Article

Journal of the Brazilian Chemical Society

https://dx.doi.org/10.21577/0103-5053.20230144

J. Braz. Chem. Soc. **2024**, *35*, 3, e-20230144, 1-10

Authentication of Specialty Coffees from the Fluminense Northwest and Caparaó Regions (Brazil) Using UV-Vis Spectroscopy and Synthetic Samples Partial Least Square Discriminant Analysis (SS-PLS-DA)

Gabriel R. F. Caldeira,[a] Tayná O. Costa,[a,b] Marcia H. C. Nascimento,[c] Patricia G. Corradini,[a] Paulo R. Filgueiras,[c] Daniel C. Ferreira[d] and Murilo de O. Souza*[*,a,b]

[a]*Laboratório de Análises Químicas e Agroambientais, Instituto Federal de Educação, Ciência e Tecnologia Fluminense, Campus Itaperuna, BR 356, km 3, 28300-000 Itaperuna-RJ, Brazil*

[b]*Programa de Pós-Graduação em Ciências Naturais, Universidade Estadual do Norte Fluminense Darcy Ribeiro, 28013-600 Campos dos Goytacazes-RJ, Brazil*

[c]*Departamento de Química, Universidade Federal do Espírito Santo, Avenida Fernando Ferrari, 514, 29060-900 Vitória-ES, Brazil*

[d]*Instituto Federal de Educação, Ciência e Tecnologia Fluminense, Campus Bom Jesus do Itabapoana, Avenue Dario Vieira Borges, 235, 28360-000 Bom Jesus do Itabapoana-RJ, Brazil*

Caparaó and the Fluminense northwest regions are nationally recognized by the important contribution on coffee production and exportation. Adulterations involving specialty coffees result in a decrease in the quality of the final product. However, obtaining many different samples from the same region is unfeasible in some cases, needing strategies to work with a limited number of samples for pattern recognition. Thus, this work is the first to use the construction of synthetic samples (SS) for analysis of coffees, and its objective is to identify adulterations in specialty coffees with bark, straw and low-quality beans, using UV-Vis spectroscopy, associated with chemometric methods. The synthetic samples partial least square discriminant analysis (SS-PLS-DA) showed better specificity, sensitivity and reliability rates than the Hard PLS-DA models. One-class methods (soft independent modeling of class analogy (SIMCA) and data driven soft independent modeling of class analogy (DD-SIMCA)) showed low specificity and reliability. The discriminant methods together with the synthetic samples proved to be adequate to identify adulterations in specialty coffees.

Keywords: specialty coffees, synthetic samples, PLS-DA, food fraud, UV-Vis spectroscopy

Introduction

The coffee tree (*Coffea* sp.) is a shrub from the Rubiaceae family. In Brazil, the main two species cultivated are *Coffea arabica* (known as Arabica coffee) and *Coffea canephora* (known as café-robusta or conilon). Coffee has a significant cultural and economic impact in Brazil, as the country is the largest producer and exporter of the grain, and has the second-highest *per* capita consumption in the world, after the United States.[1,2]

In the 19th century, the state of Rio de Janeiro was a pioneer in the cultivation of large quantities of coffee in Brazil, and today, the Fluminense northwest region is responsible for 70% of coffee production in this state.[3] The state of Espírito Santo is the second-largest producer of coffee in Brazil (with Minas Gerais state in first place). The Caparaó region, which spans both states, is where most of their coffees are grown.[4] Family farming is the primary way of growing coffee in the Fluminense northwest and Caparaó regions. Due to advances in cultivation techniques, there has been a significant increase in the production of high-quality coffees in recent years.

Specialty coffee is a rapidly growing global market, with approximately 15% growth *per* year.[5] This makes specialty coffee a great option to counter commodity price fluctuations. Specialty coffee typically commands 30 to

*e-mail: murilo.souza@iff.edu.br, m.quimic@gmail.com
Editor handled this article: Ivo M. Raimundo Jr. (Associate)
This manuscript is part of a series of publications in the Journal of the Brazilian Chemical Society by young researchers who work in Brazil or have a solid scientific connection with our country. The JBCS welcomes these young investigators who brighten the future of chemical sciences.

40% higher prices than conventional coffee, and in some auctions, it can easily exceed 100%. Some regions have requested geographical indication (IG) or denomination of origin (DO) for their coffees, such as Caparaó and Matas de Minas. The Fluminense northwest region is following this trend, and proposals have been submitted to conduct studies to verify that coffees grown in this region have their own identity (special coffees of the "Alto Noroeste Fluminense"),[6] to support the petition of IG to the National Institute of Industrial Property (INPI). This prevents coffees produced in other regions from being marketed as specialty coffees of the "Alto Noroeste Fluminense," ensuring their uniqueness. Additionally, this guarantees greater added value to these coffees, generating higher income, especially for small family-based coffee growers in the region.

It is worth noting that coffees grown in the Region of Caparaó (ES) have already received unprecedented registration for IG in 2021. Efforts are needed to prevent fraud and ensure geographic authentication and correct commercialization.

Due to the high added value of specialty coffee beans, frauds can occur when cheap materials, such as stems, sticks, shells, beans, corn and other coffee beans from different geographical origins are added to increase the final volume of coffee processed and obtain an irregular and criminal profit.[7,8] In this sense, there has been an increasing need to develop methodologies to identify these adulterations, which has been studied by the scientific community and food inspection agencies. Sensory analysis, which evaluates specialty coffees through attributes such as aroma, flavor, and acidity, has been widely used.[9,10] However, this requires rigorous training to obtain experienced professionals (Q-graders) which can be time-consuming and costly.

Different analytical techniques have been used to evaluate various adulterations in foods with high added value, such as coffee. Fourier transform infrared spectroscopy (FTIR), near infrared spectroscopy (NIRS), mass spectrometry coupled to gas chromatography (GC-MS) and high-efficiency liquid chromatography (HPLC-MS) are the most commonly used.[2,11-13] However, the high cost of these analytical techniques and their absence in educational and research institutions in the studied regions are limitations, requiring coffee samples to be sent to laboratories in other regions of Brazil at a high cost. Therefore, cheaper and more accurate methods and techniques such as UV-Vis spectrometry (which is low cost, simple and easy to operate), combined with chemometric methods, can be an excellent alternative for evaluating adulterations in coffee samples. Multivariate data analysis has been applied for the authentication of different matrices, including coffee, medicinal plants, rice, organic grapes and organic grape juices, carrots, and products with a protected designation of origin (PDO) such as honey, wine vinegar and wine.[14-23]

This study aims to build chemometrics models from results obtained by UV-Vis spectroscopy to distinguish specialty coffee produced in the Northwest Fluminense and Caparaó regions from adulterated coffee with bark, coffee straw, and low-quality coffee beans. A partial least square discriminant analysis (PLS-DA) model was built for the first time from the creation of synthetic samples (SS-PLS-DA) to evaluate the authenticity of specialty coffees. Finally, this study contributes to controlling the authenticity of coffee produced in protected origins, preventing fraud and the sale of products with incorrect origin and processing.

Experimental

Coffee samples origin

Thirty samples of different specialty coffees (*Coffea arabica*) were used in this study. Specialty coffee beans and adulterants (such as bark, sticky straw and low-quality beans) were obtained from local and regional producers, sourced from the Institute of Technical Assistance and Rural Extension of the northwest region of Rio de Janeiro (EMATER-RJ), the Laboratory of Coffee Quality at the Fluminense Federal Institute (IFF), and by the Coffee Classification and Tasting Laboratory at the Federal Institute of Espírito Santo (IFES). All samples belonged to the 2020 and 2021 crops. Table 1 provides the geographic descriptions and sensory evaluation scale of the samples, while Figure 1 shows the geographic demarcation of the two regions where specialty coffees were grown.

As mentioned in the Introduction section, the Fluminense northwest region has applied for an IG with the INPI to verify that the coffees grown in this region have their own

Table 1. Geographical origin of the specialty coffee samples studied as declared by the producers

Samples	Cities	Amount	Scores[a]
Fluminense northwest region	Bom Jesus do Itabapoana, Porciúncula, Varre-Sai	7	85-90
Caparaó region	Alegre, Alto Caparaó, Carangola, Divino de São Lourenço, Dores do Rio Preto, Espera Feliz, Guaçuí, Ibatiba, Ibitirama, Irupi, Martins Soares, São José do Calçado	23	85-94

[a]Brazilian Specialty Coffee Association (BSCA) Sensory Rating Scale.[24]

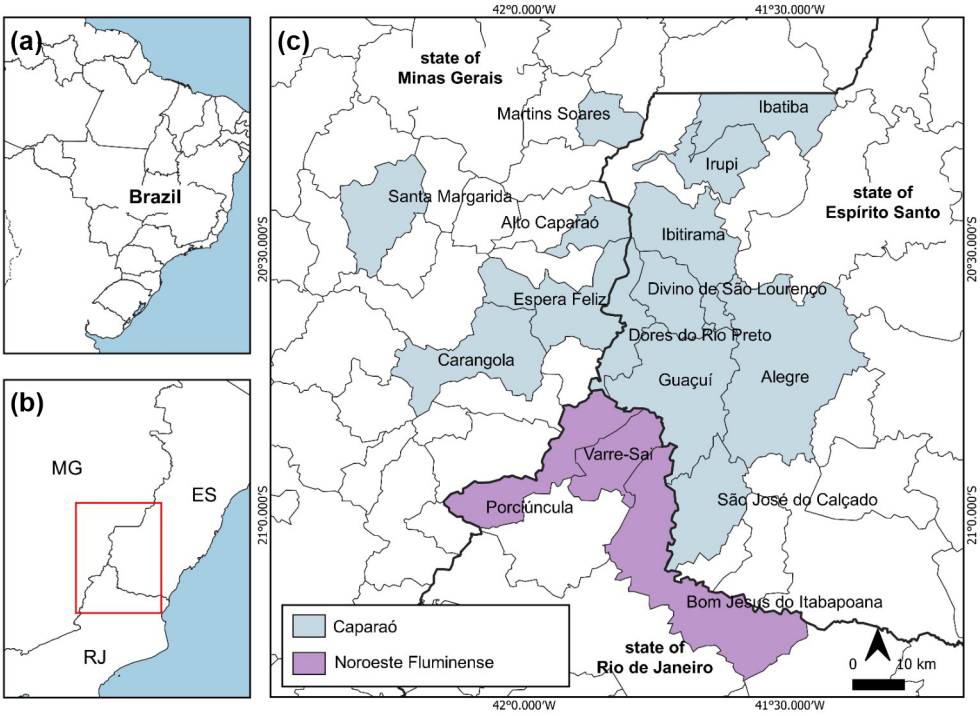

Figure 1. Geographical location of cities producing specialty coffee in the Fluminense northwest and Caparaó regions (adapted from reference 25).

unique identity (specialty coffees from the "Alto Noroeste Fluminense"). The Caparaó region (ES) has already been granted IG registration in 2021.[26]

Coffee sample preparation

Thirty samples of specialty coffees were roasted at 200 °C for 10 min using an industrial coffee roaster (Carmomaq, Torrador de Prova Tradicional, Brazil) and then ground to the smallest granulometry allowed by an electric coffee grinder (Botini, Botimetal, Brazil). The same roasting and grinding procedure was performed for the bark, sticky straw and extrinsic coffee defects (called low-quality beans).

The 30 samples of specialty coffees were adulterated with 1, 3, 7 and 10 wt.% of bark and sticky straw and with sieve bottom blends, intrinsic, and low-quality beans in the proportions of 10, 20, 30, 40, 50 and 75 wt.%. The adulterations carried out for each pure specialty coffee sample generated the following groups of adulterated samples: 120 samples (bark); 120 samples (straw) and 180 samples (low-quality beans).

The aqueous extraction of specialty coffees (pure and adulterated) was based on the literature, with some modifications.[27] Briefly, it involved of preparing coffee infusions in which 1.0000 g of each sample and filtered in 50.0 mL of distilled water at 90-98 °C. Filtration was carried out gradually (approximately 3 min) to extract the substances present in the coffee. After that, the obtained extracts were cooled to room temperature, and 500 µL of each extract was added in a 25.0 mL volumetric flask and checked with distilled water.

Spectral data acquisition using UV-Vis spectroscopy

A UV-Visible spectrometer (BioMate 3S, Thermo Fisher Scientific, USA) was used to obtain spectral data in the 200-900 nm range using the fast scan mode. The reference was measured using distilled water. Spectral bands with high noise were identified in the range of 200-230 nm and after 350 nm. Therefore, the spectral range of 230-350 nm was used for the chemometric analyses.[22,28]

Chemometric analysis

The quality of the original spectral data was improved by applying the second derivative (second-order polynomial and a 9-point window size), aiming to reduce baseline effects in the original spectra. The data were then centered on the mean. The Kennard-Stone algorithm was used to separate the pure specialty coffee samples from the adulterated coffee samples, by approximately 2/3 and 1/3 for the training and test sets, respectively, ensuring a good representation of both classes (pure specialty coffees and adulterated coffees). All processing was carried out with MATLAB R2013a (The MathWorks, Natick, MA, USA), with a few toolboxes[29-32] for modeling and in-house scripts.

Synthetic sample creation

The imbalance between the quantities of pure samples and adulterated samples can lead to errors and bias when distinguishing between each target class. To mitigate this issue, generating synthetic samples (**SS**) can be a valuable approach to address the imbalance and accurately link the predictor matrix with the response vector[33-35] In order to generate **SS**, the training set of the target class (specialty coffee) was normalized around the mean. Subsequently, the vector basis was transformed using principal component analysis (PCA) and applying singular value decomposition (SVD) (equation 1).

$$[\mathbf{T}, \mathbf{P}, \text{varexp}] = \text{pca}(\mathbf{X}\boldsymbol{m}, npc) \quad (1)$$

where $\mathbf{X}\boldsymbol{m}$ represents the set of samples of the target class centered on the mean, *npc* refers to the number of principal components, **T** is the scores matrix, **P** is the loadings matrix and varexp indicates the explained variance, all products of PCA decomposition. The above function was developed by this research group. Singular value decomposition, $\mathbf{X}\boldsymbol{m} = \mathbf{U}\cdot\mathbf{S}\cdot\mathbf{V}^{t}$; where: **U** and **V** are orthonormal, and **S** is a diagonal matrix containing the square root of the non-zero eigenvalues (explained variance) of $\mathbf{X}\boldsymbol{m}$.

The **SS** were randomly generated in the same dimension as the principal components (PCs),[30] which corresponds to the score matrix (**T**). The semi-axes of an ellipse were calculated to establish the boundaries in this PC space, employing a confidence limit of 3 standard deviations (equation 2).

$$\mathbf{semi} = \mathbf{sd} \cdot \sqrt{\text{diag}\left(\text{eig}\left(\text{cov}\left(\mathbf{T}\right)\right)\right)} \quad (2)$$

where, *sd* is the adopted limit of 3 standard deviation and diag(eig(cov(**T**))) is the main diagonal of the diagonal matrix of eigenvalues, obtained from the quadratic covariance matrix of **T**.

To create synthetic samples (**SS**) belonging to the space delimited by the ellipse, we used the product of each value of a random vector of the same dimension of PCs with the standard deviation vector of **T**, added to an average vector of **T** (equation 3).

$$\mathbf{SS} = (randn(1, \text{col}(\mathbf{T})) \times \mathbf{T}\boldsymbol{s}) + \mathbf{T}\boldsymbol{m} \quad (3)$$

The function *"randn"* generates random values from a standard normal distribution, *col*(***T***) represents the column dimension of **T**, while $\mathbf{T}\boldsymbol{s}$ corresponds to the standard deviation vector and $\mathbf{T}\boldsymbol{m}$ is the mean vector of **T**.

Subsequently, the synthetic data were returned to the vector base of the original vector space by adding the initial mean vector to the matrix product of the synthetic data and the loadings matrix (**P**) (equation 4).

$$\mathbf{SS} = \mathbf{SS} \times \mathbf{P}^{\mathbf{T}} + \mathbf{X}_{\mathbf{mean}} \quad (4)$$

where $\mathbf{P}^{\mathbf{T}}$ is the transposed loadings matrix and $\mathbf{X}_{\mathbf{mean}}$ is the mean vector of the original data.

The synthetic data were then concatenated to the training set, generating a new target class (real + synthetic samples). This approach aimed to achieve balance between the target class and the other modeled classes by generating "n" synthetic samples. After that, the models were built using the partial least square discriminant analysis (PLS-DA) applied to the dataset containing the real samples along with the **SS**.[29]

Regarding the generation of synthetic samples (SS), various methods have been proposed in the literature, such as SMOTE,[36] which is one of the best-known methods for this purpose. SMOTE generates synthetic objects along a space between the nearest neighbors of minority class objects. In this work, we used an algorithm developed by our research group, based on capturing the variance of the original data in the principal components dimension.

Hard and synthetic samples partial least square discriminant (SS-PLS-DA) models

Mean-centered data were correlated with a single **y**-vector containing 1 for pure samples (or pure + synthetic samples for SS-PLS-DA) and 2 for each type of adulteration. The number of latent variables was chosen by cross-validation, evaluating the lowest cross validation classification error (CVCE) and the lowest number of false positives (FP) and false negatives (FN) generated by the model. The test set was used to verify the performance of the model.

The limits for sample discrimination were obtained using the Bayesian threshold (TL) derived from Bayes theory.[37,38] The Bayesian threshold estimating assumes that the *y* variance predicted by the model will follow a similar distribution for all other samples to be discriminated. The *y* value is the point where two estimated distributions intersect at the selected threshold.[39] From this value that the number of FP and FN must be minimized for future predictions. The test set was used to check the performance of the model.

The soft independent modeling of class analogy (SIMCA)[22,40] and data driven soft independent modeling of class analogy (DD-SIMCA) algorithms[31] were

applied to the dataset in order to compare the performance of chemometric methods for authentication of pure and adulterated coffees.

Validation of models

The figures of merit (FOM) used to validate the models were estimated according to equations 5 to 9.[41] The FP is the number of samples that do not belong to a class but were classified as belonging to it, i.e., an adulterated sample that was classified as pure. The FN is the number of pure samples that were classified as adulterated. True positive (TP) is the number of pure samples that have been classified as belonging to this class. True negative (TN) is the number of adulterated samples that have been classified as belonging to this class. From the parameters FP, FN, TP and TN it is possible to estimate the false positive rate (FPR), the false negative rate (FNR), sensitivity (SEN), specificity (SPE) and reliability rate (RLR).[28,41]

$$FPR = \frac{FP}{FP + TN} \times 100 \quad (5)$$

$$FNR = \frac{FN}{FN + TP} \times 100 \quad (6)$$

$$SEN = \frac{TP}{TP + FN} \times 100 \quad (7)$$

$$SPE = \frac{TN}{TN + FP} \times 100 \quad (8)$$

$$RLR = 100 - (FPR + FNR) \quad (9)$$

To evaluation the performance of the SS-PLS-DA models, the contribution of the synthetic samples was excluded from the calculation to enable comparison with the models generated by Hard PLS-DA.

Results and Discussion

Figure 2 presents the spectra (Figure 2a) without pre-treating (200-500 nm), (Figure 2b) with pre-treating using the 2nd derivative with a second-order polynomial and a window size of 9 points (230-350 nm) and (Figure 2c) with the 2nd derivative combined mean centering (230-350 nm). Spectral bands with high noise were identified in the range of 200-230 nm and after 350 nm and, therefore, were not used for the chemometric analyses,[22,28] as shown in Figures 2b and 2c.

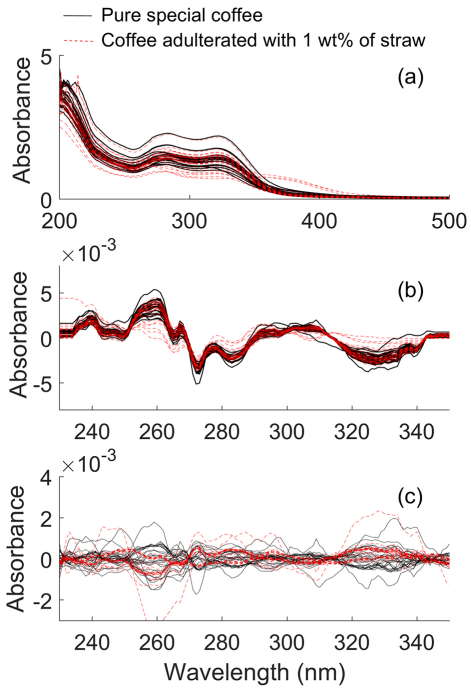

Figure 2. (a) Original spectral data, (b) pre-treated spectral data (2nd derivate) and (c) combined pre-treated spectral data (2nd derivate + mean center) of pure special coffee (solid black line) and adulterated with 1 wt.% of straw (dotted red line).

The data obtained by UV-Vis spectroscopy are numerical values of absorption units (variables) that are correlated with a vector **y** containing the prediction class. Pre-treating the data (Figure 2) helps remove non-useful information, attenuate instrumental deviations, and improve the quality of spectral data.[42] However, there is no specific rule for treating data generated by UV-Vis spectroscopy, and the best classification results for each data set must be evaluated.

Cavdaroglu and Ozen[43] evaluated the adulteration of grape vinegar with brandy vinegar and synthetic acetic acid using different pre-treating of UV-Vis spectral data, such as 1st, 2nd and 3rd derivatives, among others. The 3rd derivative associated with the orthogonal-PLS-DA model produced adequate sensitivity (100 and 85.7% training and test set, respectively) and good specificity (100 and 96.4% training and test set, respectively) for the classification of unadulterated and adulterated vinegars.

It should be noted that the raw spectra obtained from UV-Vis spectroscopy are very similar (Figure 2a), making it difficult to group and recognize patterns in the data. This high similarity of spectra without pre-treating makes it challenging to discriminate coffee samples according to their target classes. On the other hand, the second derivative excludes linear variations of the spectra, emphasizing the differences in the absorptions (Figure 2b), in addition to allowing the correction of displacements and deviations from the baseline.[42] This way, the derivative alters the

profile of the spectra, increasing the number of visible bands. This increased complexity of the derived spectra can be useful in qualitative analyses that use the spectral print for classification and discrimination of different coffee samples.[44] Finally, mean centering is a preprocessing technique used for data of the same nature and magnitude. It involves standardizing the original data and excluding the linear coefficient from the model and therefore was applied to the studied dataset (Figure 2c).

There are several chemometric models used to classify data as to their authenticity and, the most appropriate model will depend on the nature and origin of the samples studied. In this article, two chemometric approaches were evaluated: one-class (SIMCA and DD-SIMCA) and discriminant (Hard PLS-DA and SS-PLS-DA). Table 2 presents the performance of the models evaluated for the classification and discrimination of pure and adulterated specialty coffees with bark, straw, and low-quality beans. The number of factors (VLs and PCs) used to build the models (SIMCA, Hard and SS-PLS-DA) was based on the lowest CVCE. The number of PCs for building the DD-SIMCA model was chosen from the smallest number of extreme target samples (false negatives) or outliers (samples outside the threshold, with a 95% confidence probability) for the training set.[45,46] The training and test sets were chosen systematically and unbiased using the Kennard-Stone algorithm. The models presented in Table 2 were evaluated without detecting and removing outliers to allow comparison.

PLS-DA is a discriminant method that has been widely used in recent decades for supervised discrimination in food analysis.[2,10,11] The performance parameters for the discriminant models are presented in Table 2, which indicate good sensitivity (> 83.3%), specificity (> 97.2%), and high reliability rates (> 84.0%) for SS-PLS-DA models, for example. Some works[47,48] have suggested the use of class modeling methods, such as, for example, SIMCA, as the most appropriate for food authentication. However, the one-class methodologies (SIMCA and DD-SIMCA) obtained in this article showed low specificity (< 47.1%) and models with low capacity to detect samples adulterated with bark, straw, and low-quality coffee. One explanation for this behavior is the high heterogeneity of samples of pure specialty coffees (target class) produced in different cities in the Fluminense northwest and Caparaó regions (Figure 1), generating a data set with high variability.

Recent works[49,50] report that when there is a large variability of the target class, SIMCA tends to have high sensitivity (ability to detect authentic samples) at the cost of low specificity (ability to detect adulterated samples), or *vice versa*, depending on the number of PCs samples. Santos *et al.*[50] applied one-class models (SIMCA and OCPLS) to distinguish infested (target class) from non-infested (non-target class) sorghum grains. Both models showed high sensitivity (> 95%) at the cost of low specificity (< 30%) due to the high variability of the studied samples (36 different genotypes). Biancolillo *et al.*[49] applied the SIMCA and PLS-DA models to distinguish edible rice samples (target class) from those infested by storage pests

Table 2. Figures of merit for the models applied to pure special coffee and pure special coffee adulterated with bark, coffee peal and low-quality coffees

Group	Model	Training set[a,b]		Test set[a,b]		Total classification[c]		Other parameters	
		SEN / %	SPE / %	SEN / %	SPE / %	SEN / %	SPE / %	RLR / %	Number of PCs/VLs
Bark	SIMCA	85.7	49.3	100	41.9	90.0	47.1	37.2	4
	DD-SIMCA[a]	100	100	100	100	100	38.4	38.5	3
	Hard PLS-DA	100	93.2	100	90.6	100	93.3	93.3	9
	SS-PLS-DA[b]	95.2	100	88.8	100	93.3	100	93.3	12
Straw	SIMCA	71.4	43.7	100	22.5	80.0	34.1	14.2	2
	DD-SIMCA[a]	100	100	100	100	100	18.3	18.4	3
	Hard PLS-DA	95.2	91.6	88.8	80.5	90.0	91.3	81.3	10
	SS-PLS-DA	85.7	96.4	100	97.2	93.3	100	93.3	11
Low-quality grains	SIMCA	80.0	40.5	100	30.9	86.6	38.3	25.1	6
	DD-SIMCA[a]	100	100	100	100	100	26.1	23.9	3
	Hard PLS-DA	80.9	92.0	66.6	94.4	76.7	93.3	70.0	8
	SS-PLS-DA[b]	85.7	96.8	88.8	98.1	83.3	97.2	84.0	10

[a]The training and test sets for DD-SIMCA were performed with pure samples only, then the adulterated samples were inserted into the model (total classification); [b]The SEN, SPE and RLR parameters were calculated without considering the synthetic samples, comparing the generated models. [c]Total classification (training + test sets). SEN: sensitivity rate; SPE: specific rate; RLR: reliability rate; PCs: principal components; VLs: latent variables; SIMCA: soft independent modeling of class analogy; DD-SIMCA: data driven soft independent modeling of class analogy; PLS-DA: partial least square discriminant analysis; SS-PLS-DA: synthetic samples PLS-DA.

(non-target class). The SIMCA model showed low sensitivity (59.1% to test set) and high specificity (96.9% to test set) due to the high heterogeneity of samples that were collected from different countries. It is worth noting that the authors used 8 or more PCs to build the SIMCA models, which ensured adequate detection capacity for samples infested by storage pests (non-target class). Increasing the number of PCs can improve the number of hits in the non-target class (high specificity), but attention should be paid to possible overfitting of the built model. Finally, the one-class models for this work were constructed using fewer than 6 PCs, resulting in a low ability to detect samples adulterated with bark, straw, and low-quality grains (i.e., low specificity). The use of more PCs generated overfit models.

Although the discriminant models were built with more factors than the DD-SIMCA, it has been observed that neither the SS nor the PLS-DA models exhibited overfitting. Overfitting, which refers to the inclusion of irrelevant information in the model, can be identified when the performance of the training set significantly surpasses that of the test set. However, this phenomenon did not occur with the SS and PLS-DA models, as indicated in Table 2. It is important to note that the selection of the number of factors (VLs) in the SS and PLS-DA models is based on the lowest cross-validation classification error (CVCE). In contrast, for DD-SIMCA, the number of factors (PCs) used for model construction is determined from a smaller subset of extreme target samples or outliers.[31] For this reason, the number of factors in DD-SIMCA tends to be smaller compared to PLS-DA and SS-PLS-DA.

Comparison of Hard PLS-DA and SS-PLS-DA models

Figures 3a and 3b show the Hard PLS-DA and SS-PLS-DA models, respectively, for the discrimination of pure specialty coffees versus adulterated specialty coffees with bark (1 to 10 wt.%). For the Hard PLS-DA model, a good sensitivity was obtained (SEN = 100%), as all samples of pure specialty coffees were correctly discriminated, resulting in no false negatives. This high sensitivity was also observed for the SS-PLS-DA models (Figure 3b; SEN = 93.3%), indicating that the ability to discriminate pure specialty coffee samples remained when synthetic samples were added to the model. However, the Hard PLS-DA model showed a lower specificity (Figure 3a; SPE = 93.3%) when compared to the SS-PLS-DA (Figure 3b; SPE = 100%). The high specificity of the SS-PLS-DA model was attributed to a lower false positive rate (FP = 0) as compared to the Hard PLS-DA model, which had 8 false positives (FP = 8). As a result, the incorporation of synthetic samples resulted in an improvement in the specificity of the PLS-DA model. However, both models (Hard and SS-PLS-DA) had the same reliability rate (RLR = 93.3%).

Figure 4a shows the Hard PLS-DA model that was utilized to differentiate pure specialty coffees from those that were mixed with straw (ranging from 1 to 10 wt.%). This model displayed a sensitivity of 90.0% with only 3 false negatives, suggesting a strong capability of discriminating pure specialty coffee samples. On the other hand, the SS-PLS-DA model (Figure 4b) demonstrated high sensitivity (SEN = 93.3%), indicating that the addition of synthetic samples did not considerably affect the discrimination ability of pure specialty coffee samples. Additionally, the SS-PLS-DA model exhibited higher specificity (SPE = 100%) compared to Hard PLS-DA (SPE = 91.3%), showing accurate discrimination of non-target samples (mixed with straw). By balancing the target class (pure specialty coffees) through the inclusion of synthetic samples, more dependable outcomes were achieved, resulting in a reliability rate of 93.3% in contrast to Hard PLS-DA's reliability rate of 81.3%. Notably,

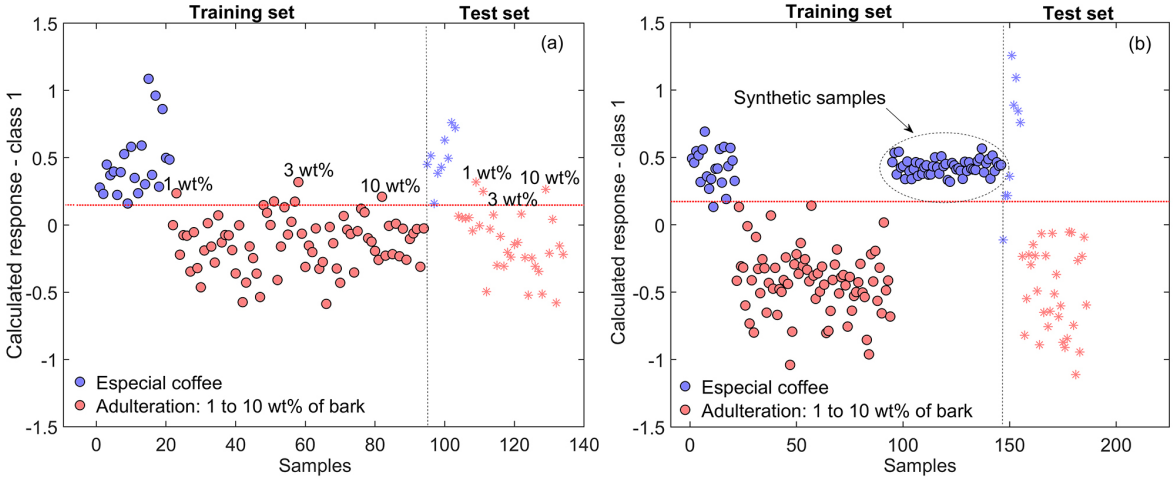

Figure 3. Scores graphs for (a) Hard PLS-DA and (b) SS-PLS-DA model with pure and adulterated specialty coffees with bark (1 to 10 wt.%).

the false positives covered the entire range of adulteration tested, suggesting that there was no tendency to discriminate only a specific percentage.

Figure 5a presents the Hard PLS-DA model built for discrimination of pure and adulterated specialty coffees with low-quality beans (10 to 75 wt.%). The model demonstrated a sensitivity of 76.7%, which can be attributed to the misclassification of 7 pure specialty coffee samples as adulterated (false negatives). The SS-PLS-DA showed a sensitivity of 83.3% due to the lower number of false negatives (FN = 5). Moreover, the SS-PLD-DA model showed a higher specificity (SPE = 97.2%) than the Hard PLS-DA model (SPE = 93.3%). Additionally, similar to the other models, the inclusion of synthetic samples improved the discrimination ability, resulting in a higher reliability rate (RLR = 84.0%) than that of the Hard PLS-DA model (RLR = 70.0%). Lastly, the false positives covered the entire range of adulteration tested, indicating no preference for discriminating a specific percentage.

Conclusions

The discriminant methods (Hard and SS-PLS-DA) showed good sensitivity rates, specificity rates and reliability rates, particularly for the SS-PLS-DA models (SEN > 83.3%, SPE > 97.2%, and RLR > 84.0%), achieving effective discrimination of adulterations made with bark, straw, and low-quality beans. The introduction of synthetic samples in the training set promoted the balancing of the target class (pure specialty coffees) and improved the model's performance. Notably, this study is the first to incorporate the creation of synthetic samples in the analysis of specialty coffees.

However, the one-class methodologies (SIMCA and DD-SIMCA) showed low specificity (< 47.1%), obtaining models with low capacity to detect samples adulterated with bark, straw, and low-quality beans. This was due to the high heterogeneity of samples of pure specialty coffees produced in different cities in the Fluminense

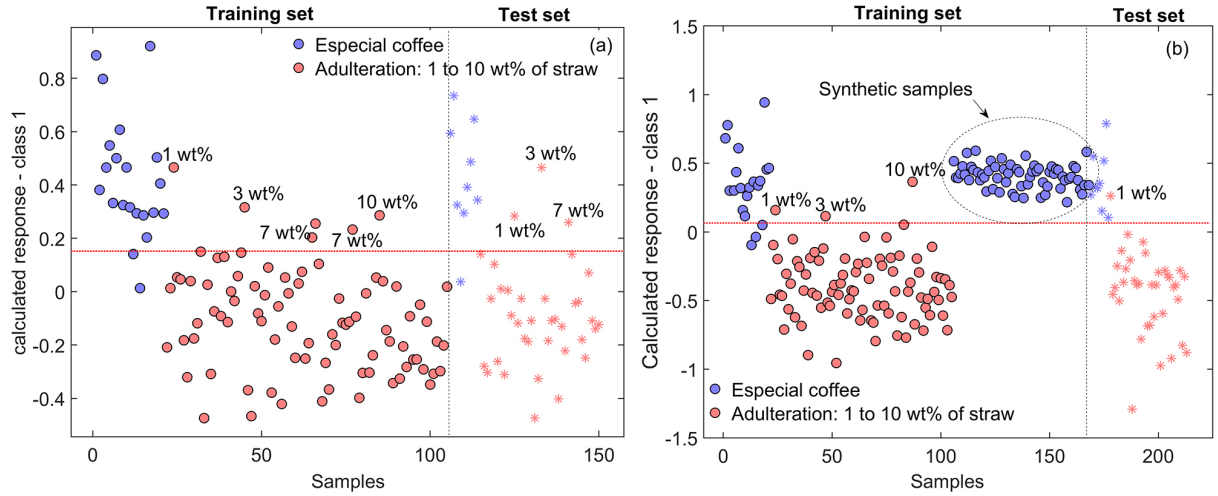

Figure 4. Graphs of scores of (a) Hard PLS-DA and (b) SS-PLS-DA model with pure and special coffees adulterated with straw (1 to 10 wt.%).

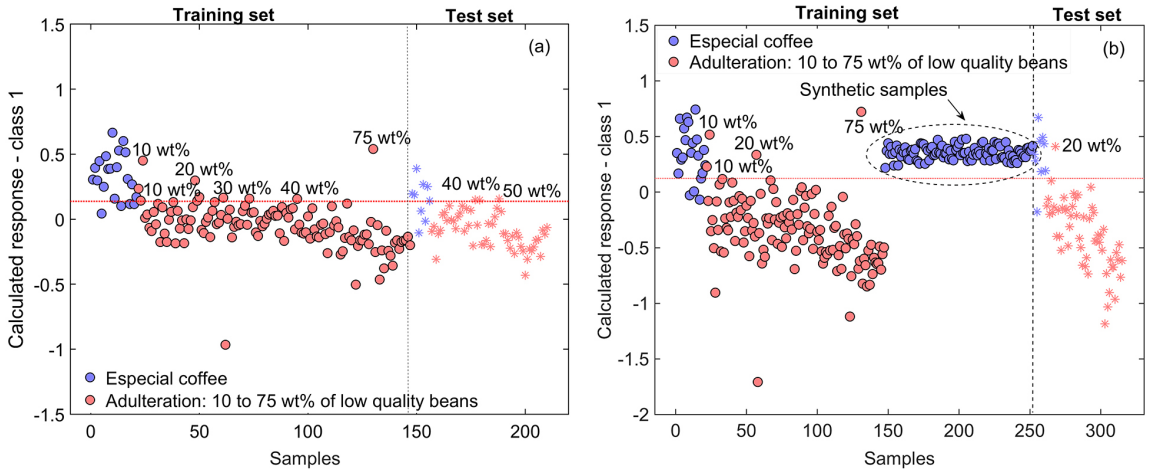

Figure 5. Scores graphs of (a) Hard Model PLS-DA and (b) SS-PLS-DA model with pure and adulterated specialty coffees with low-quality beans (10 to 75 wt.%).

northwest and Caparaó regions, generating a dataset with high variability.

Finally, this work contributes to the strengthening of coffees sold in the Fluminense northwest region, as the National Institute of Industrial Property (INPI) has been requested to provide a Geographical Indication (IG) for the coffees produced in this region. This will ensure that the coffees grown in this region have their own identity (special coffees from the "Alto Noroeste Fluminense"), guaranteeing greater added value to these coffees and generating greater income, especially for small family coffee growers in the region.

Supplementary Information

Supplementary information is available free of charge at http://jbcs.sbq.org.br as PDF file.

Acknowledgments

The authors Gabriel. R. F. Caldeira and Murilo O. Souza would like to thank the Fluminense Federal Institute of Education, Science and Technology (IFF) for the PIBIC scholarship (Edital 226/2021). The authors Tayná O. Costa and Murilo O. Souza would like to thank the Research Support Foundation of the State of Rio de Janeiro (FAPERJ) for the master's scholarship. Prof Murilo de O. Souza thanks the Technical Assistance and Rural Extension Institute of the northwest region of Rio de Janeiro (Gustavo P. Polido, EMATER-RJ), the Coffee Quality Laboratory of the Fluminense Federal Institute (Prof Daniel C. Ferreira, IFF) and to the Coffee Classification and Tasting Laboratory of the Federal Institute of Espírito Santo (Prof João B. P. Simão, IFES) for the special coffee samples provided. This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001.

Author Contributions

Gabriel R. F. Caldeira was responsible for investigation, data curation, formal analysis, writing original draft; Tayná O. Costa for data curation, formal analysis, writing original draft; Marcia H. C. Nascimento for data curation, formal analysis; Patricia G. Corradini for visualization, writing original draft, writing-review and editing; Paulo R. Filgueira for visualization, writing original draft, writing-review and editing; Daniel C. Ferreira for writing original draft, writing-review and editing; Murilo O. Souza for conception, visualization, validation, project administration, resources, writing original draft, writing-review and editing.

References

1. de Mendonça, R. F.; de Jesus Jr., W. C.; Ferrão, M. A. G.; Moraes, W. B.; Busato, L. M.; Ferrão, R. G.; Tomaz, M. A.; da Fonseca, A. F. A.; *Summa Phytopathol.* **2019**, *45*, 279. [Crossref]
2. Mendes, G. A.; de Oliveira, M. A. L.; Rodarte, M. P.; dos Anjos, V. C.; Bell, M. J. V.; *Curr. Res. Nutr. Food Sci.* **2022**, *5*, 298. [Crossref]
3. Associação dos Cafeicultores do Estado do Rio de Janeiro, https://ascarj.com.br/regiao-noroeste/, accessed in August 2023.
4. Ministério da Agricultura, Pecuária e Abastecimento, *Sumário Executivo Café*, https://estatisticas.abic.com.br/wp-content/uploads/2022/12/2022.12.SumarioCafe.pdf, accessed in August 2023.
5. Embrapa Café, https://www.embrapa.br/busca-de-noticias/-/noticia/1578252/demanda-por-cafes-especiais-do-brasil-cresce-15-ao-ano, accessed in August 2023.
6. Instituto Brasileiro de Geografia e Estatística, *Painel de Indicadores*, https://www.ibge.gov.br/indicadores, accessed in August 2023.
7. Adnan, A.; Naumann, M.; Mörlein, D.; Pawelzik, E.; *Foods* **2020**, *9*, 788. [Crossref]
8. Danezis, G. P.; Tsagkaris, A. S.; Brusic, V.; Georgiou, C. A.; *Curr. Res. Nutr. Food Sci.* **2016**, *10*, 22. [Crossref]
9. Ameca-Veneroso, C.; Sánchez-Arellano, L.; Ramón-Canul, L. G.; Herrera-Corredor, J. A.; Cuervo-Osorio, V. D.; Quetz-Aguirre, E. M.; Rodríguez-Miranda, J.; Cabal-Prieto, A.; Ramírez-Rivera, E. J.; *J. Sens. Stud.* **2021**, *36*, 12705. [Crossref]
10. Tavares, K. M.; Pereira, R. G. F. A.; Nunes, C. A.; Pinheiro, A. C. M.; Rodarte, M. P.; Guerreiro, M. C.; *Quim. Nova* **2012**, *35*, 1164. [Crossref]
11. Meng, X.; Yin, C.; Yuan, L.; Zhang, Y.; Ju, Y.; Xin, K.; Chen, W.; Lv, K.; Hu, L.; *Food Chem.* **2023**, *405*, 134828. [Crossref]
12. Greño, M.; Plaza, M.; Luisa Marina, M.; Castro Puyana, M.; *Food Chem.* **2023**, *402*, 134209. [Crossref]
13. de Araújo, T. K. L.; Nóbrega, R. O.; Fernandes, D. D. S.; de Araújo, M. C. U.; Diniz, P. H. G. D.; da Silva, E. C.; *Food Chem.* **2021**, *364*, 130452. [Crossref]
14. Magdas, D. A.; Feher, I.; Dehelean, A.; Cristea, G.; Magdas, T. M.; Puscas, R.; Marincaş, O.; *Food Chem.* **2018**, *267*, 231. [Crossref]
15. Siddiqui, A. J.; Musharraf, S. G.; Choudhary, M. I.; Rahman, A.-u.; *Food Chem.* **2017**, *217*, 687. [Crossref]
16. Paneque, P.; Morales, M. L.; Burgos, P.; Ponce, L.; Callejón, R. M.; *Food Control* **2017**, *75*, 203. [Crossref]
17. Capron, X.; Smeyers-Verbeke, J.; Massart, D. L.; *Food Chem.* **2007**, *101*, 1585. [Crossref]
18. Mussatto, S. I.; Carneiro, L. M.; Silva, J. P. A.; Roberto, I. C.; Teixeira, J. A.; *Carbohydr. Polym.* **2011**, *83*, 368. [Crossref]
19. Esquivel, P.; Jiménez, V. M.; *Food Res. Int.* **2012**, *46*, 488. [Crossref]

20. Ferreira, D. S.; Oliveira, M. E. S.; Ribeiro, W. R.; Altoé Filete, C.; Castanheira, D. T.; Rocha, B. C. P.; Moreli, A. P.; Oliveira, E. C. S.; Guarçoni, R. C.; Partelli, F. L.; Pereira, L. L.; *Agronomy* **2022**, *12*, 1885. [Crossref]
21. Nunes, K. M.; Andrade, M. V. O.; Almeida, M. R.; Sena, M. M.; *Food Anal. Methods* **2020**, *13*, 1699. [Crossref]
22. Suhandy, D.; Yulia, M.; *Int. J. Food Prop.* **2017**, *20*, S331. [Crossref]
23. Sánchez, B.; Souza, M. O.; Vilanova, O.; Canela, M. C.; *Build Environ.* **2020**, *174*, 106780. [Crossref]
24. Brazilian Specialty Coffee Association, https://brazilcoffeenation.com.br/cursos-sca, accessed in August 2023.
25. QGIS, https://www.qgis.org/en/site/, accessed in August 2023.
26. Instituto Nacional da Propriedade Industrial, https://www.gov.br/inpi/pt-br/servicos/indicacoes-geograficas, accessed in August 2023.
27. Souza, M. O.; Rainha, K. P.; Castro, E. V. R.; Carneiro, M. T. W. D.; Ferreira, R. Q.; *Quim. Nova* **2015**, *38*, 980. [Crossref]
28. Souto, U. T. C. P.; Barbosa, M. F.; Dantas, H. V.; de Pontes, A. S.; Lyra, W. S.; Diniz, P. H. G. D.; de Araújo, M. C. U.; da Silva, E. C.; *LWT - Food Sci. Technol.* **2015**, *63*, 1037. [Crossref]
29. Ballabio, D.; Consonni, V.; *Anal. Methods* **2013**, *5*, 3790. [Crossref]
30. Wold, S.; Esbensen, K.; Geladi, P.; *Chemometr. Intell. Lab. Syst.* **1987**, *2*, 37. [Crossref]
31. Zontov, Y. V.; Rodionova, O. Y.; Kucheryavskiy, S. V.; Pomerantsev, A. L.; *Chemometr. Intell. Lab. Syst.* **2017**, *167*, 23. [Crossref]
32. *Matlab* R2013a, The MathWorks, Natick, MA, USA.
33. Gosain, A.; Sardana, S.; *Farthest SMOTE: A Modified SMOTE Approach*; Springer Singapore: Singapore, 2019.
34. Zhang, X.; Li, H.; Tian, X.; Chen, C.; Su, Y.; Li, M.; Lv, J.; Chen, C.; Lv, X.; *Chemometr. Intell. Lab. Syst.* **2022**, *231*, 104681. [Crossref]
35. Torres, F. R.; Carrasco-Ochoa, J. A.; Martínez-Trinidad, J. F.; *SMOTE-D a Deterministic Version of SMOTE*; Springer International Publishing: Cham, 2016.
36. Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P.; *Int. J. Artif. Intell. Res.* **2022**, *16*, 321. [Crossref]
37. Correia, P. R. M.; Ferreira, M. M. C.; *Quim. Nova* **2007**, *30*, 481. [Crossref]
38. Zhang, J.; Yang, R.; Chen, R.; Li, Y. C.; Peng, Y.; Liu, C.; *Molecules* **2018**, *23*, 3013. [Crossref]
39. Wise, B. W.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; Windig, W.; Koch, R. S.; *PLS-Toolbox 4.0 for use with MatlabTM (Manual)*; Eigenvector Research, Inc: Wenatchee, 2006.
40. Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J.; *J. Chemom.* **2006**, *20*, 341. [Crossref]
41. Botelho, B. G.; Reis, N.; Oliveira, L. S.; Sena, M. M.; *Food Chem.* **2015**, *181*, 31. [Crossref]
42. Ríos-Reina, R.; Azcarate, S. M.; *Chemosensors* **2023**, *11*, 8. [Crossref]
43. Cavdaroglu, C.; Ozen, B.; *Food Chem.* **2022**, *379*, 132150. [Crossref]
44. Owen, T.; *Principles and Applications of UV-Visible Spectroscopy*; Agilent Technologies: Germany, 2000.
45. Manuel, M. N. B.; da Silva, A. C.; Lopes, G. S.; Ribeiro, L. P. D.; *Food Chem.* **2022**, *366*, 130480. [Crossref]
46. Pomerantsev, A. L.; Rodionova, O. Y.; *J. Chemom.* **2018**, *32*, e3030. [Crossref]
47. Rodionova, O. Y.; Titova, A. V.; Pomerantsev, A. L.; *Trends Anal. Chem.* **2016**, *78*, 17. [Crossref]
48. Oliveri, P.; *Anal. Chim. Acta* **2017**, *982*, 9. [Crossref]
49. Biancolillo, A.; Firmani, P.; Bucci, R.; Magrì, A.; Marini, F.; *Microchem. J.* **2019**, *145*, 252. [Crossref]
50. Santos, P. M.; Simeone, M. L. F.; Pimentel, M. A. G.; Sena, M. M.; *Microchem. J.* **2019**, *149*, 104057. [Crossref]

Submitted: March 31, 2023
Published online: August 24, 2023