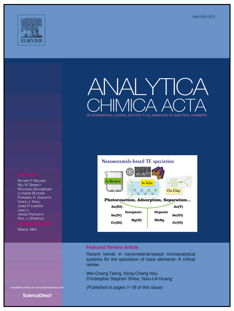


Multi-level Data Fusion Strategies for Modeling Three-way Electrophoresis Capillary and Fluorescence Arrays Enhancing Geographical and Grape variety Classification of Wines

Rocío Ríos-Reina, Silvana M. Azcarate, José M. Camiña, Héctor C. Goicoechea

PII: S0003-2670(20)30662-0

DOI: https://doi.org/10.1016/j.aca.2020.06.014

Reference: ACA 237729

To appear in: *Analytica Chimica Acta*

Received Date: 10 April 2020

Revised Date: 5 June 2020

Accepted Date: 8 June 2020

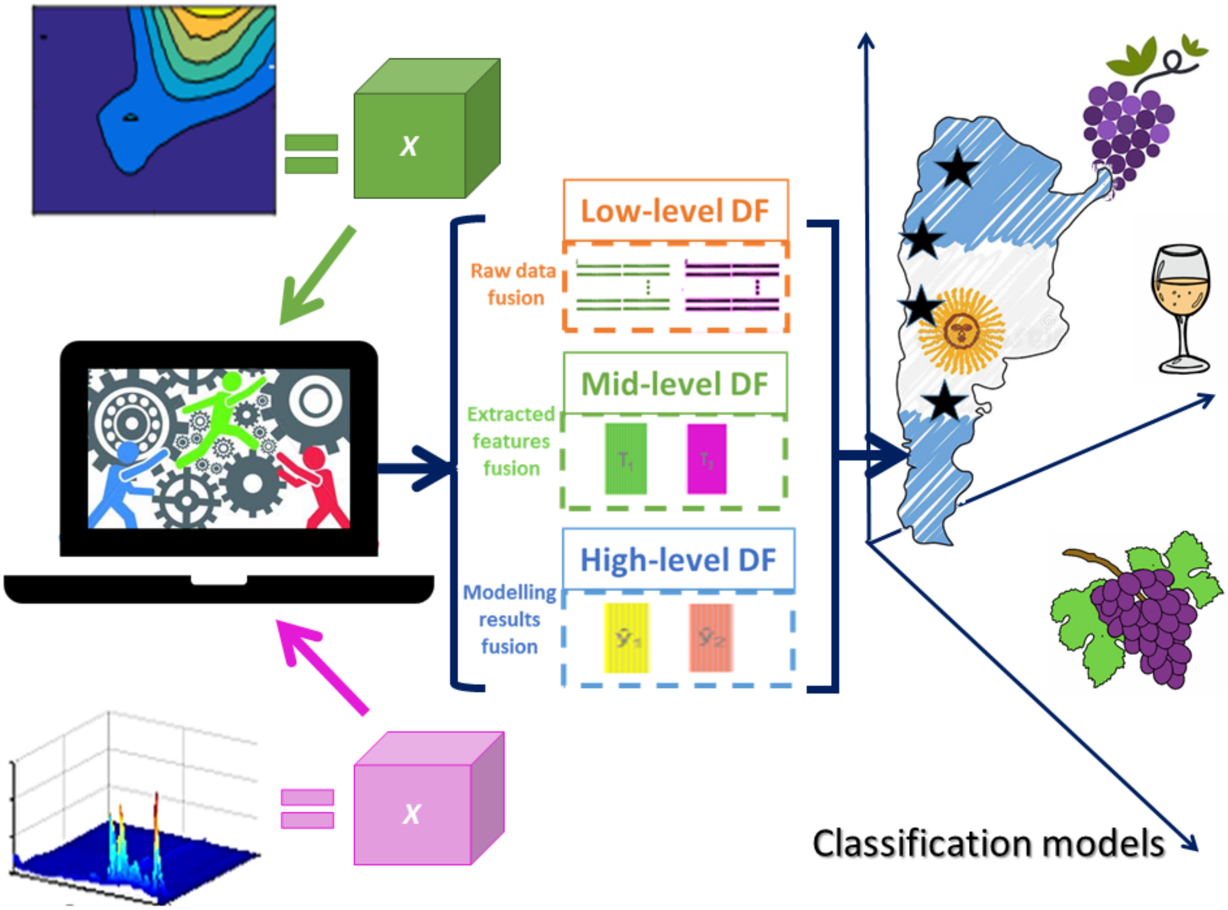
CRediT author statement

Silvana M. Azcarate: Conceptualization, Methodology, Investigation, Writing - Review & Editing.

Rocío Ríos Reina: Software, Formal analysis, Visualization.

José M. Camiña: Supervision.

Héctor C. Goicoechea: Conceptualization, Investigation, Resources, Supervision, Funding acquisition.


X
X
Low-level DF
Raw data fusion
Mid-level DF
Extracted features fusion
T_1
T_2
High-level DF
Modelling results fusion
$\hat{y}_1$
$\hat{y}_2$
Classification models

Multi-level Data Fusion Strategies for Modeling Three-way Electrophoresis Capillary and Fluorescence Arrays Enhancing Geographical and Grape variety Classification of Wines

Rocío Ríos-Reina [a], Silvana M. Azcarate [b,*], José M. Camiña [b], Héctor C. Goicoechea [c]

Affiliations

[a] Área de Nutrición y Bromatología, Fac. Farmacia, Univ. Sevilla, C/P. García González no. 2, E-41012 Sevilla, Spain.

[b] Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa-CONICET, and Instituto de Ciencias de la Tierra y Ambientales de La Pampa (INCITAP), Av. Uruguay 151 (6300) Santa Rosa, La Pampa, Argentina.

[c] Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ), Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral-CONICET, Ciudad Universitaria, Santa Fe (S3000ZAA), Argentina

*To whom correspondence should be addressed. E-mail silvanaazcarate@gmail.com

Abstract

Capillary electrophoresis with diode array detection (CE-DAD) and multidimensional fluorescence spectroscopy (EEM) second-order data were fused and chemometrically processed for geographical and grape variety classification of wines. Multi-levels data fusion strategies on three-way data were evaluated and compared revealing their advantages/disadvantages in the classification context. Straightforward approaches based on a series of data preprocessing and feature extraction steps were developed for each studied level. Partial least square discriminant analysis (PLS-DA) and its multi-way extension (NPLS-DA) were applied to CE-DAD, EEM and fused data matrices structured as two-way and three-way arrays, respectively. Classification results achieved on each model were evaluated through global indices such as average sensitivity non-error rate and average precision. Different degrees of improvement were observed comparing the fused matrix results with those obtained using a single one, clear benefits have been demonstrated when level of data fusion increases, achieving with the high-level strategy the best classification results.

Keywords: Electrophoresis capillary, Multidimensional Fluorescence Spectroscopy, Three-way data modeling, Multi-level data fusion, Classification.

1. Introduction

In multivariate classification setting, second-order data handling is producing a significant impact on the development of analytical strategies, especially for determining characteristic patterns of the analytes of interest in highly complex matrices [1, 2]. In particular, food characterization tasks are in an innovative progressive change going from the development of dedicated methods for quantification of specific compounds to the fingerprint acquisition by rapid, non-destructive and non-selective instrumental techniques [3–5].

Hence, second-order data, especially three-way arrays modeling applications, could supply interesting improvements as regards results attained when extremely complex systems should be classified. Thereby, it has been demonstrated that data analysis can be more effective when modeling second- or higher -data with multiway algorithms compared to unfolding procedures [1, 6]. Otherwise, it has been recently revealed that the ability in terms of discrimination power can be improved by using second-order data arranged in a three-way structure instead of first-order data [2].

Over the years, analytical methods and data analysis tools commonly used in food quality and process control had to be re-evaluated and modified to fit these new tasks [1]. In this progression of gathering more and better information, the multivariate statistical analysis of fused data has become a powerful tool for enhancing the reliability of the results. Being the key point how the information sources can be combined to provide the joint classification prediction of the samples, three levels of data fusion (DF) have been reported [7, 8].

Firstly, low-level DF (1-DF) implies a simple concatenation of the individual matrices to build a single array that is then used for calculating a single model for final classification. In the food authentication and quality control field, it has been the most used fusion approach to improve the results since is a common, conceptually simple, first attempt with outputs from different sources providing first-order data [9–11]. Nevertheless, out of the food analysis scope, three-way arrays have been concatenated using low-level DF strategy in a recent original report [12].

Secondly, mid-level DF (2-DF) first extracts some relevant features from each data source separately and then these outputs are concatenated to build a single array to be then processed by the desired chemometric technique. This approach has been probably the one that has purposed more challenges chiefly for second-order data analysis, where witty strategies for data compression, extraction or reduction have been explored for improving outcomes. Mainly, the attempt has been addressed to evaluate the combination of first- and second order data provided by multiple platforms since data are very different in structure, size or scale [13–18]. Otherwise, the performance of mid-level applications has commonly been compared to low-level fusion as well as to single models [19–22].

Lastly, the high-level DF (3-DF) builds separate models for the different blocks, and the individual results are then integrated into a single final response. This strategy has been lesser explored than the two mentioned above. Although several methodologies for final identity declaration by modelling the individual matrices independently have been reported [8, 23], only few of them have been inquired in food classification context. High-level DF has been mainly implemented for the comparison with the other two DF levels [24, 25].

The aim of this work was to develop multiple strategies to assess the three DF levels on two second-order arrays, with different data complexity, in order to know the correlation and analogy between both information sources for twofold classification purposes. The focus was put on the development of models able to distinguish among white wines of three different grape varieties with geographical indication (GI) from the four main wine production regions of Argentina [26]. For that, fluorescence excitation–emission matrix spectroscopy (EEM), and capillary electrophoresis with diode array detector (CE-DAD) were applied as non-target analysis in order to acquire a fingerprint to characterize the wines.

To our best knowledge, it is the first time that the multi-levels of DF strategies on second-order data are evaluated and compared revealing their advantages/disadvantages in the classification context. Thus, we developed multiple and straightforward approaches based on a series of data preprocessing and feature extraction steps, which constitutes a significant improvement in the DF analysis, and it offers a wide range of possibilities when second-order data of different nature are assessed. Finally, the challenge consisted in finding the optimal combination of data preprocessing, fused data and data modeling that would provide the best results.

2. Materials and methods

2.1. Samples

Thirty-nine samples of commercial white wine from four wine-producing origins, all belonging to provinces of Argentina (Mendoza-M, San Juan-SJ, Salta-S, and Rio Negro-RN) and three different grape varieties (Chardonnay-CH, Sauvignon Blanc-SB and Torrontés-TO), were included in this study: 14 Chardonnay wines (10 from Mendoza and 4 from San Juan), 13 Sauvignon Blanc wines (10 from Mendoza, 1 from San Juan, and 2 from Río Negro) and 12 wines from grapes of the variety Torrontés (4 from Mendoza, 1 from San Juan, 5 from Salta, and 2 from Río Negro). Wine samples were selected from the 2011 to 2013 vintages and bought from a local market. The alcoholic content ranged from 12.2 to 13.8% v/v of ethanol. These samples were analyzed in triplicate by the two techniques described below.

2.2. Data acquisition

2.2.1. EEM data

All spectrofluorimetric measures were acquired according to the method reported by Azcarate *et al.* [26] using a Cary Eclipse Fluorescence Spectrophotometer (Agilent Technologies, Waldbronn, Germany) with a 1×1 cm quartz fluorescence cell, xenon flash lamp. CaryEcplise software package was used to control the instrument, data acquisition and data analysis. Fluorescence excitation spectra were recorded by varying the wavelengths between 245 and 341 nm (increment 5 nm), and by recording the emission spectra from 300 to 500 nm (spaced by 0.5 nm interval). Fluorescence measurements were done in triplicate for each sample.

2.2.2. CE-DAD data

The electrophoretic run conditions are treated in detail in our previous work [27] and here only main analysis steps will be recalled. All electropherograms were acquired on a CE system (Agilent Technologies, Waldbronn, Germany) equipped with a DAD and an uncoated fused silica capillary of 40 cm total length (31.5 cm effective length) and 75 µm inner diameter (MicroSolv Technology Corporation, Eatontown, NJ, USA). Separation was performed by applying a voltage of 24 kV and with a typical current of approximately 80 µA. The hydrodynamic injection was performed in the positive electrode of the capillary by applying a pressure of 40 mbar for 8 s. The cartridge was maintained at 25.0°C. The electropherograms were recorded during 10 min at 0.3 s steps and recording UV spectra between 189 and 401 nm each 2 nm and samples were analyzed by triplicate.

2.3. Data analysis

In order to extract and/or merge the information presented in each data set obtained by the two different instrumental analysis of each sample, different chemometric algorithms were employed. As both techniques produced out-puts with the same data structure (i.e. three-way arrays), they were analyzed by similar algorithms in order to decompose and compress the data.

The data analysis workflow developed in this study is schematized in **Fig. 1.** In general terms, it includes: 1) building separate classification models on data obtained from the individual analytical techniques by applying 3 different approaches; and 2) building classification multiplatform models by applying different DF strategies: l-DF, 2-DF and 3-DF (assessing different approaches). Then, all the classification models obtained were assessed and compared.

Insert here Fig. 1

2.3.1. Data set and preprocessing

In order to validate the classification models, the dataset corresponding to each technique, containing 39 samples, was split into a training set of 24 samples (12 CH, 12 SB and 12 TO or 15 M, 4 SJ, 2 RN and 2 S) and a test set of 15 samples (6 CH, 5 SB and 4 TO or 9 M, 2 SJ, 2 S, 2 RN) by using the Duplex algorithm [28], keeping the triplicates in the same set (i.e. the training set contained 72 analysis and the test set 45 analysis). The split between training and test sets was done by keeping the ratio of samples of each class like in the original set, balancing the representation of each category and keeping the replicates together. Moreover, after checking by exploratory analysis that both sets spanned the whole variability domain, the same split was maintained for all the data sets (the individual and the fused data sets).

To find the optimal classification results for each class studied (variety or origin), different preprocessing options were considered in each model: both mean centering and autoscaling were used depending on the nature of the data, as well as none preprocessing.

2.3.2. Decomposition and compression methods

As can be observed in **Fig. 1**, different decomposition and compression methods (i.e. exploratory and reduction data analysis) were applied. Then, the features obtained were used for the DF models. Thus, on the one hand, the original EEM and CE-DAD three-way arrays (117×49×41 and 117×676×107, respectively, considering the samples by triplicate) were unfolded in a multiset structure via row-wise augmentation and then these new matrices (a matrix of 117×2009 for EEM unfolded data and of 117×72332 for CE-DAD unfolded data), as well as the fusion of both, were used for the classification or compressed by principal component analysis (PCA).

On the other hand, the EEM three-way array was decomposed by parallel factor analysis (PARAFAC) [29] into trilinear components, related to the main fluorophores present in the samples, whose scores (first mode loadings) were used as features for the classification, or to build a fused dataset previous to the classification process**.** A three-factor model, constrained with non-negativity in all modes, was obtained as the optimum model according to the CORe CONsistency DIAgnostic test (COR-CONDIA) [18, 30], the explained variance, the visual inspection of the profiles and residuals [26].

Therefore, the three-way array matrix was decomposed by PARAFAC to three new matrices containing the PARAFAC scores of the three fluorophores, as well as their excitation and emission loadings.

Finally, CE-DAD array decomposition was carried out by Tucker3 and the resulted **A** matrix (Tucker3 output with the concentrations) was directly used, or fused in a new data set, for the classification purpose. In this study, Tucker3 was selected due to the high complexity of the CE-DAD data, which require different number of factors in each mode. The number of factors selected was 18, 18 and 6 for each recorder mode, obtaining a model with a 95% of total explained variance. Non-negativity was imposed as unique constraint in all modes in agreement with a previous work [27] and three matrices **A**, **B** and **C** were obtained containing the concentration, electropherogram and spectra profiles, respectively, together with a ***G*** core (18×18×6) corresponding to the magnitude of the interaction among factors in different modes.

2.3.3. Classification methods

In this work, two classification techniques derived from the regression algorithm partial least squares (PLS) were used: the PLS-DA (DA for discriminant analysis) [31] for first-order data and its multi-way or multilinear extension (NPLS-DA) [32] for second-order data (three-way arrays). In order to select the proper number of latent variables (LVs), i.e. the dimensionality of the model, the minimum classification error rate in cross-validation (venetian blind) was considered. In discriminant analysis, the dependent variable, **Y**, holds the class information (as many *y*-variables as number of classes). The raw predictions from a PLS-DA model is a value of nominally zero or one. A value closer to zero indicates the new sample is not in the modeled class; a value of one indicates a sample is in the modeled class [31].

2.4. Individual and data fusion strategies

In this study, four modeling strategies were tested in order to obtain the best classification of the wine samples according to their origin and/or grape variety: classification models of individual techniques and classification models by low-, mid- and high-level data fusion approaches (**Fig. 1**).

Before the data fusion, three classification models were obtained with the individual data matrices EEM and CE-DAD: a NPLS-DA model obtained by each original three-way array, and two PLS-DA models, one with the unfolded matrices and other with the decomposed matrices by PARAFAC or Tucker3, respectively. The strategy followed is described at the top of **Fig. 1.** Each matrix was split into a training/test set (72/45) before building the classification models.

2.4.1. Low-level data fusion (1-DF)

In the 1-DF approach, the data matrices are directly concatenated to provide sample classification [8]. In this study, the unfolded data from CE-DAD and EEM matrices were concatenated before any model calculation. These single blocks were joined in a single matrix providing an overall data set with 74341 variables (2009 variables from the unfolded EEM matrix plus 72332 variables from the unfolded CE-DAD matrix). After that, the data fused matrix was split into training and test sets. Then, two different 1-DF options were tested (**Fig.1.**).

Despite there are many options to be carried out in this 1-DF approach (i.e. applying PCA on the concatenated matrix and then LDA on the scores, or the direct application of other classification algorithms), the selected option was to develop a PLS-DA model, with 6 LVs, and built directly with the concatenated data matrix, after mean centering data preprocessing, by means of the previously described validation protocol (named as Low- level DF: opt1 in **Fig.1**) in order to perform the same classification method in all the strategies of the study.

2.4.2. Mid-level data fusion (2-DF)

In general, in 2-DF strategies, the analytical data are merged at the features level [25]**.** This means that relevant features are independently extracted from each analytical data matrix, which are then concatenated into a single global matrix that is used as input to perform a classification model[8]. In comparison with 1-DF strategies, this method allows to guarantee a more balanced representation of each source of information, in the case of each analytical data matrix has a huge difference in the number of variables [25]**.** However, in this 2-DF the main issues to control are the features to retain from each model and the method to extract them as well as the preprocessing method to adopt. In this study, two different 2-DF strategies were tested, differing from the feature extraction method used.

In the first 2-DF option **(**named as Mid-level DF: opt-1 in **Fig. 1),** the relevant feature extraction was performed by the development of a PCA model for each data block. The number of principal components (PCs) chosen for each PCA model was again selected in order to give more than 90% of cumulative variance in both blocks. Thus, 7 and 4 PCs were selected for the unfolded both CE-DAD and EEM matrices (previously pre-processed by mean-centering), respectively. Then, the PCA-scores associated to the first 7 and 4 PCs for each data block (Str) were considered as extracted features and were then fused in a new matrix (with 11 variables). This fused matrix was pre-processed by auto-scaling or none-preprocessed and then modeled by means of

PLS-DA, obtaining PLS-DA models with 4 and 5 LVs according to the preprocessing method applied.

In the second 2-DF option tested in this study (named as Mid-level DF: opt-2 in **Fig. 1**)**,** the relevant features of each data block were obtained by the development of a PARAFAC and a Tucker3 model for EEM and CE-DAD matrices, respectively. These models are similar to those described above for the individual data modeling (Section 2.3.2)**.** Then, the scores associated to the 3 PARAFAC factors extracted from EEM's array were concatenated with the 18 Tucker3 scores extracted from CE-DAD's array, forming a fused matrix with 21 variables, which was autoscaled and used for building of a PLS-DA model of 6 LVs.

In all these strategies, PLS-DA models were applied to the fused score-matrices starting from the training-test set split procedure.

2.4.3. High-level data fusion (3-DF)

In 3-DF strategies, the classification of the samples is performed independently on each analytical data block, and then the predictions provided by the models calibrated on the single blocks are combined together [8]**.** In other words, the information in the different data matrices is joined at the level of the prediction obtained by each individual model into a unique solution [33]**.**

In this study, a PLS-DA and N-PLS-DA models were first independently fit for EEM and CE-DAD data matrices (data unfolded and decomposed, and original three-way arrays, respectively) and then the decisions/prediction obtained by each single-block model were fused by two different 3-DF strategies proposed in the literature [34]**:** Majority voting and Bayesian consensus with discrete probability distributions.

On the one hand, Majority voting was carried out by directly merging the predictions of the single PLS-DA or NPLS-DA models (**Fig. 1**). This 3-DF method is based on a democratic (weighted) process that combines the predictions provided by the individual classification models and classifies a sample into a class according to the most frequent class assignment. Within this method, there are three criteria deriving by applying specific limits or thresholds. The 'loose' criterion is the simplest and most intuitive, in which a sample is assigned based on the most frequent class assignment, and a sample is not classified in case of ties (frequency of assignments to a class >50%). The "intermediate" and "strict" majority voting criteria classify a sample if the agreement of predictions is higher than or equal to 75% and 100% (full prediction agreement of all the considered models), respectively [34]. In this study, as only two analytical methods are fused, only a sample is classified into a class when both techniques classify it into the same class, so the criterion used was the 'strict' (100% of frequency assignments).

On the other hand, from the confusion matrix, the Bayesian consensus estimates the probability that a sample belongs to a specific class on the basis of each analytical data block and then combines these probabilities into a joint probability used for the final assignation [34]. As Bayesian results are affected by the model sequence followed in the iteration process, all combinations of analytical sources were considered in our study (i.e. both blocks were selected as the initial block), and according to the classification results, the best order was to start with the EEM's dataset and then with the CE-DAD's dataset.

In a first step, the prior probability has been estimated as equal probability. Considering three classes according to variety and four classes according to origin, the used prior probabilities were 0.33 and 0.25, respectively. Then, likelihood conditional probabilities were estimated from the confusion matrix of each classification model being calculated by dividing the number of classes correct and incorrect predicted by the total of samples of each class. Then, once the posterior probabilities have been calculated for the first analytical block, the fusion approach proceeds iteratively, that is, the posterior probabilities of the first model were used as new prior probabilities in the second model. For that, the class predicted by the first-block classification model (with EEM array, unfolded or decomposed) was initially considered. Then, the posterior probabilities of the first model were used as new prior probabilities in the second model, where the class predicted by the second block model (i.e. those obtained from the CE-DAD array, unfolded or decomposed) was the new evidence.

Finally, this last probability obtained (i.e. the consensus probability derived from the combination of the information of both data blocks) was the one used to predict the class according to the maximum posterior probability obtained. Hence, this last posterior probability was used to accept or reject the predicted class depending on a predefined probability threshold, that in this study was defined as >50%. The corresponding equations and further details of this method can be found in the literature [34, 35].

2.4. Evaluating models

The classification models were internally validated by using venetian blind cross-validation (CV) and the final models' performance was confirmed by a test set validation (TV). For that purpose, as it was mentioned above, the dataset was split into a training and test set, with 61.5% and 38.5% of the samples, respectively, by keeping the ratio of samples of each class like in the original set and the triplicates together.

The quality of the models was assessed from the classification and prediction abilities. The optimal conditions were decided by means of primary measures related to single classes as sensitivity (Sens.), specificity (Spec.) and precision (Prec.) of the calibration and prediction, which were calculated on each class separately encoding different aspects of the classification [34]. This information can be found in the **Supporting Information** as **Table I** and **Table II** for grape variety and geographical classification results, respectively. Additionally, to provide an overall evaluation of the classification quality, the global indices derived from primary class measures such as average sensitivity (non-error rate –NER-) and average precision (PREC.) were also calculated according the recommendation of Ballabio *et al* [34].

2.5. Software

Spectra preprocessing, and low-level, mid-level and high-level DF strategies were carried out by means of hand-made routines written in Microsoft Excel v. 2016 (Microsoft Corporation, USA) and Matlab R2014b (The Mathworks, Natick, MA, USA). Decomposition and compression methods (PARAFAC, Tucker3, PCA) and PLS-DA classification models were calculated with the PLS_Toolbox 7.9.5 (Eigenvector Research Inc., Wenatchee, WA) working under MATLAB environment.

3. Results and discussion

3.1. Data visualization

The fluorescence and CE-DAD landscapes of several samples belonging to the three grape varieties of the four geographical origins are shown in **Fig. 2**. On the one hand, it could be observed that the shape of the EEM spectra varied within the same origin among varieties, as well as within the same grape variety among origins. Thus, the visual assessment of all the fluorescence features of the grape varieties pointed out a general trend for the spectral maxima to be shifted towards 450 and 350 nm of em/ex. Furthermore, similar fluorescence trend was observed for the different origins but maintaining the characteristic shape of its variety.

On the other hand, as can be seen in the CE-DAD landscapes (**Fig. 2**), they showed many overlapping peaks corresponding to the complex mixture of chemical compounds that are present in the white wines. Moreover, as was shown in a previous report [27] a remarkable peak misalignment and shape deformation in electrophoretic mode was produced. It could be also observed some differences between geographical origins and grape varieties. Thus, all the samples showed marked peaks around 3 and 5 min but with strong differences among varieties and origins. From these observations, all these differences could make possible the classification of the samples according to origins and varieties.

Insert here Fig. 2

3.2. Individual models

3.2.1. General considerations

In the first stage, excitation-emission matrices (EEM) and capillary electrophoresis (CE-DAD) data were treated separately to build the classification models. The data matrices were organized and analyzed as two- and three-dimensional arrays. Thus, three different classification approaches considering the data structures and modeling were performed: (a) three-way data by PLS-DA using factors obtained from a resolution method (PARAFAC or Tucker3); (b) three-way data by N-PLS-DA; and (c) full unfolded data using PLS-DA (schematized in **Fig. 1)**.

It should be noted that both datasets differ in the complexity of the data structure. EEMs represents a well-known illustration of bilinear data fulfilling with the so-called low rank trilinearity condition, which can be decomposed into the excitation and the emission spectra for a given fluorescent component [2]. In return, the three-dimensional array built from CE-DAD data is non-trilinear. Moreover, it presents certain drawbacks like remarkable peak misalignment and shape deformation in a data mode associated with deficiency rank in the other one. In those cases, in which multi-way data are involved for classification issues, the choice of the appropriate multi-way approach will be decisive in the validity of the solution found.

For building models, the latent variables were selected considering the lowest CV classification error rate (data not shown). The best preprocess and region (variables), together with the optimal configuration of each model, such as the number of latent variables retained, were selected as those leading to the lowest inaccuracy and highest sensitivity and specificity obtained with the prediction set.

3.2.2. Classification models for EEM data

For the approach (a) of Fig. 1, EEM dataset was arranged in a three-way data array (72 training samples, 41 emission wavelengths and 49 excitation wavelengths) and then, it was analyzed by means of PARAFAC. A three-factor model was chosen representing the best compromise between explained variance (99.5%) and core consistency (71%). The obtained model presented results that were in good agreement with works presented in literature [26] where the loadings for second (emission) and, third (excitation) modes and PCA scores have already been reported.

On the other hand, for the approach (c) of Fig 1, the EEM data array (72 × 49 × 41) was unfolded into a two-dimensional array (72 × 2009). To check the repeatability of the measurements, detect outliers and recognize patterns in the samples' distribution, PCA analyses on PARAFAC factors and unfolded matrix were performed (**Fig. IA.** and **IB. SM**). By analyzing the scores plots, it could not be observed a clear differentiation of the samples by means of geographical origin on both two- and three-way data structures, showing in both cases a similar overlapping, mainly between samples of M and SJ. The same situation was observed when the differentiation among samples was assessed by means of the three grape varieties, as it was shown in a previous report [26].

A PLS-DA model was performed on the PARAFAC factors and the unfolded matrix, which were prior preprocessed by autoscaling and mean centering, respectively. On the other side, a classification model based on NPLS-DA on the three-way data matrix was built (approach (b) of Fig 1). The obtained classification results of these data sets are reported in **Table 1** and **Table 2,** when grape variety and geographical origin were used as classifier as well as in **Fig. II** and **III. SM**, respectively. Then separate models were evaluated comparing the number of latent variables retained and the indices derived from confusion matrix (Sens., Spec., Prec.; and PREC. and NER).

All the built models for sample classification according to grape variety showed a similar performance that the obtained in a previous work [26]. However, in the present study, the best individual model for grape variety classification seemed to be the NPLS-DA model reaching the highest NER and PREC values in prediction stage being 81.1% and 82.1%, respectively (**Table 1**).

Furthermore, suitable classification results were attained according to geographical origin by the three built models. In this case, the model acquired from PLSDA on the PARAFAC factors was the optimal reaching the highest rate of well-classified samples in the prediction set and displaying the highest NER values (**Table 2**).

Insert here Table 1

Insert here Table 2

3.2.3. Classification models for CE-DAD data

For the (a) and (b) approaches (Fig 1), CE-DAD dataset was arranged in a cube structure (72 training samples, 676 times and 107 wavelengths) and it was unfolded in a matrix of size (72 × 72332) for the strategy (c). In order to avoid drawbacks, three-way array was modeled by means of Tucker3 that allows using a different number of factors in each mode. Thus, the number of eigenvalues explaining 95% of the variance of the data were 18, 18 and 6, for modes 1, 2 and 3, respectively. After that, PCA analysis on unfolded matrix and Tucker3 factors were performed. The scatter plots of these PCA analyses are shown in **Fig. IC. SM** and **Fig. ID SM**, respectively. It can be seen that the CE-DAD data showed higher variability than EEM data when the reproducibility was assessed.

In the same way, PLS-DA was performed on the Tucker3 factors and the unfolded matrix, applying autoscaling and mean centering as preprocessing, respectively. On the other side, a classification model based on NPLS-DA on the three-way data array was built (approach b). It is relevant to highlight that both (b) and (c) approaches were able to deal rank deficiency [14].

The three models for grape variety classification showed similar performances; however, for prediction set, the NPLS-DA model attained higher indices of 66.1 and 70.8 for NER and PREC, respectively (**Table 1**). Concerning geographical origin classification, despite the obtained models with the CE-DAD data achieved promising results in the calibration stages, they were not able to predict the samples correctly.

Thus, they showed in all cases NER values lower than 56.9 % (**Tabla 1**). These classification results could be also observed by looking the scores and loadings plots of the PLS-DA models reported in **Fig. II** and **III. SM.**

Thus, the application of fusion approaches was expected to increase the overall classification ability according to variety and origin classification, by integrating these different behaviors of single analytical sources.

3.3. Data fusion models

With the goal of improving the classification results according to geographical origin and grape variety, different strategies were assessed in the three data fusion levels. Thus, in the case of having second-order data, several are the strategies that could be adopted.

3.3.1. Low-level data fusion

PLS-DA was carried out in the first 1-DF option, directly on the concatenated unfolded data matrix (**Fig. 1**). The validation results on the test samples are reported in **Table 1 and Table 2.**

The models based on 1-DF achieved similar classification performances than models from individual blocks when grape variety classification was evaluated. Despite the two 1-DF options were able to correctly discriminate the same number of samples, they did not improve the prediction results respect to NPLS-DA model from EEM data, which obtained the best performance of all individual models. However, the option 1 reached better results in the calibration stage.

On the other side, the results obtained for geographical origin classification showed relevant improvements of both 1-DF approaches in comparison with the individual models, inasmuch as a total of 39 samples pf the training set were correctly predicted. In this case, this strategy seemed to provide significantly better classification results reaching 94.4% and 87.5% for NER and PREC, respectively. It is important to highlight than these results exhibited an increment of more than 5% for NER and a great improvement in PREC of more than 16%, displaying the sterling model ability to avoid wrong predictions in the classes.

As a second option, a PCA can be also applied as a necessary step to compress the information when algorithms as LDA are applied. Thus, it is important to be careful inasmuch as, as in this case, the concatenated data from 1-DF consists on an extremely large matrix where the number of irrelevant variables becomes larger than the really meaningful ones and therefore, the selection of the more relevant variables could result difficult [2, 40-41].

The scores plot of the best 1-DF models for grape variety and geographical origin classifications are shown in **Fig. 3A** and **3C**, respectively. By comparing these models to those for PLS-DA, obtained for the individual data matrices (**Fig. II** and **III SM**), the improvement in the separation of classes by the DF models is clearer.

Insert here Fig. 3

Data fusion showing better discrimination ability than individual spectroscopies have been also reported for fist-order data [36]. Indeed, most of the researches founded in the literature carried out 1-DF on first-order data [10, 37–39] due to the ease of performance together with the satisfactory results. Basically, 1-DF involves the straightforward concatenation by combining variables of the data blocks. Thus, direct first-order data concatenation is easier than second order-data concatenation since, in these last, a prior step, e.g. data unfolding, could be needed when data arrays differ in structure and complexity. In this case, fewer studies could be found in the literature [18].

3.3.2. Mid-level data fusion

As described above, two different approaches for 2-DF were evaluated. Option 1 was based on a first extraction of the relevant features by the development of a PCA model for each unfolded data block, and then, the fusion of the PCA-scores matrices obtained, being this matrix used in the development of the PLS-DA classification models. Within this option, autoscaling or none-preprocessing were tested. The option 2 was based on the feature extraction by PARAFAC and Tucker3 of the EEM and CE-DAD matrices, respectively, and then the fusion of the scores associated to the 3-PARAFAC factors from EEM array and the 18-Tucker3 scores from CE-DAD array, being the matrix used for building the PLS-DA classification models.

On the one hand, with regards to the two 2-DF options, by observing both grape variety (**Table 1**) and geographical origin (**Table 2**) classification results for calibration steps, the option 2 showed better classification results, calibrating correctly almost the total of samples of the training set. On the other hand, by assessing the prediction rate of grape variety classification (**Table 1**), it was again difficult to select the best option due to once again, the two options were able to discriminate correctly the same amount of samples for prediction set, but none of them was able to improve the performance respect to the individual models.

Otherwise, the geographical origin classification rate of the option 2 was better than the option 1 achieving 91.7 and 95.5 for NER and 79.2 and 87.5 for PREC, respectively (**Table 2**). These results can be also observable by looking the scores plots obtained by these PLS-DA models (bottom of **Fig. 3B and Fig. 3D**).

By making a general comparison of these 2-DF results with the individual classification models, they also showed relevant improvements in the prediction rates in comparison to the individual classification results (mainly for geographical origin classification). However, similar classification results were achieved in comparison with the 1-DF approaches, for both grape variety and geographical origin classification. Nevertheless, better results in classification have been reported when 2-DF approach was compared with the analysis of individual datasets or with 1-DF [14, 18, 22, 25, 42].

Despite both low- and mid- levels improved the individual classification results, being similar between them, there are different advantages and disadvantages that could be considered. For both cases, the data block obtained is then processed by the desired chemometric technique. However, on the one hand, 1-DF only implies that the matrices describing the individual blocks, after proper preprocessing, are concatenated to build a single array, being easier even more if the data has a first-order structure. However, a disadvantage of 1-DF is that typically data sets are obtained in which the number of observations is much smaller than the number of variables, which prevents to apply many multivariate data analysis techniques that are not directly applicable. The most popular way of trying to solve the problem of many variables is to reduce the dimensionality of each data matrix separately, before attempting to link them by means of DF and this is how 2-DF works.

Otherwise, there are multiple possibilities that can be applied to carry out a 2-DF strategy. The most remarkable techniques reported in the bibliography for multiway data sets have been sequential and orthogonalized partial least squares (SO-PLS) [44] and coupled matrix and tensor factorization (CMTF) [45]. Other approaches based on multiblock analysis less used but also suitable in data fusion context, is the Common Components and Specific Weights Analysis (CCSWA, also so-called ComDim) [18, 46].

3.3.3. High-level data fusion

In this level of DF, two different approaches were also developed and assessed: Majority Voting and Bayesian consensus methods. These approaches were implemented by using the classification results of the 3 individual models (NPLS-DA, and PLS-DA from the unfolded matrix and the extracted features matrix) of both analytical methods. The classification results obtained by both approaches are shown in **Table 1** and **Table 2** for grape variety and geographical origin classification, respectively.

Concerning the prediction results of Bayesian consensus DF, the model performed using the outputs of the individual NPLS-DA models for grape variety classification provided the best results. Although this model could only match the amount of samples correctly classified in the prediction set with the NPLS-DA individual model from EEM data (NER = 81.1), it was able to improve the calibration stage getting over in more than 15% in both indices NER and PREC.

Otherwise, for geographical origin classification, the Bayesian consensus model developed from the PLS-DA results obtained with PARAFAC and Tucker3 scores provided the best predictions. These results agreed with the classification results obtained by the individual classification models discussed in section 3.2.2 and 3.2.3. Therefore, in general terms, discrimination performances based on Bayesian high-level fusion approaches resulted to be better than those obtained on single analytical sources, as occurs in another work founded in the literature [25]. Hence, this improvement was better observable in the case of geographical origin classification, for which the

Bayesian consensus fusion obtained the best prediction results achieving 97.2% and 91.7% for NER and PREC. This improvement could indicate that the reliability and confidence of the final outcome are increased by the integration of heterogeneous predictions. Moreover, classification performances have been previously reported by means of Bayesian consensus 3- DF fusion achieving slightly better than those obtained in the mid- level approach [25].

However, the classification results obtained by the Majority Voting approach were worse than the results obtained by the PLS-DA models made with the individual data blocks. The main reason of that could be the fact that, considering only two analytical techniques, the criteria applied was the "strict", which means that only the samples that were perfectly classified in both techniques could be classified into a specific class. Hence, in the present study, there were many cases in which one technique classified a sample to one class and the other technique classified the same sample as another class, making that the final decision was "not-classified". For this reason, Majority Voting as 3-DF strategy should be preferably applied when three or more techniques are studied simultaneously.

In comparison to the other two DF strategies (i.e. 1-DF and 2-DF), 3-DF has not the problem of needing to adjust an adequate scaling due to each model is fitted independently with its best scaling. However, a disadvantage of 3-DF is that the order of combining the obtained predictions affects the final decisions.

As a final remark, it is important to highlight the evident improvement of the classification results as level of data fusion increases. In fact, the improvement of the DF prediction models can be linked in the level advance. Despite the combination of multiple analytical sources increases the complexity of data treatment, this is compensated by significantly better classification ability.

4. Conclusions

The proposed multi-level fusion strategies provide a useful and reliable way of improving the analytical quality of the results in second-order data for classification outcomes. The benefit of fusion is highlighted in prediction stage when samples cannot be classified from individual sources. In particular, these advantages were more evident when geographical origin classification was assessed, especially taking in account the complexity of the system presenting unbalanced classes. In addition, multi-level data fusion from multi-via modeling accomplished the best classification models. Thus, it is noteworthy that the benefits of data fusion at different levels are added to the second-order data advantage, furnishing a synergistic effect on the classification results.

Although both techniques provided good classification results separately, data fusion approaches improved the classification results and provided a larger description of the sample. Hence, the statistic mathematical integration of the information from the different analytical sources can be helpful because it leads to the minimization of the overall uncertainty due to a compensation effect among the single experimental uncertainties. This finally translates into increased reliability of the outcome, and, therefore, it can be concluded that high-level strategies are suitable approaches to obtain greater confidence on the combined (fused) analytical predictions. Notwithstanding the practical application seems to be more cumbersome insomuch as first, the independent models for each platform must be fit. However, model outputs combination is easy to implement and analyse, and it does not require higher effort to be performed.

Acknowledgements

The authors are grateful to Univ. Nacional de La Pampa, and CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project PIP-2015 Nº 0111) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Projects PICT 2017-0340 and PICT-2018-04496) for financial support.

References

[1] E. Salvatore, M. Bevilacqua, R. Bro, F. Marini, M. Cocchi, Classification Methods of Multiway Arrays as a Basic Tool for Food PDO Authentication, Compr. Anal. Chem. 60 (2013) 339-382.

[2] S.M. Azcarate, Araújo, A. Muñoz de la Peña, H.C. Goicoechea, Modeling Second-Order Data for Classification Issues : Data Characteristics, Algorithms, Processing Procedures and Applications, TRAC-Trend Anal. Chem. 107 (2018) 151-168.

[3] I. Reinholds, V. Bartkevics, I.C.J. Silvis, S.M. Ruth, S. Esslinger, Analytical Techniques Combined with Chemometrics for Authentication and Determination of Contaminants in Condiments : A Review, J. Food Compos. Anal. 44 (2015) 56-72.

[4] M.P. Callao, I. Ruisánchez, An Overview of Multivariate Qualitative Methods for Food Fraud Detection, Food Control 86 (2018) 283–293.

[5] A.M. Gómez-Caravaca, R.M. Maggio, L.Cerretani, Chemometric Applications to Assess Quality and Critical Parameters of Virgin and Extra-Virgin Olive Oil . A Review, Anal. Chim. Acta 913 (2016) 1-21.

[6] J.A. Arancibia, C.E. Boschetti, A.C. Olivieri, G.M. Escandar, Screening of Oil Samples on the Basis of Excitation-Emission Room-Temperature

Phosphorescence Data and Multiway Chemometric Techniques . Introducing the Second-Order Advantage in a Classification Study, Anal. Chem. 80 (2008) 2789–2798.

[7] A. Biancolillo, R. Boqué, M. Cocchi, F. Marini, Data Fusion Strategies in Food Analysis. Data Handl. Sci. Techn. 31 (2019) 271-310.

[8] E. Borràs, 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.

[9] 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.

[10] S. Schwolow, N. Gerhardt, S. Rohn, P. Weller, Data Fusion of GC-IMS Data and FT-MIR Spectra for the Authentication of Olive Oils and Honeys—Is It Worth to Go the Extra Mile?, Anal. Bioanal. Chem. 411 (2019) 6005–6019.

[11] C. Pizarro, S. Rodríguez-Tecedor, N. Pérez-Del-Notario, I. Esteban-Díez, J.M. González-Sáiz, Classification of Spanish Extra Virgin Olive Oils by Data Fusion of Visible Spectroscopic Fingerprints and Chemical Descriptors, Food Chem. 138 (2013) 915–922.

[12] J. Guillemant, A. Berlioz-Barbier, F. Albrieux, L.P. de Oliveira, M. Lacoue-Nègre, J.F. Joly, L. Duponchel, Low-Level Fusion of FT-ICR MS Data Sets for the Characterization of Nitrogen and Sulfur Compounds in Vacuum Gas Oils Low-Level Fusion of FT-ICR MS Data Sets for the Characterization of Nitrogen and Sulfur Compounds in Vacuum Gas Oils, Anal. Chem. 92 (2020) 2815-2823.

[13] L. Mandrile, L. Barbosa-Pereira, K.M. Sorensen, A.M. Giovannozzi, G. Zeppa, S.B. Engelsen, A.M. Rossi, Authentication of Cocoa Bean Shells by Near- and Mid-Infrared Spectroscopy and Inductively Coupled Plasma-Optical Emission Spectroscopy, Food Chem. 292 (2019) 47–57.

[14] M. Silvestri, A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M. Cocchi, A Mid Level Data Fusion Strategy for the Varietal Classification of Lambrusco PDO Wines, Chemom. Intell. Lab. Syst. 137 (2014) 181–189.

[15] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-Fusion for Multiplatform Characterization of an Italian Craft Beer Aimed at Its Authentication, Anal. Chim. Acta 820 (2014) 23–31.

[16] H.M. Santos, J.P. Coutinho, F.A.C. Amorim, I.P. Lôbo, L.S. Moreira, M.M. Nascimento, R.M. de Jesus, Microwave-Assisted Digestion Using Diluted HNO_3 and H_2O_2 for Macro and Microelements Determination in Guarana Samples by ICP OES, Food Chem. 27 (2019) 159–165.

[17] C.R. Carneiro, C.S. Silva, M.A. De Carvalho, M.F. Pimentel, M. Talhavini, I.T. Weber, Identification of Luminescent Markers for Gunshot Residues: Fluorescence, Raman Spectroscopy, and Chemometrics, Anal. Chem. 91 (2019) 12444–12452.

[18] R. Ríos-Reina, R.M. Callejón, F. Savorani, J.M. Amigo, M. Cocchi, Data Fusion Approaches in Spectroscopic Characterization and Classification of PDO Wine Vinegars, Talanta 198 (2019) 560–572.

[19] A. Bajoub, S. Medina-Rodríguez, M. Gómez-Romero, E.A. Ajal, M.G. Bagur-González, A. Fernández-Gutiérrez, A. Carrasco-Pancorbo, Assessing the Varietal

Origin of Extra-Virgin Olive Oil Using Liquid Chromatography Fingerprints of Phenolic Compound, Data Fusion and Chemometrics, Food Chem. 215 (2017) 245–255.

[20] 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.

[21] 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.

[22] Y. Li, Y. Wang, Synergistic Strategy for the Geographical Traceability of Wild Boletus Tomentipes by Means of Data Fusion Analysis, Microchem. J. 140 (2018) 38–46.

[23] T.G. Doeswijk, A.K. Smilde, J.A. Hageman, J.A. Westerhuis, F.A. Van Eeuwijk, On the Increase of Predictive Performance with High-Level Data Fusion, Anal. Chim. Acta 705 (2011) 41–47.

[24] C. Márquez, M.I. López, I. Ruisánchez, M.P. Callao, FT-Raman and NIR Spectroscopy Data Fusion Strategy for Multivariate Qualitative Analysis of Food Fraud, Talanta 161 (2016) 80–86.

[25] D. Ballabio, E. Robotti, F. Grisoni, F. Quasso, M. Bobba, S. Vercelli, F. Gosetti, G. Calabrese, E. Sangiorgi, M. Orlandi, Chemical Profiling and Multivariate Data Fusion Methods for the Identification of the Botanical Origin of Honey, Food Chem. 266 (2018) 79–89.

[26] S.M. Azcarate, A. Araújo Gomes, M.R. Alcaraz, M.C.U. Araújo, J.M. Camiña, H.C. Goicoechea, Modeling Excitation – Emission Fluorescence Matrices with Pattern Recognition Algorithms for Classification of Argentine White Wines According Grape Variety, Food Chem. 184 (2015) 214–219.

[27] S.M. Azcarate, A. Araújo Gomes, L. Vera-Candioti, M.C.U. Aráujo, J.M. Camiña, H.C. Goicoechea, Second-Order Capillary Electrophoresis Diode Array Detector Data Modeled with the Tucker3 Algorithm: A Novel Strategy for Argentinean White Wine Discrimination Respect to Grape Variety, Electrophoresis 37 (2016) 1902–1908.

[28] R.D. Snee, Validation of Regression Models : Methods and Examples, Technometrics 19 (1977) 415–428.

[29] R. Bro, Multi-Way Analysis in the Food Industry. Models, Algorithms, and Applications, Thesis manuscript, 1998.

[30] R. Bro, H.A.L. Kiers, A New Eficient Method for Determining the Number of Components in PARAFAC Models, J. Chemometrics 17 (2003) 274–286.

[31] D. Ballabio, V. Consonni, Classification Tools in Chemistry. Part 1: Linear Models. PLS-DA, Anal. Methods 5 (2013) 3790–3798.

[32] A.G.E.K. Smilde, Comments on Multilinear PLS, J. Chemom. 11 (1997) 367–377.

[33] A. Smolinska, J. Engel, E. Szymanska, L. Buydens, L. Blanchet, General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences, Data Handl. Sci. Techn. 31 (2019) 51-79.

[34] D. Ballabio, R. Todeschini, V. Consonni, Recent Advances in High-Level Fusion

Methods to Classify Multiple Analytical Chemical Data, Data Handl. Sci. Techn. 31 (2019) 129-155.

[35] A. Fernández, A. Lombardo, R. Rallo, A. Roncaglioni, F. Giralt, E. Benfenati, Quantitative Consensus of Bioaccumulation Models for Integrated Testing Strategies, Environ. Int. 45 (2012) 51–58.

[36] A. Dankowska, W. Kowalewski, Tea Types Classification with Data Fusion of UV–Vis, Synchronous Fluorescence and NIR Spectroscopies and Chemometric Analysis, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 211 (2019) 195–202.

[37] B.P. Geurts, J. Engel, B. Ra, L. Blanchet, A. Suppers, E. Szyma, J.J. Jansen, L.M.C. Buydens, Improving High-Dimensional Data Fusion by Exploiting the Multivariate Advantage, Chemometr. Intell. Lab. 156 (2016) 231–240.

[38] Z. Haddi, H. Alami, N. El Bari, M. Tounsi, H. Barhoumi, A. Maaref, N. Jaffrezic-renault, B. Bouchikhi, Electronic Nose and Tongue Combination for Improved Classification of Moroccan Virgin Olive Oil profiles, Food Res. Int. 54 (2013) 1488–1498.

[39] Q.Q. Wang, H.Y. Huang, Y.Z. Wang, Geographical Authentication of Macrohyporia Cocos by a Data Fusion Method Combining Ultra-Fast Liquid Chromatography and Fourier Transform Infrared Spectroscopy, Molecules 24 (2019) 1320-1338.

[40] A.K. Smilde, M.J. van der Werf; S. Bijlsma, B.J.C. van der Werff-van der Vat, R.H. Jellema, Fusion of Mass Spectrometry-Based Metabolomics Data, Anal. Chem. 77 (2005) 6729–6736.

[41] S. Roussel, J. Roger, P. Grenier, Authenticating White Grape Must Variety with Classification Models Based on Aroma Sensors , FT-IR and UV Spectrometry, J. Food Eng. 60 (2003) 407–419.

[42] M. Silvestri, L. Bertacchini, C. Durante, A. Marchetti, E. Salvatore, M. Cocchi, Application of Data Fusion Techniques to Direct Geographical Traceability Indicators, Anal. Chim. Acta 769 (2013) 1–9.

[43] M. Cocchi, Introduction: Ways and Means to Deal With Data From Multiple Sources, Data Handl. Sci. Techn. 31 (2019) 1-26.

[44] A. Biancolillo, T. Næs, R. Bro, I. Måge, Extension of SO-PLS to multi-way arrays: SO-N-PLS, Chemom. Intell. Lab. Syst. 164 (2017) 113-126.

[45] E. Acar, R. Bro, A.K. Smilde, Data Fusion in Metabolomics Using Coupled Matrix and Tensor Factorizations, P. IEEE 103 (2015) 1602-1620.

[46] M. Hanafi, G. Mazerolles, E. Dufour, E.M. Qannari, Common components and specific weight analysis and multiple co-inertia analysis applied to the coupling of several measurement techniques, J. Chemometrics 20 (2006) 172–183.

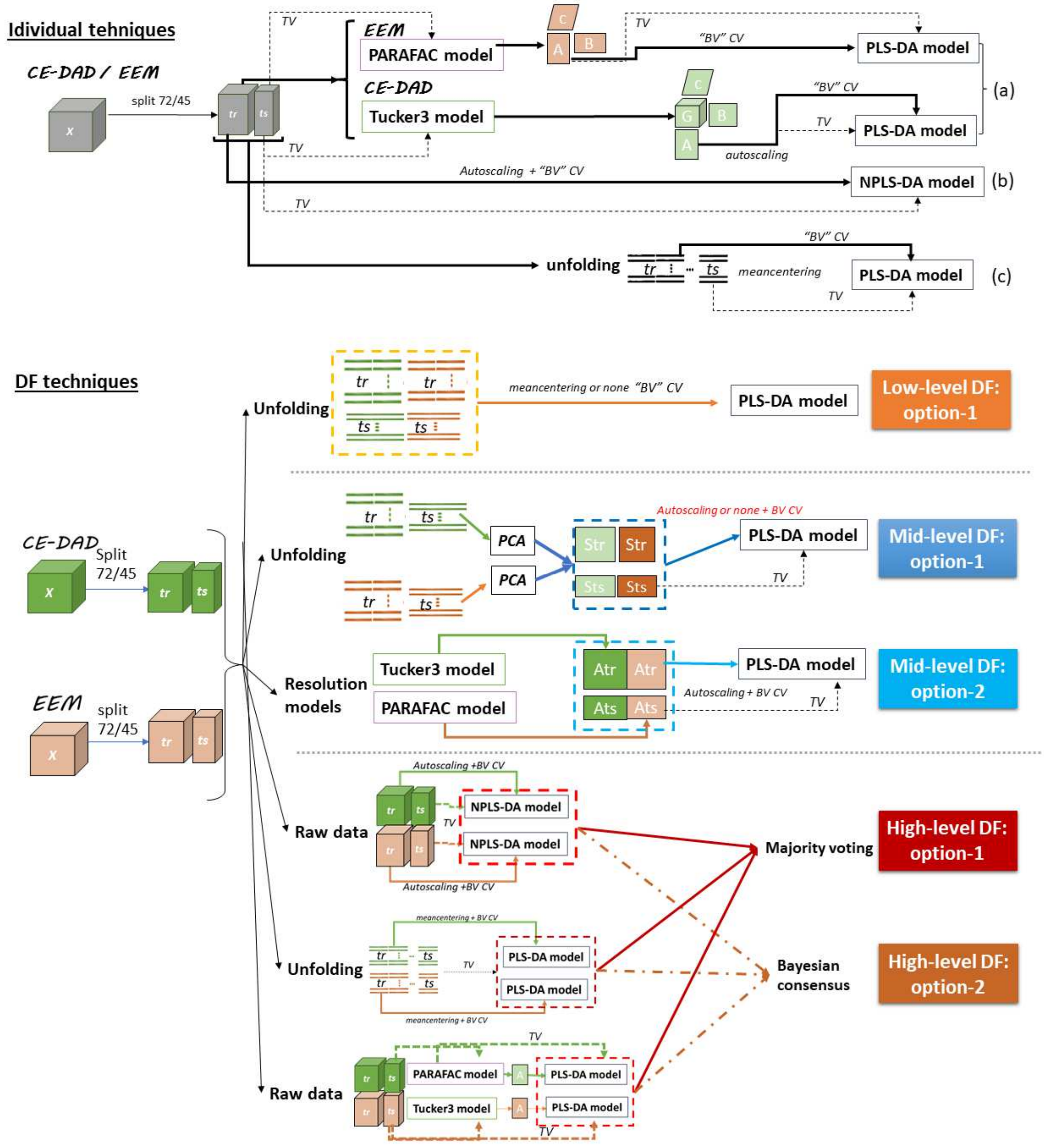
Figure captions

Figure 1. Schematic representation of data analysis workflow

Figure 2. Typical landscapes of (A) EEM and (B) CE-DAD data for a wine sample showing within each geographical origin –Mendoza (M), San Juan (SJ), Río Negro (RN) and Salta (S)- each grape variety -Chardonay (CH) Sauvignon blanc (SB) and Torrontés (T)-

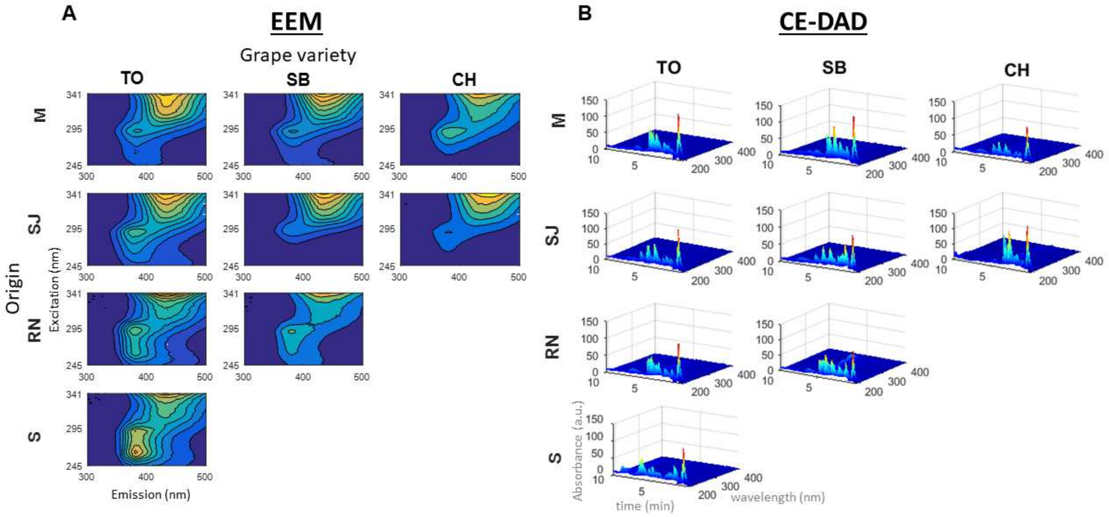
Figure 3. Scores plots for the first three LVs exhibiting the best classification results obtained from (A and C) 1-DF and (B and D) 2-DF models showing the differentiation among wines from (A and B) grape variety and (C and D) geographical origin classifications. 95% confidence ellipses for each class are plotted in 3D in each scores plot.

Figure 1


Idividual tehniques
CE-DAD / EEM
X
split 72/45
tr
ts
TV
EEM
PARAFAC model
CE-DAD
Tucker3 model
TV
C
A
B
TV
"BV" CV
PLS-DA model
C
G
B
A
"BV" CV
TV
PLS-DA model
autoscaling
(a)
Autoscaling + "BV" CV
NPLS-DA model
(b)
TV
unfolding
tr
ts
meancentering
"BV" CV
PLS-DA model
TV
(c)
DF techniques
Unfolding
tr
tr
ts
ts
meancentering or none "BV" CV
PLS-DA model
Low-level DF: option-1
CE-DAD
Split 72/45
X
tr
ts
Unfolding
tr
ts
tr
ts
PCA
PCA
Autoscaling or none + BV CV
Str
Str
Sts
Sts
PLS-DA model
TV
Mid-level DF: option-1
EEM
split 72/45
X
tr
ts
Resolution models
Tucker3 model
PARAFAC model
Atr
Atr
Ats
Ats
PLS-DA model
Autoscaling + BV CV
TV
Mid-level DF: option-2
Raw data
Autoscaling +BV CV
tr
ts
TV
tr
ts
NPLS-DA model
NPLS-DA model
Autoscaling +BV CV
Majority voting
High-level DF: option-1
Unfolding
meancentering + BV CV
tr
ts
tr
ts
TV
PLS-DA model
PLS-DA model
meancentering + BV CV
Bayesian consensus
High-level DF: option-2
Raw data
TV
tr
ts
tr
ts
PARAFAC model
A
PLS-DA model
Tucker3 model
A
PLS-DA model
TV

832 **Figure 2**

833



834

835

836

837

838

839

840

841

842

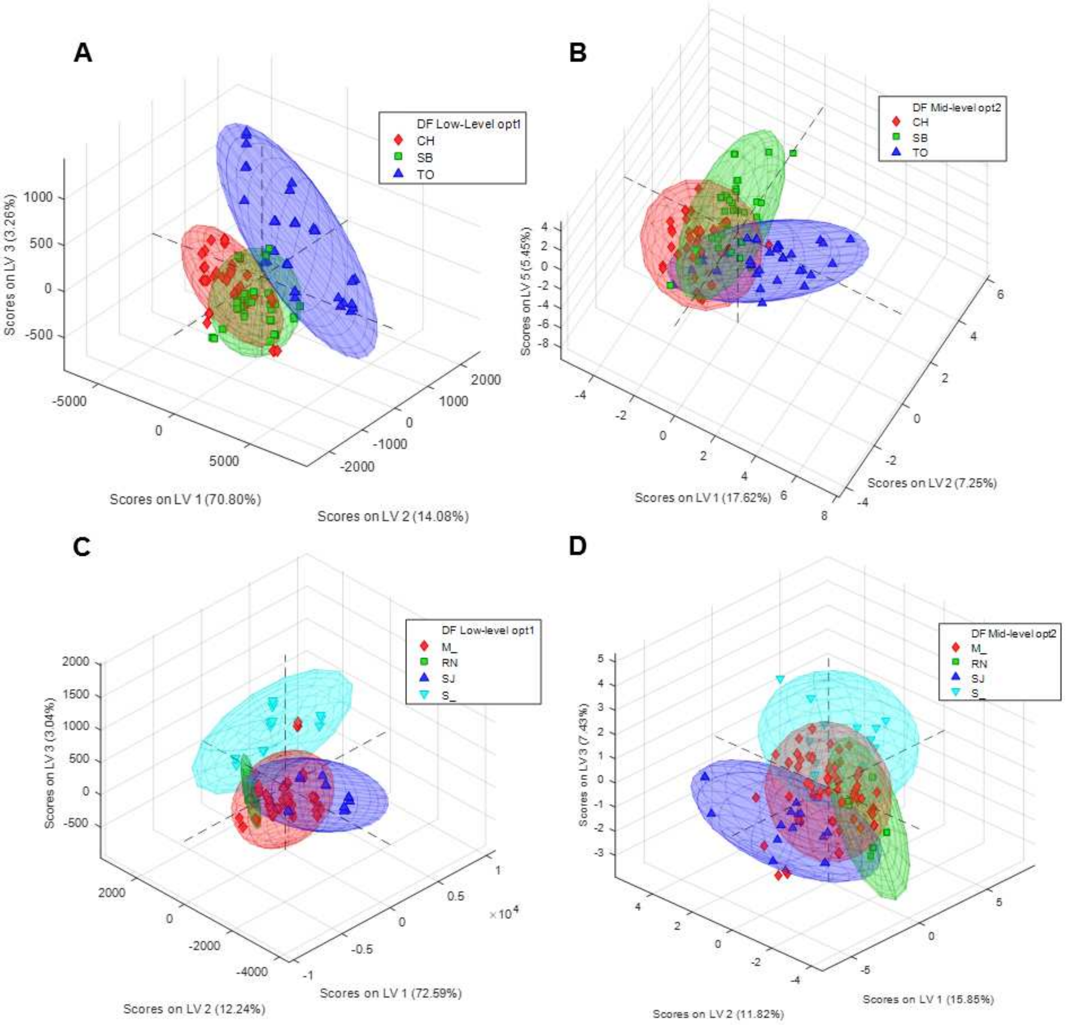
843

844

845

846

Figure 3


A
DF Low-Level opt1
CH
SB
TO
Scores on LV 3 (3.26%)
1000
500
0
-500
-5000
0
5000
Scores on LV 1 (70.80%)
2000
1000
0
-1000
-2000
Scores on LV 2 (14.08%)
B
DF Mid-level opt2
CH
SB
TO
Scores on LV 5 (5.45%)
4
2
0
-2
-4
-6
-8
-4
-2
0
2
4
6
8
Scores on LV 1 (17.62%)
6
4
2
0
-2
-4
Scores on LV 2 (7.25%)
C
DF Low-level opt1
M_
RN
SJ
S_
Scores on LV 3 (3.04%)
2000
1500
1000
500
0
-500
2000
0
-2000
-4000
Scores on LV 2 (12.24%)
1
0.5
0
-0.5
-1
$\times 10^4$
Scores on LV 1 (72.59%)
D
DF Mid-level opt2
M_
RN
SJ
S_
Scores on LV 3 (7.43%)
5
4
3
2
1
0
-1
-2
-3
4
2
0
-2
-4
Scores on LV 2 (11.82%)
5
0
-5
Scores on LV 1 (15.85%)

Tables

Table 1. Classification results according to grape variety (Chardonnay -CH-, Sauvignon blanc –SB-, and Torrontés -TO-) obtained in the calibration and prediction stage from individual data blocks (EEM and CE-DAD) and fused data (EEM-CE-DAD) evaluating 1-DF, 2-DF and 3-DF. For each model it is displayed the number of samples correctly classified, not error rate (NER) and average precision (PREC) for both calibration (CAL) and prediction (PRED) sets of each evaluated model.

GRAPE VARIETY CLASSIFICATION	DATA STRUCTURE		Correct classified samples CAL/PRED			NER (CAL)	NER (PRED)	PREC (CAL)	PREC (PRED)
			CH 24/18	SB 24/15	TO 24/12				
EEM	Unfolded data		24/9	18/9	18//6	83.3	53.3	86.1	64.3
	Three-way data		18/15	15/9	18/12	70.8	81.1	71.0	82.1
	3-factors PARAFAC scores		15/15	15/9	12/6	58.3	64.4	61.9	70.8
CE-DAD	Unfolded data		24/15	24/3	24/9	100.0	59.4	100.0	66.7
	Three-way data		18/15	21/6	24/9	87.5	66.1	87.8	70.8
	Tucker3 scores		24/12	24/6	24/9	100.0	60.6	100.0	66.7
EEM-CE-DAD	LOW-LEVEL	Opt. 1	24/15	24/6	24/9	100.0	66.1	100.0	74.1
	MID-LEVEL	Opt. 1	24/12	18/9	24/6	91.7	58.9	93.3	69.0
		Opt. 2	24/15	24/3	24/9	100.0	59.4	100.0	66.7
	HIGH-LEVEL	Bayesian consensus	24/18	21/0	24/9	95.8	58.3	100.0	58.3
			18/15	21/9	24/12	87.5	81.1	87.8	82.1
			24/15	24/9	24/9	100.0	72.8	100.0	77.1
		Majority voting	24/6	18/0	18/3	83.3	19.4	86.1	40.7
			18/12	12/6	18/9	66.7	60.6	67.2	65.7
			15/9	15/6	12/6	58.3	50.0	61.9	52.2

Table 2. Classification results according to geographical origin (Mendoza –M-, Río Negro –RN- San Juan –SJ-, and Salta –S-) obtained in the calibration and prediction stage from individual data blocks (EEM and CE-DAD) and fused data (EEM-CE-DAD) evaluating 1-DF, 2-DF and 3-DF. For each model it is displayed the number of samples correctly classified, not error rate (NER) and average precision (PREC) for both calibration (CAL) and prediction (PRED) sets of each evaluated model.

GEOGRAPHICAL ORIGIN CLASSIFICATION	DATA STRUCTURE		Correct classified samples CAL/PRED				NER (CAL)	NER (PRED)	PREC (CAL)	PREC (PRED)
			M 45/27	RN 6/6	SJ 12/6	S 9/6				
EEM	Unfolded data		36/18	6/3	12/3	9/3	95.0	54.2	85.0	60.4
	Three-way data		24/15	3/3	9/3	9/3	69.6	51.4	62.0	47.9
	3-factors PARAFAC scores		36/15	6/6	3/6	3/6	59.6	88.9	63.3	70.8
CE-DAD	Unfolded data		45/24	6/0	12/3	9/0	100.0	34.7	100.0	27.9
	Three-way data		33/21	6/0	9/3	9/6	87.1	56.9	77.5	50.8
	Tucker3 scores		33/18	6/3	9/6	6/0	78.8	54.2	66.7	51.7
EEM-CE-DAD	LOW-LEVEL	Opt. 1	39/21	6/6	9/6	9/6	90.4	94.4	85.7	87.5
	MID-LEVEL	Opt. 1	33/18	6/6	6/6	6/6	72.5	91.7	65.5	79.2
		Opt. 2	36/21	6/6	12/6	9/6	95.0	95.5	85.0	87.5
	HIGH-LEVEL	Bayesian consensus	45/21	6/6	12/6	9/3	100.0	81.9	100.0	84.4
			39/15	6/6	12/6	9/6	96.7	88.9	91.7	75.0
			39/24	6/6	12/6	9/6	96.7	97.2	86.7	91.7
		Majority voting	36/18	6/0	12/0	9/0	95.0	16.7	85.0	13.6
			21/12	3/0	6/0	9/3	61.7	23.6	55.9	25.0
			27/12	6/3	3/6	3/0	54.6	48.6	46.3	35.1

Highlights

1) Second-order data were fused and chemometrically processed.

2) Multiple strategies for multi-levels data fusion were evaluated.

3) Straightforward approaches for classification purposes are presented.

4) Different degrees of improvement were observed on the results.

5) High-level strategy provided the best classification results.

Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: