Talanta 84 (2011) 829–833

Contents lists available at ScienceDirect

Talanta

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

^{1}H NMR and UV-visible data fusion for determining Sudan dyes in culinary spices

Carolina V. Di Anibal, M. Pilar Callao, Itziar Ruisánchez*

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

ARTICLE INFO

Article history:
Received 15 November 2010
Received in revised form 1 February 2011
Accepted 14 February 2011
Available online 24 February 2011

Keywords:
Variable level data fusion
Decision level data fusion
UV-visible
^{1}H NMR
Fuzzy aggregation connectives
Food adulteration

ABSTRACT

Two data fusion strategies (variable and decision level) combined with a multivariate classification approach (Partial Least Squares-Discriminant Analysis, PLS-DA) have been applied to get benefits from the synergistic effect of the information obtained from two spectroscopic techniques: UV-visible and ^{1}H NMR. Variable level data fusion consists of merging the spectra obtained from each spectroscopic technique in what is called "meta-spectrum" and then applying the classification technique. Decision level data fusion combines the results of individually applying the classification technique in each spectroscopic technique. Among the possible ways of combinations, we have used the fuzzy aggregation connective operators. This procedure has been applied to determine banned dyes (Sudan III and IV) in culinary spices. The results show that data fusion is an effective strategy since the classification results are better than the individual ones: between 80 and 100% for the individual techniques and between 97 and 100% with the two fusion strategies.

1. Introduction

Nowadays, food quality and safety characterization is still one of the key issues whenever dealing with foodstuffs and great effort has been devoted to the detection of hazardous additives. The application of spectroscopic techniques has become a usual tool in food analysis [1] and requires the use and development of chemometrics tools in order to display and interpret vast amounts of data. There are some approaches that couple or merge data from two or more analytical techniques to improve multivariate data interpretation. This procedure goes under different names: data fusion, analysis of coupled or linked data, multiset or multiblock data analysis and integrative data analysis [2], among others. In this manuscript, we shall use the term data fusion as it is the term most commonly used by data analysts.

Data fusion has been applied in a variety of fields: for example, the combination of electronic noses and spectroscopic techniques to authenticate olive oil [3] and white grape must [4], and to determine sensory attributes in red wines [5]. Spectroscopic techniques have also been fused to identify pigments in works of art [6,7] and cultivars of extra virgin olive oils [8]. Most fused data comes from Infrared (NIR, FT-IR, MIR), UV-visible, Raman, Fluorescence, and Mass Spectrometry. One of the emerging research areas in which data fusion is applied is known as "metabolomics" or "metabonomics", the goal of which is to obtain information from such highly complex samples as biofluids, cells and tissues [9,10].

Previous studies have shown that the UV-visible [11] and High-Resolution ^{1}H Nuclear Magnetic Resonance (^{1}H NMR) [12] spectroscopic techniques coupled with multivariate classification techniques are well suited for determining the possible adulteration of commercial spices with Sudan dyes (I–IV). Based on those results, the main objective of the present study is to evaluate the combination of both techniques to improve the classification results. In addition, as different concentrations of Sudan dyes can be found in adulterated spices [13], in this paper we will explore the classification ability when the samples have lower concentration levels than those studied in the papers mentioned above. Also, as most of the classification errors were obtained with samples adulterated with Sudan III and IV, the present study focuses on these two dyes.The idea is to get benefits of the possible synergism that two techniques as UV-visible and NMR could have each other; due to the fact that both are based in different fundaments and give different analytical signals, which allows thinking that the information provided by each one could be complementary.

The data provided by the two spectroscopic techniques have been processed separately and jointly by two data fusion strategies: variable and decision level data fusion. So, the overall performance of the classification process is evaluated through the well known classification technique Partial Least Squares-Discriminant Analysis (PLS-DA) for each individual spectroscopic technique and the fusion process.

* Corresponding author at: Department of Analytical and Organic Chemistry, Rovira i Virgili University, Campus Sescelades, Marcel·lí Domingo s/n, 43007 Tarragona, Spain. Tel.: +34 977558490; fax: +34 977558446.
E-mail address: itziar.ruisanchez@urv.cat (I. Ruisánchez).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.02.014

2. Materials and methods

2.1. Samples

A total of 35 spices from different common markets were studied. For UV-visible analysis, each spice was extracted with acetonitrile and the obtained extract was twice filtered. For NMR, samples were dissolved in deuterated chloroform and once filtered. Samples contaminated with Sudan III and IV were prepared by spiking the non-contaminated samples at three concentration levels: 1.4, 3.6 and 7.1 g/kg. In the end, then, three classes were defined: class 1 contained the 35 non-adulterated spices, class 2 contained a total of 105 samples adulterated with Sudan III corresponding to the three concentration levels (35 samples each one) and class 3 contained 105 samples corresponding to the three concentration levels adulterated with Sudan IV. More details about the sample treatment and experimental section can be found in previous studies [11,12].

2.2. Spectrometric techniques and dataset

^{1}H NMR spectra were acquired at 600.13 MHz on a Bruker Avance III-600 spectrometer, equipped with an inverse TCI 5 mm cryoprobe®. One dimensional pulse experiments were carried out using a 90° pulse sequence (zg). For each sample, eight scans of 9.6 kHz of spectral width were collected at 300 K into 64 K data points. A recycling delay time of 15 s was applied between scans to ensure fully relaxation. Exponential line broadening of 0.3 Hz was applied before Fourier transformation and the NMR spectra acquired were phased, baseline-corrected (5th order polynomial adjustment) and calibrated by setting the $CDCl_3$ peak at 7.27 ppm (TopSpin 2.1, Bruker Biospin, Rheinstetten, Germany). UV-visible measurements were made by an Agilent 8453 UV-visible spectrophotometer (Agilent Technologies Inc., Palo Alto, CA, USA) equipped with a diode array detector (DAD) and ChemStation Software (ChemStation Rev. A. 08.03). Each sample was measured against solvent as a blank in a 1-cm pathlength quartz cell and with a spectral resolution of 1 nm.

The UV-visible spectra were acquired between 260 and 600 nm and had a total of 341 variables. The spectral region for NMR is located between 0.5 and 8.9 ppm and a range corresponding to the solvent signal (centred at 7.26 ppm) was removed, so finally, there were a total of 5698 variables.

The dataset (245 samples) is divided into a training and test set. The test set was generated by leaving out a 14% of the samples from class 1 and from each of the three concentration levels included in classes 2 and 3. The selection criterion is based on the PCA scores plot distribution from both UV-visible and NMR data. Finally, the training set consisted of 210 samples and the test set of 35.

3. Chemometrics tools

3.1. Software

All chemometrics treatment was made with Matlab 6.5 Software (The MathWorks, Natick, MA) and PLS_Toolbox 3.5 (Eigenvector Research Incorporated).

3.2. Partial Least Squares-Discriminant Analysis

PLS-DA is the classical PLS regression technique adapted to a supervised classification task. A regression model is calculated that relates the independent variables (e.g. spectra) to an integer "*y*" that designates the class of the sample, with a binary response encoded as {1 0 0} meaning that a sample belongs to class 1; {0 1 0} to class 2 and {0 0 1} to class 3. The model predicts the class for each sample based on a value from zero to one. A value close to zero indicates that the sample is not in the modelled class, while a value closer to one indicates that it is. A threshold between 0 and 1 (above which a sample is considered part of the class) is calculated using Bayesian statistics [14]. The Bayesian threshold assumes that the "*y*" PLS predicted values are normally distributed and the threshold is selected at the *y* value at which the number of false positives and false negatives is minimized. More details of the PLS-DA technique can be found in the literature [15,16].

The optimal number of latent variables (LVs) to be included in each model was chosen using leave-one-out cross-validation to minimize the root mean square-cross validation error (RMSECV) for each class. At the end, this number is selected through a compromise between the optimal value for each class.

3.3. Data fusion

Two levels of data fusion architectures are investigated in this paper: variable and decision level data fusion.

3.3.1. Variable level data fusion

Variable level fusion concatenates the variables into a single vector, which is called a "meta-spectrum". Data must be balanced (all variables in the same scale) prior to the fusion process, so in our particular case only the NMR variables are normalized since the UV-visible intensity values are already between 0 and 1. If the number of concatenated variables is quite high, a variable selection is required. Of the various selection approaches, interval Partial Least Squares (iPLS) was used here [17]. We are not going to describe the iPLS methodology in detail, merely point out that it investigates the influential zones of the spectra that contain the most discriminating predictors, and calculates local PLS-DA models in pre-fixed narrow intervals.

3.3.2. Decision level data fusion

Decision level data fusion combines the classification results obtained from each individual technique. In this study, the PLS-DA classification results are fused using the fuzzy set theory which implements fuzzy aggregation connective operators. Fuzzy theory, introduced by Zadeh [18], is a powerful and general technology for processing information.

A fuzzy set allows membership values between 0 and 1, so in our case the PLS-DA class assignation values are normalized to the interval [0,1] through a simple rescaling such as the following Eq. [19]:

$$m_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})}$$

where ($m_{i,j}$) is the normalized class assignation value, $x_{i,j}$ is the PLS-DA class assignation value for the *i*th sample in the *j*th class and min and max are the minimum and maximum class assignation values in each *j*th class considering all samples, respectively.

Of the wide range of fuzzy connectives and aggregation operators that are available, we chose four aggregation connective operators which belong to the class of context independent constant behavior (CICB) operators [20]. The aggregation operators are: Minimum, Maximum, Product and Arithmetic Average. Considering the PLS-DA normalized assignation values obtained with each technique, the minimum and maximum are identified and the product and average are calculated. For the sake of clarity, two examples will be shown in the results section. To obtain the "ensemble decision" of each operator, the maximum value of the three possible classes is chosen [21]. The sample is finally assigned by the majority vote provided by all the fuzzy operators in the "ensemble decision".

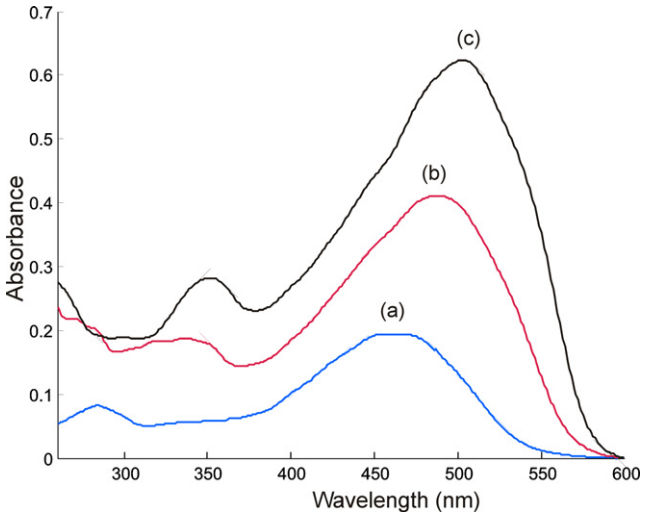

Fig. 1. UV-visible spectra of a random paprika spice: (a) unadulterated, (b) spiked with Sudan III and (c) spiked with Sudan IV. Both Sudan dyes are at 5 mg l^{-1} (7.1 g kg^{-1}).

4. Results and discussion

4.1. Spectra characterization

The only difference that exists between the chemical structures of the two Sudan dyes is the presence of two methyl groups that Sudan IV has. Figs. 1 and 2 show the UV-visible and NMR spectra, respectively, of (a) a random unadulterated paprika, (b) the same spice spiked with Sudan III and (c) the same spice spiked with Sudan IV. The UV-visible spectra show that the absorption maximum from both adulterated samples shift slightly towards a longer wavelength respect to the unadulterated one, and that the sample adulterated with Sudan IV dye has the highest sensitivity (higher absorbance values). In the NMR spectra, a detailed comparison is not so easy, although it is evident that the aromatic zone in the samples containing Sudan dyes has more signals than the unadulterated sample.

4.2. Selection of test samples

Fig. 3 shows, as an example for class 1, the PCA scores plots used for selecting the test set samples. In this case, 5 out of 35 samples are selected from both UV-visible and NMR scores plots, to cover in the most representative way the PCA sample's distribution. For the other two classes, the same criterion has been applied but considering the concentration level, so 15 out of the 105 samples from class 2 and 15 from out of the 105 samples from class 3 have been selected (5 from each concentration level).

4.3. Independent decision-making

First of all, the UV-visible and NMR data are pre-treated separately: the UV-visible spectra are mean centred while the NMR spectra are autoscaled. In addition, before PLS-DA is applied as the classification technique, a NMR variable selection is carried out by means of the iPLS algorithm. The intervals selected by iPLS are depicted as solid-line rectangles (Fig. 4) and it can be seen that these intervals are in both the aromatic zone and several aliphatic zones. The final number of selected variables is 777.

Table 1 presents the PLS-DA misclassification results obtained from UV-visible and NMR data independently. It can be seen that the overall error trend for both techniques is that samples are assigned to more than one class, one of which is the true class. It can also be seen that the misclassifications provided by the two

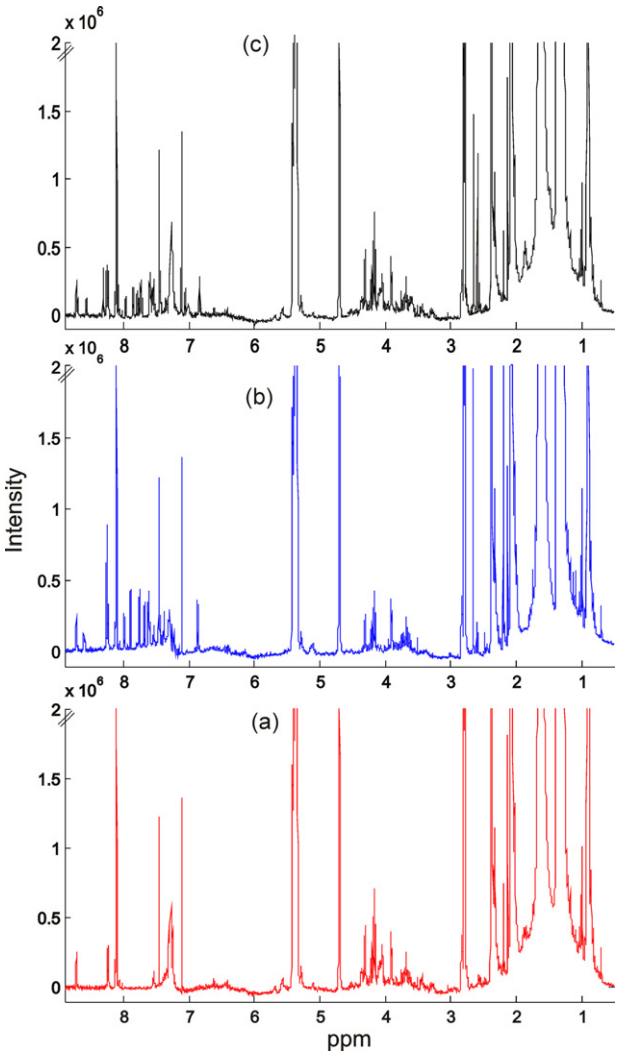

Fig. 2. ^{1}H NMR spectra of a random paprika spice: (a) unadulterated, (b) spiked with Sudan III and (c) spiked with Sudan IV. Both Sudan dyes are at 50 mg l^{-1} (7.1 g kg^{-1}).

techniques do not match each other, thus demonstrating the complementarities existing between the different types of information. A closer look at the table shows that for the training set, most of the UV-visible errors occur for samples from the unadulterated class (class 1), which also have a high probability (evaluated from the Bayesian distribution) of being assigned to class 3 (samples adulterated with Sudan IV). This probability varies between 99 and 100% for class 1 and between 79 and 100% for class 3. This type of error represents an economic risk since they must be withdrawn from commercial markets until they have been confirmed. Similarly, most of the NMR errors occur in samples from class 2 (adulterated with Sudan III), which are also assigned to the unadulterated class (class 1). To a lesser extent, the opposite also occurs: samples from class 1 are also assigned to class 2. The implications of these errors are the same as the ones discussed above, as they are samples assigned to more than one class.

Following the discussion of the table, only one sample is not assigned to any class with UV-visible (sample 39) and two with NMR (samples 29 and 57), shown in hyphenated line (see Table 1). In addition, two samples are wrongly assigned to another class by NMR (9 and 33). Of these, sample 33 is an example of the most serious error possible, as it has implications for the consumer's health: a spice is being considered safe for human consumption when in fact it is not. Finally, if the proportion of samples in each data set is taken into account, the prediction results for the test set behave in

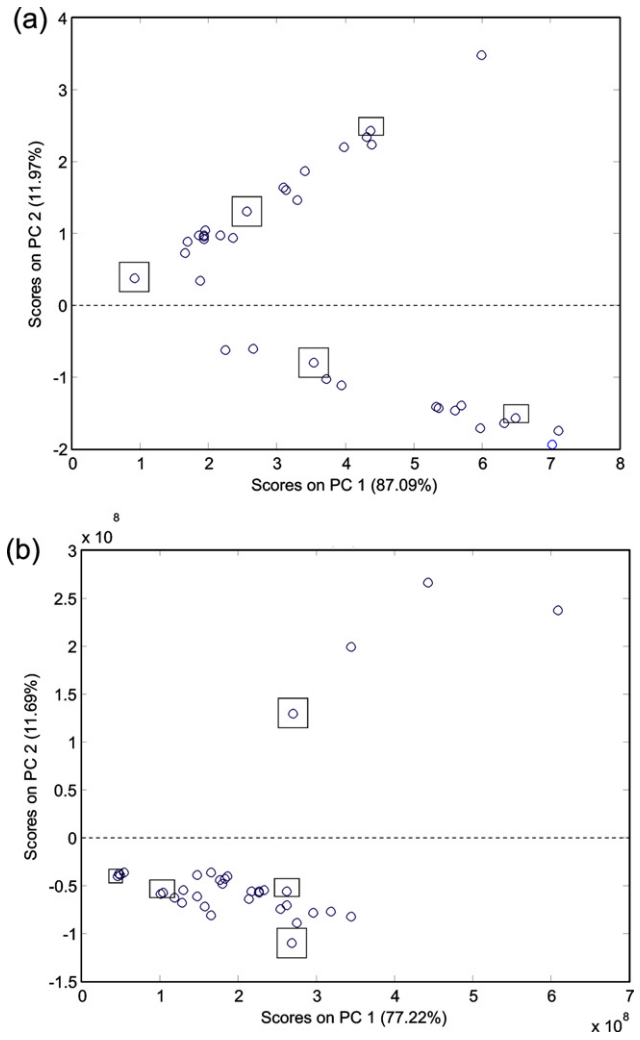

Fig. 3. PCA scores plots from class 1: (a) UV-visible and (b) NMR. Selected samples are marked with squares.

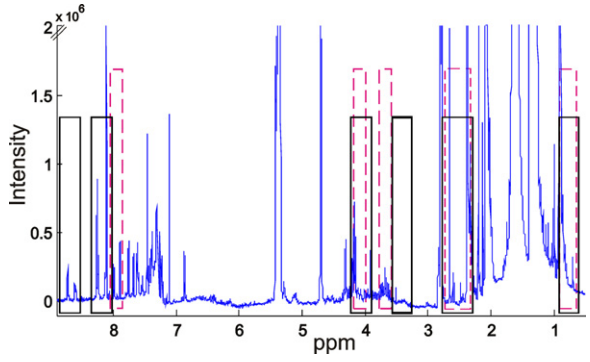

Fig. 4. iPLS selected intervals for the NMR spectrum from a random spice spiked with Sudan III. The solid line represents the iPLS selected intervals when NMR is used individually. The dotted line represents the iPLS selected intervals when UV-visible and NMR are fused.

a similar fashion to those of the training set. It should be mentioned that all the misclassified samples from classes 2 and 3 (Table 1) are adulterated at the lowest Sudan dye concentration.

4.4. Data fusion

4.4.1. Variable level fusion

Secondly, UV-visible and NMR data are fused. For the variable level data fusion, the 341 and 5698 raw variables from UV-visible and NMR, respectively, are concatenated into a meta-spectrum. Due to the fact that the number of variables is really high, three consecutive iPLS models were made. The final intervals selected include all the UV-visible variables as well as some NMR zones, which are depicted in Fig. 4 as dotted line rectangles. At this point, 826 final variables are selected, 341 of which are UV-visible and 485 NMR. 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.

Table 1
PLS-DA misclassification results when UV-visible or NMR are used individually.

	PLS-DA class assignation			
	Sample	True class	UV-visible	NMR
Training Set	1	1	1,3	
	3	1	1,3	
	4	1	1,3	
	5	1	1,3	
	6	1	1,3	
	7	1	1,3	
	9	1		2
	11	1		1,2
	14	1	1,2	
	18	1		1,2
	29	1		-----
	30	1		1,2
	32	2		2,1
	33	2		1
	36	2		2,1
	38	2		2,1
	39	2	-----	
	40	2		2,1
	50	2		2,1
	55	2		2,1
	57	2		-----
	127	3	2,3	
Test Set	211	1		1,2
	217	2		2,1
	219	2		2,1

4.4.2. Decision level fusion

In the decision level data fusion, the individual class assignation results provided by PLS-DA are fused by means of the four fuzzy aggregation connective operators (minimum, maximum, product and average) and the majority vote rule, as described in the theory section. As an example, Table 2 shows the classification results in terms of numerical values when the different operators are used. The maximum values for each fuzzy operator are shown in bold. Sample 18 (class 1) is indistinctly assigned to class 1 and 2 by NMR, but after the decisions of all the fuzzy operators it is correctly assigned. Similarly, sample 32 (class 2) is also indistinctly assigned to class 1 and 2 by UV-visible and NMR spectroscopic techniques, as they give scaled classification results very close each other. However, in this case the final sample classification is the same as when the individual techniques are used.

4.5. Comparison of the different strategies

Table 3 shows the PLS-DA misclassification results for the variable and decision level data fusion. Overall, it can be seen that there is a great improvement in the results provided by the two fusion strategies, since the number of errors is lower than when the individual techniques are used (Table 1). With variable level data fusion no wrong assignations to another class are obtained. These results are positive because, from a practical point of view, samples that are assigned to more than one class or assigned to any class should be confirmed by an alternative technique. This is preferable to assigning a sample wrongly. With decision level data

Table 2
Assignation of classes by decision level data fusion.

Sample 18	Scaled PLS-DA class assignation values			Ensemble decision
	Class 1	Class 2	Class 3	
UV-visible	0.68	0.42	0.23	
NMR	0.51	0.51	0.30	
Minimum	**0.51**	0.42	0.23	Class 1
Maximum	**0.68**	0.51	0.30	Class 1
Product	**0.34**	0.21	0.07	Class 1
Average	**0.59**	0.46	0.26	Class 1
Majority vote				**Class 1**
Sample 32	Scaled PLS-DA class assignation values			Ensemble decision
	Class 1	Class 2	Class 3	
UV-visible	0.46	0.45	0.19	
NMR	0.48	0.50	0.33	
Minimum	**0.46**	**0.45**	0.19	Class 2,1
Maximum	**0.48**	**0.50**	0.33	Class 2,1
Product	**0.22**	**0.23**	0.06	Class 2,1
Average	**0.47**	**0.48**	0.26	Class 2,1
Majority vote				**Classes 2,1**

Table 3
PLS-DA misclassification results for variable and decision level data fusion.

	PLS-DA class assignation			
	Sample	True class	Variable fusion	Decision fusion
Training Set	3	1	1,3	
	32	2		2,1
	33	2	-----	1
	57	2		1
	127	3	-----	
Test Set	235	3	-----	

fusion, two adulterated samples (33 and 57) pose a serious problem because, as mentioned above, they can affect the consumer's health. The other misclassified sample (32) become in a suspicious sample that has to be subsequently confirmed as mentioned above. As in the previous case, the test set follows the same pattern as the training set.

Finally, Table 4 shows the correct classification percentages obtained with the individual spectroscopic techniques and with the two fusion strategies. The global classification percentages obtained for class 1 with variable and decision level fusion strategies are clearly better than those obtained with UV-visible and NMR. For class 2, the fusion strategies improve the classification results obtained with the NMR technique while for class 3 all four strategies give satisfactory and comparable classification values. A detailed look at the results considering the concentration levels (classes 2 and 3) indicates that lower percentages are obtained for samples at the lowest concentration level, being the worst case the percentage obtained for class 2 with the individual NMR technique (71.4%). For class 2, the NMR classification result is improved by the two fusion strategies which give a 97 and 91.4% of correct classification with variable and decision fusion respectively, which are

Table 4
Global correct classification percentages for each class (bold values). For classes 2 and 3, the results corresponding to each concentration level (from lower to upper) are shown in brackets.

	Class 1	Class 2	C lass 3
UV-visible	**80.0**	**99.0** (97.1; 100; 100)	**99.0** (97.1; 100; 100)
NMR	**82.8**	**90.5** (71.4; 100; 100)	**100** (100; 100; 100)
Variable fusion	**97.1**	**99.0** (97.1; 100; 100)	**98.1** (94.3; 100; 100)
Decision fusion	**100**	**97.1** (91.4; 100; 100)	**100** (100; 100; 100)

similar to the UV-visible one. For class 3, the percentages obtained with the individual techniques and fusion strategies are similar although the variable selection strategy gives lower values than the individual ones. On the other hand, with the other two concentration levels, the maximum classification ability is obtained (100%) in all cases (individual techniques and data fusion strategies).

5. Conclusions

Fusing data from UV-visible and ^{1}H NMR instruments is a powerful tool for detecting banned Sudan dyes at three different concentrations levels in commercial spices destined for human consumption. None of the adulterated samples with Sudan III and Sudan IV were misclassified to each other. Some samples at the lowest concentration level were assigned to their own class and also to the non adulterated class.

Tacking into account that nowadays many laboratories have a variety of analytical equipment any data fusion strategy is a feasible way to deal with multivariate approach. Decision level data fusion has the extra-advantage that it can be applied to all types of measurements, since it combines the individual multivariate results. Fuzzy aggregation connectives have been demonstrated to be a good and simple tool to implement for classification analysis.

The benefits of the data fusion methodology in the present study are clear because the classification results are better than the ones obtained individually with UV-visible and NMR techniques, thus demonstrating that the information obtained from the two spectroscopic techniques has a synergistic effect.

Acknowledgment

The authors would like to thank the Agency for the Administration of University and Research Grants of the Catalan Government (AGAUR) for providing Carolina Di Anibal with a doctoral fellowship.

References

[1] C. Cordella, I. Moussa, A.C. Martel, N. Sbirrazzuoli, L. Lizzani-Cuvelier, J. Agric. Food Chem. 50 (2002) 1751–1764.
[2] I.V. Mechelen, A. Smilde, Chemom. Intell. Lab. Syst. 104 (2010) 83–94.
[3] M. Casale, C. Casolino, P. Olivieri, M. Forina, Food Chem. 118 (2010) 163–170.
[4] S. Roussel, V. Bellon-Maurel, J.M. Roger, P. Grenier, J. Food Eng. 60 (2003) 407–419.
[5] D. Cozzolino, H.E. Smyth, K.A. Dattey, W. Cynkar, L. Janik, R.G. Dambergs, I. Leigh Francis, M. Gishen, Anal. Chim. Acta 563 (2006) 319–324.
[6] P. Ramos, I. Ruisanchez, K. Andrikopoulos, Talanta 75 (2008) 926–936.
[7] P. Ramos, I. Ruisanchez, Anal. Chim. Acta 558 (2007) 274–282.
[8] M. Casale, N. Sinelli, P. Oliveri, V. Di Egidio, S. Lanteri, Talanta 80 (2010) 1832–1837.
[9] J. Forshed, H. Idborg, S.P. Jacobsson, Chemom. Intell. Lab. Syst. 85 (2007) 102–109.
[10] A.K. Smilde, M.J. van der Werf, S. Bijlsma, B.J.C. van der Werff-van der Vat, R.H. Jellema, Anal. Chem. 77 (2005) 6729–6736.
[11] C.V. Di Anibal, M. Odena, I. Ruisanchez, M.P. Callao, Talanta 79 (2009) 887–892.
[12] C.V. Di Anibal, I. Ruisanchez, M.P. Callao, Food Chem. 124 (2011) 1139–1145.
[13] K. Mishra, S. Dixit, S.K. Purshottam, R.C. Pandey, M. Das, S.K. Khanna, Int. J. Food Sci. Technol. 42 (2007) 1363–1366.
[14] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[15] M. Barker, W. Rayens, J. Chemom. 17 (2003) 166–173.
[16] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Kich, PLS_Toolbox Version 3.5 for Use with Matlab™, Eigenvector Research, Inc., Manson, WA, USA, 2005, pp. 185–189.
[17] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413–419.
[18] L.A. Zadeh, Inf. Control 12 (1968) 94–102.
[19] P. Ramos, M.P. Callao, I. Ruisanchez, Anal. Chim. Acta 584 (2007) 360–369.
[20] I. Bloch, IEEE Trans. Syst. Man Cybern. Part A 26 (1996) 52–67.
[21] L. Lam, S.Y. Suen, IEEE Trans. Syst. Man Cybern. Part A 27 (1997) 553–568.