Variable selection for multivariate classification aiming to detect individual adulterants and their blends in grape nectars

Carolina Sheng Whei Miaw[a,b,c], Marcelo Martins Sena[d], Scheilla Vitorino Carvalho de Souza[a], Itziar Ruisanchez[c*], and Maria Pilar Callao[c]

[a] *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.*

[b] *CAPES Foundation, Ministry of Education of Brazil, 70040-020, Brasília, DF, Brazil.*

[c] *Chemometrics, Qualimetric and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n, 43007 Tarragona, Spain.*

[d] *Department of Chemistry, 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.*

* Corresponding author: itziar.ruisanchez@urv.cat

Abstract

During the quality inspection control of fruit beverages, some types of adulterations can be detected, such as the addition or substitution with less expensive fruits. To determine whether grape nectars were adulterated by substitution with apple or cashew juice or by a mixture of both, a methodology based on attenuated total reflectance Fourier transform mid infrared spectroscopy (ATR-FTIR) and multivariate classification methods was proposed. Partial least squares discriminant analysis (PLS-DA) and soft independent modeling of class analogy (SIMCA) models were developed as multi-class methods (classes unadulterated, adulterated with cashew and adulterated with apple) with the full-spectra. PLS-DA presented better performance parameters than SIMCA in the classification of samples with just one adulterant, while poor results were achieved for samples with blends of two adulterants when using both classification methods. Three variable selection methods were tested in order to improve the effectiveness of the classification models: interval partial least squares (iPLS), variable importance in projection scores (VIP scores) and a genetic algorithm (GA). Variable selection methods improved the performance parameters for the SIMCA and PLS-DA methods when they were used to predict samples with only one adulterant. Only PLS-DA coupled with iPLS was able to classify samples with blends of two adulterants, providing sensitivity values between 100 and 83% at 100% specificity for the three studied classes.

Keywords

Variable selection, multi-class methods, PLS-DA, SIMCA, grape nectar, food fraud

1. Introduction

Adulteration or possible food fraud is a problem that affects many food products and has an important economic impact. In the case of fruit beverages, the most frequent types of adulteration include the addition of water or syrup, acidification, addition or substitution with cheaper fruits and addition of colorants or flavors. Unfortunately, fruit beverages are one of the easiest products to adulterate because of their complex chemical composition and the wide natural variation of fruits [1].

Specifically, fruit nectars are unfermented beverages intended for direct consumption, which are formulated by dilution of the edible part of the fruits or their extracts with water and added sugars [2]. The most commonly consumed flavor of fruit nectar in Brazil, and one of the most expensive, is grape. Often, consumers choose grape nectars looking for a more nutritional product, since this fruit is rich in phenolic compounds, mainly flavonoids [3]. Due to their sensory characteristics and lower cost, apple and cashew juices are likely to be used as adulterants in some of the more expensive fruit nectars. Additionally, apple and cashew are fruits being suspected to be used as fillers by fraudulent industries based on denunciations received by the Ministry of Agriculture, Livestock, and Supply – MAPA, and some evidence in fiscal activities. Since the commercialization of nectars containing more than one fruit has been expanded in the market, the declared presence of each and every single fruit must be confirmed in order to guarantee the authenticity of the single fruit nectars and to prevent adulterations [4].

Recent methods have been developed to identify and detect different fruits in fruit beverages employing various analytical techniques. The use of ultra-performance liquid chromatography–quadrupole time of flight mass spectrometry (UPLC–QToFMS) [5], high-performance liquid chromatography [6], conventional [7] and real time polymerase chain reaction [7,8] are some examples. Unfortunately, all of these techniques are laborious and expensive, consume reagents and/or solvents, and generate a considerable amount of residue.

Vibrational spectroscopic techniques, such as Raman and near and mid infrared (NIR and MIR) spectroscopy, are simpler, faster and less expensive alternative techniques, which require little or no sample pretreatment. Methods developed based on these techniques are in accordance with the principles of green chemistry, being more environmentally friendly [9]. These techniques generate spectra that demand subsequent analysis with multivariate classification methods. Recently, the combination of vibrational spectroscopic techniques and multivariate qualitative methods has provided good results in the detection of food frauds [10,11]. The use of multivariate classification methods involves the assignment of a sample to a class previously established. In the case of food adulterations, one class corresponds to the non-adulterated food, and different additional classes are established depending on the number of adulterants to be detected. The ideal result assigns samples to the classes to which they actually belong.

Given that in multi-class classification methods the number of predefined classes corresponds to the number of known adulterants present in the samples plus one (authentic class), the developed multivariate models are commonly built to detect the presence or absence of only a single adulterant per sample. Although less common in the literature, adulteration with blends of two or more adulterants is a possibility in some real situations. Therefore, it is important to check whether the developed methods are able to correctly classify samples adulterated with more than one adulterant. These samples should be assigned to all the classes established for the respective adulterants [12]. Since this problem is rarely addressed in the literature, the present study represents an important contribution to the food science community and can be easily extended to other types of frauds involving other products or matrices.

The amount of information generated with mid-infrared (MIR) analysis often comprises hundreds or thousands of variables, and a number of them may be noisy and/or irrelevant to the problem under study. The selection of a limited number of informative predictors/variables can improve the effectiveness of the classification models, reduce their complexity and increase their robustness [13]. Therefore, prior to the application of the developed classification methods, variable selection strategies will be employed to optimize these methods. Recent analytical methods have been found in the literature applying variable selection with Fourier transform infrared (FT-IR) spectra in classification problems [14,15]. There are different methods of variable selection for removing noisy and irrelevant information and other methods aiming to select the most discriminatory variables when working with different groups of samples. Due to the broad diversity of variable selection methods, choosing the most appropriate method is not a simple matter [16].

Multi-class methods using partial least squares discriminant analysis (PLS-DA) and soft independent modeling of class analogy (SIMCA) were previously developed by the authors, showing good performance for classification of nectar samples with only one adulterant [17].

Thus, the objective of this work was to apply the developed strategy to detect blends of two adulterants in grape nectar using total reflectance attenuated Fourier transform mid infrared spectroscopy (ATR-FTIR). To improve the ability of these methods to detect apple, cashew and blends of both of these adulterants in grape nectars, three variable selection methods, namely, interval partial least squares (iPLS) [18], variable importance in projection scores (VIP scores) [19] and a genetic algorithm (GA) [20], were implemented. The strategy proposed in this work is described in Fig. 1, which schematically shows its different steps, including sampling, classification methodology (PLS-DA and SIMCA), analytical validation and variable selection strategies.

2. Materials and methods

2.1 Formulation of nectar samples

In this study, we chose to manufacture all of the analyzed nectar samples starting from reliable raw materials and rigorously meeting the established regulations [21-23]. This is due to the lack of the reliability about information provided on the composition of commercial Brazilian nectars related to the minimum required amounts of pulps.

Isabel grapes were supplied by EMBRAPA (Brazilian Agricultural Research Corporation) Grape & Wine, located in Petrolina, PE, Brazil. Red cashews and Fuji apples were purchased at Minas Gerais Supply Center (CEASA), located in Contagem, MG, Brazil. The fruits without physical and phytopathological damage were stored in the refrigerator (4–7ºC) until the preparation of the nectars.

The components in the final produced nectars were grape juice, pulps of the respective adulterants (apple, cashew or both), sugar syrup (water and sugar) and permitted additives (added within the permitted limits) [23], such as ascorbic acid, citric acid and guar gum (Pryme Foods, Sorocaba, SP, Brazil). Eq. (1) was applied to calculate the quantities of main fruit, adulterant fruit(s) and syrup:

$$\frac{a\times A}{100}+\frac{b\times B}{100}+\frac{c\times C}{100}+\frac{d\times D}{100}=\frac{m\times(A+B+C+D)}{100} \quad (1)$$

where “a”, ”b”, ”c” and “d” denote the Brix of the main fruit, syrup, fruit adulterant 1 and fruit adulterant 2, respectively; “A”, “B”, “C” and “D” represent the percentages of the main fruit, which is established as 50% for grape nectars [21,22], syrup, fruit adulterant 1 and fruit adulterant 2, respectively; "m" is the final nectar’s Brix; and “A + B + C+ D” is equal to 100 [17]. If no adulterant is added, this formula will only include “A” and “B”. If only one adulterant is added, this formula will include “A”, “B” and “C”.

Grape nectar samples were prepared for each of the three studied classes as follows:

- unadulterated (UN) - formulated with 50% grape juice, 50% sugar syrup and additives;
- adulterated with cashew (CAS) - formulated with 40% grape juice, 10% cashew pulp, 50% sugar syrup and additives;
- adulterated with apple (APP) - formulated with 40% grape juice, 10% apple pulp, 50% sugar syrup and additives;

Additionally, two external data sets were prepared, in which samples were adulterated with blends of cashew and apple at two different concentration levels:

- external data set 1 - adulterated with cashew and apple (CAS + APP) - formulated with 40% grape juice, 5% cashew pulp, 5% apple pulp, 50% sugar syrup and additives;
- external data set 2 - adulterated with cashew and apple (CAS + APP) - formulated with 30% grape juice, 10% cashew pulp, 10% apple pulp, 50% sugar syrup and additives;

The adulteration level has been fixed considering that too low concentrations (lower than 10%) are not economically advantageous for the fraudulent industry and too high concentrations could be perceived by sensorial evidences.

Fig. 2 shows a schematic of the experimental design for preparing the samples, resulting in 42 representative samples of each class, as well as 15 samples for the external data set 1 and 42 for the external data set 2, totaling 183 samples. Spectra from each class were split into training (28 samples) and test sets (14 samples) using the Kennard-Stone algorithm, selecting representative samples distributed homogeneously into the multivariate space [24].

2.2 Instrumentation and software

The analysis was conducted at controlled temperature (20.0 ± 0.5 ºC). ATR-FTIR spectra were obtained using an IRAffinity-1 FTIR (Shimadzu, Kyoto, Japan) spectrophotometer with a DLATGS detector (Deuterated Triglycine Sulfate Doped with L-Alanine). It was equipped with a horizontal ATR accessory composed of a ZnSe prism (PIKE Technologies, Madison, WI, USA) with twenty internal reflections.

The spectrum of each sample was recorded three times with 16 scans at a resolution of 4 cm^{-1}, from 4000 to 650 cm^{-1}. An air background correction was applied after each measurement to avoid atmospheric interference and to reduce instrumental noise. Spectral region from 938 to 650 cm^{-1} was removed, as it presents too noisy variables. Spectra were pre-processed by multiplicative scatter correction (MSC) [25] to eliminate non-linear baseline deviations and trends.

Multi-class models based on SIMCA and PLS-DA were calculated by using software MATLAB, version 8.0.0.783 - R2012b (Natick, MA, USA), and PLS Toolbox 7.0.2 (Eigenvector Research Inc., Wenatchee, WA, USA).

2.3 Variable selection methods

Interval Partial Least Squares method (iPLS) builds local PLS-DA models after the full spectra are divided into equal subintervals. Interference and noisy spectral regions are removed, and the best model containing the most relevant information is obtained [26]. The prediction performance of these local models and the global (full-spectra) model are compared, mainly by means of the validation parameter RMSECV (root mean squared error of cross-validation).

Genetic algorithm (GA) process is based on Darwin's evolution theory and involves several steps: i) generation of random variable subsets; ii) evaluation of each individual subset for fitness to predict Y; iii) elimination of half of the worst individuals; iv) breed of the remained individuals; v) mutation; vi) repetition of steps 2–5 until a subset of variables providing the optimal classification performance in comparison to the previous iteration is found. Each

individual corresponds to a sample and the variables are considered as genes forming a chromosome, which will be tested for its ability to survive, i.e., the ability to provide correct predictions. More details about GA can be found elsewhere [20]. Its performance was evaluated by means of cross-validation and the lowest RMSECV values [13].

Variable importance in the projection scores (VIP scores) is the squared function of the PLS weights taking into account the amount of explained **y** variance in each dimension. The VIP scores value is calculated for each variable. Therefore, VIP scores provide information about the significance of each variable on the latent variables (LV). The greater the VIP scores, the more important the corresponding variable is. VIP scores values greater than 1.0 are used as a threshold criterion for variable selection [19], since the respective variables are considered to be the most influential in the model.

2.4 Multivariate classification methods

SIMCA is a class modeling method. Therefore, it models each target class individually, without taking into account samples of the other classes. Multi-class SIMCA was implemented for the three defined classes (UN, APP and CAS). SIMCA assignations were obtained considering the distance value "*d''*, Eq. (2),

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

where T_r^2 and Q_r are the reduced statistics of *Hotelling* T^2 and Q, which are the boundaries of the classes and are calculated for each predefined class at the α significance level [27]. For a sample to be considered "within the class model", it must present values of d lower than 1.0 [28].

PLS-DA is a discriminant method that defines delimiters between two or more classes and works by splitting the variables' hyperspace in a number of regions corresponding to the number of classes. Multi-class PLS-DA was applied by establishing a linear regression between matrix **X** of independent variables and matrix **Y** of dependent/dummy variables. Binary dummy variables are defined as 0 or 1, such that 1 indicates samples belonging to the target class and 0 indicates samples not belonging to the target class. Since values predicted by PLS regression are not exactly 1.0 or 0.0, a threshold needs to be estimated. Bayesian statistics were used to calculate each class threshold, selecting the y value that presents the minimal number of false-classifications [29].

2.4.1 Evaluation of performance

To assess the quality of the classification model, sensitivity, specificity and inconclusive ratio were defined [28, 30]. Their calculations are based on true assignations (true positive and

true negative) and false assignations (false positive and false negative) and no conclusive assignations.

Sensitivity (*SEN*) of a class "j" represents the fraction of samples from the class of interest that was truly assigned to their own class, Eq. (3):

$$SEN_j = {}^{TP}/_{n^{\underline{o}}\, S_j} \quad (3)$$

where "$n^o\ S_j$" are the total number of samples in class "j", and *TP* are true positive predictions.

Specificity (*SPE*) of a class "j" represents the fraction of samples not belonging to class "j" that were correctly not assigned to this class by the model, Eq. (4):

$$SPE_j = {}^{TN}/_{n^{\underline{o}}\, S_{not\ j}} \quad (4)$$

where "$n^o\ S_{not\ j}$" is the total number of samples that do not belong to class "j" and *TN* are true negative predictions.

The number of inconclusive samples (*IN*) indicates the number of samples that cannot be undoubtedly assigned to class" j", and thus considers no assignation to any class and the multiple assignation Eq. (5):

$$IN_j = (NA_j + MA)/n^{\underline{o}}\, S_j \quad (5)$$

where "NA_j" means unassigned samples (samples that are from class "j" that are not assigned either to class "j" or to any other class); "*MA*" means multiple assignation samples (samples from class "j" assigned to more than one class); and "n^oS_j" means the total number of samples that really belong to class "j".

In the particular case in which samples adulterated with blends have to be assigned, it should be kept in mind that if the MA is due to assignations to both adulterated classes, the sample is not considered as inconclusive.

3. Results and discussion

MIR spectra can be divided into four regions, which are related to X-H stretching (4000–2500 cm^{-1}), triple bonds (2500–2000 cm^{-1}), double bonds (2000–1500 cm^{-1}), and the so called fingerprint region (1500–400 cm^{-1}). Bending and skeletal vibrations characterize the fingerprint region. Bands associated with skeletal vibrations normally reflect a pattern of the molecule as a whole and not only a specific group within the molecule. Thus, analysis of a vibrational molecular fingerprint can increase the potential to detect adulterations [31,32]. Spectra of all of the samples analyzed in this study are shown in Fig. 1S (Supplementary material). The main vibrations of the MIR spectra of fruit nectars have been already assigned in our previous papers in general, including different fruit nectars (grape, orange, peach, passion fruit) [33], and for grape nectars

in particular [17]. The most intense signals in these spectra are a broad band centered at 3300 cm^{-1} and a sharp band approximately 1640 cm^{-1}, both assigned to O-H water vibrations. Nevertheless, the spectral region most rich in information seems to be the fingerprint region, mainly below 1200 cm^{-1}, where C-H, C-C and C-O stretching and bending vibrations associated with phenols, carboxylic acids and carbohydrates are observed [17,33].

Three classes were established with the training samples and modeled by means of SIMCA and PLS-DA, a class modeling and a discriminant method, respectively. To evaluate the classification performance of the methods, the test set and two external data sets (1 and 2, previously described in Section 2.1) were used for prediction.

Each class was independently modeled by SIMCA, taking into account the lowest errors of leave-one-out cross validation. A total of 3 principal components were necessary to build each class model, accounting for 95.2%, 93.8% and 90.6 % of variance for the classes UN, CAS and APP, respectively. For PLS-DA, the best number of LVs was chosen based on the smallest cross validation classification error. Six LVs were selected, accounting for 95.1% of the variance in **X** block and 82.9% of the variance in **Y** block. The threshold values set for each class were 0.25 for UN class, 0.14 for CAS class and 0.09 for class APP.

Table 1 shows the number of samples assigned to each class by the two classification methods for the four different data sets. SIMCA was not able to properly model the UN and CAS classes. Considering the samples of the training and test sets, a large number of inconclusive samples were observed. The APP class presented better results, since 22 out of 28 samples from the training set and all of the test samples were properly assigned as APP. For APP, no wrong assignations were observed, and 6 out of 28 training samples were not assigned to any class.

Regarding the prediction of the two external sets, containing blends of the two adulterants, it was expected in an ideal situation that all of the samples were assigned to both adulterated classes, CAS and APP, independent of the percentages of adulteration. Samples from the external set with higher concentrations (each adulterant at 10%) were mostly classified into the CAS class (38 out of 42), but not in the APP class (only one out of 42). Samples from the external set with the lowest concentration of blend adulterants (each adulterant at 5%) were not assigned to any class.

Predictions for training and test samples with PLS-DA (Table 1) show that almost all of the samples were properly assigned to their own class. No wrong assignations were obtained, and only one sample solely adulterated with cashew was simultaneously assigned to the CAS and APP classes. In addition, a few samples (1 out of 28 in the training set and 3 out of 14 in the test set) were not assigned to any class.

In relation to the external set 1, 5 out of 15 were correctly multiple assigned, with 11 samples assigned to CAS and 7 to the APP class. Only one sample from external set 2 was correctly multiple assigned, while all of the rest were only assigned to the CAS class. In summary,

a PLS-DA model built with all of the variables (full-spectra) was a good choice for detecting samples containing only one adulterant (as samples from the training and test sets), but it failed to classify samples containing blends of two adulterants (as is the case for the external data sets).

To improve the previous models, a variable selection optimization was performed. The objective was to select the most discriminant variables/wavenumbers among the three classes, which might lead to a more robust and accurate model able to correctly classify samples containing one or two adulterants. Three variable selection methods were tested, iPLS, GA, and VIP scores.

For the iPLS variable selection, full-spectra were divided into 20 subintervals of 75 variables. The RMSECV values were obtained for each interval. The best predictive ability was observed for the intervals 1 and 2, which correspond to 150 variables/wavenumbers between 1514 and 937 cm^{-1}. These wavenumbers can be associated with the presence of sugars and organic acids present in the fingerprint region of fruit nectar MIR spectra [31,33-35].

For GA variable selection, the following parameters were applied: population of 64, window width of 1, initial terms of 10, maximum generations of 100, convergence of 50 %, mutation rate of 0.005 and double crossing over. Cross-validation parameters were set randomly with 20 splits and 3 iterations. Selected variables, 112, were distributed across the entire spectrum.

For VIP scores variable selection, the squared function of the PLS weights was calculated for each variable, considering the models of the three studied classes. A total of 524 wavenumbers presented VIP scores higher than 1.0. Therefore, they were retained in the optimized models.

SIMCA and PLS-DA models were built in combination with the variable selection methods. The results obtained for SIMCA are shown in Table 2. Variable selection with iPLS significantly improved the predictions of the training and test sets compared to those obtained using full-spectra (Table 1), providing few inconclusive samples and no errors in the assignments. However, for the prediction of the two external data sets, no sample was correctly assigned to the two classes to which they belong. No significant improvement was observed in the models constructed with the variables selected by GA and VIP scores with respect to the models built with the full-spectra.

Table 3 shows the results obtained with PLS-DA combined with the variable selection methods. In general, classification was improved in relation to the full-spectra model (Table 1). All of the samples in the training and test sets were correctly classified with iPLS-DA. For the GA and VIP scores variable selection, the results were highly satisfactory, with only two out of 42 samples not assigned to any class, and no wrong assignments.

For the external set 2, only iPLS provided satisfactory results for classifying samples in both adulterated classes (39 out 42). When variables selected by GA and VIP-scores were used, samples were mostly assigned only to the CAS class. It should be noted that none of the adulterated sample was predicted as unadulterated. For the external set 1, no improved results

were obtained with the variables selected by the three selection techniques, and the results were similar to the ones obtained using the full-spectra.

The main difference between the variable selection strategies implemented lies in the fact that while iPLS provides a continuous block selection, GA and VIP-scores selects individual discrete variables. Thus, a more exhaustive optimization was performed by applying GA and VIP scores variable selection for building PLS-DA models with the 150 wavenumbers selected by iPLS. The same parameters previously described for GA were used, except for the number of iterations, which was 1.

The number of variables selected by combining iPLS with GA and VIP scores were 16 and 84, respectively. A slightly better result was obtained with iPLS-GA, since all of the samples from the external set 2 were correctly assigned to both CAS and APP classes and 10 out of 15 samples from the external set 1.

In general, PLS-DA presented a classification performance that was consistently better than SIMCA. Thus, Table 4 reports a comparison between the full spectra and five different variable selection PLS-DA models based on the appropriate figures of merit. SEN and SPE were estimated as global parameters, considering all of the analyzed data sets. An SPE of 100% was observed for all of the cases except for the UN class when predicted by the full-spectra and the VIP scores PLS-DA models (SPE of 98 and 96%, respectively). This result (SPE of 100%) indicates that no sample was assigned to a class to which it does not belong. All of the models of Table 4 provided an SEN of 100% for the UN class and above 92% for the CAS class. Nevertheless, only models based on iPLS presented an acceptable SEN for APP. The best SENs were 83% (iPLS) and 93% (iPLS-GA). These results demonstrate the difficulty of predicting samples containing both of the adulterants as belonging to the APP class. In other words, it was only possible to efficiently detect apple adulteration in the simultaneous presence of cashew as an adulterant by employing iPLS variable selection.

4. Conclusions

ATR-FTIR spectroscopy was employed jointly with multivariate classification methods, SIMCA and PLS-DA, for developing a screening analytical method able to specifically detect adulterations in grape nectars with cashew and apple juices. Because of their sensorial characteristics, these two fruits have been commonly used as adulterants in other nectars manufactured with more expensive fruits, such as grapes. The developed method is direct, rapid and requires no sample pretreatment.

Discriminant PLS-DA presented better results than class modeling SIMCA, since this last method provided a high number of inconclusive predictions. Full spectra PLS-DA was able to correctly classify samples from the training and test sets, with good performance parameters.

However, this same model was not able to correctly assign samples adulterated simultaneously with cashew and apple specifically to both their respective classes. This task was possible only for blends at a higher concentration level (external set 2) by employing variable selection.

Therefore, when dealing with blends, it will be interesting to implement different strategies that would allow for improving the results at a low concentration level. In that sense, knowing the lower level that can be detected will also be the aim in further studies.

The best model was obtained combining the two most predictive wavenumber intervals selected by iPLS. A slightly improvement was obtained when iPLS selection was combined with the genetic algorithm discrete variable technique. The most predictive wavenumbers were located in the fingerprint region, mainly between 1200 and 950 cm^{-1}, and can be related to organic acids and sugars. Thus, these components of the fruit nectars may be suggested as discriminants between different fruits used for manufacturing and adulterating this type of beverage.

From the analytical point of view, the most interesting aspect of this study is to provide a strategy to build robust multivariate classification models able to specifically detect more than one adulterant in the presence of a mixture of them. This subject is relatively underestimated in the chemometric literature. This study showed the limitations of current classification methods to face this task if they had not been specifically designed for this purpose. In addition to food and beverages, this strategy can be applied to the detection of frauds and adulterations in other types of matrices, such as pharmaceutical products, fuels and forensic samples.

Acknowledgements

The authors acknowledge CAPES for providing the sandwich PhD scholarship, Biofuels Laboratory (Escola de Engenharia, UFMG, Belo Horizonte, Brazil) for allowing the use of the ATR-FTIR Spectrophotometer and Giuliano Elias Pereira from EMBRAPA Grape & Wine (Bento Gonçalves, Brazil) for providing the Isabel grapes.

References

[1] P. Rinke, Tradition Meets High Tech for Authenticity Testing of Fruit Juices. In: G. Downey (Ed.), Advances in Food Authenticity Testing, Elsevier, Amsterdam, 2016, pp. 625-665.
[2] Brasil. Ministério da Agricultura, Pecuária e Abastecimento. Decreto nº 6.871, de 04 de junho de 2009. Regulamenta a Lei n. 8.918, de 14 de julho de 1994. Brasília, 2009.
[3] J. Kanner, E. Frankel, R. Granit, B. German, J.E. Kinsella. Natural antioxidants in grapes and wines, J. Agric. Food Chem. 42 (1994) 64-69.
[4] A. Soria, A. Ruiz-Matute, M. Sanz, I. Martínez-Castro, Chromatographic Technique: Gas Chromatography (GC). In: D. W. Sun (Ed.), Modern Techniques for Food Authentication, Academic Press, Cambridge (USA), 2008, pp. 321-360.
[5] Z. Jandrić, D. Roberts, M.N. Rathor, A. Abrahim, M. Islam, A. Cannavan, Assessment of fruit juice authenticity using UPLC–QToF MS: a metabolomics approach, Food Chem. 148 (2014) 7-17.

[6] F.R. Spinelli, S.V. Dutra, G. Carnieli, S. Leonardelli, A.P. Drehmer, R. Vanderlinde, Detection of addition of apple juice in purple grape juice, Food Control 69 (2016) 1-4.
[7] J. Han, Y. Wu, W. Huang, B. Wang, C. Sun, Y., Ge, Y. Chen, PCR and DHPLC methods used to detect juice ingredient from 7 fruits, Food Control 25 (2012) 696-703.
[8] M.A. Pardo, Evaluation of a dual-probe real time PCR system for detection of mandarin in commercial orange juice, Food Chem. 172 (2015) 377-384.
[9] A. Fernández-González, J.M. Montejo-Bernardo, H. Rodríguez-Prieto, C. Castaño-Monllor, R. Badía-Laíño, M.E. Díaz-García, Easy-to-use analytical approach based on ATR–FTIR and chemometrics to identify apple varieties under Protected Designation of Origin (PDO), Comput. Electron. Agric. 108 (2014) 166-172.
[10] M. Pilar Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Control 86 (2018) 283-293.
[11] D. Cozzolino, The role of vibrational spectroscopy as a tool to assess economically motivated fraud and counterfeit issues in agricultural products and foods, Anal. Methods 7 (2015) 9390-9400.
[12] B.G. Botelho, N. Reis, L.S. Oliveira, M.M. Sena, Development and analytical validation of a screening method for simultaneous detection of five adulterants in raw milk using mid-infrared spectroscopy and PLS-DA, Food Chem. 181 (2015) 31-37.
[13] C.M. Andersen, R. Bro, Variable selection in regression - a tutorial, J. Chemom. 24 (2010) 728-737.
[14] X. Li, S. Wang, W. Shi, Q. Shen, Partial least squares discriminant analysis model based on variable selection applied to identify the adulterated olive oil, Food Anal. Methods 9 (2016) 1713-1718.
[15] M. Manfredi, E. Robotti, F. Quasso, E. Mazzucco, G. Calabrese, E. Marengo, Fast classification of hazelnut cultivars through portable infrared spectroscopy and chemometrics, Spectrochim. Acta A 189 (2018) 427-435.
[16] C.V. Di Anibal, M. Pilar Callao, I. Ruisánchez, 1H NMR variable selection approaches for classification. A case study: the determination of adulterated foodstuffs, Talanta 86 (2011) 316-323.
[17] C.S.W. Miaw, M.M. Sena, S.V.C. de Souza, M. Pilar Callao, I. Ruisánchez, Detection of adulterants in grape nectars by attenuated total reflectance Fourier-transform mid-infrared spectroscopy and multivariate classification strategies, Food Chem. 266 (2018) 254-261
[18] A. Höskuldsson, Variable and subset selection in PLS regression, Chemom. Intell. Lab. Syst. 55 (2001) 23-38.
[19] I.-G. Chong, C.-H. Jun, Performance of some variable selection methods when multicollinearity is present, Chemom. Intell. Lab. Syst. 78 (2005) 103-112.
[20] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review, J. Chemom. 15 (2001) 559-569.
[21] MAPA. Secretaria de Defesa Agropecuária, Ministério da Agricultura, Pecuária e Abastecimento (MAPA), Instrução Normativa No. 12, Brasília, Brazil, 2003.
[22] MAPA. Secretaria de Defesa Agropecuária, Ministério da Agricultura, Pecuária e Abastecimento (MAPA), Instrução Normativa No. 42, Brasília, Brazil, 2013.
[23] ANVISA. Agência Nacional de Vigilância Sanitária, Resolução da Diretoria Colegiada n° 8, de 06 março de 2013, Brasília, Brazil, 2013.
[24] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137-148.
[25] A. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC-Trends Anal. Chem. 28 (2009) 1201-1222.
[26] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413-419.
[27] A. Rius, M. Pilar Callao, F.X. Rius, Multivariate statistical process control applied to sulfate determination by sequential injection analysis, Analyst 122 (1997) 737-741.

[28] C. Márquez, M. Isabel López, I. Ruisánchez, M. Pilar Callao, FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud, Talanta 161 (2016) 80-86.
[29] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemom. 28 (2014) 213-225.
[30] M. Isabel López, M. Pilar Callao, I. Ruisánchez, A tutorial on the validation of qualitative methods: From the univariate to the multivariate approach, Anal. Chim. Acta 891 (2015) 62-72.
[31] J.D. Kelly, G. Downey, Detection of sugar adulterants in apple juice using Fourier transform infrared spectroscopy and chemometrics, J. Agric. Food Chem. 53 (2005) 3281-3286.
[32] B.H. Stuart, Infrared Spectroscopy: Fundamental and Applications. John Wiley & Sons, New York, 2004.
[33] C.S.W. Miaw, C. Assis, A.R.C.S. Silva, M.L. Cunha, M.M. Sena, S.V.C.de Souza, Determination of main fruits in adulterated nectars by ATR-FTIR spectroscopy combined with multivariate calibration and variable selection methods, Food Chem. 254 (2018) 272-280.
[34] S. Bureau, D. Ruiz, M. Reich, B. Gouble, D. Bertrand, J.-M. Audergon, C.M. Renard, Application of ATR-FTIR for a rapid and simultaneous determination of sugars and organic acids in apricot fruit, Food Chem. 115 (2009) 1133-1140.
[35] J. He, L.E. Rodriguez-Saona, M.M. Giusti, Mid infrared spectroscopy for juice authentication-rapid differentiation of commercial juices, J. Agric. Food Chem. 55 (2007) 4443–4452.

Figure captions

Fig. 1. Overview of the steps to develop a multivariate classification method with variable selection.
ATR-FTIR: total reflectance attenuated Fourier transform mid infrared spectroscopy; APP: adulterated with apple class; CAS: adulterated with cashew class; CAS+APP: adulterated with blend of cashew and apple sets; GA: genetic algorithm; iPLS: interval partial least squares; PLS-DA: partial least squares discriminant analysis; SIMCA: soft independent modeling of class analogy; UN: unadulterated class; VIP scores: variable important in projection scores.

Fig. 2. Experimental design for the formulation of the nectars.
S: samples.

Fig. 1S. Spectra of all analysed samples after pre-processing with multiplicative scatter correction.