Microchemical Journal 181 (2022) 107816

Contents lists available at ScienceDirect

Microchemical Journal

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

In-depth chemometric strategy to detect up to four adulterants in cashew nuts by IR spectroscopic techniques

Glòria Rovira [a], Carolina Sheng Whei Miaw [b], Mário Lúcio Campos Martins [b], Marcelo Martins Sena [c,d], Scheilla Vitorino Carvalho de Souza [b], Itziar Ruisánchez [a,*], M. Pilar Callao [a]

[a] Chemometrics, Qualimetric and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n, 43007 Tarragona, Spain
[b] Department of Food Science, Faculty of Pharmacy (FAFAR), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG, Brazil
[c] Chemistry Department, Institute of Exact Sciences (ICEX), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG, Brazil
[d] Instituto Nacional de Ciência e Tecnologia em Bioanalítica (INCT-Bio), Campinas, SP 13083-970, Brazil

ARTICLE INFO

Keywords:
Untargeted chemometrics
One-class SIMCA
NIR
ATR-FTIR
ROC curve
High-level data fusion

ABSTRACT

An untargeted strategy was developed to determine cashew nuts adulteration with Brazilian nuts, pecan nuts, macadamia nuts and peanuts. A one-class SIMCA model was developed for the cashew non-adulterated samples by means of two spectroscopic techniques: Near-Infrared (NIR) and Attenuated Total Reflection-Fourier Transform Infrared (ATR-FTIR). Receiver operating characteristic (ROC) curves have been proved to be useful to optimize class limits, both for the NIR and ATR-FTIR models, allowing to balance the values of the performance parameters. An increase in the sensitivity of the training and test set has been obtained from 79% with NIR and 85% with ATR-FTIR to 93% in both cases. As a result, the specificity has slightly decreased from 100% with NIR and a range of 90–98% with ATR-FTIR to a range of 82–98% and 84–96%, respectively. The implementation of high-level data fusion to the classification results obtained from NIR and ATR-FTIR, considering the limit value optimized by ROC curves, allowed the improvement of the performance parameters of the untargeted strategy. Obtaining sensitivity values for the training and test set of 100% and 93%, respectively. Specificity values of 100% were obtained for the detection of Brazilian nuts, macadamia nuts and peanuts, while for pecans it was 98%.

1. Introduction

The consumption of nuts and peanuts is widespread throughout the world not only due to their high organoleptic value, but also due to their beneficial effects on human health [1]. This food class includes a wide range of products such as almonds, Brazil nuts, cashew nuts, hazelnuts, macadamia nuts, peanuts, pecans, among others. Some of them present medium or high risk of food fraud due to adulteration, usually by adding cheaper and lower quality products [2]. Particularly, Brazil is among the major world producers of cashew nuts, which are also one of the preferred nuts of Brazilian consumers due to their pleasant taste. Thus, they are potential target of frauds. The market prices of this food products vary substantially. Currently in Brazil, cashew nuts are sold for R$51–65, while pecan, Brazil nuts and peanuts are sold for R$25–50, R $37–45 and R$4–10; macadamia nuts, which is primarily imported, present prices comparable to cashew nuts [3].

Detecting food adulteration is important for economic reasons but it is especially important when the non-declared substance involves a health risk. Food adulteration is always a concern due to the high complexity of food and the difficulty in detecting the presence of an adulterant in an easy and rapid way. In such scenario, it is of great interest the development of screening methods since it generally implies low time analysis, permitting a high throughput of samples at low cost, thus making them suitable for routine analysis. Today, the application of multivariate instrumental techniques together with chemometrics is a consolidated strategy in the field of food fraud detection. Multivariate supervised classification models might provide a binary response of the type there is / there is no adulteration. Some examples of these strategies

* Corresponding author.
E-mail address: itziar.ruisanchez@urv.cat (I. Ruisánchez).

https://doi.org/10.1016/j.microc.2022.107816
Received 22 June 2022; Received in revised form 21 July 2022; Accepted 24 July 2022
Available online 27 July 2022

for detecting nut fraud due to the addition of adulterants have been recently referenced [4], such as adulteration of almonds [5–8], pistachio [9] and hazelnut [10–12].

There is a wide range of both instrumental techniques and classification models that can be implemented to solve a given adulteration problem. Methods developed for nuts adulteration have included a study on hazelnut paste adulterated with almonds and chickpea four, which applied near-infrared spectroscopy (NIRS) combined with target and non-target modelling [10]; another article has detected hazelnut paste adulterated with almonds by applying data fusion of NIR and Raman spectra combined with class modelling [12]. A very recent article has applied and compared several multivariate classification models to short-wave infrared hyperspectral images aiming to detect contaminants in edible pistachio nuts, such as inedible pistachio nuts, pistachio shells, husks, twigs, and stones [13]. Among the analytical techniques most used to detect nut adulteration, it can be cited inductively coupled plasma optical emission spectroscopy (ICP-OES) [5], gas chromatography with flame ionization detector (GC-FID) [6], high performance liquid chromatography (HPLC) with fluorescence and UV detection [7,8], FT-Raman hyperspectral imaging [9], NIR [10,12], mid-infrared [11] and FT-Raman [12] spectroscopies. Regarding chemometric models, principal component analysis (PCA) has been used mainly as a data exploration tool, while the main supervised classification methods have been soft independent modelling of class analogies (SIMCA) [10–12], partial least squares discriminant analysis (PLS-DA) [7,8], support vector machine (SVM) [5,6] and linear discriminant analysis (LDA) [6]. Among chemometricians, the most used are SIMCA as a class modelling model and PLS-DA as a discriminant model.

A key step in the development of a screening method based on multivariate supervised classification is the choice of a suitable chemometric strategy. This imply a proper data pre-processing, model parameter optimization and the possible exploitation of the synergies between different data sources using data fusion, if more than one instrumental technique was used.

Signal pre-processing is applied to correct/remove the contribution of undesired phenomena ranging from stochastic measurement noise to various sources of systematic errors. Different possibilities have been critically discussed [14,15].

Regarding the classification technique, a choice must be performed between discriminating and class modelling. While discriminant methods establish a delimiter between two (or more) classes and split the hyperspace in a number of regions corresponding to the number of classes, class modelling methods model each class individually, irrespectively of the others. In view of this, some authors [16,17] have recently criticized the predominant use of discriminant methods in the literature, arguing that class modelling methods are more robust than discriminant ones considering the practical limitations in acquiring a sample set representative of all possible types of adulterations [18]. Class modelling methods offer the possibility of building just one class (untargeted modelling or one-class approach), being this strategy considered one of the most appropriate for food fraud detection [10,16,19]. If that the case, untargeted/one-class approach builds a class from the non-adulterated samples and detects which new samples resembles them no matter which adulterant under study is present. In fact, this approach is not a novelty, but its application has recently been incremented. The model can be optimized by selecting suitable class limits that define whether a sample is considered unadulterated or not (fits the model). Studies of this type has been recently published [11,16,18–21]. In the case of the present study, it is important to note that the use of one-class modelling ensures the representativeness of the model, avoiding the need to expand the comprehensiveness of the samples by incorporating other types of nuts or adulterants as non-authentic classes. This is in contrast with the use of discriminant models, such as PLS-DA. When using one-class modelling, the decision regarding compliance is not at all influenced by out-of-class samples. Therefore, this avoids any need to collect a representative data set that includes all the possible sources of out-of-specification variations [10,22].

Classification results can be improved by implementing data fusion, which allows to obtain a single result from more than one source [18,23]. In fact, data fusion has been growing used in the last years for detecting adulterations and frauds in food matrices such as meat and nuts [12,24,25]. Currently, this is a consolidated chemometric strategy, although it requires extra experimentation. There are three types of data fusion: low-, mid- and high-level data fusion. Details of the basis of each one of these levels can be found in the literature [12,23].

In this paper an untargeted strategy was developed to determine the possible adulteration of cashew nuts with other types of nuts (Brazilian nuts, pecan nuts, macadamia nuts and peanuts). A one-class SIMCA model was developed for cashew non-adulterated/authentic samples by means of two spectroscopic techniques, near-infrared (NIR) and attenuated total reflection-Fourier transform infrared (ATR-FTIR). Aiming to optimize the performance parameters of the method, different tools have been sequentially applied. First, different signal processing methods were studied. Second, model optimal limits have been determined by developing receiver operating characteristic (ROC) curves. Finally, high-level data fusion has been applied and compared with the result obtained from the models established with the two individual techniques (NIR and ATR-FTIR). In spite of the recent publication of many articles developing multivariate qualitative methods to detect adulterations, particularly on food analysis, very few of them have exhaustively explored all the chemometric possibilities available today.

2. Materials and methods

2.1. Samples

Commercial batches of each nut (Cashew nut, Brazilian nut, Macadamia nut, Peanut and Pecan) were acquired from certified producers. They were crushed in a sample processor (Arno Magiclean WWBC Blender), homogenized, sieved to size 40 mesh using calibrated tamis, packed in polyethylene packaging, sealed, and kept at room temperature (25 ± 3 °C) until preparation of the formulated batches. This processing aimed to simulate ground nuts product. Unadulterated/authentic samples of cashew nuts were composed of seven formulated batches prepared in eight variations, giving a total of 56 samples. Adulterated samples were prepared from batches of the corresponding adulterant nut (Brazilian nut, Macadamia nut, Peanut and Pecan) plus different amounts of the seven formulated batches of the unadulterated samples (8 levels of adulteration: 10.0; 5.0; 2.5; 1.3; 0.6; 0.3; 0.2 and 0.1 % w/w). The total of adulterated samples was 224 (4×56 samples). The experimental design used to formulate the samples were previously described [26]. The 56 samples of non-adulterated cashew nuts were systematically divided into training (42 samples) and test (14 samples) sets by employing the Kennard-Stone algorithm [27].

2.2. NIR spectroscopy

NIR analysis was conducted using a portable MicroNIR® 1700 equipment from Viavi Solution (San Jose, CA, USA) in the diffuse reflectance mode. Its dispersive element is a linear variable filter (LVF), and its detector is a 128-pixel InGaAs photodiode array. This detector is a variable-band semiconductor with excellent optical properties. Spectralon was used as a reflectance standard reference. A sample portion, previously homogenized, was placed on a Petri dish (3.5 cm in diameter $\times$ 1.2 cm in height) until complete covered. Then, the plate was placed on the MicroNIR®, a reading was performed with 20 scans at a resolution of 6.25 nm and spectra were recorded in the wavelength range from 908 to 1676 nm (11013–5967 cm^{-1}). Readings were performed randomly, under repeatability conditions. The reflectance values were converted into pseudo-absorbance, log(1/R) prior to data processing.

2.3. ATR-FTIR spectroscopy

ATR-FTIR analysis was carried out using a Perkin Elmer Frontier spectrophotometer (Waltham, MA, USA) equipped with a deuterated triglycine sulphate detector and a single-reflection diamond crystal ATR accessory. A sample portion, previously homogenized, was placed on the ATR crystal. Sample was pressed at a constant pressure level with a metallic tip accessory. The spectrum of each sample was recorded with 16 scans at a resolution of 4 cm^{-1}, from 4000 to 650 cm^{-1}. Readings were performed randomly, under repeatability conditions. The reflectance values were converted into pseudo-absorbance, log(1/R), prior to data processing.

2.4. Software

Recorded data were processed and models were built by using MATLAB software, version 8.0.0.783 – R2012b (Natick, MA, USA) and PLS_Toolbox version 7.0.2 (Eigenvector Research Inc., Wenatchee, WA, USA).

2.5. Simca

SIMCA is a multivariate supervised class-modelling technique that models each class independently from all the others [28]. Assignation of unknown samples has evolved from the first criteria proposed by Wold et al. in 1976, but whatever the criteria, it is always related to calculating a distance to the model [28]. One of the SIMCA modifications implies defining the limits of the two scalar statistics, Hotelling T^2 and Q residues (Hotelling T^2_{lim} and Q_{lim}) at a specific significance level (α), normally set at 0.05. Once defined, there are several criteria to assign or classify a sample in a certain class. One criterion is that a sample should have values of both statistic parameters lower than the two statistic limits to be considered as belonging to the class model.

Another criterion of sample assignation is based on calculating the distance of a sample from the class. The distance of a sample *i* from the class *j* (d_{ij}) is a combination of its reduced statistic parameters expressed as in the following equation (Eq. (1)).

$$d_{ij} = \sqrt{\left(Q_{r,i}\right)^2 + \left(T^2_{r,i}\right)^2} \quad (1)$$

where "*r*" stands for the ratio between the statistics of sample "*i*" (T_i^2 and Q_i) and the corresponding class frontiers (T^2_{lim} and Q_{lim}).

Once the distance value is calculated, a sample could be assigned to a class model when its distance value is lower than 1 [29,30], $\sqrt{2}$ [28,31] or an optimized distance class limit calculated by means of the ROC curves [19]. In our study, the first criterion, distance limit of 1.0 for the target class, was initially adopted, and then the model's limit were optimized using more robust criteria based on ROC curves.

2.6. Receiver operating curves (ROC)

ROC curve represents sensitivity versus 1-specificity for a considered parameter (score) used as a criterion to classify. Therefore, it allows visualizing, organizing, and selecting classifiers (scores) on the basis of their performance [17,32,33]. As it has been previously stated, ROC curves can be implemented to optimize class limit values of a class.

2.7. High-level data fusion

High level data fusion combines the assignation results obtained from the classification models of each individual data source. In this work, the fusion has been performed following the fuzzy set theory by choosing as operators minimum, maximum, average and product values. The final decision (*ensemble decision*) is obtained by the majority vote provided by all the fuzzy operators [12,34,35].

3. Results and discussion

Fig. 1a and Fig. 2a shows the average original spectra for each class, the non-adulterated cashew nut samples and the adulterated ones with Brazilian nuts (BN), pecan nuts (PN), macadamia nuts (M) and peanuts (P), recorded with the two instrumental techniques employed, NIR and ATR-FTIR, respectively.

By observing Fig. 1a, the largest NIR bands are present approximately between 8550 and 7690 cm^{-1} and 7140–6670 cm^{-1}. The first band can be assigned to the second overtone of the C—H stretching, while the second band to the first overtone of the O—H stretching. Particularly, the spectral band centred around 8330 cm^{-1} has been reported as discriminant of nuts in relation to other food materials (wheat, milk, and cocoa) [36]. The smaller band between 7245 and 7090 cm^{-1} can be assigned to the combination band of C—H vibrations [37]. By observing Fig. 2a, the most important absorption regions were observed in the fingerprint region (1750–1050 cm^{-1}) assigned to carbonyl ester stretching, -C—N amide II and III stretching, and -C—H symmetric stretching vibration modes; around 2850 cm^{-1}, related to C—H asymmetric and symmetric stretching vibrations of long-chain fatty acids; and 3500–3000 cm^{-1}, related to axial bending of OH and NH bonds. The peak around 1750 cm^{-1} is assigned to the carbonyl group of fatty acid esters in fats, while absorption around 1560 cm^{-1} is associated with C—N stretching and N—H bending modes in proteins. The region between 1300 and 1100 cm^{-1} presents C—H stretching vibrations of carbohydrates [25]. Thus, the discrimination between different nuts provided by NIR and FTIR spectra in combination with chemometrics might be related to their different contents of components, such as proteins, lipids and carbohydrates.

Before chemometric modelling, some pre-processing was necessary aiming to eliminate non-linear baseline deviations caused by multiplicative scatter and to improve the signal-to-noise ratio. After some trails (supervised models), the first derivative followed by mean centering was applied to NIR spectra. For ATR-FTIR spectra, smoothing with Savitsky-Golay algorithm with a window width of 5, followed by multiplicative scatter correction (MSC) and generalized least squares weighting (GLSW) with an alpha of 0.01 were applied [12,24,38,39].

It is always advisable to apply PCA previously to developing the supervised classification. Thus, PCA was applied to each pre-treated spectroscopic dataset, aiming at observing any possible discriminating trend between cashew nuts authentication and adulteration. Fig. 3 shows PC1 versus PC2 score plots for each dataset. For NIR spectra (Fig. 3a), the first two PC accounted for 88.1 % of the total variance. PC2 (6.7 %) clearly showed a discriminating trend with most of the non-adulterated cashew nuts presenting positive scores, in contrast with adulterated samples, whose scores were predominantly negative. As expected, the comparison between the average pre-treated spectra (Fig. 1b) and the PCA loadings values of the second PC (Fig. 1c) shows shape similarities. More specifically around 6100 cm^{-1}, 7200 cm^{-1} and 8600 cm^{-1}, which in a certain way indicate their relevance in the discrimination of non-adulterant related to adulterate samples.

For ATR-FTIR spectra (Fig. 3b), variance was more partitioned between several PC, and the first two accounted for only 27.8 % of the total variance. In this case, PC1 (19.9 %) clearly discriminated non-adulterated samples in its positive part. In contrast, almost all the samples adulterated with Brazilian, pecan and macadamia nuts showed negative scores on PC1, while samples adulterated with peanuts presented scores in an intermediate region. This intermediate behaviour of samples adulterated with peanuts was already observed along PC2 in the NIR PCA model (Fig. 3a). When performing the ATR-FTIR spectra pretreatment (Fig. 2b), the differences between the average signal of the non-adulterated samples with respect to the average signals of the adulterated samples were magnified. Likewise, it can be seen that the loadings of the first PC (Fig. 2c) present shape similarities with the pretreated average spectrum of the non-adulterated samples (green line, Fig. 2b).

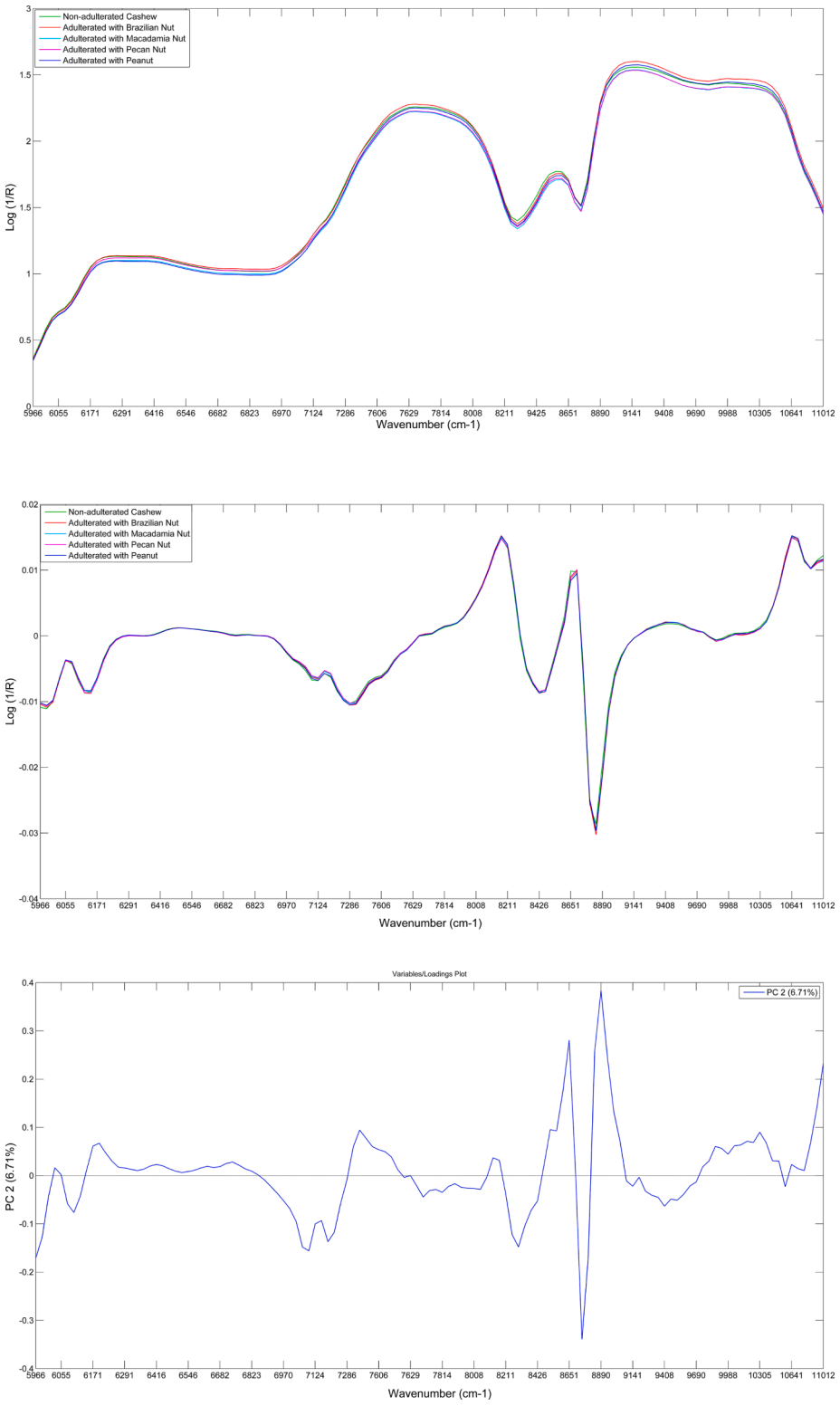

Fig. 1. NIR spectra of non-adulterated and adulterated samples. (a) the average original spectra, (b) the average pre-treated spectra and (c) PCA loadings values of the second PC obtained by NIR data. Color code: green for non-adulterated cashew nuts, red for Brazilian nuts, pink for pecan nuts, light blue for macadamia nuts and dark blue for peanuts.

For both instrumental techniques, no differences were observed in PCAs score plots related to the adulteration levels in any of the adulterants. Thus, these two PCA score plots suggested that it is possible to split variances related to cashew nuts authentication and adulteration by utilizing one-class modelling.

In the sequence, SIMCA models were built individually for each technique. These models were obtained using 42 training samples from non-adulterated cashew nuts. For the validation/test set, 14 non-adulterated samples were used together with all samples containing the four adulterants (BN, PN, M and P). Thus, two one-class classification models were constructed, one with NIR spectra and another with ATR-FTIR spectra. Leave-one-out cross-validation on the training

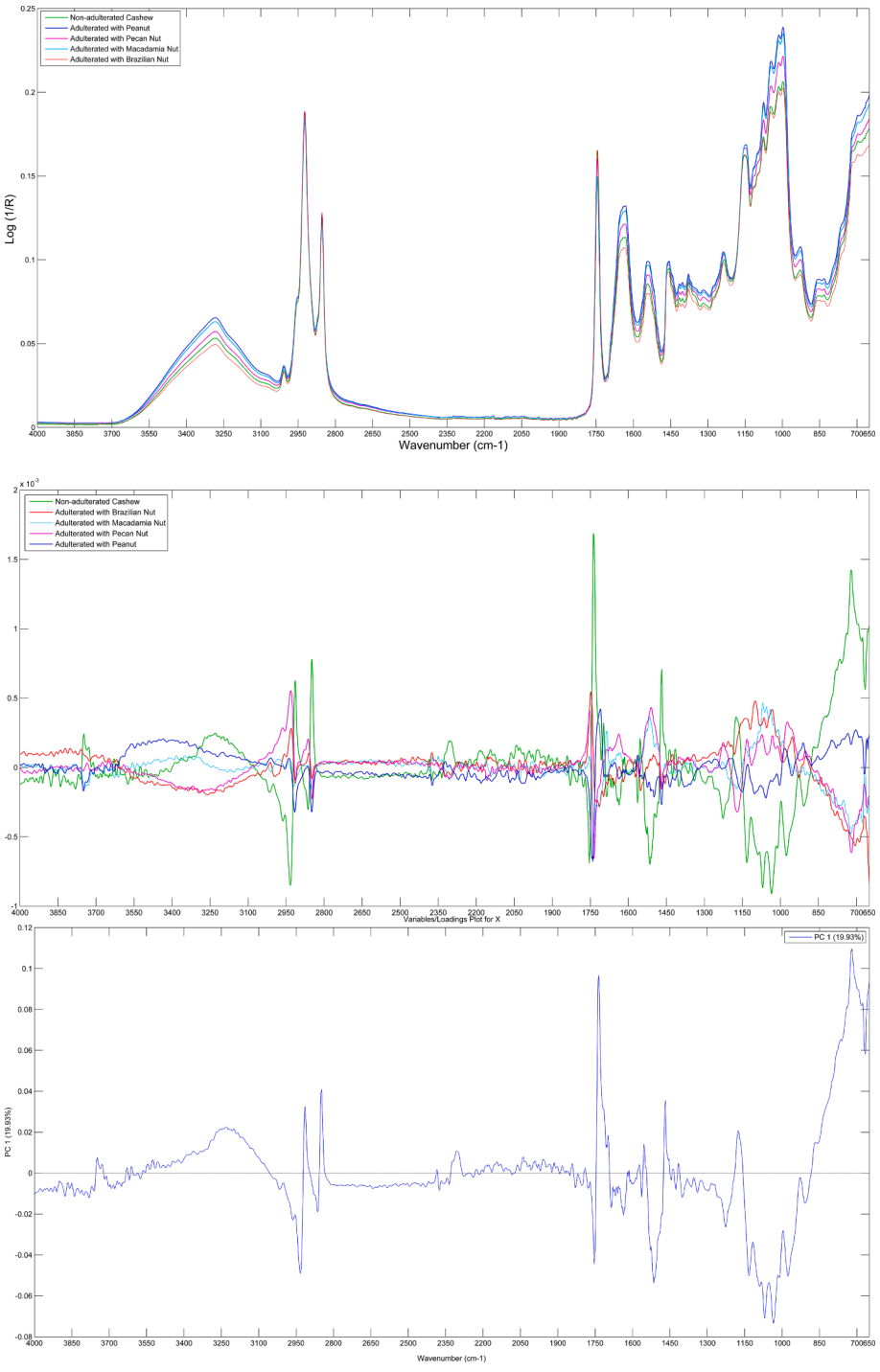

Fig. 2. ATR-FTIR spectra of non-adulterated and adulterated samples. (a) the average original spectra, (b) the average pre-treated spectra and (c) PCA loadings values of the first PC obtained by ATR-FTIR data. Color code: green for non-adulterated cashew nuts, red for Brazilian nuts, pink for pecan nuts, light blue for macadamia nuts and dark blue for peanuts.

samples has been used to decide the number of retained PCs for each model, based on the lowest cross-validation classification error (CVCE). For NIR SIMCA model, 5 PC and were chosen accounting for a 98.3 % of the spectral variance. For ATR-FTIR SIMCA model, 10 PC were selected representing a cumulative spectral variance of 91.8 %.

Sample assignation to authentic class was performed using the distance value according to Eq. (1). Initially, the criterion adopted to assign samples to the target class was a distance value lower than 1.0. Performance parameters for these models are shown in Table 1, which include efficiency as a global parameter calculated as the ratio between the number of the true assignments (TP + TN) and the total number of samples. One-class SIMCA results obtained for NIR model indicated a training sensitivity (rate of true positives) of 95 %, which means that the model properly assigned the own samples used to build it. While the test set sensitivity was lower, around 79 %. A specificity (rate of true negatives) of 100 % was achieved for all adulterants, meaning that the model properly recognized them as adulterated. The results obtained for ATR-FTIR model showed lower performance than NIR model, with the sensitivity of the training and test sets of 83 and 86 %, respectively. The specificity values for this model were slightly lower than 100 % (91–98

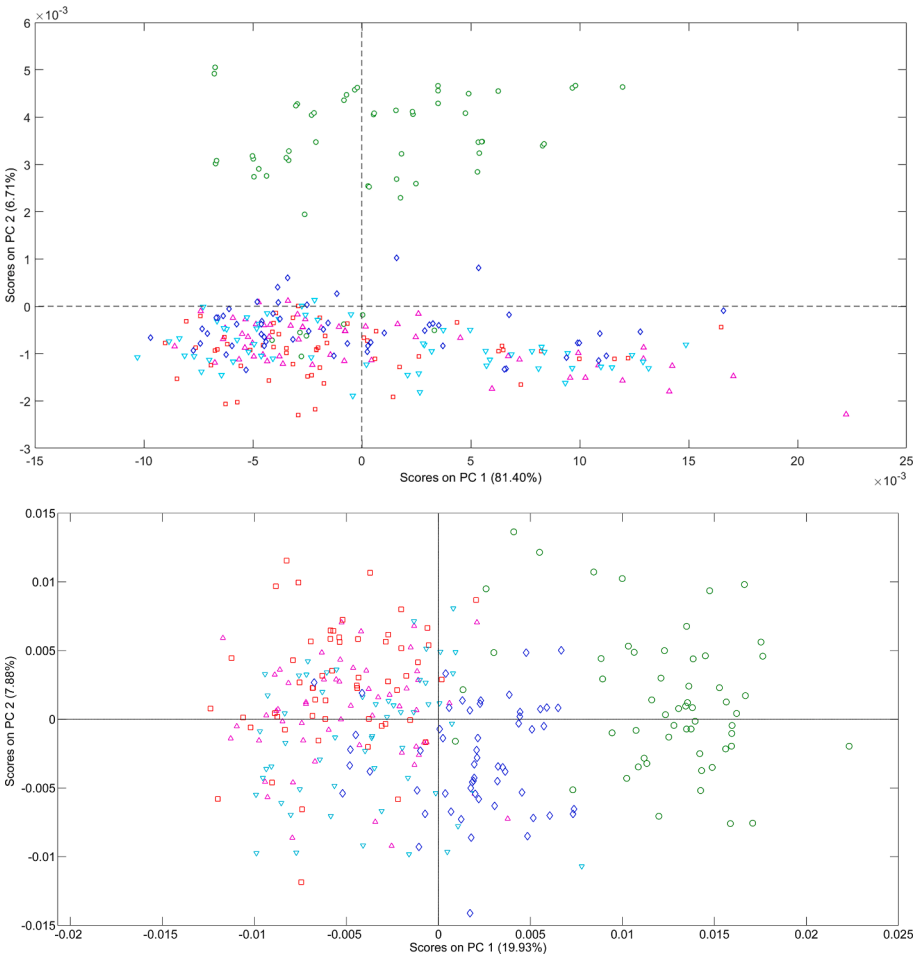

Fig. 3. Score plot of PC1 vs PC2 for (a) NIR data and (b) ATR-FTIR data. Color code: green circles for non-adulterated cashew nuts, red squares for Brazilian nuts, pink triangles for pecan nuts, light blue triangles for macadamia nuts and dark blue diamonds for peanuts.

Table 1
Performance parameters of one-class SIMCA models for NIR and ATR-FTIR data. NA: non adulterated cashew nuts; BN: Brazilian nuts; PN: Pecan nuts; M: Macadamia nuts; P: Peanuts and BN/PN/M/P: all adulterants.

		NA	BN	PN	M	P	BN/PN/M/P
NIR	Sensitivity training	95.24 %					
	Sensitivity test	78.57 %					
	Specificity		100 %	100 %	100 %	100 %	100 %
	Efficiency		95.71 %	95.71 %	95.71 %	95.71 %	98.74 %
ATR-FTIR	Sensitivity training	83.33 %					
	Sensitivity test	85.71 %					
	Specificity		98.21 %	98.21 %	91.07 %	92.86 %	95.09 %
	Efficiency		95.71 %	95.71 %	90.00 %	91.43 %	94.54 %

%).

Both individual models showed an ability of detecting adulterated samples higher than the rate of recognition of non-adulterated samples. This behaviour suggests that a change in the class limit could lead to higher sensitivity values or a more suitable balance between both parameters (sensitivity and specificity).

In order to optimize distance class limits for each SIMCA model, ROC curves were constructed. Fig. 3 shows ROC curves obtained for one-class classification models built with NIR (Fig. 4a) and ATR-FTIR (Fig. 4b) spectra. ROC curves were estimated using the distance defined in Eq. (1) as the basis (score) for calculating the performance parameters. Specifically, the distances of the test samples (14 non-adulterated samples, and 56 samples for the presence of each adulterant) were used. The optimal distance is the one closest to the point (0, 1), which corresponds to both sensitivity and specificity equal to 100 %. In Fig. 3, optimal distances are marked with blue points, corresponding to 1.44 and 1.11, for the NIR and ATR-FTIR models, respectively. These distances were therefore the class limits, which means that samples with distances lower or equal than 1.44 in the NIR model and 1.11 in the ATR-FTIR model will be considered as belonging to the target class. Reciprocally, samples with larger distances will be considered as not belonging to the target class.

Figures of merit obtained by applying the optimal distances estimated based on ROC curves are shown in Table 2. For the NIR model, sensitivities of both training and test sets were improved. The most remarkable improvement was observed for the test set, from 79 % to 93 %. Specificity and efficiency reasonably decreased, regardless of the adulterant, presenting values between 82 % and 98 %. For the ATR-FTIR

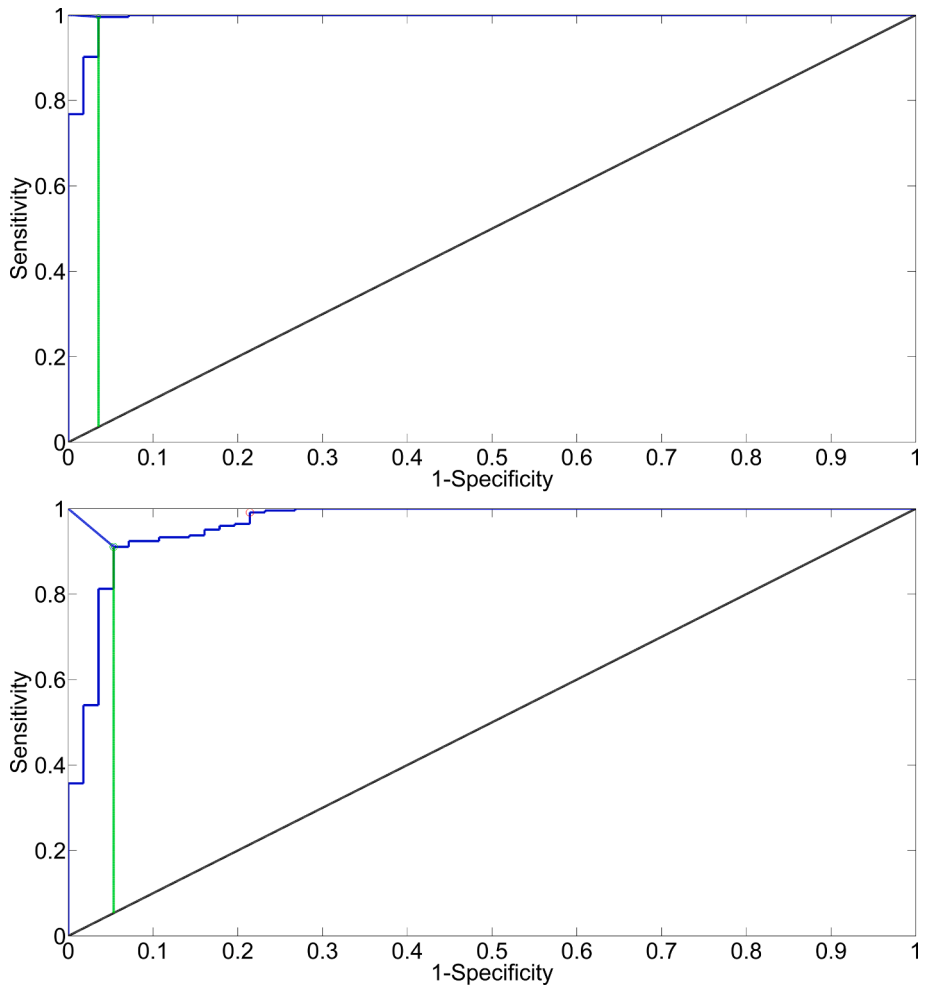

Fig. 4. Receiver operating characteristic (ROC) curves estimated for one-class SIMCA models built with (a) NIR data set and (b) ATR-FTIR data set.

Table 2
Performance parameters of one-class SIMCA models optimized with optimal distances based on ROC curves for NIR and ATR-FTIR data. NA: non adulterated cashew nuts; BN: Brazilian nuts; PN: Pecan nuts; M: Macadamia nuts; P: Peanuts and BN/PN/M/P: all adulterants.

		NA	BN	PN	M	P	BN/PN/M/P
NIR	Sensitivity training	100 %					
	Sensitivity test	92.86 %					
	Specificity		85.71 %	82.14 %	94.64 %	98.21 %	90.18 %
	Efficiency		87.14 %	84.29 %	94.29 %	97.14 %	90.34 %
ATR-FTIR	Sensitivity training	92.86 %					
	Sensitivity test	92.86 %					
	Specificity		96.43 %	96.43 %	91.07 %	83.93 %	91.96 %
	Efficiency		95.71 %	95.71 %	91.43 %	85.71 %	92.02 %

model, sensitivities of the training and the test were also improved, increasing from 83 to 86 % to 93 %. Following a trend similar to the NIR model, specificity slightly decreased for all adulterants except for peanuts, which showed a greater decrease, from 93 % to 84 %. Similarly, the efficiency did not change or slightly increased for all adulterants except for Peanut.

It was observed that implementing class limits based on ROC curves balanced sensitivity and specificity values of SIMCA models. In this study, choosing minimum distances using ROC curves helped to maximize the sensitivity, which is equivalent to increasing the ability of the model to recognize target samples (non-adulterated/authentic). In contrast, specificity decreased, while efficiency remained the same. This statement can be clearly realized by evaluating the total specificity value, that is, the specificity regardless of the type of adulterant used (from 100 % to 90 % and from 95 % to 92 %, for NIR and ATR-FTIR models, respectively).

In the sequence, high-level data fusion was implemented aiming to improve the results obtained with SIMCA models built with individual techniques. In order to build a high-level fusion model, samples misclassified by the NIR model but correctly classified by the ATR-FTIR model or vice-versa, were selected. In total, 45 samples were chosen as susceptible to apply the fuzzy operators. Before applying them, distance values obtained from one-class SIMCA models of each instrumental technique were normalized to 1.0. In such a way, the contributions of both models were balanced.

Figures of merit obtained for the high-level data fusion model (Table 3) clearly demonstrated the improvement over NIR and ATR-FTIR individual models. Training sensitivity was 100 %, the same

Table 3
Performance parameters of the optimized high-level data fusion one-class SIMCA model. NA: non adulterated cashew nuts; BN: Brazilian nuts; PN: Pecan nuts; M: Macadamia nuts; P: Peanuts and BN/PN/M/P: all adulterants.

		NA	BN	PN	M	P	BN/PN/M/P
Optimized	Sensitivity training	100 %					
	Sensitivity test	92.86 %					
	Specificity		100 %	98.21 %	100 %	100 %	99.55 %
	Efficiency		98.57 %	97.14 %	98.57 %	98.57 %	99.16 %

value as for the NIR model based on a ROC curve (Table 2) and higher than for the ATR-FTIR model (93 %). Sensitivity for the test set obtained with high-level data fusion was the same as in Table 2. However, data fusion provided simultaneously specificities close to 100 % (above 97 % for all adulterants), much better than SIMCA models built with individual vibrational techniques (Tables 1 and 2). The efficiency was also improved (between 97 % and 99 %) as compared to both the situations, using the class distances obtained by ROC curve (between 85 % and 97 %) or equal to 1.0 (between 91 % and 96 %). Thus, the use of a high-level data fusion strategy was justified considering the improvement of the classification results in comparison with individual spectral data.

4. Conclusions

An untargeted strategy to the rapid detection of adulteration in cashew nuts using portable NIR and ATR-FTIR spectroscopies jointly with one-class SIMCA models was developed. The specificity against four types of nuts (Brazilian nuts, pecan nuts, macadamia nuts and peanuts) was established. The implementation of the receiver operating characteristic (ROC) curves allows optimizing target class limits, that is, the limit of authentic cashew nuts class below which a sample will be considered as non-adulterated. As a result, it has been proved that the proposed strategy allowed to balance the values of the performance parameters (sensitivity and specificity), providing information on the probability of success in the assignments of non-adulterated and adulterated samples.

A high-level data fusion model based on Fuzzy operators was constructed resulting in an improvement of the performance parameters as compared to models based on individual techniques. It should be emphasized that the development of a high-level data fusion implied the need of sample measurements by at least two instrumental techniques. However, once the classification model was built and optimized, no additional effort is necessary as in the case of applying low- or mid-level data fusion.

The proposed strategy of one-class modelling is preferred when the target class to be modelled is the non-adulterated/authentic, regardless of the possible adulterants. Given the need for constant improvements in food fraud detection, this study represents a contribution to the food scientific community that can easily be extended to other types of nut fraud or involving other products/matrices. The developed analytical methodology is simple, rapid, green (does not consume reagents or solvents nor generates chemical waste) and non-destructive, thus being considered suitable for screening analysis.

CRediT authorship contribution statement

Glòria Rovira: Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Carolina Sheng Whei Miaw:** Formal analysis, Investigation, Methodology, Writing – review & editing. **Mário Lúcio Campos Martins:** Formal analysis, Methodology. **Marcelo Martins Sena:** Resources, Supervision, Writing – review & editing, Funding acquisition. **Scheilla Vitorino Carvalho de Souza:** Conceptualization, Resources, Supervision, Writing – review & editing, Funding acquisition. **Itziar Ruisánchez:** Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – review & editing, Funding acquisition. **M. Pilar Callao:** Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgements

This study was supported by the research program “Program of research activity (2020PMF-PIPF) at the Rovira i Virgili University, Tarragona, Spain. The acquisition of a portable NIR spectrometer (Viavi MicroNIR® 1700) was supported by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) through project APQ03457-16.

References

[1] R.G.M. de Souza, R.M. Shincaglia, G.D. Pimentel, J.F. Mota, Nuts and human health outcomes: A systematic review, Nutrients 9 (2017) 1311, https://doi.org/10.3390/nu9121311.
[2] P. Visciano, M. Schirone, Food frauds: Global incidents and misleading situations, Trends Food Sci. Technol. 114 (2021) 424–442, https://doi.org/10.1016/j.tifs.2021.06.010.
[3] M.F. Rural, Castanhas à venda com preço, Mercado Físico Rural, Marília, Brazil, 2022 https://www.mfrural.com.br/produtos/2-681/alimentos-castanhas, (accessed in May 2022.
[4] A. Valdés, A. Beltrán, C. Mellinas, A. Jiménez, M.C. Garrigós, Analytical methods combined with multivariate analysis for authentication of animal and vegetable food products with high fat content, Trends Food Sci. Technol. 77 (2018) 120–130, https://doi.org/10.1016/j.tifs.2018.05.014.
[5] M. Esteki, Y.V. Heyden, B. Farajmand, Y. Kolahderazi, Qualitative and quantitative analysis of peanut adulteration in almond powder samples using multi-elemental fingerprinting combined with multivariate data analysis methods, Food Contr. 82 (2017) 31–41, https://doi.org/10.1016/j.foodcont.2017.06.014.
[6] M. Esteki, Y.V. Heyden, B. Farajmand, Y. Kolahderazi, Cromatographic fingerpinting with multivariate data analysis for detection and quantification of apricot kernel in almond powder, Food Anal. Methods 10 (2017) 3312–3320, https://doi.org/10.1007/s12161-017-0903-5.
[7] G. Campmajó, G.J. Navarro, N. Nuñez, L. Puignou, J. Saurina, O. Nuñez, Non-Targeted HPLC-UV Fingerprinting as chemical descriptors for the classification and authentication of nuts by multivariate chemometric methods, Sensors 19 (2019) 1388, https://doi.org/10.3390/s19061388.
[8] G. Campmajó, R. Saez-Vigo, J. Saurina, O. Nuñez, High-performance liquid chromatography with fluorescence detection fingerprinting combined with chemometrics for nut classification and the detection and quantitation of almond-based product adulterations, Food Contr. 114 (2020), 107265, https://doi.org/10.1016/j.foodcont.2020.107265.
[9] H. Eksi-Kocak, O. Mentes-Yilmaz, I.H. Boyaci, Detection of green pea adulteration in pistachio nut granules by using Raman hyperspectral imaging, Eur. Food Res. Technol. 242 (2016) 271–277, https://doi.org/10.1007/s00217-015-2538-3.
[10] M.I. López, E. Trullols, M.P. Callao, I. Ruisánchez, Multivariate screening in food adulteration: Untargeted versus targeted modelling, Food Chem. 147 (2014) 177–181, https://doi.org/10.1016/j.foodchem.2013.09.139.
[11] M.I. López, N. Colomer, I. Ruisánchez, M.P. Callao, Validation of multivariate screening methodology. Case study: Detection of food fraud, Anal. Chim. Acta 827 (2014) 28–33, https://doi.org/10.1016/j.aca.2014.04.019.
[12] C. Márquez, M.I. López, I. Ruisánchez, M.P. Callao, FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud, Talanta 161 (2016) 80–86, https://doi.org/10.1016/j.talanta.2016.08.003.
[13] G. Bonifazi, G. Capobianco, R. Gasbarrone, S. Serranti, Contaminant detection in pistachio nuts by different classification methods applied to short-wave infrared

hyperspectral images, Food Contr. 130 (2021), 108202, https://doi.org/10.1016/j.foodcont.2021.108202.
[14] J.M. Roger, J.C. Boulet, M. Zeaiter, F. Marini, Pre-processing Methods, in: S. D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Chemical and Biochemical Data Analysis, 2020, pp. 1–75.
[15] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, TrAC Trends Anal. Chem. 132 (2020), 116045, https://doi.org/10.1016/j.trac.2020.116045.
[16] O. Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemometr. Intell. Lab. Syst. 159 (2016) 89–96, https://doi.org/10.1016/j.chemolab.2016.10.002.
[17] P. Oliveri, Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues - A tutorial, Anal. Chim. Acta 982 (2017) 9–19, https://doi.org/10.1016/j.aca.2017.05.013.
[18] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Contr. 86 (2018) 283–293, https://doi.org/10.1016/j.foodcont.2017.11.034.
[19] I. Ruisánchez, A.M. Jiménez-Carvelo, M.P. Callao, ROC curves for the optimization of one-class model parameters A case study: Authenticating extra virgin olive oil from Catalan protected designation of origin, Talanta 222 (2021), 121564, https://doi.org/10.1016/j.talanta.2020.121564.
[20] R. Vitale, F. Marini, C. Ruckebusch, SIMCA modeling for overlapping classes: fixed or optimized decision limit? Anal. Chem. 90 (2018) 10738–10747, https://doi.org/10.1021/acs.analchem.8b01270.
[21] B. Quintanilla-Casas, J. Bustamante, F. Guardiola, D.L. García-González, S. Barbieri, A. Bendini, T.G. Toschi, S. Vichi, A. Tres, Virgin olive oil volatile fingerprint and chemometrics: Towards an instrumental screening tool to grade the sensor quality, LWT – Food Sci Technol. 121 (2020), 108936, https://doi.org/10.1016/j.lwt.2019.108936.
[22] P. Oliveri, G. Downey, Multivariate class modeling for the verification of food-authenticity claims, TrAC Trends Anal. Chem. 35 (2012) 74–86, https://doi.org/10.1016/j.trac.2012.02.005.
[23] E. Borrás, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment- A review, Anal. Chim. Acta 891 (2015) 1–14, https://doi.org/10.1016/j.aca.2015.04.042.
[24] C. Alamprese, M. Casale, N. Sinelli, S. Lanteri, E. Casiraghi, Detection of minced beef adulteration with turkey meat by UV-vis, NIR and MIR spectroscopy, LWT – Food Sci, Technol. 53 (2013) 225–232, https://doi.org/10.1016/j.lwt.2013.01.027.
[25] D.P. Aykas, A. Menevseoglu, A rapid method to detect green pea and peanut adulteration in pistachio by using portable FT-MIR and FT-NIR spectroscopy combined with chemometrics, Food Contr. 121 (2021), 107670, https://doi.org/10.1016/j.foodcont.2020.107670.
[26] C.S.M. Miaw, M.L.C. Martins, M.M. Sena, S.V.C. de Souza, Screening method for the detection of other allergenic nuts in cashew nuts using chemometrics and a portable near-infrared spectrophotometer, Food Anal, Methods 15 (2022) 1074–1084, https://doi.org/10.1007/s12161-021-02184-0.
[27] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148, https://doi.org/10.1080/00401706.1969.10490666.
[28] M. Bevilaqua, R. Bucci, A.D. Magrì, R. Nescatelli, F. Marini, Classification and class-modelling in: F. Marini (Editor), Data handling in science and Technology, Elsevier 28 (2013) 171-233. https://doi.org/10.1016/B978-0-444-59528-7.00005-3.
[29] C.S.M. Miaw, M.M. Sena, S.V.C. de Souza, M.P. Callao, I. Ruisánchez, Detection of adulterants in grape nectars by attenuated total reflectance Fourier-transform mid-infrared spectroscopy and multivariate classification, Food Chem. 266 (2018) 254–261, https://doi.org/10.1016/j.foodchem.2018.06.006.
[30] C.S. Gondim, R.G. Junqueira, S.V.C. de Souza, I. Ruisánchez, M.P. Callao, Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies, Food Chem. 230 (2017) 68–75, https://doi.org/10.1016/j.foodchem.2017.03.022.
[31] C. Durante, R. Bro, M. Cocchi, A classification tool for N-way array based on SIMCA methodology, Chemometr. Intell. Lab. Syst. 106 (2011) 73–85, https://doi.org/10.1016/j.chemolab.2010.09.004.
[32] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (2006) 35–46, https://doi.org/10.1016/j.patrec.2005.10.010.
[33] M. De Figueiredo, C.B.Y. Cordella, D.J.R. Bouveresse, X. Archer, J.M. Bégué, D. N. Rutledge, A variable selection method for multiclass classification problems using two-class ROC analysis, Chemometr. Intell. Lab. Syst. 177 (2018) 35–46, https://doi.org/10.1016/j.chemolab.2018.04.005.
[34] Y. Li, Y. Huang, J. Xia, Y. Xiong, S. Min, Quantitative analysis of honey adulteration by spectrum analysis combined with several high-level data fusion strategies, Vib. Spectrosc. 108 (2020), 103060, https://doi.org/10.1016/j.vibspec.2020.103060.
[35] C.V. di Anibal, M.P. Callao, I. Ruisánchez, ^{1}H NMR and UV-visible data fusion for determining Sudan dyes in culinary spices, Talanta 84 (2011) 829–833, https://doi.org/10.1016/j.talanta.2011.02.014.
[36] S. Ghosh, P. Mishra, S.N.H. Mohamad, R.M. de Santos, B.D. Iglesias, P.B. Elorza, Discrimination of peanuts from bulk cereals and nuts by near infrared reflectance spectroscopy, Biosyst. Eng. 151 (2016) 178–186, https://doi.org/10.1016/j.biosystemseng.2016.09.008.
[37] H.E. Genis, S. Durna, I.H. Boyaci, Determination of green pea and spinach adulteration in pistachio nuts using NIR spectroscopy, LWT – Food Sci, Technol. 136 (2021), 110008, https://doi.org/10.1016/j.lwt.2020.110008.
[38] M.G. Nespeca, W.D. Pavini, J.E. Oliveira, Multivariate filters combined with interval partial least square method: A strategy for optimizing PLS models developed with near infrared data of multicomponent solutions, Vib. Spectrosc. 102 (2019) 97–102, https://doi.org/10.1016/j.vibspec.2019.05.001.
[39] O. Anjos, A.J.A. Sanots, L.M. Estevinho, I. Caldeira, FTIR-ATR spectroscopy applied to quality control of grape-derived spirits, Food Chem. 205 (2016) 28–35, https://doi.org/10.1016/j.foodchem.2016.02.128.