FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud

Cristina Márquez, M. Isabel López, Itziar Ruisánchez[*]*, M. Pilar Callao*

Chemometrics, Qualimetric and Nanosensors Grup, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n, 43007 Tarragona, Spain

Corresponding author: Tel.: +34 977558299; fax: +34 977558446

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

ABSTRACT

Two data fusion strategies (high- and mid-level) combined with a multivariate classification approach (Soft Independent Modelling of Class Analogy, SIMCA) have been applied to take advantage of the synergistic effect of the information obtained from two spectroscopic techniques: FT-Raman and NIR. Mid-level data fusion consists of merging some of the previous selected variables from the spectra obtained from each spectroscopic technique and then applying the classification technique. High-level data fusion combines the SIMCA classification results obtained individually from each spectroscopic technique. Of the possible ways to make the necessary combinations, we decided to use fuzzy aggregation connective operators. As a case study, we considered the possible adulteration of hazelnut paste with almond. Using the two-class SIMCA approach, class 1 consisted of unadulterated hazelnut samples and class 2 of samples adulterated with almond. Models performance was also studied with samples adulterated with chickpea. The results show that data fusion is an effective strategy since the performance parameters are better than the individual ones: sensitivity and specificity values between 75 and 100% for the individual techniques and between 96-100% and 88-100% for the mid- and high-level data fusion strategies, respectively.

Keywords: Mid- and high-level data fusion, two-class SIMCA, FT-Raman, NIR, food adulteration, hazelnut adulteration

1. Introduction

Food fraud is becoming increasingly sophisticated due to the use of unconventional or synthetic adulterants. To guarantee food safety and quality, most analytical strategies are based on the knowledge of the contaminants [1-2]. Because of the ever-increasing range of analytes that can be used in food fraud together with the impossibility of covering them all, there is a demand for the development of fast, easy-to-use and low-cost analytical methods to test for adulteration. Multivariate qualitative methodologies that combine multivariate data with classification techniques are used to detect anomalous samples in food fraud, both, adulteration and authentication [3-7] of a wide variety of foods.

Spectroscopic instrumental techniques are the most commonly used because there is almost no need for sample pretreatments and they can be applied to a wide range of analytes. As is well known, food composition is complex and contains compounds with functional groups having nonselective spectral bands. Univariate analysis is based on selecting a discrete wavelength (i.e. absorbance at one wavelength), so the lack of selectivity is the main disadvantage when the selected wavelength does not respond only to a single compound. This is not a disadvantage in multivariate analysis as the whole spectra is used. Among other techniques, food has been analyzed by ultraviolet (UV) [8-9], fourier transform- infrared (FT-IR) [10-11], fluorescence [12-13], nuclear magnetic resonance (NMR) [14-15], Raman [16-17] and near-infrared (NIR) [18-19].

Due to the increasing ease in which data can be obtained, data fusion is an expanding trend. The main goal is to optimize the information ~~[20]~~ obtained in order to exploit the synergies of individual information provided by different techniques [20]. There are various strategies for carrying out data fusion: low-, mid- and high- level data fusion. In low-level data fusion, raw data from more than one source are directly fused (concatenated), taking into account that data must be correctly balanced (all the variables must be on the same scale) before they are combined. In mid-level data fusion, some of the raw variables are selected and then fused. In both approaches (low- and mid-level), only one classification model is implemented over the fused variables. Finally, in high-level data fusion the classification results obtained from individual classification models are fused. In this case, a classification approach is implemented for each data source.

Fusion techniques have proven to be useful in fields such as metabolomics [21-22], pigments in artworks [23-24], dye degradation processes [25] and food [26 -28].

The aim of this study is to evaluate the performance of two types of data fusion: high- and mid-level data fusion. We propose to implement high-level data fusion because although it is not used too often, it has advantages: it is easy to use and easy to expand since when new data source becomes available, its classification results can be added to the classification decision rule, thus increasing the versatility of the decision process.

We also used mid-level data fusion, which involves selecting or reducing variables before establishing the classification model. The most common approach for variable reduction is to fuse the information from a reduced number of latent variables obtained independently from the signals of each instrument. Usually scores from principal component analysis (PCA) or partial-least squares discriminant analysis (PLS-DA) are fused [20]. Of the various strategies available for selecting variables (iPLS, genetic algorithms, etc.) [29], in this study we use a variable selection strategy that identifies the most different variables between classes. It is based on a transformation of the original variables by means of a normalization calculation, x*diff* [14, 30].

As a case study, hazelnut adulteration problem is considered. Hazelnuts and their derivatives (oils and pastes) are widely used as ingredients in many desserts, ice creams and chocolates. The price of hazelnuts depends on parameters such as geographical origin, abundance of harvest, etc. In case of unfavorable economic conditions, the price can be reduced by adding such other ingredients as almond, because of its similarity. However, other more unexpected products might be added, for example chickpea. NIR and FT-Raman data are used to determine whether the synergism between them can be exploited. These data are processed through a SIMCA model, separately for each technique and after applying two data fusion strategies (mid- and high-level data fusion).

2. Material and methods

2.1. Samples

The unadulterated set consists of 24 hazelnut pastes (*Corylus avellana*) from different geographical origins (Spain, Italy, Georgia and Azerbaijan). The two adulterated data sets consists of 26 hazelnut pastes adulterated with almond paste and 27 hazelnut pastes adulterated with chickpea flour, all of them at 7% (w/w), because experience indicates that it is the most common percentage of adulteration. In previous studies [18] no trends related to level of adulteration were found in neither of adulterants.

Hazelnut and almond were provided by *La Morella Nuts S.A.U*. and the chickpea flour were an ecological product from a commercial supplier. Details about hazelnuts, adulterants and sample preparation can be found in our previous study [18].

2.2 Instrumentation and software

NIR spectra were recorded by a Bruker VECTOR 22/N spectrophotometer, working in diffuse reflectance conditions. The spectral profile of each sample was acquired as the mean of 32 scans recorded during rotation, and in the spectral range 3650-12000cm^{-1} at 8 cm^{-1}.

FT-Raman measurements were obtained from a Thermo Nicolet 5700 FT-IR spectrometer equipped with a FT-Raman module NXR with an InGaAs detector, CaF_2 beamsplitter. A 1064 nm radiation from a Nd:YAG laser with a laser power of 0.5W was used for excitation. The spectral profile of each sample was acquired as the mean of 256 scans in the spectral range 290-3200cm^{-1} at a resolution of 4cm^{-1}.

Both NIR and FT-Raman spectral data were exported to Matlab [31] and treated with PLS toolbox [32].

3. Data Analysis

3.1. Classification technique: Soft Independent Modelling of Class Analogy (SIMCA)

SIMCA is a modelling technique based on Principal Component Analysis (PCA) in which each class is modelled independently from all others [33]. Each sample is characterized by two scalar statistics, Hotelling T^2 and Q, which measures the information from each sample included or not included in the model, respectively.

Class frontiers (Hotelling T^2_{lim} and Q_{lim}) are calculated for each pre-defined class (class model), at a specific significance level (α), usually set at 0.05 [34]. For the sake of simplicity, samples are assigned by means of the reduced statistics values (Hotelling T^2_r and Q_r) which are the ratio between the statistic of sample *i* and the corresponding class limit. A sample must have values lower than 1 for both reduced statistics to be considered "within the class model". Another criterion for sample assignation is the distance of a

sample *i* from class *j* (d_{ij}) which is defined as a combination of its reduced statistic (equation 1):

$$d_i = \sqrt{(Q_{r,i})^2 + (T^2_{r,i})^2} \quad \text{(eq. 1)}$$

In a two-class system, the assignment of a sample gives four types of output: sample belongs to class 1; sample belongs to class 2; sample belongs to both classes; sample does not belong to any class. From these results, the main performance parameters - sensitivity, specificity and inconclusive ration - are calculated for each class using equations 2 to 4.

$$\text{Sensitivity}_j = TP_j / n^oS_j \quad \text{(eq. 2)}$$

where *j* indicates the class under study, TP_j means true positives (samples from class *j* that have been properly predicted by the model as belonging to class *j*), and n^oS_j is the total number of samples that really belong to class *j*. Therefore, sensitivity indicates the likelihood of recognizing truly positive samples.

$$\text{Specificity}_j = TN_j / n^oS_{not\ j} \quad \text{(eq. 3)}$$

where *TN* means true negatives (samples that are not from class *j* and have been predicted as not belonging to class *j*), and $n^oS_{not\ j}$ means the total number of samples that really do not belong to class *j*. Therefore, specificity indicates the likelihood of recognizing samples that are truly different from the class.

$$\text{Inconclusive ratio}_j = (NA_j + MA) / n^oS_j \quad \text{(eq. 4)}$$

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*.

3.2. Data fusion

The two levels of data fusion architectures studied in this paper are summarized in Fig. 1.

---------- Fig. 1 ----------

High-level data fusion: decision fusion [23,24,36]

Decision level data fusion combines the classification results obtained from each individual technique. (Fig. 1a). In this study, the SIMCA classification results obtained separately for each instrument data are fused. The parameters (results) considered were the individual distances (Eq. 1) of each sample from each model. Fusion was implemented using the fuzzy set theory. Among the several available operators, the Minimum, Maximum, Averange and Product fuzzy aggregation connective operators have been chosen because of its conceptual simplicity and ease implementation. The final decision (*ensemble decision*) was obtained by the majority vote provided by all the aggregation operators.

Given the nature of the operators, the concordant results for both techniques give the same result after the fusion so, for practical purposes, only non-concordant results are fused.

Mid-level data fusion: variable selection

In mid-level fusion, a previous variable selection step was independently performed ~~in~~ over the data (spectra) obtained from each technique, so only the most relevant variables were fused. The original variables were transformed to give the normalized differences between the mean spectrum of a class considered the reference and the spectra of each sample ($\mathbf{X}_{diff}$). As we are dealing with an adulteration problem, we have set the unadulterated class as the reference class. The hypothesis is that variables with different intensities will have reinforced x_{diff} values, so only variables that differentiate between the classes are selected [14, 30].

For each data set, NIR and FT-Raman spectra, the raw variables are first transformed by calculating the corresponding x_{diff} values [29,35] in accordance with Eq. 5

$$x_{diff,ij} = \frac{\left| x_{ij} - \overline{x_i} \right|}{\sigma_i} \qquad \text{(eq.5)}$$

where x_{ij} is the i^{th} variable (i.e. NIR intensity value at frequency i) for the j^{th} sample and $\overline{x_i}$ and σ_i are the mean and standard deviation, respectively, calculated from each i^{th} variable of the reference class (unadulterated class).

The $\mathbf{X}_{diff}$ matrix was then obtained from all the samples and all variables for both NIR and FT-Raman spectra. Its magnitude is indicative of the variables characteristic of the adulterant. Then, a threshold value was defined from the x_{diff} values calculated for the

reference class and only those variables in the adulterated class with x_{diff} values higher than the set threshold were selected.

The raw variables selected from each data source were then fused (concatenated), and SIMCA was performed for both classes (unadulterated and adulterated), as shown in Fig. 1b.

4. Results and discussion

NIR and FT-Raman data were pre-treated separately. A Savitzky-Golay smoothing was applied to FT-Raman data (1510 variables) using a 15 data point window and a first¬ order polynomial to suppress the instrumental noise. Then, a baseline correction was applied by using first-order polynomial. Fig. 2a shows the corrected FT-Raman spectra. It can be seen that the most intense bands appear in the range of 800-1800 cm-1 and 2700-3100 cm-1.

An offset correction was applied to the raw NIR data (2166 variables) to eliminate any vertical shift by subtracting the absorbance value at 10538 cm-1. Fig. 2b shows the corrected NIR spectra. In this case, the most intense bands appear in four regions: 3650-4750 cm-1 (combination bands), 5600-5900 cm-1 (first overtone), 8000-9000 cm-1 (second overtone) and 10000-11000 cm-1 (third overtone).

---------- Fig. 2 ----------

First, independent two-class models were built for each data source studied. In both cases, class 1 was built from unadulterated samples and class 2 from samples adulterated with almond. Model performance was also studied with samples adulterated with chickpea. NIR data were autoscaled and FT-Raman data were mean-center. Both models were validated by leave-one-out cross-validation and in both cases the optimal number of PC's used to build the SIMCA models were selected on the basis of the RMSECV. For FT-Raman and NIR models, the first three PCs were considered for the unadulterated class and the first four PCs for the adulterated class.

Table 1 shows the assignations using the SIMCA models built individually with NIR and FT-Raman data. Samples were assigned to one class (unadulterated or adulterated with almond), to both classes (multiple assignation) or to no class. Overall, both spectroscopic techniques have similar abilities to recognize their own samples (sensitivity) and different

samples (specificity). A closer look at the table shows that most of the NIR and FT-Raman errors occur for unadulterated samples that were not properly recognized by their own model: 6 out of 24 for both, NIR and FT-Raman models. The adulterated models, on the other hand, were more able to recognize their own samples: two samples were not recognized by the NIR model and just one by the FT-Raman model.

---------- Table 1 ----------

It should be pointed out that most of the misclassifications were due to inconclusive assignations (multiple or not assigned). This means that a 13% of all analyzed samples needed to be submitted to a confirmatory analysis (inconclusive assignation) as they were assigned to both classes and just one unadulterated sample was not assigned to any class by the NIR data. The inconclusive assignations cannot be considered as "~~real~~ true" errors, since from the practical point of view they will be submitted to a confirmatory analysis. So a screening strategy enables them to be identified and action can be taken.

With FT-Raman models, unadulterated and adulterated with almond samples were not wrongly classified since there were not any unadulterated samples assigned as adulterated as well as vice versa. However, there were samples with multiple assignation. With the NIR models there were two wrong assignations. One unadulterated sample was wrongly assigned as adulterated, an error that represents an economic risk since it would have to be withdrawn from commercial markets since it was identified as adulterated even though it was not. And one adulterated sample was wrongly assigned as unadulterated. This type of error is a fraud since these samples will not be identified as adulterated even though they are. In this particular case, it has no health implications, but it does represent an economic fraud to the final consumer. Inconclusive outputs in both techniques, were mostly due to multiple assignations of unadulterated samples and just one adulterated sample.

All of the samples adulterated with chickpea were recognized as not belonging to any class by the NIR models, while two out of the 27 samples were assigned as unadulterated by the FT-Raman models. As has been discussed this type of error is an economic fraud for the final consumer.

The discussed miss-assignations (performance parameters) indicate that the classification results could be improved by a data fusion strategy.

In the high-level data fusion, the individual classification results provided by SIMCA were fused by the four fuzzy aggregation connective operators (minimum, maximum, product and average) and the majority vote rule, as described in the theory section. Each fuzzy operator was applied to the distances values of the sample from each class using equation 1. A sample was assigned to a class when the majority of its fuzzy values (ensemble decision) was lower than 1 (and not assigned for values higher than 1). Table 2 shows the sample distances obtained from the individual models (NIR and FT-Raman) and the corresponding fuzzy numerical values when the different operators were used to classify the samples that had been wrongly or inconclusively assigned by the individual models.

---------- Table 2 ----------

Overall 9 out of the 11 non-concordant assignations were solved by the fusion. A closer look at the table shows that the two wrong NIR assignations were partly solved. Sample nº1(U), with multiple assignation by the RAMAN models and wrongly assigned by NIR, cannot be clarified and continues as double assigned (class 1 and 2). The same can be said for sample nº16(A): it was wrongly assigned as an unadulterated sample by the NIR and multiple assigned by the FT-Raman models, and the ensemble decision set it as multiply assigned. Although they were not properly assigned to their own classes (unadulterated and adulterated, respectively), the multiple assignation is much convenient than a wrong assignation.

Sample nº25(U) had a distance value equal 1 for the unadulterated NIR model, so strictly it should not be considered to fit the unadulterated model, but the final ensemble decision of unadulterated is clear. The other two samples adulterated with chickpea (nº18 (C) and 28 (C)) that were wrongly assigned by the FT-Raman model as unadulterated were not assigned to any model by the ensemble decision. The rest of the conflictive (multiple) assignations were solved after the decisions of all the fuzzy operators.

These results are summarized in table 3, which shows the final performance parameters. It can be seen that the inconclusive assignations have been reduced to 12% for the unadulterated class while the percentage for the adulterated class is the same. The same can be stated regarding the sensitivity and specificity values, which have increased considerably in the unadulterated class and stays or slightly improve in the adulterated class.

---------- Table 3 ----------

For the variable level data fusion, the $\mathbf{X}_{diff}$ matrix is computed as described in the theory section. Figure 3 shows the $\mathbf{X}_{diff}$ matrix representation for class 1 and class 2 for FT-Raman and NIR data. It can be seen from the unadulterated samples (defined as the reference class) that x_{diff} values minimize the difference between the samples and variables, all of which are in a narrow interval. For the adulterated class, however, the $\mathbf{X}_{diff}$ matrix had higher values in some of the zones. To select the optimal threshold value, several values were checked, no significant differences were observed in the prediction ability and finally a threshold of 1.25 and 1, respectively, for FT-Raman and NIR were selected.

---------- Fig. 3 ----------

Fig. 4 shows the 822 selected variables, 307 of which were from the 1510 FT-Raman spectra and 515 from the 2166 NIR spectra. This was a significant reduction in variables (around five times fewer). In the FT-Raman spectra, the selected variables are localized at 800-1800 cm^{-1} spectral band (mainly corresponding to simple bonds of C-X (X = Cl, S, O) and double bonds of C = X (X = S, C, O) and N=N. Variables are also selected in the 2700-3100 cm^{-1} range (corresponding to C-H and =C-H bonds). In the NIR spectra, the selected variables are localized in the combination bands region (3650-4750 cm^{-1}, corresponding to the spectral bands of amine, alcohol, amide and generic carbon bonds), in the first overtone region (5900-5600 cm^{-1}, simple bonds S-H and C-H) and in the third overtone region (10000-11000 cm^{-1}, simple bonds C-H and O-H). No variables from the second overtone region were selected. The variables selected were concatenated and autoscaled before SIMCA was implemented. Two SIMCA models were built, the unadulterated class keeping the first four PCs and the adulterated class the first six PC's.

---------- Fig. 4 ----------

The performance parameters of the mid-level data fusion are summarized in table 3. It can be seen they all increase significantly: for the unadulterated class all values (sensitivity and specificity) were 100%, there were no inconclusive assignations. And for the adulterated class values were between 96-100% with 4% of inconclusive assignations, which corresponds to one sample adulterated with almond that was not assigned to any of the two models. From the practical point of view, this error has no real

implications since samples that are not assigned to any class should be confirmed by an alternative technique. This is preferable to assigning a sample wrongly.

Overall, it can be seen that there is a great improvement in the results provided by the two fusion strategies, since there are fewer errors than when individual techniques are used, and there are significantly fewer inconclusive assignations. With the mid-level data fusion, results were slightly better than with the high-level data fusion, since no samples were wrongly assigned to another class. The fact that the variables are selected from both spectroscopic techniques is an indication of the synergy between them: they both provide information that is important for discriminating the classes considered.

5. Conclusions

Since laboratories of these days have a variety of analytical equipment, any data fusion strategy is a feasible way of dealing with multivariate approach. The advantages of applying data fusion strategies using complementary instrumental information has been demonstrated. On the other, it has to be consider that although spectroscopic measurements have a low cost, measuring by more than one technique represents an additional cost.

The benefits of the data fusion methodology in the present study are clear because the classification results are better, especially with respect to unadulterated class, than those obtained individually with NIR and FT-Raman techniques, thus demonstrating that the information obtained from the two spectroscopic techniques has a synergistic effect.

High (decision-level) data fusion has the extra-advantage that it can be applied to all types of measurements, since it combines individual multivariate results (assignations). Fuzzy aggregation connectives have proven to be a good and simple tool for classification analysis.

Acknowledgements

This work was supported by project 2009 SGR 549 of the Catalan Government. M.I López thanks the Rovira i Virgili University for providing her doctoral fellowship (2011BRDI-06-09). We would like to thank LaMorellaNuts S.A.U. for providing the samples.

References

1. A. Garrido, R. Romero-González, M. del Mar Aguilera-Luiz, Comprehensive analysis of toxics (pesticides, veterinary drugs and mycotoxins) in food by UHPLC-MS, Trends Anal. Chem. 63 (2014) 158-169.

2. V. Scognamiglio, F. Arduini, G. Palleschi, G. Rea, Biosensing technology for sustainable food safety, Trends Anal. Chem. 62 (2014) 1-10.

3. P. Oliveri, G. Downey, Multivariate class modeling for the verification of food-authenticity claims, Trends Anal. Chem. 35 (2012) 74-86.

4. J.M. Bosque-Sendra, L. Cuadros-Rodriguez, C. Ruiz-Samblas, A. P. de la Mata, Combining chromatography and chemometrics for the characterization and authentication of fats and oils from triacylglycerol compositional data - A review, Anal. Chim. Acta 724 (2012) 1-11.

5. J.M. Camiña, Pellerano, R.G. Pellerano & E.J. Marchevsky, Geographical and Botanical Classification of Honeys and Apicultural Products by Chemometric Methods. A review, Curr. Anal. Chem. 8 (2012) 408-425.

6. E. Cubero-Leon, R. Peñalver, A. Maquet, Review on metabolomics for food authentication, Food Res. Int. 60 (2014) 95-107.

7. J. M. Cevallos-Cevallosa, J.I. Reyes-De-Corcueraa, E. Etxeberria, M.D. Danyluk, G.E. Rodrick, Metabolomic analysis in food science: a review, Trends in Food Sci. Tech. 20 (2009) 557-566.

8. C.V. Di Anibal, M. Odena, I. Ruisánchez, M.P. Callao, Determining the adulteration of spices with Sudan I-II-II-IV dyes by UV–visible spectroscopy and multivariate classification techniques, Talanta 79 (2009) 887-892.

9. R. Boggia, M.C. Casolino, V. Hysenaj, P. Oliveri, P.Zunin, A screening method based on UV-Visible spectroscopy and multivariate analysis to assess addition of filler juices and water to pomegranate juices, Food Chem. 140 (2013) 735-741.

10. N.A. Fadzlillah, A. Rohman, A.Ismail, S. Mustafa, A. Khatib, Application of FTIR-ATR Spectroscopy Coupled with Multivariate Analysis for Rapid Estimation of Butter Adulteration, J. Oleo Sci. 62 (2013) 555-562.

11. M. De Luca, W. Terouzi, G. Ioele, F. Kzaiber, A. Oussama, F. Oliverio, R. Tauler, G. Ragno, Derivative FTIR spectroscopy for cluster analysis and classification of morocco olive oils, Food Chem. 124 (2011) 1113-1118.

12. L. Lenhardt, I. Zekovic, T. Dramicanin, M.D. Dramicánin, R. Bro, Determination of the Botanical Origin of Honey by Front-Face Synchronous Fluorescence Spectroscopy, Appl. Spectrosc. 68 (2014) 557-563.

13. A. Dankowska, M. Maecka, W. Kowalewski, Detection of plant oil addition to cheese by synchronous fluorescence spectroscopy, Dairy Sci. Technol. 95 (2015) 413-424.

14. C.V. Di Anibal, I. Ruisánchez, M.P. Callao, High-resolution ^{1}H Nuclear Magnetic Resonance spectrometry combined with chemometric treatment to identify adulteration of culinary spices with Sudan dyes, Food Chem. 124 (2011) 1139-1145.

15. R. Consonni, L.R. Cagliani, C. Cogliati, NMR based geographical characterization of roasted coffee, Talanta 88 (2012) 420-426.

16. X. Feng, Q. Zhang, P. Cong, Z. Zhu, Preliminary study on classification of rice and detection of paraffin in the adulterated samples by Raman spectroscopy combined with multivariate analysis, Talanta 115 (2013) 548-555.

17. C.V. Di Anibal, L.F. Marsal, M.P. Callao, I. Ruisánchez, Surface Enhanced Raman Spectroscopy (SERS) and multivariate analysis as a screening tool for detecting Sudan I dye in culinary spices, Spectrochim, Acta A 87 (2012) 135-141.

18. M.I. López, E. Trullols, M.P. Callao, I. Ruisánchez, Multivariate screening in food adulteration: Untargeted versus targeted modelling, Food Chem. 147 (2014) 177-81.

19. X. Lu, S. Peng-Tao, Y. Zi.Hong, Y. Si-Min, Y, Xiao-Ping, Rapid analysis of adulterations in Chinese lotus root powder (LRP) by near-infrared (NIR) spectroscopy coupled with chemometric class modeling techniques, Food Chem. 141 (2013) 2434-9.

20. 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.

21. J. Boccard, D.N. Rutledge, A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion, Anal. Chim. Acta 769 (2013) 30-39.

22. B. Diémé, S. Mavel, H. Blasco, G. Tripi, F. Bonnet-Brilhault, J. Malvy, C. Bocca, C.R. Andres, L. Nadal-Desbarats, P. Emond, Metabolomics Study of Urine in Autism Spectrum Disorders Using a Multiplatform Analytical Methodology, J. Proteome Res. 14 (2015) 5273-5282.

23. P.M. Ramos, I. Ruisánchez, Data fusión dual-domain classification analysis of pigments studied in Works of art, Anal. Chim. Acta 558 (2007) 274–282.

24. P.M. Ramos, M.P. Callao, I. Ruisánchez, Data fusion in the wavelet domain by means of fuzzy aggregation connectives, Anal. Chim. Acta 584 (2007) 360-369.

25. C. Fernández, M.P. Callao, M.S. Larrechi, UV-visible-DAD and 1H-NMR spectroscopy data fusion for studying the photodegradation process of azo-dyes using MCR-ALS, Talanta 117 (2013) 75-80.

26. M.S.Godinho, M.R. Blanco, F.F. Gambarra Neto, L.M. Lião, M.M. Sena, R. Tauler, A.E. de Oliveira, Evaluation of transformer insulating oil quality using NIR, fluorescence, and NMR spectroscopic data fusion, Talanta 129 (2014) 143-149.

27. E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, A. Calvo, O. Busto, Olive oil sensory defects classification with data fusion of instrumental techniques and multivariate analysis (PLS-DA), Food Chem. 203 (2016) 314-322.

28. K.M. Nunes, M.V.O. Andrade, A.M.P. Santos Filho, M.C. Lasmar, M.M. Sena, Detection and characterisation of frauds in bovine meat in natura by non-meat ingredient additions using data fusion of chemical parameters and ATR-FTIR spectroscopy, Food Chem. 205 (2016) 14-22.

29. C.V. Di Anibal, M.P. Callao, I. Ruisánchez, ^{1}H NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs, Talanta 86 (2011) 316-323.

30. A.J. Charlton, P. Robb, J.A. Donarski, J. Godward, Non-targeted detection of chemical contamination in carbonated soft drinks using NMR spectroscopy, variable selection and chemometrics, Anal. Chim. Acta 618 (2008) 196-203

31. T. Mathworks, MATLAB Version 7.0 (n.d.).

32. E.R. Incorporated, PLS_Toolbox Version 7.0.2 (n.d.).

33. M. Bevilacqua, R. Bucci, A.D. Magri, A.L. Magri, R. Nescatelli, F. Marini, Classification and Class-Modelling, F. Marini, Data Handling in Science and Technology, Chemometrics in Food Chem. 28 (2013) 171-233.

34. A. Rius, M.P. Callao, F.X. Rius, Multivariate statistical process-control applied to sulfate determination by sequential injection analysis, Analyst 122 (1997) 737-741.

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.

Figure Captions

Fig 1. Scheme of the data fusion process: a) high- level data fusion and b) mid- level data fusion

Fig 2. Raw spectra of hazelnut paste samples: a) FT-Raman spectra and b) NIR spectra

Fig 3. x_{diff} values: a) unadulterated FT-Raman spectra, b) adulterated FT-Raman spectra, c) unadulterated NIR spectra and d) adulterated NIR spectra

Fig 4. Variables selected by the x_{diff} criteria: a) FT-Raman variables and b) NIR variables

Figure 1.

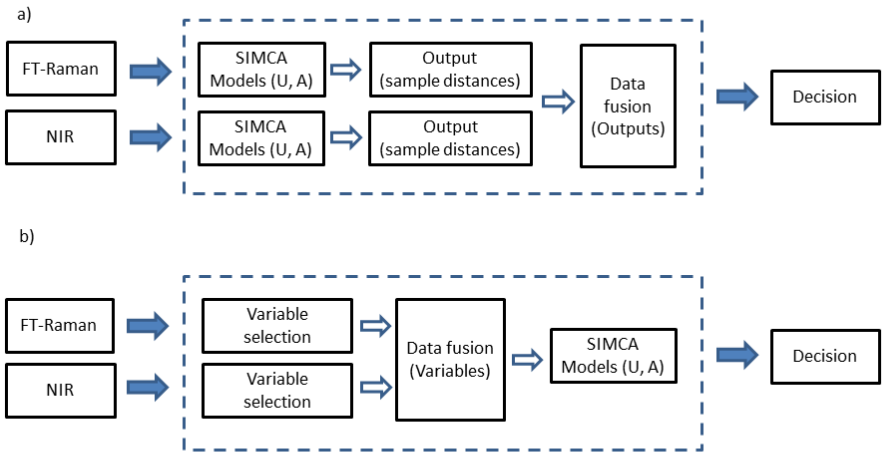

Figure 2.

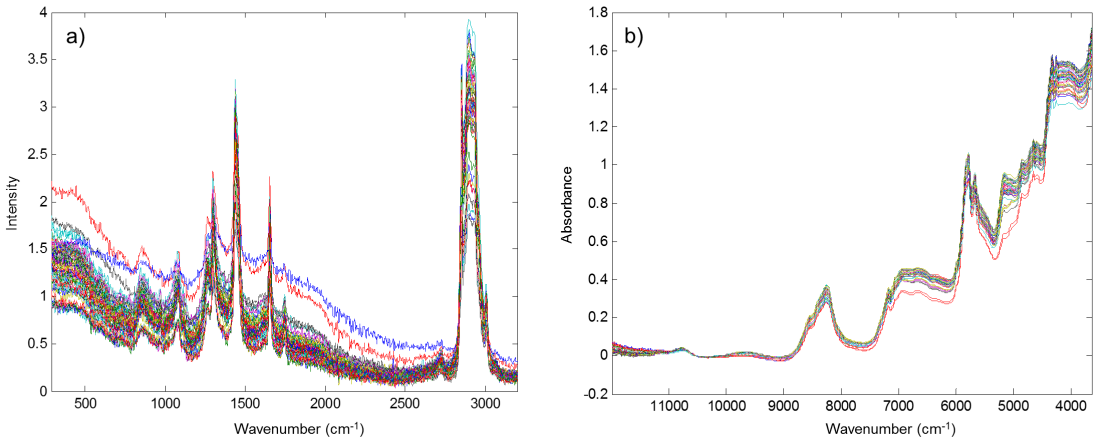

Figure 3.

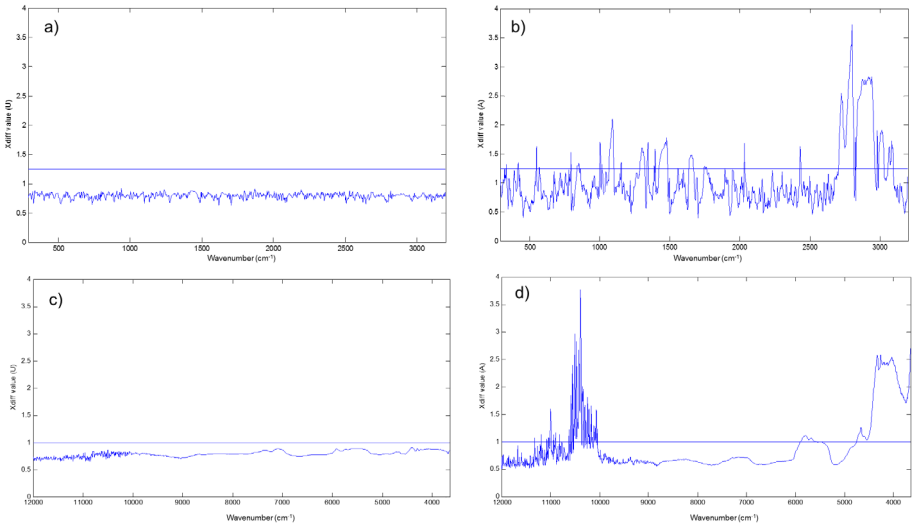

Figure 4

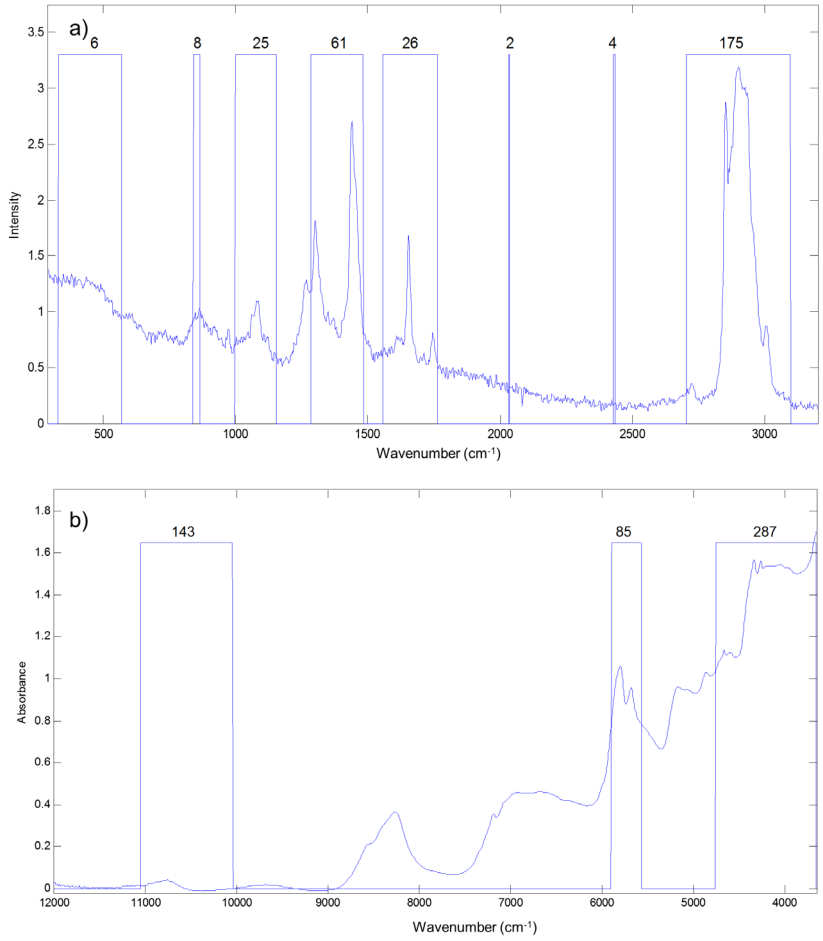

Table 1. Number of samples classified by SIMCA models using FT-Raman and NIR individually

		Assignation			
		Unadulterated class	Adulterated class	Multiple	None
Data	**Number of samples**				
FT-Raman	24 Unadulterated	18	0	6	0
	26 Adulterated almond	0	25	1	0
	27 Adulterated chickpea	2	0	0	25
NIR	24 Unadulterated	18	1	4	1
	26 Adulterated almond	1	24	1	0
	27 Adulterated chickpea	0	0	0	27

Table 2. Class assignation by high-level data fusion for samples misclassified using FT-Raman and NIR techniques. The four fuzzy operators applied are minimum (min), maximum (max), product (prod) and average (avg) for samples unadulterated (U), adulterated with almond (A) and adulterated with chickpea (C). For each operator the majority votes (bold values) were chosen to obtain the ensemble decision

	Unadulterated (U)						Adulterated (A)						
Sample Real class	**Model distance**		**Fusion operators**				**Model distance**		**Fusion operators**				**Ensemble decision**
	FT-Raman	NIR	Min	Max	Prod	Avg	FT-Raman	NIR	Min	Max	Prod	Avg	
1 (U)	0.71	1.02	**0.71**	1.02	**0.72**	**0.88**	0.95	0.78					U, A
10 (U)	0.52	0.48					0.95	2.07	0.95	**2.07**	**1.97**	**1.51**	U
15 (U)	0.69	0.85					0.91	2.49	0.91	**2.49**	**2.27**	**1.70**	U
16 (U)	0.62	0.76					0.87	2.25	0.87	**2.25**	**2.09**	**1.47**	U
19 (U)	0.71	0.42					2.33	0.94	0.94	**2.33**	**2.19**	**1.64**	U
25 (U)	0.82	1.00	**0.82**	1.00	**0.82**	**0.96**	1.63	2.52					U
27 (U)	0.75	0.52					1.81	0.94	0.94	**1.81**	**1.70**	**1.38**	U
6 (A)	2.20	0.85	0.85	**2.20**	**1.87**	**1.53**	0.70	0.50					A
16 (A)	0.94	0.91					0.70	1.29	0.70	**1.29**	**0.90**	**0.99**	A, U
18 (C)	0.98	1.82	0.98	**1.82**	**1.78**	**1.40**	1.77	1.87					-
28 (C)	0.91	1.52	0.91	**1.52**	**1.38**	**1.22**	1.13	1.33					-

Table 3. Performance parameters (%) for samples unadulterated (U), adulterated with almond (A) and adulterated with chickpea (C), using individual techniques and data fusion strategies (high- and mid-level)

	Unadulterated (U)				Adulterated (A)			
	Sensitivity	Specificity (A)	Specificity (C)	Inconclusive	Sensitivity	Specificity (U)	Specificity (C)	Inconclusive
FT-Raman	75	100	93	25	96	100	100	4
NIR	75	96	100	21	92	96	100	4
High Level	88	100	100	12	96	100	100	4
Mid Level	100	100	100	0	96	100	100	4