Chemometrics and Intelligent Laboratory Systems 137 (2014) 120–127

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Parallel factor analysis and multivariate curve resolution as data fusion tools to supervise a stream

Alejandro G. García-Reiriz

Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario, S2002LRK, Argentina

ARTICLE INFO

Article history:
Received 20 January 2014
Received in revised form 24 June 2014
Accepted 25 June 2014
Available online 2 July 2014

Keywords:
Chemometrics
Multivariate curve resolution
Parallel factor analysis
Environmental monitoring

ABSTRACT

In this work, a new method is proposed to monitor the distribution, evolution and correlation of dissolved organic matter on the superficial water of a stream with respect to physicochemical variables that characterize the basin and season sampling of each campaign. The method is based on measuring fluorescence emission–excitation matrices and some physicochemical parameters of water samples through both time and space. In a first phase, parallel factor analysis (PARAFAC) or multivariate curve resolution with alternating least-squares (MCR-ALS) were applied to extract the information on the relative proportions of each fluorophore on each sample. Then, MCR-ALS was applied again to the entire database, in order to study the spatial and time distribution. This methodology was used to study the behavior of a basin stream that is significantly modified by anthropic activities.

1. Introduction

Any global monitoring plan should include measurements of concentrations of various analytes of environmental interest for the system under study and some physicochemical variables that characterize it. Dissolved organic matter (DOM) is a simple parameter, indirectly measured by fluorescence excitation–emission matrices (EEM), which provides valuable information of anthropogenic activities in the watershed, because the DOM has different fluorescent properties depending on its origin [1].

The aim of this paper is to propose a new strategy to combine the information that can be extracted by chemometric methods from fluorescence data matrices with specific measurements of physicochemical variables and/or analytes, in order to study their possible relationships and their distribution in space and time. The final objective is to detect possible contamination sources.

The system under study is the Ludueña stream. It is located in the Santa Fe Province of Argentina, in the Rosario Department. Its basin is about 800 km^2. Before its confluence with the Parana River, it flows inside a tube for along 1.5 km. In the higher areas, it has an earthen dam that helps to slow the water runoff during the rainy season, and also contributes to collect water from two channels: the Ibarlucea and the Salvat channels (Fig. 1).

The Ludueña stream watershed is currently in constant modification by human activities. This is because big cities exist in its margins that contribute to sealing large areas of soils; for this reason, its caudal increases dramatically during periods of rainfall. Currently, several private and open neighborhoods are being developed in its vicinity. Also, dense and irregular settlements exist in its margins, generating clandestine channels which provide both stormwater and sewage effluents.

In natural aquatic environments, DOM is composed of a great variety of organic substances, mainly arising from two different origins: 1) autochthonous, stemming from the chemical and biological activity of microorganisms, and 2) allochthonous, due to anthropogenic activities, such as industrial wastewaters or sewage discharges. The first group of compounds comprises some humic-likes substances; their structure and composition allow one to characterize the water quality in which they are dissolved. The second group, in contrast, may be composed of proteic substances, i.e., amino acids arising from dissolved peptides and proteins, when they stem from sewage discharges [2–4]. However, other compounds of allochthonous origin may be present in natural water samples, such as pesticides, hydrocarbons, or the so-called emergent contaminants (human or veterinary pharmaceuticals), etc.

Fluorescence spectroscopy has allowed to characterize DOM in samples of different origins, to monitor the contamination level of polluted areas, and to distinguish anthropically impacted regions from less affected ones [5–7]. Natural waters usually contain a mixture of fluorophors which makes their identification difficult by means of unidimensional fluorescence spectra [1]. An excellent analytical alternative is to measure fluorescence EEMs, which allow one to obtain much richer information related to the presence and type of dissolved fluorophors. EEMs began to be studied in the decade of 1990, with the distinction of humic and non-humic-like compounds in natural waters [1,8,9].

In order to extract the chemical component information from the registered EEM data, several methodologies can be applied. One of the most popular ones is to build a three-way signal array from the recorded

E-mail address: garciareiriz@iquir-conicet.gov.ar.

http://dx.doi.org/10.1016/j.chemolab.2014.06.016
0169-7439/

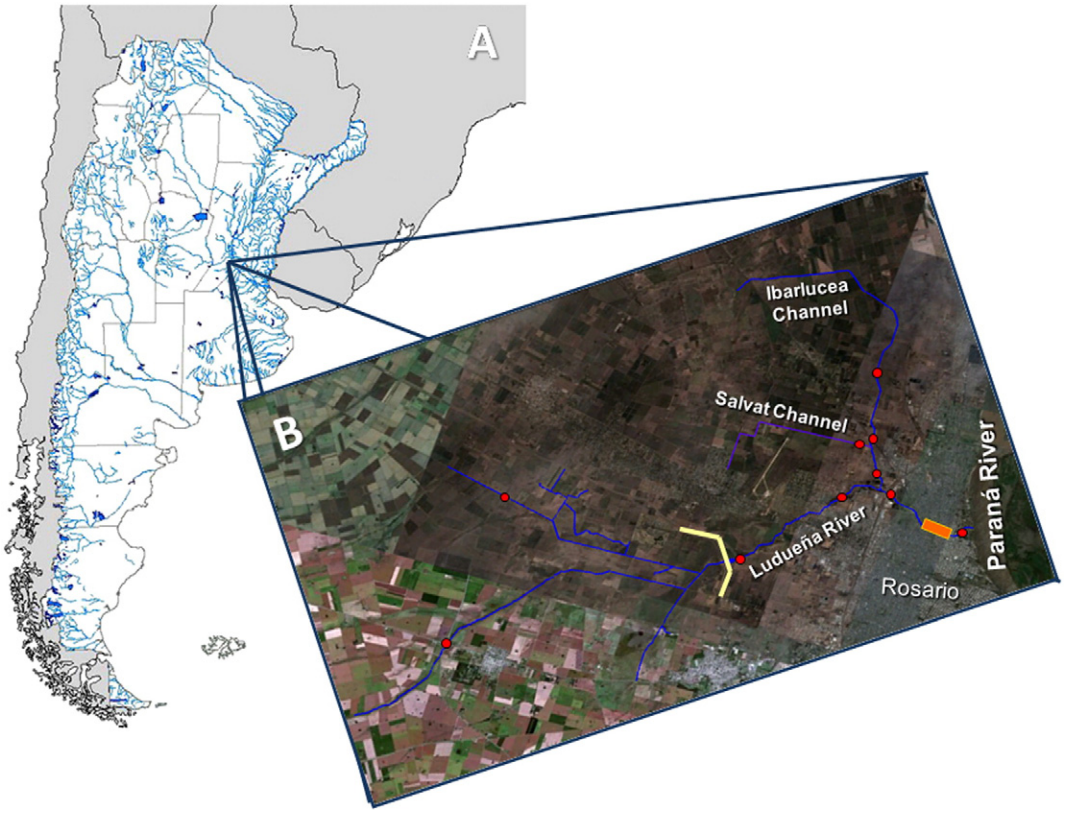

Fig. 1. Location of the study area; **A**, in Argentina; and **B**, inside Rosario Department.

EEMs, followed by parallel factor analysis (PARAFAC) [10]. This algorithm allowed to identify humic-like and proteic-like substances in water samples [11], to characterize the DOM present in lakes and soils [12], to identify anthropogenic contaminants and metal traces in waters [13], to study the discharge of effluents into rivers and marine waters associated with ranges of salinity and nutrients [14], to detect fulvic acids and tryptophan-rich proteins in sewage discharges [15], and to classify water samples based only on the content of humic acids [16]. A related algorithm, multivariate curve resolution coupled to alternating least-squares (MCR-ALS) was also employed for similar purposes [17]. Therefore, both methods were applied to process the EEMs and to compare their results. MCR-ALS works similarly to PARAFAC when it is applied with the trilinearity constraint. However, instead of working with a three-way array, an augmented data matrix is built by placing each EEM below the other one (see specific details below).

To study the distribution of each fluorophore obtained by PARAFAC together with other variables in time and space, a new MCR-ALS approach can be applied [18]. This allows to group the variables with similar behavior and provides spatial and time distribution. In this way, it was possible to make an interpretation about their origin and interrelation.

MCR-ALS is a powerful chemometric tool with an increasing application for the analysis of environmental monitoring data sets [19,20]. Other chemometric methods have also been applied to the investigation of environmental data, such as partial least-squares (PLS) [21,22], PARAFAC and Tucker3 models [23]. The use of multivariate factor analysis, such as those proposed in the present work, has also been discussed in several books [24,25]. Additionally, other recent examples exist proposing similar approaches for the resolution and interpretation of major contamination sources of surface waters operating in several river basins over the world [26].

Thus there are three main objectives in this work: 1) the possibility of combining data of different complexity to obtain a better characterization of the system, 2) the investigation of main long-term diffuse contamination sources of organic contaminants in the Ludueña stream basin area, and 3) the estimation of their geographical distribution, in order to contribute to the evaluation of the environmental health of the surface waters of the region under study. To achieve these three goals, multivariate data methods of analysis based on combinations of PARAFAC and MCR-ALS were applied.

In order to get useful environmental information from the data, the application of modern chemometric methods based in multivariate factor analysis tools is proposed [27]. The basic assumption of these methods when they are applied to environmental data tables is that each value of a measured variable in a particular sample is due to the sum of contributions from individual independent sources of different origin. Each of these sources is characterized by a particular chemical composition profile and is distributed among samples in a different way. As a result of the application of chemometric methods, the main point and diffuse sources of contamination in the environment and their origin may be identified and their distribution profiles among samples (geographical, temporal, among environmental compartments) are characterized.

2. Material and methods

2.1. Equipment

Fluorescence spectral measurements were performed on a fast-scanning Varian Cary Eclipse fluorescence spectrophotometer, equipped with two Czerny–Turner monochromators and a xenon flash lamp, and connected to a PC microcomputer via an IEEE 488 (GPIB) serial interface. Excitation–emission data arrays were recorded in a 10 mm quartz cell, in the following ranges: excitation, 220–481 nm each 3 nm; emission, 280–600 nm each 5 nm. Thus, the size of each data matrix was $88 \times 65 = 5720$ data points. The wavelength scanning speed used was 12,000 nm/min. The detector voltage was fixed at 700 V.

2.2. Sample collection

Ten sampling points of the stream were selected to represent the different branches and thus they represent the overall state of the stream according to the activities in its vicinity (Fig. 1). At least five days elapsed since the last rain were taken into account that before each campaign, to ensure that conditions were reproducible as much as possible. The samples were collected approximately every 45 days, in the period between September 2010 to July 2011 (five campaigns). The parameters pH and conductivity were directly measured. In order to select an appropriate dilution to avoid the phenomenon of inner filter on fluorescence measurements, the samples were filtered through a cellulose acetate filter of 22 μm pore and UV-visible spectra were measured. The absorption spectra were measured between 220 and 485 nm. An appropriate dilution was selected for each sample based on the absorption maximum being smaller than 0.01 AU. Finally, the samples were stored at 8 °C. All measurements were performed within 24 h of collection.

3. Theory

3.1. EEM pre-processing

Rayleigh and second-order harmonic signals are not bilinear, i.e., they cannot be described in terms of the combination of single excitation and emission spectral profiles, and should be removed before successful data processing. In order to remove these unwanted contributions from each of the registered EEMs, several procedures have been described: 1) digital subtraction of the EEM for blank samples [28], 2) digital removal of the spectral regions where the dispersion signals appear, 3) replacement of the dispersion values by missing values and PARAFAC modeling the resulting data array [29,30], 4) use of a so-called weighted PARAFAC model [31,32], and 5) non-linear interpolation [33]. The latter method was preferred, which involves the following steps: 1) at each excitation wavelength, the wavelength range at which the dispersion signal appeared is located (either for the Rayleigh or second-order harmonic), 2) the fluorescence signal is removed at these wavelengths, and replaced by polynomic interpolating values, using as reference for estimating the polynomic constants the points before and after the removed window. A built-in MATLAB code (The Mathworks, Natick, Massachussets, USA, 2007) was employed for this purpose, selecting the cubic spline option for interpolation [34].

3.2. PARAFAC

After measuring second-order EEM data for a set of samples, each of them as a $J \times K$ matrix (J is the number of data points in the excitation mode and K the number of data points in the emission mode), the I matrices $\boldsymbol{X}_i$ are joined into a three-way data array $\boldsymbol{X}$, whose dimensions are $I \times J \times K$. Provided $\boldsymbol{X}$ follows a trilinear PARAFAC model, it can be written in terms of three vectors for each responsive component, designated as $\mathbf{a}_n$, $\mathbf{b}_n$ and $\mathbf{c}_n$, and collecting the relative concentrations or scores ($I \times 1$) for component n, and the profiles in both modes ($J \times 1$) and ($K \times 1$) respectively. The specific expression for a given element of $\boldsymbol{X}$ is [35]:

$$\boldsymbol{X}_{ijk} = \sum_{i=1}^{N} a_{in} b_{jn} c_{kn} + \boldsymbol{E}_{ijk} \tag{1}$$

where N is the total number of responsive components or fluorophores, a_{in} is the relative concentration of component n in the i^{th} sample, and b_{jn} and c_{kn} are the intensities at channels j and k, respectively. The values of E_{ijk} are the elements of the array $\boldsymbol{E}$, which is a residual error term of the same dimensions as $\boldsymbol{X}$. The column vectors $\mathbf{a}_n$, $\mathbf{b}_n$ and $\mathbf{c}_n$ are collected into the corresponding score matrix $\mathbf{A}$ and loading matrices $\mathbf{B}$ and $\mathbf{C}$ ($\mathbf{b}_n$ and $\mathbf{c}_n$ are usually normalized to unit length).

The model described by Eq. (1) defines a decomposition of $\boldsymbol{X}$ which provides access to profiles in both data modes ($\mathbf{B}$ and $\mathbf{C}$) and relative concentrations ($\mathbf{A}$) of individual components in the I mixtures, whether they are chemically known or not. The decomposition is usually accomplished through an alternating least-squares minimization scheme [10,36].

Issues relevant to the application of the PARAFAC model to three-way data are: 1) initializing the algorithm, 2) establishing the number of responsive components, 3) constraints of the least-squares fit in order to obtain physically interpretable profiles, 4) identifying specific components from the information provided by the model and, for our case, 5) employing the scores for sample classification.

Initializing PARAFAC for the study of three-way arrays can be done using: 1) loadings provided by the direct trilinear decomposition (DTLD) [37], 2) spectral profiles which are known in advance for pure components, or 3) loadings giving the best fit after small PARAFAC runs involving both DTLD and several sets of random loadings. These options are all implemented in Bro's PARAFAC package [38].

Several constraints are available in order to be imposed during the alternating least-squares PARAFAC fitting. They may serve different purposes, for example to retrieve physically recognizable component profiles. Non-negativity constraint in all three modes serves this purpose, allowing the fit to converge to the minimum with physical meaning from the several minima which may exist for linearly dependent systems.

The number of responsive components (N) can be estimated by several methods. A useful technique is CORCONDIA, a diagnostic tool considering the PARAFAC internal parameter known as core consistency [37,39]. The core consistency analysis involves the study of the structural model based on the data and the estimated parameters of gradually augmented models. A model is considered to be appropriate if adding other combinations of components does not improve the fit considerably, i.e., when the core consistency parameter drops from a value of ca. 50. Another useful technique is the consideration of the PARAFAC residual error, i.e., the standard deviation of the elements of the array $\mathbf{E}$ in Eq. (1) [10]. Usually this parameter decreases with increasing N, until it stabilizes at a value compatible with the instrumental noise (the latter can be assessed by blank replicate measurements). A reasonable choice for N is thus the smallest number of components for which the residual error is not statistically different than the instrumental noise. Still another possibility is split-half analysis [40] which involves the consideration of the profiles retrieved when the data set is randomly divided in two sub-sets and decomposed using an increasing number of PARAFAC components. This latter method, however, is preferred when the sample composition is homogeneous; it may not be the best choice in the present case, where some of the samples may be unique or contain chemical constituents which are absent in the remaining ones.

Identification of the chemical constituents under investigation is done with the aid of the estimated profiles, and comparing them with those for the known pure components, provided they are available in pure form or from the literature. This is required since the components obtained by decomposition of $\boldsymbol{X}$ are sorted according to their contribution to the overall spectral variance, and this order is not necessarily maintained when the unknown sample is changed.

3.3. MCR-ALS

Multivariate techniques are methods of analysis generally recognized as very useful tools to study environmental problems. From these methods, MCR-ALS [41,42] has been selected as one of the most advantageous to study our system. The merged and scaled data matrix $\mathbf{D}$ allows MCR to analyze the data in space and time modes. MCR-ALS is a bilinear based method which can be basically described as a matrix decomposition:

$$\mathbf{D} = \mathbf{S}\mathbf{L}^{\mathbf{T}} + \mathbf{E} \tag{2}$$

where **D** is the data matrix, **S** is the scores matrix related to the objects and **L** is the loadings matrix related to the variables. Every vector of **S** is associated with a vector of **L** through a product that represents a component. It is supposed that each component represents a kind of source (or combination of similar sources) which contributes to the overall state of the system. The bilinear model in Eq. (2) assumes that the major sources of the experimental data variance can be explained by a small number of components defining the two reduced-size factor matrices (scores and loadings). The model described by this equation assumes that the variables (or measured concentrations of contaminants) in a particular sample are the sum of a reduced number of contributions of this contaminant coming from different sources. It is therefore a mixture analysis problem with unknown sources which have to be estimated from the analysis. Since the solution of Eq. (2) is ambiguous, the matrix decomposition in this equation has to be performed under some constraints. The decomposition of Eq. (2) is similar to Principal Components Analysis (PCA), but PCA decomposition is performed under orthogonal constraints, loadings normalization and maximum explained variance for the successive extracted components. Under these constraints, PCA provides unique solutions. However, these solutions are an abstract linear combination of the true experimental variance sources and, although they are very useful for data exploration and summary, in many cases they can be too complicated in terms of environmental interpretation. Although there are many good textbooks about PCA, we refer the interested reader to Jolliffe [43]. Unlike PCA, the matrix bilinear decomposition performed by MCR-ALS uses softer natural constraints and as a result, the interpretation of loading and score profiles is easier and more reasonable from an environmental point of view [1,20]. Constraints used in this work during the MCR-ALS bilinear matrix decomposition were non-negativity and normalization of loadings to equal length as those used in previous works [20].

MCR-ALS was applied in this work in two different ways. In a first step, MCR-ALS (with trilinearity constraint) and PARAFAC were both used as tools that provide information to be merged with other data into a single global matrix. This is because PARAFAC and MCR-ALS allow to summarize the specific fluorescence matrix information into lower-order data suitable for data fusion. In the last stage, MCR-ALS is applied without the trilinearity constrain to this overall data set. Fig. 2 shows a diagram of the overall data flow.

3.4. Software

All calculations were made using MATLAB 7.0 (The Mathworks, Natick, Massachussets, USA, 2007). For the removal of the dispersion signals, the routine described by Zepp was employed [33]. PARAFAC was implemented using the MATLAB routines provided by Bro in the webpage [38]. In order to apply MCR-ALS, the codes available on internet were implemented [18,44]. To make the spatial representation ArcGis was employed to georeference the data and overlap them with the images (ESRI Headquarters, New York, USA).

4. Results and discussion

4.1. PARAFAC and MCR-ALS fluorescence decomposition

In order to analyze the information globally, the first step is to estimate the proportion of each fluorophore at each site and in each campaign from the EEM data. For this purpose, the data can be conveniently processed with PARAFAC or MCR-ALS [10]. Prior to these analyses, individual fluorescence matrices were corrected for the Rayleigh signal using method of Zepp [33], since the latter does not respond to the trilinear decomposition. Once this signal was corrected, all EEMs from all campaigns were staked one above the other into a single three-way data array, thereby forming a three-way arrangement of size 88 × 65 × 50 (excitation × emission × sampling_site-campaign). Subsequently, the three-way array was analyzed by PARAFAC. On the other hand, for the application of MCR-ALS, all EEMs were appended one behind the other one, building a so-called augmented matrix of size 88 × 3250 (excitation × emission-sampling_site-campaign). This means that the matrix was augmented in the emission mode. It also was proved to work with the augmented matrix in the excitation mode and similar results were obtained because the samples have a similar level of spectral overlap in both modes.

Humic acids are not a single compound, but a complex mixture of structurally related components. Specific fluorescence information on humic-like fluorophors shows that they have a fluorescence emission maximum concentrated in the range 420–450 nm, but excitation spectral profiles distributed between two differentiated regions: 230–260 nm (humic-like A), and 320–350 nm (humic-like C). As regards the proteic substances, they can be distinguished in B fluorophors (tyrosine-rich amino acids), with excitation in the range 225–237 nm and emission at 309–321 nm, and T fluorophors (tryptophan-rich amino acids), with excitation in the range 225–237 nm and emission at 340–381 nm [45].

The decomposition of the fluorescence signals from natural samples containing a variety of responsive components having overlapped spectra constitutes a difficult task. In this sense, the recording of excitation–emission fluorescence matrices provides a wealth of information, which

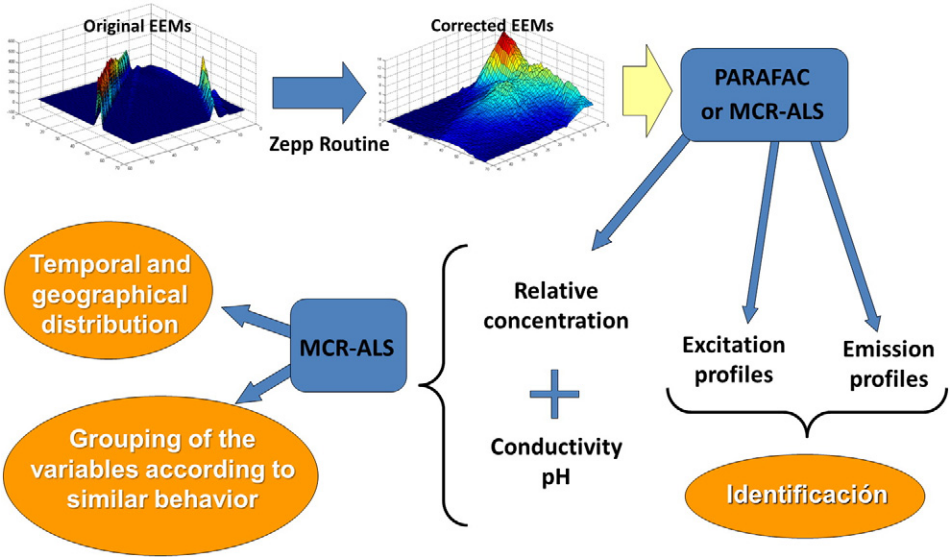

Fig. 2. Schematic of the overall data processing.

however has to be adequately processed by suitable algorithms in order to reach a successful deconvolution into the individual contribution of the several sample components.

The most important steps to obtain success in the PARAFAC analysis are the initialization strategy, the setting of the number of responsive components, and the application of constraints during the least-squares fitting phase. The best initialization method was found to be the best run of a series of small runs starting from DTLD values or from randomly chosen numbers. Non-negativity constraint was applied during the PARAFAC least-squares fitting phase. Both of these options are available in the PARAFAC package available on the internet, and are easily applicable.

As for PARAFAC, constraints and initialization options are very important in MCR-ALS analysis. For this kind of data, the best choice was to apply non-negativity on all modes and trilinerization of the decomposition (because MCR-ALS is a bilinear model). For initialization, spectral estimates obtained from the measured data in the 'purest' wavelengths were employed, using the procedure described in ref. [46].

In relation to the number of PARAFAC components, values of residual standard deviation were computed as a function of increasing number of PARAFAC components. The residual fit decreases significantly in going from the first to the fifth component, and then continues to decrease, but more slowly (1.38, 1.09, 0.80, 0.69, 0.57, 0.49 and 0.47 arbitrary fluorescence units or AFU are the respective residuals fit of PARAFAC models from one to seven components). The CORCONDIA test also was applied (100, 97.0, 84.2, 66.8 and 52.9 were computed from one to five components). Thus, a five-component PARAFAC model was selected. On the other hand, the selection of the number of MCR-ALS components was made by singular value decomposition (SVD) [47]. The corresponding SVD plots for the fluorescence data allowed to select five components for the analysis, because selecting more than five did not significantly change the singular values (3309, 384, 273, 189, 111, 71, 54, 48, 42 and 38 were computed as singular values in arbitrary units, from 1 to 10 components respectively). This analysis agrees with the selection of PARAFAC components, and it confirms that the samples contain mainly five different fluorophores.

From these analyses it was possible to obtain the proportion of each fluorophore in each sample, and also their specific excitation and emission profiles (Fig. 3). Their identification was possible through bibliographic data [1].

The profiles for PARAFAC component 1 (equivalent to MCR-ALS component 4) corresponds to the spectra of humic-like substances of type A (excitation at 237–260 nm and emission at 400–500 nm, see Fig. 3), and its presence is associated with organic matter generated within the stream by decomposition of organic matter. It is a characteristic of environments that are not impacted, and it is the major proportion of DOM humic substance.

On the other hand, the profiles for PARAFAC component 2 (equivalent to MCR-ALS component 1) agree with the spectral characteristics of humic-like C fluorophores (excitation 300–370 nm and emission at 400–500 nm, see Fig. 3). These are characteristic of low impacted environments, because they are humic-like substances of allochthonous origin. They are typical of waters that have been in contact with mud used during the treatment of wastewater effluent, and they are associated with the presence of organic material of allochthonous origin, coming from the soil in contact with the edges of the water channel, which is swept away by the rains. Humic-like C substances are not generated within the body of water, but are carried by natural factors.

PARAFAC component 3 (equivalent to MCR-ALS component 5) can be associated with T fluorophores; these are characteristic of environments with high anthropogenic impact. Because they have a high protein fraction corresponding to amino acid tryptophan (225–237/275 excitation and emission at 340–381 nm, see Fig. 3). They are protein substances or non-humic specifically tryptophan-rich proteins [1]. They are of anthropogenic origin and their presence in natural waters is associated with organic matter from industrial effluents and/or

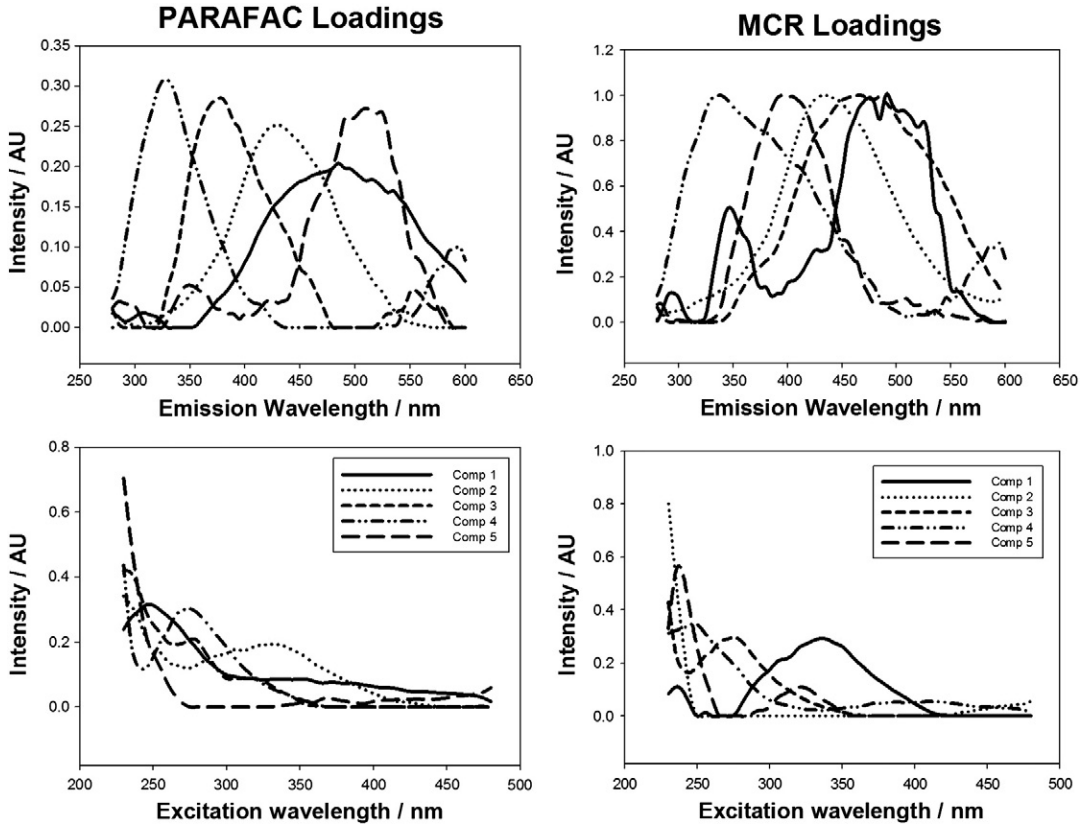

Fig. 3. Profiles of emission and excitation of fluorescence obtained by PARAFAC at left and by MCR-ALS with trilinearity at right.

untreated sewage. The presence of the amino acid tryptophan is related to anthropogenic influence in the waters of the bays, estuaries and coastal areas with high productivity, bacterial activity or effluent discharges.

PARAFAC component 4 (equivalent to MCR-ALS component 3) has the spectral characteristics of the B flurorophores. These are also characteristic of environments with high anthropogenic impact, with the difference that the largest protein fraction corresponds to the amino acid tyrosine (225–237/275 excitation and emission at 309–321 nm, see Fig. 3).

Finally, PARAFAC component number 5 (equivalent to MCR-ALS component 2), can be tentatively associated with fluorescent whitening agents and/or surfactants (260/430 nm 260/540 nm and 400/460 nm are the respective peak excitation/emission for this class of compounds) [48]. In Table 1 the correspondence between the PARAFAC components and MCR-ALS components and their respective fluorophore assignment are summarized.

Once fluorophores corresponding to each PARAFAC or MCR-ALS scores were identified, they were ordered to simplify comparison and they were combined into a single data matrix, along with conductivity and pH values, to be analyzed again by MCR-ALS [44], as if each PARAFAC or MCR-ALS component was a single measured variable. In order to analyze all variables together and to avoid that some variables have more importance than others because of their different scales, an appropriate data scaling was applied (see next section).

4.2. Scaling and data fusion

Once the EEMs were analyzed by PARAFAC or MCR-ALS, the proportions of each fluorophore in each sample were obtained. This information is contained within the **A** matrix of PARAFAC scores or in the **S** matrix of MCR-ALS scores. The rows of these matrices correspond to each sampling site in each campaign, and the columns to each particular fluorophore (50×5). First each row was corrected by its sampling dilution factor. In order to add the information from the other variables measured, extra columns were added. In our case, only two columns were required, so the merged matrix now has size 50×7. Once all data of fluorophores, variables and/or analytes were collected into a single array, the scale was corrected, because the distributions of the variables are not statistically normal, and the different variables have very different scales, artificially giving greater importance to larger scales.

The best scaling for this kind of data was MinMax of the logarithm of data transformation. The specific expression for the MinMax transformation is:

$$x_{\text{transf}} = \frac{x - \min(\log(\mathbf{x}))}{\max(\log(\mathbf{x})) - \min(\log(\mathbf{x}))} \quad (1)$$

where $\mathbf{x}$ is a vector with the values of one variable, $\max(\log(\mathbf{x}))$ and $\min(\log(\mathbf{x}))$ are the maximum and minimum of logarithmic transformation of $\mathbf{x}$ respectively, and x and x_{transf} are the raw and transformed elements. The matrix obtained after MinMax pre-processing is the **D** augmented matrix for the MCR-ALS model, since it has the information for each campaign one below the other.

Table 1
Correspondence between PARAFAC components and MCR-ALS components and their fluorophore assignment.

PARAFAC component	MCR-ALS component	fluorophores assignment
1	4	humic-like type A
2	1	humic-like type C
3	5	type T
4	3	type B
5	2	whitening agents and/or surfactants

4.3. MCR-ALS global results

MCR-ALS analysis was applied to study the spatial and temporal distribution of all variables together [41]. This is based on the hypothesis that variables that have the same behavior or origin will be joined within the same 'group' by MCR-ALS, obtaining information on the composition of potential pollution sources, punctual or diffuse, and the geographical and temporal distribution. Fig. 2 shows a diagram summarizing the overall data flow.

To summarize the results of this section, the data here reported only correspond to the global MCR-ALS analysis with PARAFAC scores (PARAFAC/MCR-ALS), because similar results were obtained by analyzing the MCR-ALS scores (MCR-ALS/MCR-ALS). The results of applying MCR-ALS/MCR-ALS are only shown in Figs. 3 and 4. It can be corroborated that these are very similar to PARAFAC/MCR-ALS results, but in a different order. For this reason, Table 1 shows the correspondence of PARAFAC components with MCR-ALS components. So, before MCR-ALS global analysis the columns of scores obtained by PARAFAC or MCR-ALS were ordered in the same way to obtain comparable results.

MCR-ALS allowed grouping the different fluorophores together with the measured variables within MCR-ALS groups, according to their location, origin and time evolution. One important step in MCR-ALS modeling, as before, is the choice of the number of components or groups and constrains. SVD [47] plots for the scaled data allowed to select four components for the analysis because more than four components do not significantly affect the singular value (5.82, 2.24, 1.38, 1.25, 0.78 and 0.75; values in arbitrary units from 1 to 6 components respectively). Similar results were obtained with MCR-ALS scores of fluorescence data.

In the case of this latter MCR-ALS analysis, only non-negativity constraint was required, in order to obtain a simpler interpretation of the results. The explained variance was 98.2%. Fig. 4 shows how the different variables were grouped into MCR-ALS scores (for both initial options, i.e., either PARAFAC or MCR-ALS) and Fig. 5 shows the spatial distribution of MCR-ALS scores from an average of all measurement campaigns conducted for each group of variables, since the sampling campaigns were performed during a single year (only for PARAFAC results). Therefore, no conclusions can be drawn regarding their behavior over time, because the information is insufficient to be able to study this mode.

Figs. 4 and 5 show what variables correspond to each MCR-ALS group, and where they are located within the watershed of the stream. Humic-like A and C substances, and in a lesser proportion the T fluorophores, were grouped together in a first group (60.6% variance explained). A normal behavior can be seen along the stream, due to drag of rain and of the own river.

The second group (23.0% variance explained) explains the distribution of the pH variable. It can be seen that there are alkalinized water areas in the sampling points above the basin.

The third group (8.5% variance explained) justifies the behavior of the variable conductivity. An area with a very high conductivity in the middle of the stream can be observed, due to the deposition of salts observed in the soil adjacent to the tract, because these zones were highly floodable areas before the construction of the retarding dam.

The last group (6.1% variance explained) mainly models fluorophores B and T, and also whitening agents (or surfactants) being found principally in the Ibarlucea channel. Significant amounts of xenobiotics, compounds of anthropogenic origin and different types of surfactants and/or detergents were observed. This area is characterized by irregular settlements, making the existence of channels where clandestine sewage and gray water are thrown without any processing highly likely.

These results are consistent by observing the origin of each variable in the groups, since MCR-ALS collects the B and T fluorophores with whitening agents (which are of anthropogenic origin), and on the other hand the humic-like A and C (which are natural), leaving the pH and conductivity separated, since their distributions are not related to the behavior of other variables.

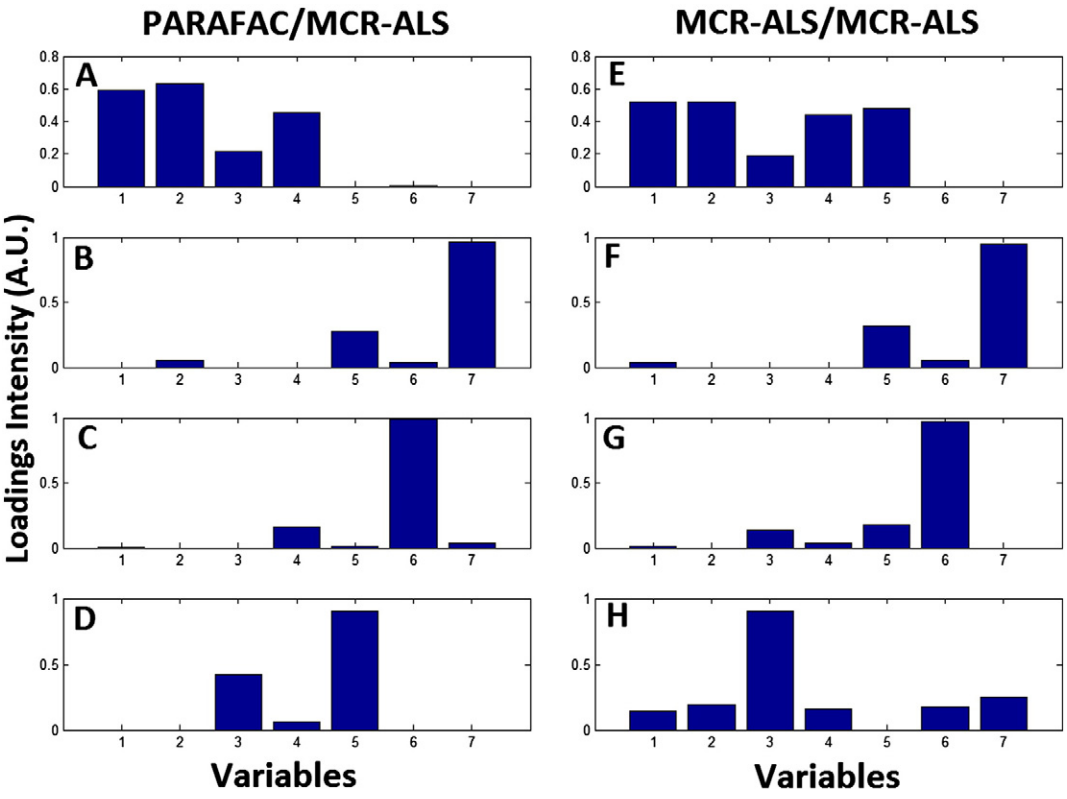

Fig. 4. Grouping of different variables formed within MCR-ALS scores. Variables: **1,** fluorophore humic-like type A; **2,** fluorophore humic-like type C; **3**, fluorophore type T; **4,** fluorophore type B; **5**, whitening agents and/or surfactants; **6,** Conductivity; and **7**, pH. **A**, **B**, **C** and **D** are the several PARAFAC/MCR-ALS groups; and **E**, **F**, **G** and **H** are the several MCR-ALS/MCR-ALS groups.

Fig. 5 was made in ArcGis (ESRI Headquarters, New York, USA). First, a satellite image was loaded and it was georeferenced with known control points with GPS coordinates. Then, four raster layers were created with the MCR-ALS scores (one for each group). These layers were overlaid with the satellite image for its georeferencing.

5. Conclusions

Two chemometric algorithms, PARAFAC and MCR-ALS, were successfully combined to model information of different complexity. The ability to fuse information of different modes provides a more complete

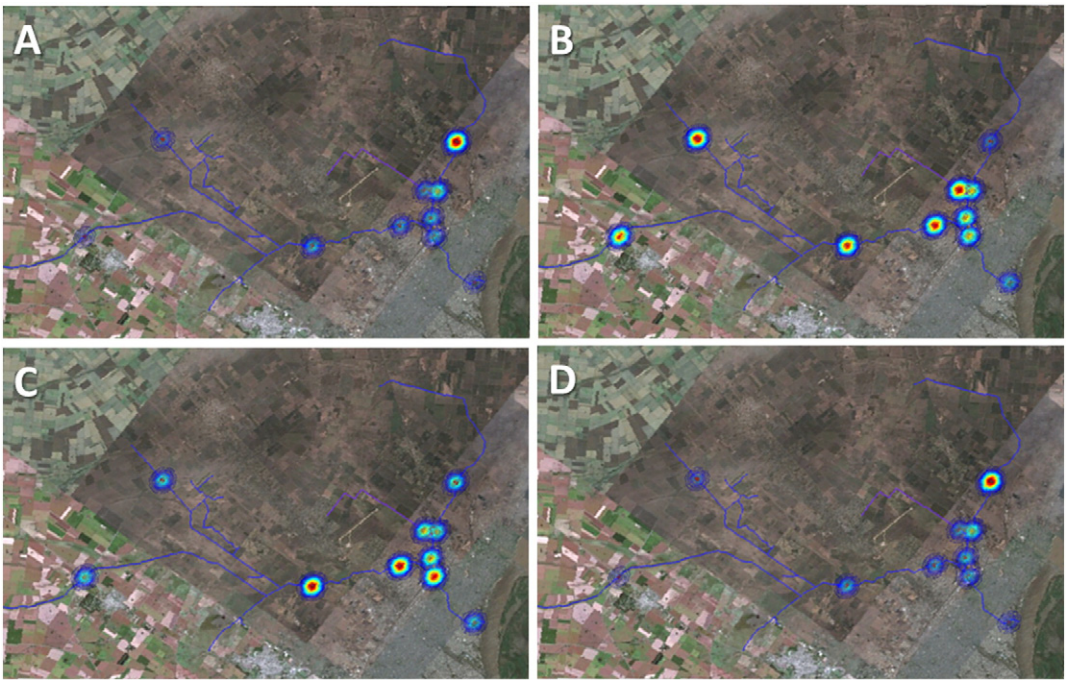

Fig. 5. Spatial distribution of the different MCR-ALS groups of variables. **A**, **B**, **C** and **D** are the several MRC-ALS groups.

analysis and additional possibilities to continue adding other techniques to determine different variables, or analytes to make a better characterization of a system. Thus this methodology allows to conduct a global monitoring of fluorescence data with other physicochemical variables together. Non-impacted areas of Ludueña stream are characterized by the presence of humic-like A and C fluorophores in a higher proportion. In the Ibarlucea channel a high proportion of tryptophan and tyrosine appears. These are products of biodegradable materials, common in the waters with anthropogenic influence and in areas with industrial effluent discharges and/or sewage.

Conflict of interest

There is no conflict of interest.

Acknowledgments

The following institutions are gratefully acknowledged for financial support: Universidad Nacional de Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, PICT-2011-0033).

References

[1] P.G. Coble, Characterization of marine and terrestrial DOM in seawater using excitation emission matrix spectroscopy, Mar. Chem. 51 (1996) 325–346.
[2] K.M. Elkins, D.L. Nelson, Fluorescence and FT-IR spectroscopic studies of Suwannee river fulvic acid complexation with aluminum, terbium and calcium, J. Inorg. Biochem. 87 (2001) 81–96.
[3] N. Patel-Sorrentino, S. Mounier, Y. Lucas, J.Y. Benaim, Effects of UV-visible irradiation on natural organic matter from the Amazon basin, Sci. Total Environ. 321 (2004) 231–239.
[4] C.A. Stedmon, S. Markager, R. Bro, Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy, Mar. Chem. 82 (2003) 239–254.
[5] A. Baker, Spectrophotometric discrimination of river dissolved organic matter, Hydrol. Process. 16 (2002) 3203–3213.
[6] W.K.L. Cammack, J. Kalf, Y.T. Prairie, E.M. Smith, Fluorecent dissolved organic matter in lakes: relationship with heterotrophic metabolism, Limnol. Oceanogr. 49 (2004) 2034–2045.
[7] C.D. Clark, J. Jimenez-Morais, G. Jones, E. Zanardi-Lamardo, C.A. Moore, R.G. Zika, A time-resolved fluorescence study of dissolved organic matter in a riverine to marine transition zone, Mar. Chem. 78 (2002) 121–135.
[8] P.G. Coble, K. Mopper, C.A. Schultz, Fluorescence contouring analysis of DOC Intercalibration Experiment samples: a comparison of techniques, Mar. Chem. 41 (1993) 173–178.
[9] M.M. de Souza-Sierra, O.X.F. Donard, M. Lamotte, C. Bellin, M. Ewald, Fluorescence spectroscopy of coastal and marine waters, Mar. Chem. 47 (1994) 127–144.
[10] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171.
[11] P. Kowalczuk, M.J. Durako, H. Young, A. Kahn, W. Cooper, M. Gonsior, Characterization of dissolved organic matter fluorescence in the South Atlantic Bight with use of PARAFAC model: interannual variability, Mar. Chem. 113 (2009) 182–196.
[12] J. Fellman, M. Miller, R. Cory, D. D'amore, D. White, Characterizing dissolved organic matter using PARAFAC modeling of fluorescence spectroscopy: a comparison of two models, Environ. Sci. Technol. 43 (2009) 6228–6234.
[13] R.K. Henderson, A. Baker, K.R. Murphy, A. Hambly, R.M. Stuetz, S.J. Khan, Fluorescence as a potential monitoring tool for recycled water systems: a review, Water Res. 43 (2009) 863–881.
[14] L. Gao, D. Fan, D. Li, J. Cai, Fluorescence characteristics of chromophoric dissolved organic matter in shallow water along the Zhejiang coasts, southeast China, Mar. Environ. Res. 69 (2010) 187–197.
[15] K.M.G. Mostofa, F. Wu, C.Q. Liu, W.L. Fang, J. Yuan, W.L. Ying, L. Wen, M. Yi, Characterization of Nanming River (southwestern China) sewerage-impacted pollution using an excitation–emission matrix and PARAFAC, Limnology 11 (2010) 217–231.
[16] G.J. Hall, J.E. Kenny, Estuarine water classification using EEM spectroscopy and PARAFAC–SIMCA, Anal. Chim. Acta. 581 (2007) 118–124.
[17] J.C.C.G. Esteves da Silva, M.J.C.G. Tavares, R. Tauler, Multivariate curve resolution of multidimensional excitation–emission quenching matrices of a Laurentian soil fulvic acid, Chemosphere 64 (2006) 1939–1948.
[18] http://www.mcrals.info/.
[19] R. Tauler, D. Barceló, E.M. Thurman, Multivariate correlation between concentrations of selected herbicides and derivatives in outflows from selected US midwestern reservoirs, Environ. Sci. Technol. 34 (2000) 3307–3314.
[20] M. Terrado, D. Barceló, R. Tauler, Quality assessment of the multivariate curve resolution alternating least squares (MCR-ALS) method for the investigation of environmental pollution patterns, Environ. Sci. Technol. 43 (2009) 5321–5326.
[21] U. Dietze, T. Braunbeck, W. Honnen, H.R. Köhler, J. Schwaiger, H. Segner, Chemometric discrimination between streams based on chemical, limnological and biological data taken from freshwater fishes and their interrelationships, J. Aquat. Ecosyst. Stress. Recover. 8 (2001) 319–336.
[22] S.P. Mujunen, P. Minkkinen, B. Holmbom, A. Oikari, PCA and PLS methods applied to ecotoxicological data: ecobalance project, J. Chemometr. 10 (1996) 411–424.
[23] R. Tauler, S. Lacorte, M. Guillamon, R. Cespedes, P. Viana, D. Barceló, Chemometric modeling of main contamination sources in surface waters of Portugal, Environ. Toxicol. Chem. 23 (2004) 565–575.
[24] E.D. Malinowski, Factor analysis in chemistry, 3rd ed., John Wiley & Sons, New York, 2002.
[25] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of chemometrics and qualimetrics, Elsevier, Amsterdam, 1998.
[26] I.M. Fharnham, A.K. Singh, K.J. Stetzenbach, K.H. Lohannesson, Treatment of nondetects in multivariate analysis of groundwater geochemistry data, Chemom. Intell. Lab. Syst. 60 (2002) 265–281.
[27] A. Smilde, R. Bro, P. Zeladi, Multi-way analysis with applications in the chemical sciences, John Wiley & Sons Ltd, New York, 2004.
[28] D.M. McKnight, E.W. Boyer, P.K. Westerhoff, P.T. Doran, T. Kulbe, D.T. Andersen, Spectrofluorometric characterization of dissolved organic matter for indication of precursor organic material and aromaticity, Limnol. Oceanogr. 46 (2001) 38–48.
[29] J. Christensen, V.T. Povlsen, J. Sørensen, Application of fluorescence spectroscopy and chemometrics in the evaluation of processed cheese during storage, J. Dairy Sci. 86 (2003) 1101–1107.
[30] Thygesen, et al., Stabilizing the PARAFAC decomposition of fluorescence spectra by insertion of zeros outside the data area, Chemometr. Lab. 71 (2004) 97–106.
[31] R.D. JiJi, K.S. Booksh, Mitigation of Rayleigh and Raman spectral interferences in multiway calibration of excitation–emission matrix fluorescence spectra, Anal. Chem. 72 (2000) 718–725.
[32] Andersen Rinnan, Handling of first-order Rayleigh scatter in PARAFAC modelling of fluorescence excitation–emission data, Chemometr. Lab. 76 (2005) 91–99.
[33] R. Zepp, W.M. Sheldon, M.A. Moran, Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating Rayleigh and Raman scattering peaks in excitation–emission matrices, Mar. Chem. 89 (2004) 15–36.
[34] C. de Boor, A practical guide to splines, Applied mathematical sciences, 27, Springer-Verlag, Berlin, 1978. 392.
[35] S. Leurgans, R.T. Ross, Multilinear models: applications in spectroscopy, Stat. Sci. 3 (1992) 289–319.
[36] P. Paatero, Monitoring the kinetics of the ion-exchange resin catalysed esterification of acetic acid with ethanol using near infrared spectroscopy with partial least squares (PLS) model, Chemom. Intell. Lab. Syst. 38 (1997) 223–242.
[37] R. Bro, Multi-way analysis in the food industry, University of Amsterdam, Netherlands, 1998. (Doctoral Thesis).
[38] http://www.models.kvl.dk/algorithms.
[39] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemometr. 17 (2003) 274–286.
[40] R.A. Harshman, M.E. Lundy, The PARAFAC model for three-way factor analysis and multidimentional scaling, in: H.G. Law, C.W. Snyder Jr., J. Hattie, R.P. McDonald (Eds.), Research methods for multimode data analysis, Praeger, New York, 1984, p. 122.
[41] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1) (1995) 133–146.
[42] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, 3-way data analysis and ambiguity in multivariate curve resolution, J. Chemometr. 9 (1) (1995) 31–58.
[43] T. Jolliffe, Principal component analysis, Springer- Verlag, New York, 2002. 26.
[44] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemom. Intell. Lab. Syst. 76 (2005) 101–110.
[45] P.G. Coble, C.E. del Castillo, B. Avril, Distribution and optical properties of CDOM in the Arabian Sea during the 1995 Southwest Monsoon, Deep-Sea Res. II 45 (1998) 2195–2223.
[46] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (1991) 1425.
[47] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numer. Math. 14 (1970) 403–420.
[48] P. Westerhoff, W. Chen, M. Esparza, Fluorescence analysis of a standard fulvic acid and tertiary treated wastewater, J. Environ. Qual. 30 (2001) 2037–2046.