The peaks at 1283 cm^{-1} , 1360 cm^{-1} , and 1457 cm^{-1} (Figure 1) were related to the O-C-H, C-C-H, and C-O-H bending modes, respectively. C-H deformation of the ring was observed at 882 cm^{-1} , and C-OH of the phenyl group was observed at 1143 cm^{-1} and 1517 cm^{-1} . The alkene stretching vibrations (C = C) at 1663 cm^{-1} and 1620 cm^{-1} can be attributed to axial deformation of the N-H group of the aromatic ring due to the possible presence of alkaloids such as caffeine and theobromine, which usually show signals in the range 1750-1600 cm^{-1} (Rojas et al., 2020). Additionally, the vibrations at 1645-1544 cm^{-1} can be attributed to the C-C stretching of the aromatic ring (Batista et al., 2016). The bands at 2922 cm^{-1} , 2852 cm^{-1} , and 1743 cm^{-1} are related to asymmetric and symmetric CH_2 stretching, as well as to the C=O stretching group of triglycerides adjacent to the C-O group in esters (Batista et al., 2016; Craig et al., 2018; Barrios et al., 2021). Finally, the wavenumber at 3003 cm^{-1} was associated with the stretching vibration of the cis-olefinic double bond (Sánchez-Reinoso et al., 2017), and the bands associated with the phenol group at 3562-3322 cm^{-1} were attributed to O-H stretching (Batista et al., 2016).

As can be seen from the Figures 1 and 2E, these spectra reports were similar to the results verified in the present study and are coherent with the composition of dried cocoa

beans, which mainly comprise water (6-7%), followed by carbohydrates such as fructose (1.61%), glucose (0.75%), sucrose (0.05%) and mannitol (0.62%), fiber (22.85%), and ash (3.04%) (Barrientos et al., 2019), and compounds responsible for the taste and aroma such as theobromine (2-3%), caffeine (0.2%), and traces of theophylline (Rojas et al., 2020).

The standard spectrum was recorded in the wavenumber range of 1800-1400 cm^{-1} . In Figures 2B and D, the signals at

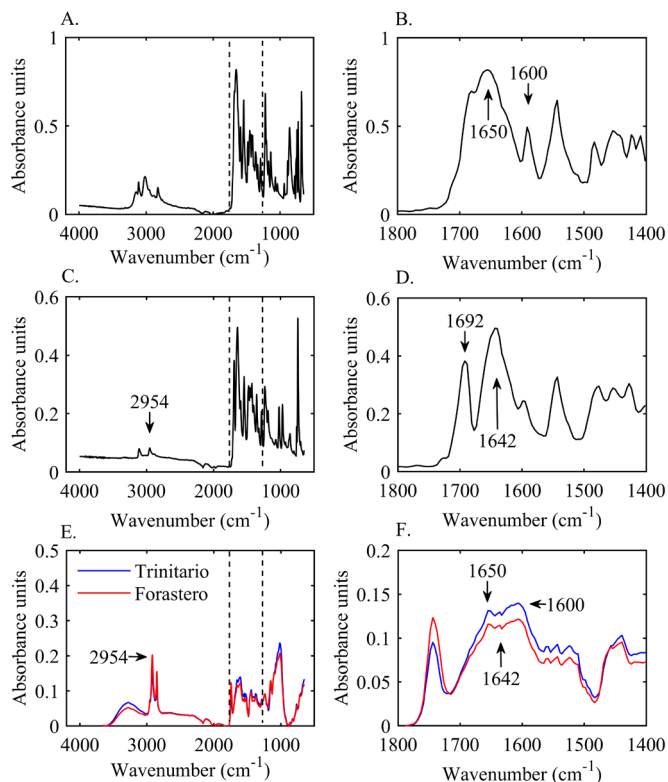


Figure 2. Full mean ATR-FTIR standard spectra of theobromine (A); theobromine enlargement, 1800-1400 cm^{-1} (B); caffeine standard (C); caffeine enlargement 1800-1400 cm^{-1} (D); FTIR-ATR spectrum of dry cocoa from 'Trinitario' and 'Forastero' (E); region of the spectrum with higher theobromine and caffeine intensity (F)

1650 and 1600 cm^{-1} confirmed the presence of theobromine, and those at 1692 and 1642 cm^{-1} confirmed the presence of caffeine in the cocoa samples (Figure 2F). These signals resembled those reported by Rojas et al. (2020), who argued that the vibration of these alkaloids typically shows signals within a wavenumber range of 1750-1600 cm^{-1} . The ATR-FTIR spectrum of the caffeine standard (Figure 2C) corresponded to that reported by Bahamon et al. (2018). Moreover, the vibrational band at 2954 cm^{-1} was attributed to the C-H stretching of caffeine. Nugrahani et al. (2019) determined the specific spectral area of caffeine at 2967.27-2930.51 cm^{-1} and its association with the -N-H amide functional group. To confirm the presence of methylxanthine (alkaloids) in the ATR-FTIR spectra of the cocoa samples, theobromine (Figure 2A) and caffeine standards were analyzed according to a previously reported approach (Figure 2C).

An exploratory PCA was performed with all spectral data (with baseline correction) using different processing techniques, such as SNV, MSC, and first and second derivatives. The best clustering results were obtained for MSC (Figure 3). As was mentioned previously, the chemical compounds, theobromine and caffeine, obtained using HPLC were adequate for identifying the cocoa bean variety. These were found in the MIR spectral measurements of the samples at a wavelength spectral range between 1700 and 1600 cm^{-1} , suggesting that these chemical compounds can also be identified via ATR-FTIR in the biochemical fingerprint region.

As shown in Figure 3, the first two principal components summarized 68.55% of the spectral variability, indicating that almost all the variability of the MIR information was explained by these components (Figure 3A). The first principal component (PC1) contributed the most (43.12%), followed by the second (PC2; 25.43%). The score plot indicated that Trinitario and Forastero could be separated into two groups. Trinitario cocoa samples were located on the negative axis of PC1 and were distributed throughout PC2, whereas Forastero samples were generally located on positive PC1 values and had similar distributions on PC2. Moreover, differentiation

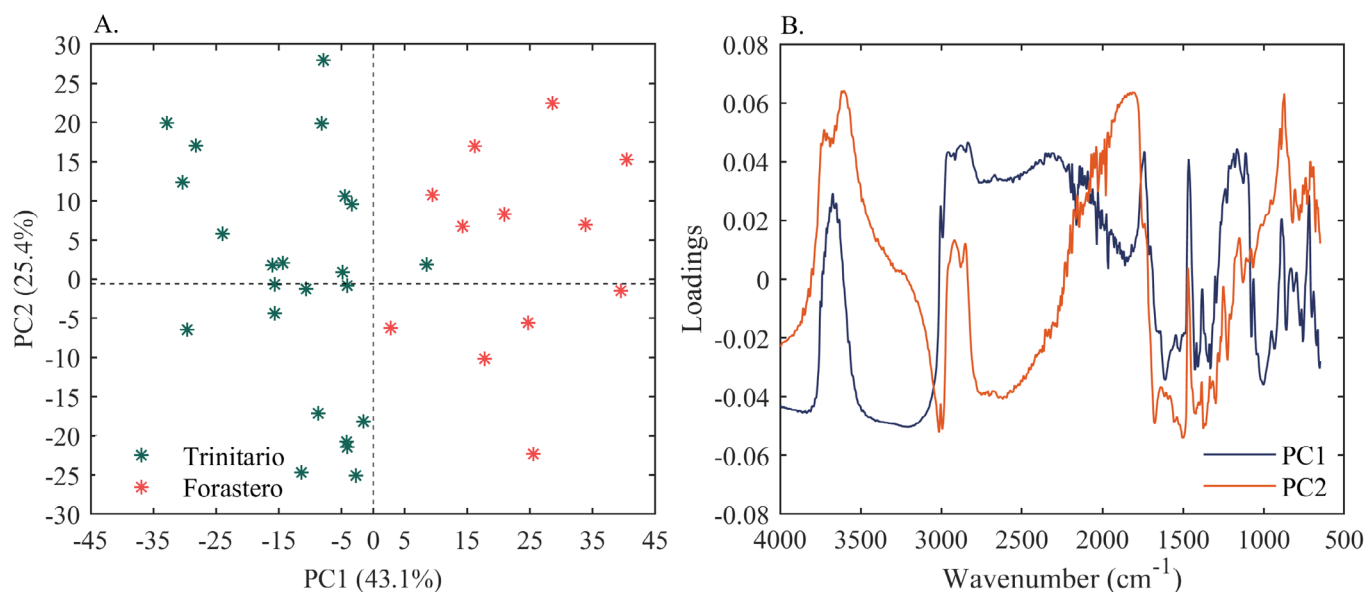


Figure 3. PCA scatter plot of Fourier transform infrared spectroscopy by means of attenuated total reflectance (ATR-FTIR) preprocessed by multiplicative dispersion correction (MSC) normalization (A) and loading plot of MSC spectral information (B)

between the spectral features of cocoa was observed. From the grouping behavior, important differences were observed in the infrared spectra of the Forastero and Trinitario varieties, and the PCA analysis explained the differences in the MIR profile and the grouping of the samples.

To understand the discriminatory effect of PCA, Figure 3B shows the highest loadings for the first two principal components. The loadings of PC2 showed a maximum positive value at 1744 cm^{-1} , which correlated with a decrease in the same area for PC1. This matches the wave number observed in Figure 1, suggesting that PC2 influences the separation of cocoa varieties owing to the ester functional group ($\text{C}=\text{O}$) observed in this spectral region.

The stretching of this functional group can be attributed to the lipid content of the cocoa samples (Cortés et al., 2019). This suggests that the differences in the lipid concentrations of each variety contributed to their differentiation. Similarly, the negative loading in the wavelength range 1670-1500 cm^{-1} is related to the stretching of the cis functional group $\text{C}=\text{C}$ (1654 cm^{-1}) and broadening of the NH amide group (1531 cm^{-1}) identified in the FTIR spectrum (Craig et al., 2018; Cortés et al., 2019). These functional groups can be associated with theobromine and caffeine contents (Figure 2). The peak at 1167 cm^{-1} did not correspond to a specific signal in the infrared spectrum (Figure 1). However, this signal was related to the C-OH groups associated with the polyphenol content in chocolate (Hu et al., 2016). In PC1 (positive) and PC2 (negative), the influence of the 2922 and 2852 cm^{-1} bands related to the asymmetric and symmetric stretching of CH_2 and the $\text{C}=\text{O}$ stretching group of the triglycerides, was evident. A strong contribution of the wavenumber at 3003 cm^{-1} associated with the stretching vibration of the cis-olefinic double bond, was also observed (Sánchez-Reinoso et al., 2017).

According to the results obtained, analyzing the loadings coupled with infrared spectral information obtained from chemical standards and samples constitutes a valuable tool for identifying the MIR features that facilitate differentiation between the Trinitario and Forastero cocoa varieties.

To discriminate between these cocoa varieties, LDA and PLS-DA were performed on the preprocessed spectral information; the results are shown in Table 4.

Using both chemometric models, all Trinitario and Forastero cocoa samples were satisfactorily classified into datasets belonging to the pre-established varieties that were differentiated using HPLC (Table 3). The overall accuracy percentage, sensitivity, specificity, precision, recall, and F Score of the models were high for both the calibration and validation sets. As is shown in Table 4, the two supervised classification methods resulted in an overall accuracy higher than 90%. However, the ANOVA-based LSD intervals highlighted that PLS-DA was significantly better than LDA for classifying the observations from the validation test set, the PLS-DA model showed the highest overall accuracy (96.1 \pm 2.4%), highest recall (0.98 \pm 0.06), and highest F Score (0.97 \pm 0.05), and to our knowledge, could be the first study to correlate FTIR-ATR spectra and primary data of theobromine/caffeine ratio quantification using HPLC in dry cocoa samples to discriminate

Table 4. Classification results and overall accuracy using the LDA and PLS-DA chemometric model for classifying cocoa varieties differentiated using HPLC

Linear discriminant analysis (LDA)			
Calibration set (n = 27)			
	Trinitario	Forastero	Overall accuracy (%)
Predicted Trinitario	17 \pm 1	0	98.2 \pm 1.8
Predicted Forastero	0	10 \pm 1	
Sensitivity	0.98 \pm 0.02		
Specificity	0.99 \pm 0.01		
Precision	0.99 \pm 4.76 \times 10 ⁻³		
Recall	0.98 \pm 0.02		
F-Score	0.99 \pm 0.01		
Validation set (n = 9)			
	Trinitario	Forastero	Overall accuracy (%)
Predicted Trinitario	5 \pm 1	0 \pm 1	92.3 \pm 7.7a
Predicted Forastero	0 \pm 1	3 \pm 1	
Sensitivity	0.94 \pm 0.1a		
Specificity	0.91 \pm 0.2a		
Precision	0.95 \pm 0.1a		
Recall	0.94 \pm 0.1a		
F-Score	0.94 \pm 0.07a		
Partial least square regression-discriminant analysis (PLS-DA)			
Calibration set (n = 27)			
	Trinitario	Forastero	Overall accuracy (%)
Predicted Trinitario	17 \pm 1	0	96.9 \pm 1.7
Predicted Forastero	1	9 \pm 1	
Sensitivity	0.99 \pm 0.01		
Specificity	0.92 \pm 0.04		
Precision	0.95 \pm 0.02		
Recall	0.99 \pm 0.01		
F-Score	0.98 \pm 0.01		
Validation set (n = 9)			
	Trinitario	Forastero	Overall accuracy (%)
Predicted Trinitario	6 \pm 1	0	96.1 \pm 2.4b
Predicted Forastero	0	3 \pm 1	
Sensitivity	0.98 \pm 0.06b		
Specificity	0.93 \pm 0.15a		
Precision	0.96 \pm 0.07a		
Recall	0.98 \pm 0.06b		
F-Score	0.97 \pm 0.05b		

Different letters indicate statistically significant differences ($p \leq 0.05$) obtained using Fisher's test between the chemometric models for the validation dataset. Results are expressed as mean \pm standard deviation

the Trinitario and Forastero varieties using chemometric models, thus providing a satisfactory classification tool for differentiating the cocoa varieties in the two classes evaluated.

Additionally, the results showed that the LDA model was overfitted for the training dataset owing to decreased evidence (Table 4) of the goodness of classification metrics with respect to those obtained with the test set.

Concerning the PLS-DA model, slight decreases in the classification metrics were observed, indicating the improved ability of PLS-DA to predict the cocoa variety of an unknown observation based on its spectral information. Hence, the ATR-FTIR spectroscopy technique, coupled with chemometric models, reported satisfactory and reliable classification performance for the varieties under study. Using the PLS-DA, the highest overall accuracy percentage of 96.1 \pm 2.4% was obtained. Thus, PLS-DA is the most suitable chemometric model for solving the classification problem and can be

considered a valuable tool for distinguishing between these cocoa varieties.

CONCLUSIONS

1. Attenuated total reflectance-Fourier transform infrared spectroscopy proved to be valuable for characterizing and identifying functional groups associated with chemical compounds such as theobromine and caffeine.

2. The results obtained using discriminant chemometric models, LDA, and PLS-DA indicated that the proposed ATR-FTIR spectroscopy technique can be regarded as a promising alternative, which is appropriate for identifying Trinitario and Forastero cocoa bean varieties.

3. The partial least squares regression-discriminant analysis model provided the most adequate and useful information for predicting the dry bean cocoa varieties.

LITERATURE CITED

- Álvarez, C.; Pérez, E.; Cros, E.; Lares, M.; Assemat, S.; Boulanger, R.; Davrieux, F. The use of near infrared spectroscopy to determine the fat, caffeine, theobromine and (-)-epicatechin contents in unfermented and sun-dried beans of Criollo cocoa. *Journal of Near Infrared Spectroscopy*, v.20, p.307-315, 2012. <https://doi.org/10.1255/jnirs.990>
- Bahamon, M. A. F.; Parrado, L. X.; Gutiérrez-Guzmán, N. ATR-FTIR for discrimination of espresso and americano coffee pods. *Coffee science*, v.13, p.550-558, 2018. <https://doi.org/10.25186/cs.v13i4.1499>
- Baker, J. M.; Trevisan, J.; Bassan, P.; Bhargava, R.; Butler, J. H.; Dorling, M. K.; Fielden, R. P.; Forgarty, W. S.; Fullwood, J. N.; Heys, A. K.; Hughes, C.; Lash, P.; Martin-Hirsch, L. P.; Obinaju, B.; Sockalingum, D. G.; Sulé-Suso, J.; Strong, J. R.; Walsh, J. M.; Wood, R. B.; Gardner, P.; Martin, L. F. Using Fourier-transform infrared spectroscopy to analyze biological materials. *Nature protocols* v.9, p.1771-1791, 2014. <https://doi.org/10.1038/nprot.2014.110>
- Barbin, D. F.; Maciel, L. F.; Bazoni, C. H. V.; Ribeiro, M. da S.; Carvalho, R. D. S.; Bispo, E. da S.; Miranda, M. da P. S.; Hirooka, E. Y. Classification and compositional characterization of different varieties of cocoa beans by near infrared spectroscopy and multivariate statistical analyses. *Journal of Food Science and Technology*, v.55, p.2457-2466, 2018. <https://doi.org/10.1007/s13197-018-3163-5>
- Barrientos, P. L. D.; Oquendo, T. J. D.; Garzón, G. M. A.; Álvarez, M. O. L. Effect of the solar drying process on the sensory and chemical quality of cocoa (*Theobroma cacao* L.) cultivated in Antioquia, Colombia. *Food Research International*, v.115, p.259-267, 2019. <https://doi.org/10.1016/j.foodres.2018.08.084>
- Barrios, R. Y. F.; Gutiérrez, G. N.; Girón, H. J. Effect of the postharvest processing method on the biochemical composition and sensory analysis of arabica coffee. *Engenharia Agrícola, Jaboticabal*, v.40, p.177-183, 2020. <https://doi.org/10.1590/1809-4430-Eng.Agric.v40n2p177-183/2020>
- Barrios, R. Y. F.; Reyes, C. A. R.; Campos, J. S. T.; Girón-Hernández, J.; Rodríguez-Gamir, J. Infrared spectroscopy coupled with chemometrics in coffee post-harvest processes as complement to the sensory analysis. *LWT- Food Science and Technology*, v.145, p.1-7, 2021. <https://doi.org/10.1016/j.lwt.2021.111304>
- Bartella, L.; Donna, D. L.; Napoli, A.; Siciliano, C.; Sindona, G.; Mazzotti, F. A rapid method for the assay of methylxanthines alkaloids: Theobromine, theophylline and caffeine in cocoa products and drugs by paper spray tandem mass spectrometry. *Food Chemistry*, v.278, p.261-266, 2019. <https://doi.org/10.1016/j.foodchem.2018.11.072>
- Batista, N. N.; Andrade, P. D. de; Ramos, L. C.; Dias, R. D.; Schwan, F. R. Antioxidant capacity of cocoa beans and chocolate assessed by FTIR. *Food Research International*, v.90, p.313-319, 2016. <http://dx.doi.org/10.1016/j.foodres.2016.10.028>
- Belchior, V.; Botelho, G. B.; Oliveira, S. L.; Franca, S. A. Attenuated Total Reflectance Fourier Transform Spectroscopy (ATR-FTIR) and chemometrics for discrimination of espresso coffees with different sensory characteristics. *Food Chemistry*, v.273, p.178-185, 2019. <https://doi.org/10.1016/j.foodchem.2017.12.026>
- Borja Fajardo, J. G., Horta Tellez, H. B., Peñaloza Atuesta, G. C., Sandoval Aldana, A. P., & Mendez Arteaga, J. J. Antioxidant activity, total polyphenol content and methylxanthine ratio in four materials of *Theobroma cacao* L. from Tolima, Colombia. *Heliyon*, v.8, p.1-6, 2022. <https://doi.org/10.1016/j.heliyon.2022.e09402>
- Carrillo, C. L.; Londoño-Londoño, J.; Gil, A. Comparison of polyphenol, methylxanthines and antioxidant activity in *Theobroma cacao* beans from different cocoa-growing areas in Colombia. *Food Research International*, v.60, p.273-280, 2014. <http://dx.doi.org/10.1016/j.foodres.2013.06.019>
- Christou, C.; Agapiou, A.; Kokkinofa, R. Use of FTIR spectroscopy and chemometrics for the classification of carobs origin. *Journal of Advanced Research*, v.10, p.1-8, 2018. <https://doi.org/10.1016/j.jare.2017.12.001>
- Cortés, V.; Talens, P.; Barat, J. M.; Lerma-García, M. J. Discrimination of intact almonds according to their bitterness and prediction of amygdalin concentration by Fourier transform infrared spectroscopy. *Postharvest Biology and Technology*, v.148, p.236-241, 2019. <https://doi.org/10.1016/j.postharvbio.2018.05.006>
- Craig, A. P.; Botelho, B. G.; Oliveira, L. S.; Franca, A. S. Mid infrared spectroscopy and chemometrics as tools for the classification of roasted coffees by cup quality. *Food Chemistry*, v.245, p.1052-1061, 2018. <https://doi.org/10.1016/j.foodchem.2017.11.066>
- Craig, A. P.; Franca, A. S.; Oliveira, L. S.; Irudayaraj, J.; Ileleji, K. Fourier transform infrared spectroscopy and near infrared spectroscopy for the quantification of defects in roasted coffees. *Talanta*, v.134, p.379-386, 2015. <http://dx.doi.org/10.1016/j.talanta.2014.11.038>
- Durak, T.; Depciuch, J. Effect of plant sample preparation and measuring methods on ATR-FTIR spectra results. *Environmental and Experimental Botany*, v.169, p.1-13, 2020. <https://doi.org/10.1016/j.envexpbot.2019.103915>
- El Darra, N.; Rajha, N. H.; Saleh, F.; Al-Oweini, R.; Maroun, R. G.; Louka, N. Food fraud detection in commercial pomegranate molasses syrups by UV-VIS spectroscopy, ATR-FTIR spectroscopy, and HPLC methods. *Food Control*, v.78, p.132-137, 2017. <http://dx.doi.org/10.1016/j.foodcont.2017.02.043>
- Hanson, B. A. ChemoSpec: Exploratory Chemometrics for Spectroscopy. R package version 6.1.3, 2022. <https://CRAN.R-project.org/package=ChemoSpec>

- Hernández-Hernández, C.; Fernández-Cabanás, V. M.; Rodríguez-Gutiérrez, G.; Fernández-Prior, Á.; Morales-Sillero, A. Rapid screening of unground cocoa beans based on their content of bioactive compounds by NIR spectroscopy. *Food Control*, v.131, p.1-9, 2022. <https://doi.org/10.1016/j.foodcont.2021.108347>
- Hu, Y.; Pan, J. Z.; Liao, W.; Li, J.; Gruget, P.; Kitts, D. D.; Lu, X. Determination of antioxidant capacity and phenolic content of chocolate by attenuated total reflectance-Fourier transformed-infrared spectroscopy. *Food Chemistry*, v.202, p.254-261, 2016. <http://dx.doi.org/10.1016/j.foodchem.2016.01.130>
- Juybar, M.; Khorrami, K. M.; Garmarudi, B. A. FTIR/PLS and SVM multivariate calibrations to determination of the coke amount into the deactivated catalysts and the product of the methanol to gasoline conversion. *Infrared Physics and Technology*, v.105, p.1-10, 2020. <https://doi.org/10.1016/j.infrared.2020.103229>
- Kongor, J. E.; Hinneh, M.; Van de Walle, D.; Afoakwa, O. E.; Boeckx, P.; Dewettinck, K. Factors influencing quality variation in cocoa (*Theobroma cacao*) bean flavour profile- A review. *Food Research International*, v.82, p.44-52, 2016. <http://dx.doi.org/10.1016/j.foodres.2016.01.012>
- Nugrahani, I.; Manosa, Y. L.; Chintya, L. FTIR-derivative as a green method for simultaneous content determination of caffeine, paracetamol, and acetosal in a tablet compared to HPLC. *Vibrational Spectroscopy*, v.104, p.1-10, 2019. <https://doi.org/10.1016/j.vibspec.2019.102941>
- Quiroz-Reyes, C. N.; Fogliano, V. Design cocoa processing towards healthy cocoa products: The role of phenolics and melanoidins. *Journal of Functional Foods*, v.45, p.480-490, 2018. <https://doi.org/10.1016/j.jff.2018.04.031>
- Riswahyuli, Y.; Rohman, A.; Setyabudi, M. C. S. F.; Raharjo, S. Indonesian wild honey authenticity analysis using attenuated total reflectance-fourier transform infrared (ATR-FTIR) spectroscopy combined with multivariate statistical techniques. *Heliyon*, v.6, p.1-7, 2020. <https://doi.org/10.1016/j.heliyon.2020.e03662>
- Rojas, M.; Chejne, F.; Ciro, H.; Montoya, J. Roasting impact on the chemical and physical structure of Criollo cocoa variety (*Theobroma cacao* L). *Journal of Food Process Engineering*, v.43, p.1-15, 2020. <https://doi.org/10.1111/jfpe.13400>
- Samaniego, I.; Espín, S.; Quiroz, J.; Ortiz, B.; Carrillo, W.; García-Viguera, C.; Mena, P. Effect of the growing area on the methylxanthines and flavan-3-ols content in cocoa beans from Ecuador. *Journal of Food Composition and Analysis*, v.88, p.1-9, 2020. <https://doi.org/10.1016/j.jfca.2020.103448>
- Sánchez-Reinoso, Z.; Osorio, C.; Herrera, A. Effect of microencapsulation by spray drying on cocoa aroma compounds and physicochemical characterisation of microencapsulates. *Powder Technology*, v.318, p.110-119, 2017. <http://dx.doi.org/10.1016/j.powtec.2017.05.04>
- Vásquez-Vuelvas, O. F.; Chávez-Camacho, F. A.; Meza-Velázquez, J. A.; Mendez-Merino, E.; Ríos-Licea, M. M.; Contreras-Esquivel, J. C. A comparative FTIR study for supplemented agavin as functional food. *Food Hydrocolloids* v.103, p.1-10, 2020. <https://doi.org/10.1016/j.foodhyd.2020.105642>
- Żyżelewicz, D.; Budryn, G.; Oracz, J.; Antolak, H.; Kręgiel, D.; Kaczmarska, M. The effect on bioactive components and characteristics of chocolate by functionalization with raw cocoa beans. *Food Research International*, v.113, p.234-244, 2018. <https://doi.org/10.1016/j.foodres.2018.07.017>