Microchemical Journal 142 (2018) 208–218

Contents lists available at ScienceDirect

Microchemical Journal

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

Development of a fast and inexpensive method for detecting the main sediment sources in a river basin

Marianela Batistelli, Andrés R. Martínez Bilesio, 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

Keywords:
Chemometrics
Near infrared spectroscopy
Sediments sources

ABSTRACT

The sediments are a key link to understand the dynamics of a water basin, since they are the main transport and retention agent of nutrients and contaminants. Therefore, a method to identify the main sediment sources and study their distribution through the basin over time without the need to use reference standards was developed.

Water samples with suspended sediments were collected for one year from different sites of the Ludueña stream basin (Argentina). The sediments were deposited in fiberglass filters and their near infrared spectrum was measured. The data obtained were processed by bilinear and trilinear methods. Principal component analysis (PCA), multivariate curves resolution with alternating least squares (MCR-ALS) with and without trilinearity; and parallel factor analysis (PARAFAC) were applied. These algorithms allow identifying the three main sediment sources. The spectral loadings that characterize each source and the scores that reflect their distribution along the different sampling sites were obtained. It was concluded that the main contribution of sediments is given by the plant tissue developed by photosynthetic organisms that have a seasonal behavior. The remaining sources are characterized by being soil particles of the region under study (incorporated into the system by wind and rain) and the development of organisms with anoxygenic photosynthesis (their growth is favored due to the contribution of fluid discharges from anthropogenic activities).

1. Introduction

Sediments are the main transport agent of nutrients within a watershed [1]. Therefore, they have a great influence on the eutrophication pathways and on the chemistry of the rivers. They can also participate as transport agents and retention of organic and metallic pollutants [2, 3]. Therefore, the study of the contributions and behavior of the sediments is of great interest in the environmental area, since the distribution in space and the evolution in time of sediments is a critical element to determine the dynamics of a basin.

One of the main steps to be able to determine a correct model to understand the behavior of the sediments is the determination of its sources that give origin and/or contribute to the different sediments of a basin [4]. This information is of great importance, not only to understand the behavior of sediments, but also to create models of erosion or optimize the interactions between terrestrial and aquatic ecosystems [5, 6]. It is very complex to determine the sources that contribute to the different sediments and trace their location in the space based on visual observations, since there is a great diversity of possible sources [7–9]. It must be considered that the sediments are not composed of a single pure compound, but they are a mixture of particles from multiple origins and, additionally, they have point and/or diffuse sources which make it even more difficult to determine their origin, since the sediments collected from a basin are the sum of all these sources [10].

Near infrared spectroscopy (NIR) is a fast, inexpensive and non-destructive technique that has been previously used in soil and sediment studies [11, 12]. NIR reflectance spectroscopic measurements can be carried out very quickly in sieved and dried samples. There are several studies that have investigated the use of near infrared spectroscopy, along with chemometric methods, such as partial least squares (PLS) analysis, to evaluate the physical, chemical and even biological properties of soils [13–16].

In this work, we use the fingerprint methodology to identify and trace the origin of the sediments. This methodology uses the most distinctive characteristics of the sediment spectrum originated by a certain source. Therefore, it does not use a particular NIR band to identify a source, but the full spectrum that characterizes it. There are previous works where the fingerprint methodology is similarly used [17–19]. In all of these works this methodology is used to determine or predict a particular characteristic of the sediments or the water body

* Corresponding author.
E-mail address: garciareiriz@iquir-conicet.gov.ar (A.G. García-Reiriz).

https://doi.org/10.1016/j.microc.2018.06.040
Received 22 February 2018; Received in revised form 27 June 2018; Accepted 29 June 2018
Available online 02 July 2018

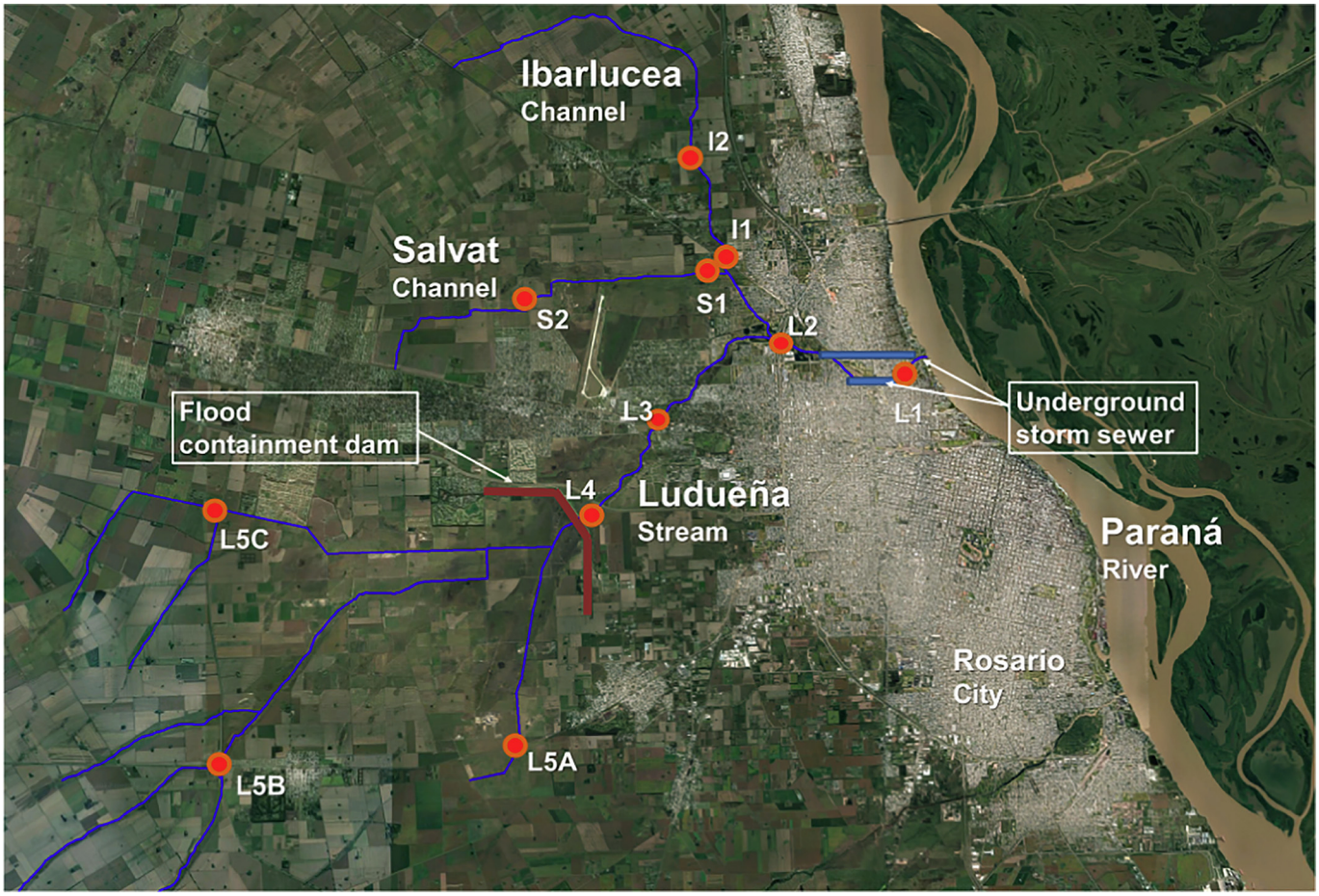

Fig. 1. Detailed sight of the sector under study showing sampling points and main geographical features.

using calibration samples. There is a single work where the authors quantify the contribution of the different sources of sediments in a basin [20]. They used samples from known sources of sediment to build a calibration model. The aim of our work is to determine these sources without the use of patterns or references samples, which is a major challenge to the algorithm for processing the data. However, the fact of not having reference samples offers the advantage that the results obtained will not be restricted to the possible sources of sediments used to build the calibration model.

In the fingerprint methodology, the NIR spectral data cannot be assigned to a specific chemical compound, and therefore multivariate statistical methods must be used to evaluate the spectral data. In this study, bilinear and trilinear methodologies are applied to solve the measured data. The bilinear methods obtain the main sediment sources only taking into account the spectral variations (second mode) of all the samples. On the other hand, trilinear methods also take into account the third mode of the data, i.e., resolve the data contemplating that the same sampling site was measured over time.

Principal component analysis (PCA) is a bilinear model used to obtain the main patterns in sediment composition from the NIR spectral data. PCA is a multivariate technique to replace a larger number of covarying variables with fewer independent (orthogonal) variables, i.e., principal components. Another bilinear algorithm is multivariate curve resolution coupled to alternating least squares (MCR-ALS). It was employed for similar purposes [21]. Therefore, both methods were applied to process the NIR spectral data and compare their results. MCR-ALS works in a similar way to PCA, but its constraints provide the results with physicochemical meaning and thus it is simpler to interpret (see specific details below).

MCR-ALS is a chemometric tool with an increasing application for the analysis of environmental monitoring data sets [22, 23]. 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 [24]. Another algorithm popularly used to analyze NIR data is PLS [25], but MCR-ALS and PCA were chosen because they do not require concentrations or calibration samples to perform the decomposition of the measured signal.

In order to apply trilinear methods, the data were analyzed again with MCR-ALS, but this time with the trilinearity constraint, and with parallel factor analysis (PARAFAC). Both methods allow obtaining the sediment sources by means of different strategies. These decomposing the data measured in three matrices, each of these matrices represents the behavior of sediment sources (components) in the different modes of the data (sampling sites-campaigns-spectra). In bibliography there are a large number of works where both algorithms are applied to different problems of a trilinear nature [26].

All the algorithms mentioned before decompose the signal in loadings and scores. In our case the loadings are related to the NIR spectra that characterize each component or sediment source, so they contain qualitative information. On the other hand, the values of the scores calculated with appropriate constrains are directly proportional to the concentration of specific sediment, i.e., they measure their relative concentrations. For this reason, this paper innovates in using these values without needing to obtain the absolute concentration value of specific sediments as in previous works [20]. Thus, the objective of this work is to use this advantage to be able to identify the different sediments sources and study their distribution throughout space-time without reference samples to optimize the model.

2. Material and methods

2.1. Study area

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 two tubes 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 gutters: 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 neighborhoods are being developed in its vicinity. Also, dense and irregular settlements exist in its margins, generating clandestine channels which provide storm water and sewage effluents.

2.2. Sample collection and preparation

Eleven sampling points (Fig. 1 and Table 1) of the stream were selected to represent the different branches and thus they represent the overall states of the stream according to the activities in its vicinity. At least five days were taken into account elapsed since the last rain before each campaign, to ensure that conditions were reproducible as much as possible. The campaigns were done approximately every 45 days, in the period between November 2016 and September 2017 (seven campaigns). (07/11/16, 13/12/16, 16/02/17, 24/04/17, 06/06/17, 24/07/17 and 19/09/17).

The samples were collected by duplicate for each site and they were stored in 1 l caramel colored glass bottles. The samples were filtered on mesh of 4 mm pore to eliminate traces of crustaceans, waterweeds and wastes. The sediments of a fraction of 200 ml of each sample were deposited in a Munktell micro-glass fiber MG/C filter of 47 mm diameter and 1.2 μm pore, by vacuum filtration. The filters were placed in glass plates and heated at 50 °C during 1 h, then dried in desiccators at room temperature and Vis-NIR spectra were measured. The Vis-NIR specters were measured as described below. All measurements were performed within 48 h of sample collection.

2.3. Equipment

Near infrared absorbance measurements were performed in reflectance mode using a NIRS DS2500 spectrometer (FOSS, Hilleroed, Denmark) equipped with pre-dispersive monochromator. This equipment allows measurements to be made starting from the visible to the NIR region of the spectrum. It converts the reflectance measurements in absorbance, so it allowed acquiring the full absorbance spectra in the Vis-NIR range from 550 to 2500 nm with a step of 0.5 nm. The equipment is connected to a PC where the spectra were stored. The spectra of the samples were obtained by placing the previously dried filters inside the cup provided by the equipment manufacturer. Each sample was scanned twice in different positions by manual rotation at an angle of 180°. The average spectrum was taken as the sample spectrum. The blank was obtained from the average spectrum of several dry and clean fiberglass filters. The recorded spectral data were processed and stored as absorbance units.

2.4. Data pretreatment

The first step in the data pretreatment was to subtract the average signal for the fiberglass filters to the spectra of each sample. This step was carried out in order to obtain the signal only from the sediments deposited on the filter.

The goal of this work is to locate and identify the sediments sources. For this reason, the Vis-NIR spectra of all samples must be analyzed together. The data arrangement of matrix D was built with the spectra of all samples. A vector of 3901 data points was obtained by sample; each vector was arranged one below each other. The size of the matrix D was 154 × 3901, where the first mode includes all samples taken in different sites and times by duplicate and the second mode are the measured wavelengths (see Fig. 2).

Since in this work sediment sources apportionment was studied, data were not initially mean centered or derived, because these data pretreatments cause some negative values that cannot be processed with the non-negativity constrain in the MCR-ALS algorithm. The MinMax pretreatment was preferred, because it scales the data between zero and one, maintaining the original magnitude of the data dispersion. The specific expression for the MinMax transformation is:

$$D_{MM} = \frac{D - \min(D)}{\max(D) - \min(D)} \quad (1)$$

where D is the data matrix with the original Vis-NIR absorbance values of all samples at all wavelengths, max(D) and min(D) are the global maximum and minimum of D respectively, and D_{MM} are transformed or scaled elements where the subscript 'MM' corresponds to the pretreatment selected. Scaled data are showed in Fig. 3A.

This procedure can also be applied individually for each wavelength. When done in this way, it has the advantage of giving equal importance to the magnitude of all the values of the spectrum. However, the disadvantage is that the original shape of the spectra is lost. It was corroborated that with both pre-processing strategies the results are very similar, and for this reason we chose to present the results with the first option because they are easier to interpret.

Table 1
Values of latitude and longitude of the location of the different sampling sites with their corresponding nomenclature dependent on the tributary to which they belong.

Sampling site identification		Tributary of the stream	Latitude	Longitude
Number	Code			
1	L1	Ludueña Stream	32°54′32.7″ South	60°40′53.1″ West
2	L2	Ludueña Stream	32°54′0.67″ South	60°43′21.3″ West
3	L3	Ludueña Stream	32°55′27.1″ South	60°45′43.9″ West
4	L4	Ludueña Stream	32°57′02.5″ South	60°47′19.7″ West
5	L5A	Ludueña Stream	33°00′42.2″ South	60°48′27.0″ West
6	L5B	Ludueña Stream	33°01′05.9″ South	60°54′15.0″ West
7	L5C	Ludueña Stream	32°56′54.0″ South	60°54′19.9″ West
8	S1	Salvat Channel	32°53′03.6″ South	60°44′51.0″ West
9	S2	Salvat Channel	32°53′37.3″ South	60°48′18.6″ West
10	I1	Ibarlucea Channel	32°52′50.6″ South	60°44′27.8″ West
11	I2	Ibarlucea Channel	32°51′07.9″ South	60°45′07.1″ West

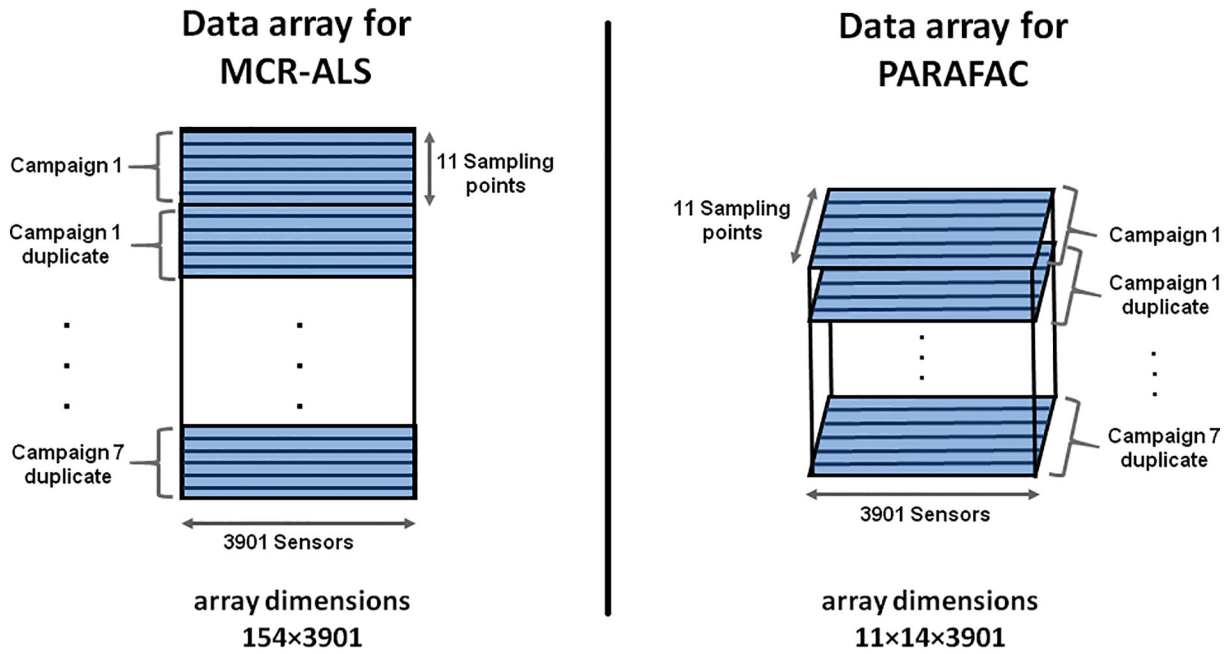

Fig. 2. Arrangement of the data for the MCR-ALS and PARAFAC analyses.

3. Theory

3.1. PCA

PCA assumes a bilinear model to explain the observed data variance using a reduced number of components only under the orthogonality constraint. For a detailed description of this well-known methodology in chemometrics and other multivariate statistical data analysis methods see references [27, 28]. The bilinear decomposition may be written by the element wise equation:

$$d_{ij} = \sum_{n=1}^{N} x_{in} y_{jn} + e_{ij} \tag{2}$$

where d_{ij} is one of the entries of the experimental data matrix (intensity of Vis-NIR reflectance in a particular sample to a specific wavelength) from the ith row (a particular sample) and the jth column (a specific wavelength), x_{in} is an scalar element corresponding to the nth score of the i sample, y_{jn} is an scalar element corresponding to the nth loading at the j variable and e_{ij} is the residual not modeled by the sum of N components or contributions. The same bilinear equation can be written in matrix form as:

$$\mathrm{D} = \mathrm{X}\mathrm{Y}^{\mathrm{T}} + \mathrm{E} \tag{3}$$

where D is the experimental data array expressed as a data matrix. Eq. (3) describes the decomposition (matrix factorization) of matrix D into two matrices, the loading matrix Y^{T} and the score matrix X. The loading matrix Y^{T} identifies the main sources of the data variance by means of their Vis-NIR spectra (spectral loadings), which eventually may be related to the main sediments sources. The score matrix X provides sample scores for these data variance patterns, indicating their geographical and temporal distribution in each sampling point because they are directly related to sediment concentrations. PCA solves Eq. (3) under orthogonal constraints. Each successively extracted principal component explains maximum variance. The determination of the complexity of the model in PCA (i.e. the number of principal components) is performed as a compromise between several goals: model simplicity (few components), maximum variance explained by the model (more components), and model interpretability.

3.2. MCR-ALS

MCR-ALS [29, 30] works with the data array arranged in the so-called column-wise augmented data matrix $\mathrm{D}_{\mathrm{aug}}$ (Fig. 2), which is the same data array built for PCA (described in the previous section, matrix D). The bilinear decomposition of the augmented matrix $\mathrm{D}_{\mathrm{aug}}$ is performed according to the same expression already given for PCA [i.e. Eq. (3)], but for its resolution an iterative procedure is applied called alternating least squares (ALS). In contrast to PCA, however, during the ALS optimization phase of MCR-ALS some other constrains can be applied. The selected constraints were non-negativity for profiles in both modes (for the augmented score mode and for the loadings, i.e. concentrations and spectral modes) and the loadings normalization to equal length.

3.3. MCR-ALS for trilinear models

Trilinear model can be implemented iteratively as a constraint during ALS optimization in the MCR-ALS method [31, 32]. The application of MCR-ALS using this constraint should not be considered to be equal to a standard bilinear decomposition of the augmented two-way matrix $\mathrm{D}_{\mathrm{aug}}$. This is because the profiles of the components must be the same in all the samples like in bilinear models, but in addition, trilinearity requests that their distribution in the sites must be the same in all the campaigns. During the ALS optimization, each individual profile of the augmented scores matrix X is constrained to fulfill the trilinearity condition independently and iteratively. The same procedure used previously for the recovery of the loadings in the three modes from the augmented scores matrix obtained by PCA or MCR-ALS is applied now inside/during the ALS optimization instead of at the end of the optimization as in PCA. Each column of the X matrix is appropriately folded at each ALS iteration step to give a matrix with a number of rows equal to the number of sampling sites (eleven) and seven columns corresponding to each campaign. Singular value decomposition (SVD) of this folded scores matrix gives the loadings in the second and third modes for the considered component. These two loadings describe the common variation captured by ALS in the two modes (sampling sites and campaigns) for that particular component. The Kronecker product [33–35] of these two new loading vectors gives the new augmented scores vector which substitutes the corresponding column of the X

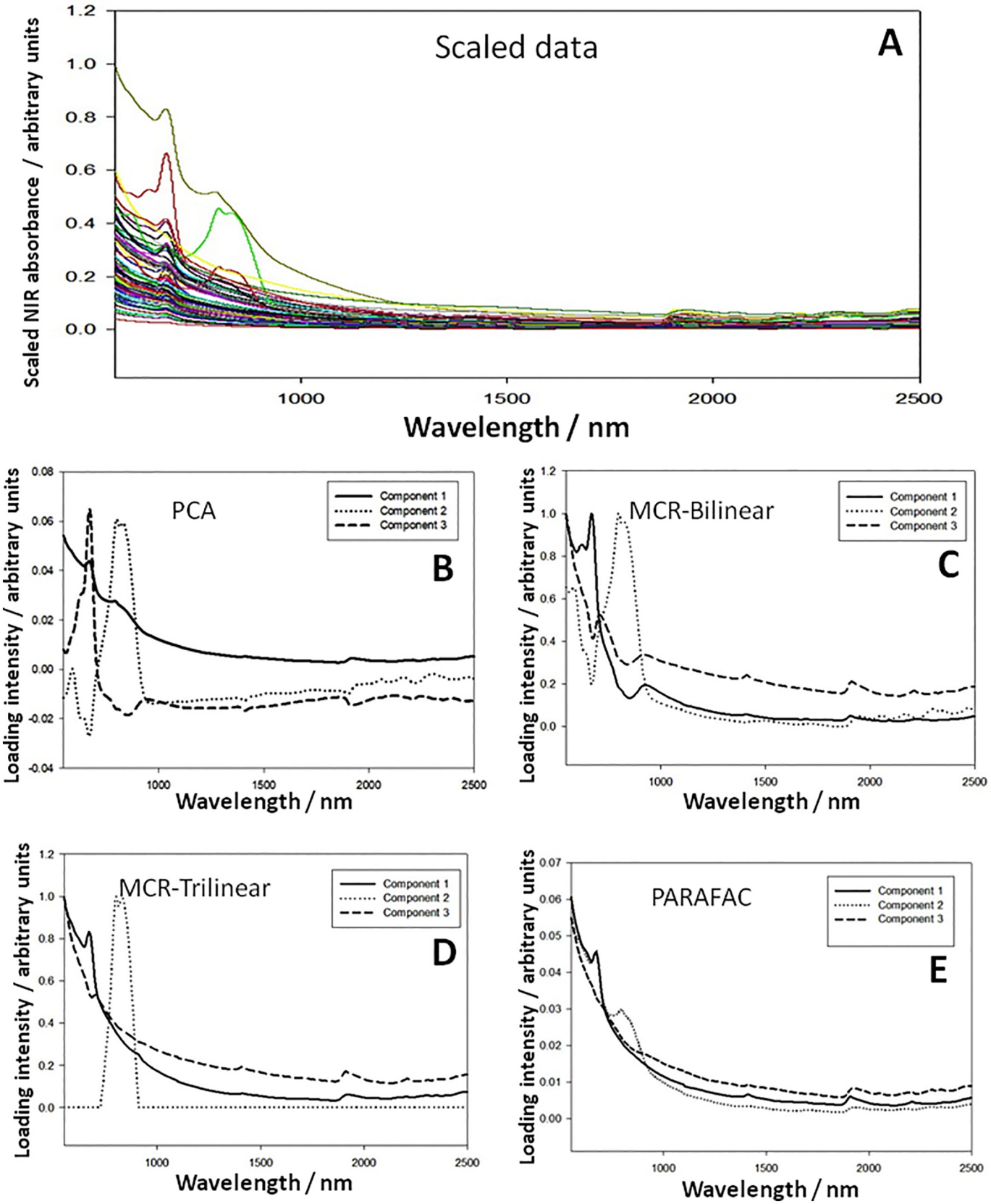

Fig. 3. A. Scaled data of all samples. B. Loadings profiles of the 3 components of PCA. C. Loadings profiles obtained by MCR-ALS bilinear model. D. Loadings profiles obtained by MCR-ALS trilinear model. E. Loadings profiles obtained by PARAFAC model.

scores matrix. When this constraint is inserted during each step of the ALS iterative optimization procedure, it forces the shape of the loadings vector in the second mode (describing the sampling site variation of the considered component) to be the same for the seven campaigns. Moreover, it captures the intensity (scale) variation of this component in the loadings of the third mode, showing the scale differences of this component among the seven campaigns.

3.4. PARAFAC

After measuring first-order data for a set of samples, each of them as a $1 \times K$ vector (K is the number of sensor on the spectral mode) of the I sampling sites at the J campaigns are joined into a three-way data array D, whose dimensions are $[I \times J \times K]$ (see Fig. 2). Provided D follows a trilinear PARAFAC model, it can be written in terms of three vectors for each responsive component or sediment source, designated as a_n, b_n and c_n, and collecting the relative concentrations $[I \times 1]$ for component n and the profiles in the two modes $(J \times 1)$ and $(K \times 1)$, respectively. The specific expression for a given element of D is [36]:

$$d_{ijk} = \sum_{n=1}^{N} a_{in} b_{jn} c_{kn} + e_{ijk} \quad (4)$$

where N is the total number of responsive components or sediment sources, a_{in} is the relative concentration of component n in the ith sample, b_{jn} is the sediment intensity in the campaign j and c_{kn} is the absorbance intensity at the wavelength k, respectively. The values of e_{ijk} are the elements of the array E, which is a residual error term of the same dimensions as D. The column vectors a_n, b_n and c_n are collected into the corresponding loading matrices A, B and C (b_n and c_n are usually normalized to unit length).

The model described by Eq. (4) defines a decomposition of D, which provides access to sediment evolutions through campaigns (B) and sediment spectral profiles (C), and relative concentrations (A) of individual components in the I samples, whether they are chemically known or not. The decomposition is usually accomplished through an alternating least-squares minimization scheme [37, 38].

Issues relevant to the application of the PARAFAC model on three-way data are as follows: (1) initializing the algorithm, (2) establishing the number of responsive components.

Initializing PARAFAC for the study of three-way arrays can be done by the loadings giving the best fit after small PARAFAC runs involving several sets of orthogonal random loadings.

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 [39, 40]. Another useful technique is the consideration of the PARAFAC residual error, i.e., the standard deviation of the elements of the array E in Eq. (4) [37]. Usually this parameter decreases with increasing N, until it stabilizes at a value compatible with the instrumental noise. 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.

3.5. Software

All calculations were made using MATLAB 7.0 (The Mathworks, Natick, Massachusetts, USA, 2007). In order to apply MCR-ALS, the codes available on the internet were implemented [41, 42]. The PARAFAC package codes are available thanks to Bro [43].

4. Results and discussion

4.1. Bilinear decomposition

4.1.1. PCA results

The first approximation to estimate the number of components was obtained by PCA, which indicates the number of possible major independent sediments sources affecting the measured data. The number of components was estimated by examining the size of the changes in explained variance in PCA as a function of the number of principal components and the values of the intensity of each eigenvalues. Three components were proposed to model the MinMax pre-processed data matrix, which allowed explaining 99.2% of the overall variance.

In Fig. 3B, loadings obtained by PCA are shown. It can be observed that the first component (94.3% of the variance explained) describes the average spectra sediments affecting the geographical region under study over the investigated period of time, and the other two components are describing the contrast with more specific sediments sources. The second component (2.8% of the explained variance) highlights the different sediment spectral region between 750 and 900 nm and it has at large wavelengths (1700–2500 nm) the characteristic shape of biomass rich in plant tissues with cellulose and lignin [44, 45]. Finally, the third component (2.1% of the explained variance) describes the different spectral behavior mainly in the wavelengths region between 600 and 720 nm. In addition, in this component, the peaks of the water that is retained in the hygroscopic material of the sediments can be observed in inverted form (1400 and 1900 nm).

The corresponding PCA scores describe the geographical and temporal distribution of these sediments patterns. These show the sites with high general sediment concentration (PC1 scores), the sites affected by a specific sediment source (PC2 scores) and the sites related with a seasonal behavior of sediments (PC3 scores). Because PCA defines the same vector space as the one obtained by MCR-ALS decomposition using the same number of components (see below), PCA scores plots are available in the Supplementary material. An advantage of MCR-ALS over PCA is the possibility of applying natural constraints like non-negativity, making easier the physical interpretation of the results. For this reason, the discussion about the possible sources or patterns was mainly focused on the MCR-ALS results.

4.1.2. Bilinear MCR-ALS results

MCR-ALS was first applied with only non-negativity constraint for scores and loadings. The trilinearity constraint was not applied in this case to build a more flexible model. The explained variance was 99.7% for three components.

In the Fig. 3C are shown the bilinear MCR-ALS loadings obtained and in the Fig. 4 are shown the scores of the three major sediments patterns represented in bar plots and its geographical distribution in each campaign. The graphics of the geographical distribution were constructed with a routine designed by our working group written in MATLAB. This routine works with two superimposed georeferenced layers. The background layer is a simplified image of the watercourses of the basin and the superior layer contains gaussians of different heights that represent the intensity of the scores in the different sampling sites. These maps allow interpreting the results in a more agile way since they allow relating the results with the geography of the area. These three components are interpreted in environmental terms as follows.

The first component (Figs. 3C and 4) (98.0% of the total variance explained) has a loading spectra signal in the same region than the main light absorbing protein complex of photosystem II [46]. This signal is consistent with organism with chlorophyll-*a*, which is responsible of the oxygenic photosynthesis and absorbs light at 680 nm. In its geographical and temporal distribution, it can be seen that this component has a cyclical behavior dominated by the seasonal period (higher intensities at the beginning of the monitoring, corresponding to spring

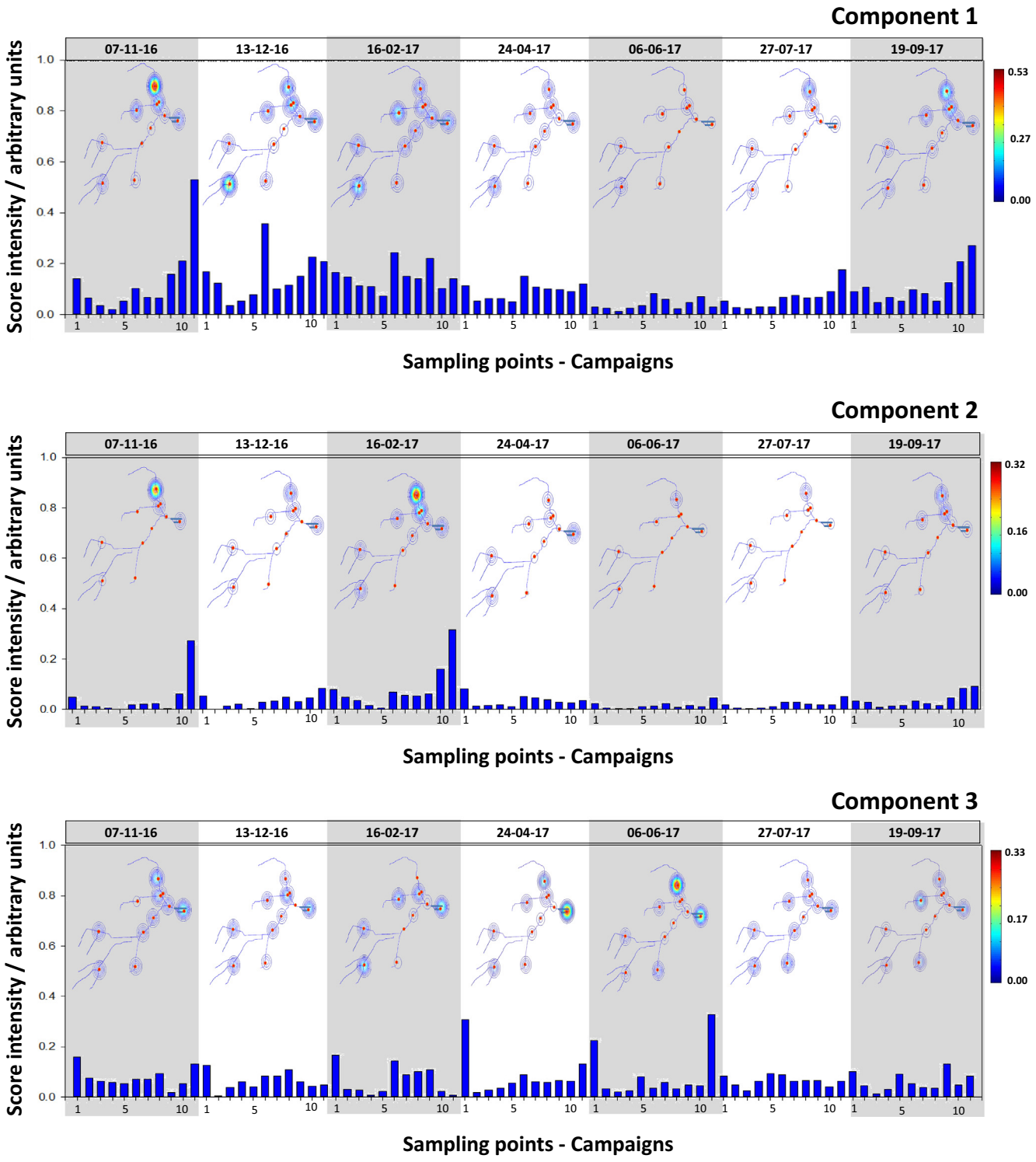

Fig. 4. Bar plots and spatial distributions of the bilinear MCR-ALS scores of the three components in an all sampling sites of all campaigns. Sampling sites identification: 1-L1, 2-L2, 3-L3, 4-L4, 5-L5A, 6-L5B, 7-L5C, 8-S1, 9-S2, 10-I1 and 11-I2.

and summer in the southern hemisphere). This component represents the sedimentary material contributed by oxygenic photosynthesizing organisms that proliferate in periods of high temperatures in areas with high load of nitrates and phosphates.

The second component (Figs. 3C and 4) (1.0% of the total variance explained) has two signal regions in its loading spectra. One is dominated by the characteristic signal of biomass rich in plant tissues with cellulose and lignin at large wavelengths [44, 45], and the other between 750 and 900 nm is consistent with the spectra of bacteriochlorophyll [46], because the main light absorbing complex in anoxygenic photosynthesis [47] of purple bacteria, green bacteria, heliobacteria and chlorobacteria absorb in the region between 770 and 870 nm. The biomass material comes from the vegetable tissue that develops due to the high nutrient load of the water (eutrophication of the basin). Such profile is specifically located on the zone of Ibarlucea Channel which presents a highly impacted zone by anthropological

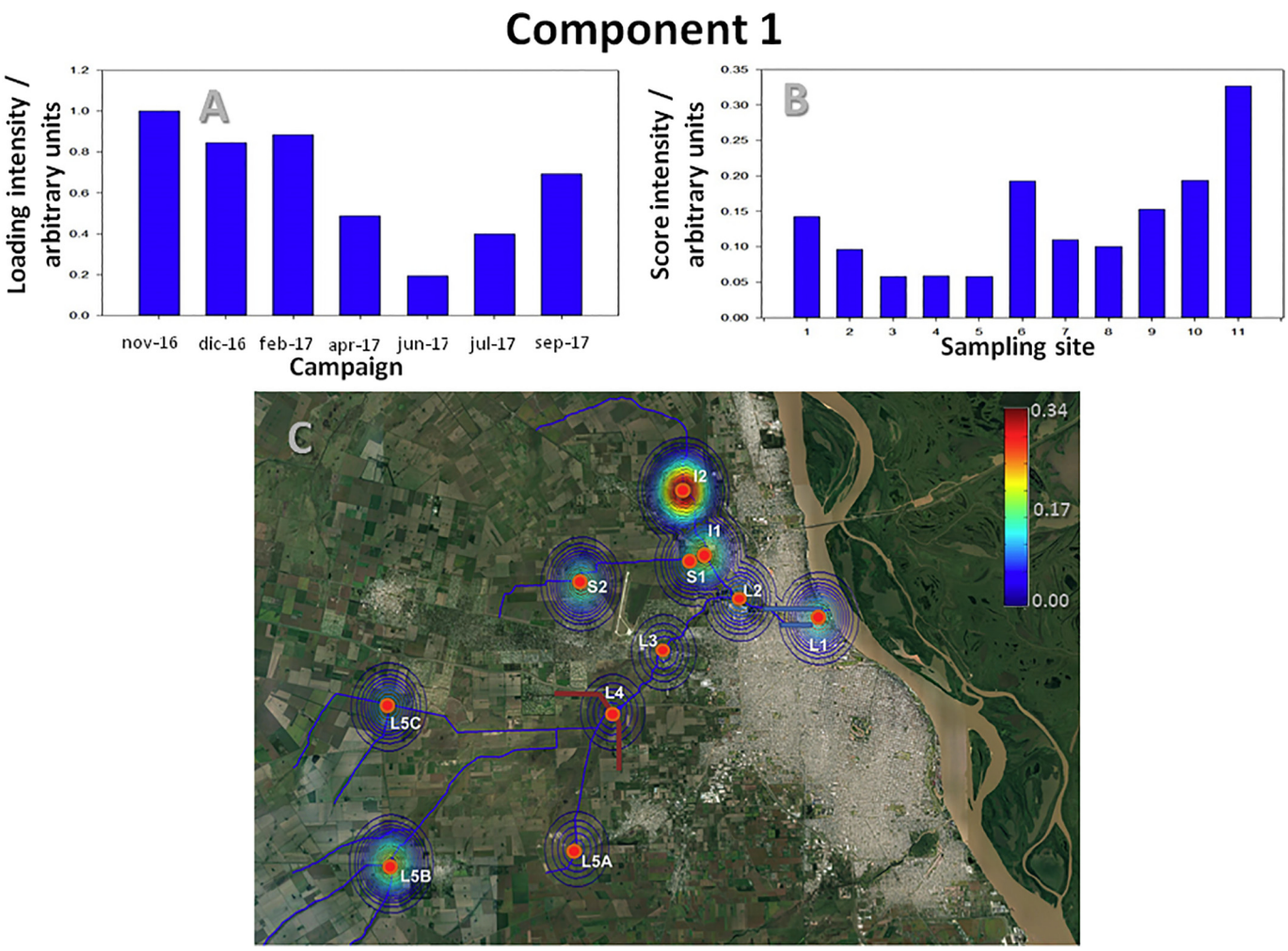

Fig. 5. Trilinear MCR-ALS scores of component 1. A. Temporal distribution in each campaign. B. Representation in a bar plot of the distribution of the scores in each sampling site. Sampling sites identification: 1-L1, 2-L2, 3-L3, 4-L4, 5-L5A, 6-L5B, 7-L5C, 8-S1, 9-S2, 10-I1 and 11-I2. C. Map representation of the geographical distribution of the scores in each sampling site.

activities. In this area a sanitary landfill is located with leachate treatment lagoons and also a large amount of local denounce about untreated sewage discharges through clandestine connections or the emptying of tanker truck exists.

The loading spectrum of the third component (0.7% of the total variance explained) (Figs. 3C and 4) corresponds to the Vis-NIR spectra of the particles of the soil of basin incorporated into the system by the wind and the flow stream. Only to corroborate this conclusion, the loading was compared with the spectra of several sieved and dried soil samples and it always has the same shape.

4.2. Trilinear decomposition

4.2.1. Trilinear MCR-ALS results

MCR-ALS was then applied with non-negativity again for scores and loadings, and trilinearity constraint [48, 49]. This latter constraint is more restrictive, leading to a decreased percentage of explained variance, but it has the advantage of separating the components of the patterns between the campaigns and the time. It demands that all campaigns have the same number of sampling sites. This is not always possible in environmental databases because not all sampling points are taken in the monitoring plans, but it is possible for our data base. In terms of explained variance, the results obtained by trilinear MCR-ALS analyses were slightly worse than those obtained with bilinear model (90.5% with this approach), but resulted rather similar in relation to the resolved components. This suggested that the data could be approximated by the trilinear model, giving more easily interpretable component profiles, especially in terms of the geographical distribution representation (mapping) of the resolved components describing the different sediment sources under study.

Fig. 3D shows the results corresponding to the non-negativity/trilinearity constrained MCR-ALS study. Three different patterns or source groups were identified (total explained variance of 90.5%): (1) the first component (89.1% of the total variance explained) is dominated by the signal at 680 nm of organisms with oxygenic photosynthesis, characteristic of chlorophyll-*a*; (2) the second component (0.7% of the total variance explained) is dominated by the spectra of anoxygenic photosynthetic organisms with bacteriochlorophyll, and (3) the third component (0.6% of the total variance explained) is dominated by the sediment signal coming from the particles of soil. Once identified the spectral characteristics of the main sediments patterns, the localization of these patterns and the corresponding possible sources are investigated.

In Figs. 5A, 6A and 7A, the temporal loadings of each trilinear MCR-ALS component are shown. In Figs. 5B, 6B and 7B, the scores intensity in each sampling site are shown in bar plots; and Figs. 5C, 6C and 7C show the same scores but in a map representation to analyze their geographical distribution by each component respectively. These last figures were built as mentioned previously for Fig. 4, but it using a georeferenced satellite image as background.

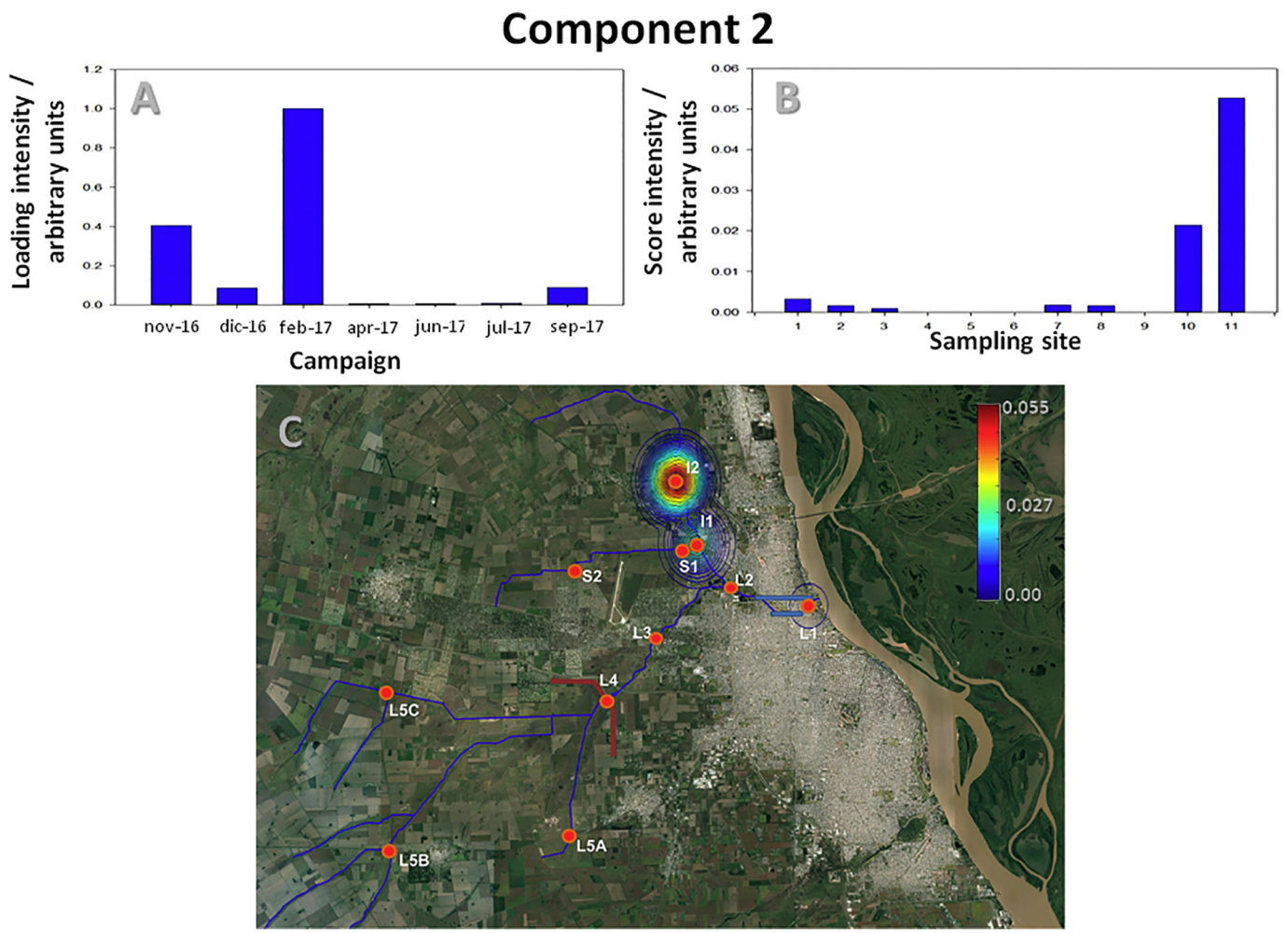

Fig. 6. Trilinear MCR-ALS scores of component 2. A. Temporal distribution in each campaign. B. Representation in a bar plot of the distribution of the scores in each sampling site. Sampling sites identification: 1-L1, 2-L2, 3-L3, 4-L4, 5-L5A, 6-L5B, 7-L5C, 8-S1, 9-S2, 10-I1 and 11-I2. C. Map representation of the geographical distribution of the scores in each sampling site.

The first MCR-ALS component or sediment source (89.1%) defined by the first spectra (see above) is localized throughout the basin (see Fig. 5B and C), but is mainly located in the highest areas of the basin which have lower flow, allowing a greater proliferation of algae and others oxygenic photosynthetic organisms. This component explains the behavior of sediments with a natural origin and has a highest seasonal or cyclical behavior in the spring and summer period (Fig. 5A).

The second component (0.7%) is focused on regions near the Ibarlucea Channel and has high concentrations in the first and third campaigns. This behavior is typical of the impact of anthropic activities, since it does not present a particular trend (Fig. 6A). Its spectrum is characteristic of bacteriochlorophyll coming from organisms with anoxygenic photosynthesis. The bacteriochlorophyll can only be generated by purple bacteria, green bacteria, heliobacteria and chlorobacteria [47]. These organisms develop in environments with high load of organic matter and presence of sulfur (Fig. 6B and C).

Finally, the third component (0.6%) is present in all sampling points, similar to the first component, but is highest at the Ludueña outlet site to the Paraná River (Fig. 7B and C). The water in this site is a mixture coming from both rivers. The water form Paraná River is characterized by having high concentrations of colloidal soil particles. As regarding the time evolution of this component (Fig. 7A), it can be concluded that it is present in all campaigns but with greater intensity in the fourth and fifth campaign (24-04-17 and 06-06-17, autumn and winter). These campaigns correspond to the months with less rainfall recorded by the Argentine National Meteorological Service [50], so the particulate material from the soil can be in corporate more easily to the stream by the wind.

Results obtained with trilinearity agree with previous results obtained by modeling the dataset with bilinear MCR-ALS algorithm. Again, three MCR-ALS components were used to justify the observed data variance. Interpreting the composition and location of each component, we can conclude that the first component can be associated with the development of algae and others oxygenic photosynthetic organisms since its spectrum corresponds to the proteins of photosystem II and has a seasonal behavior, being higher in the months of higher temperature. The second component can be associated with an anthropological origin because it has a random behavior related with clandestine discharges of untreated fluids and its spectrum is associated with the signal of bacteriochlorophyll due to the development of organisms with photosynthesis anoxygenic because of the high load of organic matter and sulfur. The third component profile can be specifically related to the particulate material from the soil of the area that is incorporated into the system by the wind or resuspended by the flow of the stream.

4.2.2. *PARAFAC results*

As explained in the Theory section, PARAFAC is a trilinear method; therefore it requires a linear behavior in the three measured modes (sampling sites, campaigns and spectra). The previously scaled data was used to apply this algorithm, but grouped together in a three-dimensional array (as shown in Fig. 2). Each sampling campaign generates a

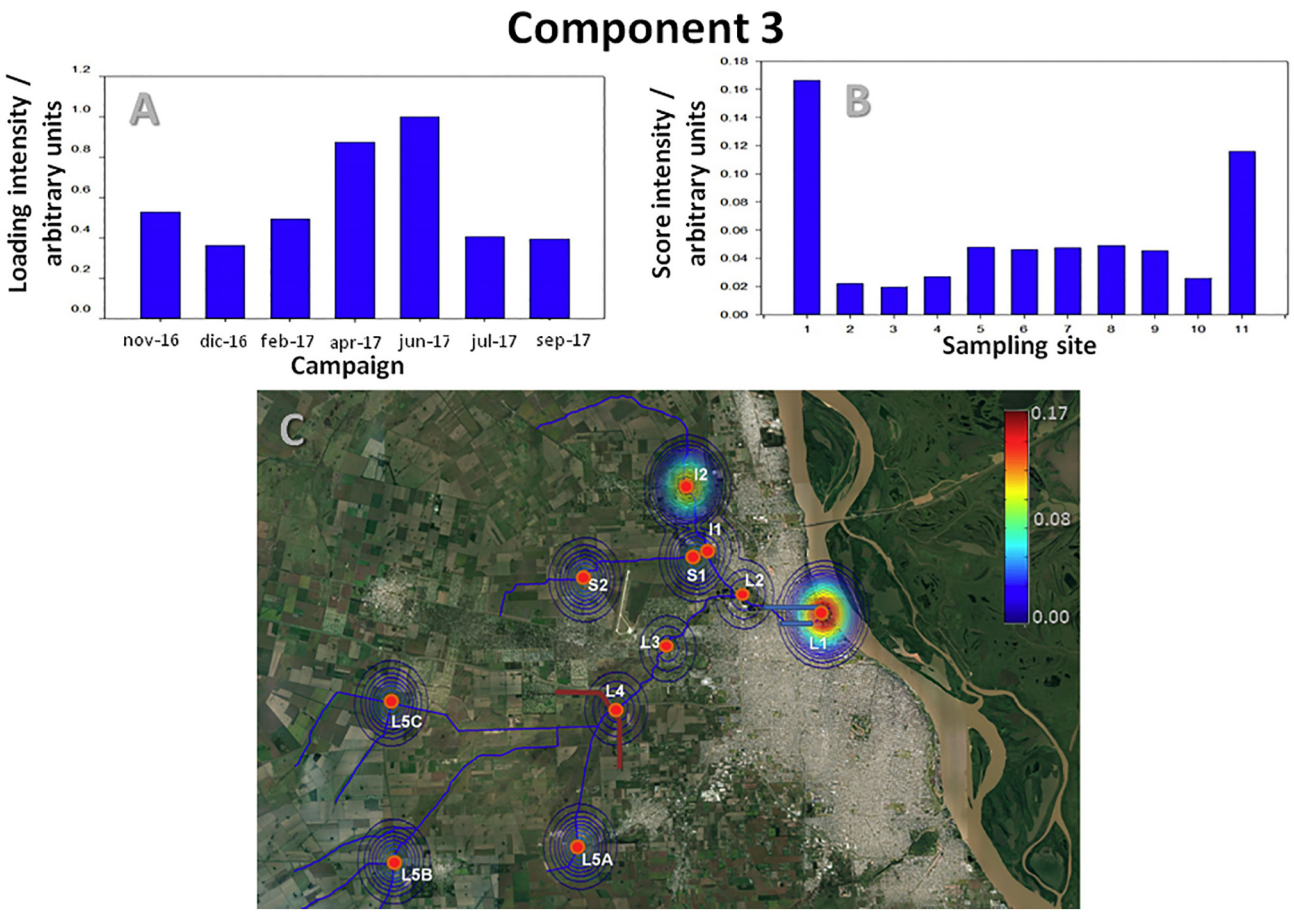

Fig. 7. Trilinear MCR-ALS scores of component 3. A. Temporal distribution in each campaign. B. Representation in a bar plot of the distribution of the scores in each sampling site. Sampling sites identification: 1-L1, 2-L2, 3-L3, 4-L4, 5-L5A, 6-L5B, 7-L5C, 8-S1, 9-S2, 10-I1 and 11-I2. C. Map representation of the geographical distribution of the scores in each sampling site.

two-way matrix of dimensions 11 × 3901, which correspond to the 11 sampling sites and the 3901 wavelengths. 7 sampling campaigns were carried out in duplicate throughout this work, thus having a total of 14 matrices. These matrices were stacked one on top of the other to obtain a three-way arrangement data of 11 × 14 × 3901.

As previously mentioned, the best-fitting models of several models fitted using a few iterations was used to initialize the algorithm. Similar to MCR-ALS model, PARAFAC was applied with no-negativity constraint for the three modes. The quantity of components used was selected by the CONCORDIA test. Values of the core consistency of 100, 48.9, 60.1, 21.4 and −4.2 were obtained for 1 to 5 components respectively. The component selection criteria is the value of this parameter falls below 50 units means that the components are not necessary. Consistent with the previous analyses, 3 components were chosen. This model explains a total of 97.2% of the variance (89.1% corresponds to the first component, and 6.2% and 1.3% correspond to the second and third component respectively).

The PARAFAC spectral loadings are shown in Fig. 3E. These are similar to those obtained with trilinear MCR-ALS model. The PARAFAC components order is the same obtained by bi- and trilinear MCR-ALS. The spectral loadings of components 1 and 3 have similar shape in both methods. The component 2 is slightly different but it has signal at the region between 750 and 900 nm typical of bacteriochlorophyll. For this reason, it can be concluded these represent the same sediment sources previously found. It was corroborated studying their distribution in the sampling sites and their behavior throughout the campaigns. Since the distribution and temporal behavior of the components of PARAFAC is the same to the results obtained by trilinear MCR-ALS, only the results obtained by the latter are shown to summarize the figures of the work (they are available in the Supplementary material).

5. Conclusions

In this work, PCA, MCR-ALS and PARAFAC were applied to investigate main sediments sources affecting a river basin of a particular geographical region over several monitoring campaigns and analysis. PCA allows determining the quantity of sediment sources, being only three components enough to justify the variability of the data. PARAFAC and MCR-ALS with non-negativity and with or without trilinearity constraints resulted to be more efficient tools to resolve the major sediments sources explaining the measured data variance. Three major patterns were detected, which were respectively related to: organism with oxygenic photosynthesis with a seasonal behavior; organism with anoxygenic photosynthesis characteristic of sites with high load of organic matter and sulfur; and particles coming from the soil of the zone. It was possible to study their temporal evolution and, by mapping representations, their geographical distribution. All these results were obtained without reference or calibration samples; this is a very important advantage above the methods available in bibliography.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.microc.2018.06.040.

Acknowledgments

The following institutions are gratefully acknowledged for financial support: Universidad Nacional de Rosario and CONICET (Consejo

Nacional de Investigaciones Científicas y Técnicas).

References

[1] A. Sharpley, Soil phosphorus dynamics: agronomic and environmental impacts, Ecol. Eng. 5 (1995) 261–279.
[2] I. Haag, U. Kern, B. Westrich, Erosion investigation and sediment quality measurements for a comprehensive risk assessment of contaminated aquatic sediments, Sci. Total Environ. 266 (2001) 249–257.
[3] N. Warren, I.J. Allan, J.E. Carter, W.A. House, A. Parker, Pesticides and other micro-organic contaminants in freshwater sedimentary environments-a review, Appl. Geochem. 18 (2003) 159–194.
[4] A.L. Collins, D.E. Walling, Documenting catchment suspended sediment sources: problems, approaches and prospects, Prog. Phys. Geogr. 28 (2004) 159–196.
[5] D.E. Walling, P.N. Owens, B.D. Waterfall, G.J. Leeks, P.D. Wass, The particle size characteristics of fluvial suspended sediment in the Humber and Tweed catchments, UK, Sci. Total Environ. 251-252 (2000) 205–222.
[6] D.E. Walling, A.L. Collins, The catchment sediment budget as a management tool, Environ. Sci. Pol. 11 (2008) 136–143.
[7] M.R. Peart, D.E. Walling, Techniques for establishing suspended sediment sources in two drainage basins in Devon, UK: a comparative assessment (IAHS Publ.), in: M.P. Bordas, D.E. Walling (Eds.), Sediment Budgets, Proc. Porto Alegre Symp, IAHS Press, Wallingford, UK, 1988, pp. 269–279.
[8] I.D.L. Foster, J.A. Lees, P.N. Owens, D.E. Walling, Mineral magnetic characterization of sediment sources from an analysis of lake and flood plain sediments in the catchments of the Old Mill reservoir and Slapton Ley, South Devon, U.K, Earth Surf. Process. Landf. 23 (1998) 685–703.
[9] A.L. Collins, D.E. Walling, Selecting fingerprint properties for discriminating potential suspended sediment sources in river basins, J. Hydrol. 261 (2002) 218–244.
[10] D.E. Walling, Tracing suspended sediment sources in catchments and river systems, Sci. Total Environ. 344 (2005) 159–184.
[11] A. Gredilla, S. Fdez-Ortiz De Vallejuelo, N. Elejoste, A. de Diego, J.M. Madariaga, Non-destructive spectroscopy combined with chemometrics as a tool for green chemical analysis of environmental samples: a review, TrAC Trends Anal. Chem. 76 (2016) 30–39.
[12] D. Cozzolino, Near infrared spectroscopy as a tool to monitor contaminants in soil, sediments and water—state of the art, advantages and pitfalls, Trends Environ. Anal. Chem. 9 (2016) 1–7.
[13] R.A. Viscarra Rossel, D.J.J. Walvoort, A.B. McBratney, L.J. Janik, J.O. Skjemstad, Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties, Geoderma 131 (2006) 59–75.
[14] D.J. Brown, K.D. Shepherd, M.G. Walsh, M.D. Mays, T.G. Reinsch, Global soil characterization with VNIR diffuse reflectance spectroscopy, Geoderma 132 (2006) 273–290.
[15] A.B. McBratney, B. Minasny, R.A. Viscarra Rossel, Spectral soil analysis and inference systems: a powerful combination for solving the soil data crisis, Geoderma 136 (2006) 272–308.
[16] G.N. Elliott, H. Worgan, D. Broadhurst, J. Draper, J. Scullion, Soil differentiation using fingerprint Fourier transform infrared spectroscopy, chemometrics and genetic algorithm-based feature selection, Soil Biol. Biochem. 39 (2007) 2888–2896.
[17] T. Tiecher, L. Caner, J.P. Gomes Minella, M.A. Bender, D.R. dos Santos, Tracing sediment sources in a subtropical rural catchment of southern Brazil by using geochemical tracers and near-infrared spectroscopy, Soil Tillage Res. 155 (2016) 478–491.
[18] P.N. Owens, W.H. Blake, L. Gaspar, D. Gateuille, A.J. Koiter, D.A. Lobb, E.L. Petticrew, D.G. Reiffarth, H.G. Smith, J.C. Woodward, Fingerprinting and tracing the sources of soils and sediments: Earth and ocean science, geoarchaeological, forensic, and human health applications, Earth Sci. Rev. 162 (2016) 1–23.
[19] A.L. Collins, S. Pulley, I.D.L. Foster, A. Gellis, P. Porto, A.J. Horowitz, Sediment source fingerprinting as an aid to catchment management: a review of the current state of knowledge and a methodological decision-tree for end-users, J. Environ. Manag. 194 (2017) 86–108.
[20] J. Poulenard, Y. Perrette, B. Fanget, P. Quetin, D. Trevisan, J.M. Dorioz, Infrared spectroscopic mapping of sediment sources in a small rural basin (French Alps), Sci. Total Environ. 407 (2009) 2808–2819.
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Control 86 (2018) 283–293.
[26] R.L. Pérez, G.M. Escandar, Experimental and chemometric strategies for the development of Green Analytical Chemistry (GAC) spectroscopic methods for the determination of organic pollutants in natural waters, Sustain. Chem. Pharm. 4 (2016) 1–12.
[27] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 2003.
[28] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.
[29] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, 3-way data analysis and ambiguity in multivariate curve resolution, J. Chemometrics 9 (1995) 31–58.
[30] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146.
[31] R. Tauler, I. Marques, E. Casassas, Multivariate curve resolution applied to three-way trilinear data: study of a spectrofluorimetric acid-base titration of salicylic acid at three excitation wavelengths, J. Chemom. 12 (1998) 55–75.
[32] A. de Juan, R. Tauler, Comparison of three-way resolution methods for non-trilinear chemical data sets, J. Chemometrics 15 (2001) 749–771.
[33] A. Smilde, R. Bro, P. Geladi (Eds.), Multi-way Analysis, John Wiley & Sons Ltd., Chichester, England, 2004.
[34] D.S. Burdick, An introduction to tensor-products with applications to multiway data-analysis, Chemom. Intell. Lab. Syst. 28 (1995) 229–237.
[35] H.A.L. Kiers, Towards a standardized notation and terminology in multiway analysis, J. Chemom. 14 (2000) 105–122.
[36] S. Leurgans, R.T. Ross, Multilinear models: applications in spectroscopy, Stat. Sci. 7 (1992) 289–319.
[37] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171.
[38] P. Paatero, A weighted non-negative least squares algorithm for three-way 'PARAFAC' factor analysis, Chemom. Intell. Lab. Syst. 38 (1997) 223–242.
[39] R. Bro, Multi-way Analysis in the Food Industry (Doctoral Thesis), University of Amsterdam, The Netherlands, 1998.
[40] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemom. 17 (2003) 274–286.
[41] http://www.mcrals.info.
[42] 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.
[43] http://www.models.kvl.dk/source/.
[44] L. Xiaoli, S. Chanjun, Z. Binxiong, H. Yong, Determination of hemicellulose, cellulose and lignin in Moso bamboo by near infrared spectroscopy, Sci. Rep. 5 (2015) (Article number: 17210).
[45] L. Galvez-Sola, R. Moral, M.D. Perez-Murcia, A. Perez-Espinosa, M.A. Bustamante, E. Martinez-Sabater, C. Paredes, The potential of near infrared reflectance spectroscopy (NIRS) for the estimation of agroindustrial compost quality, Sci. Total Environ. 408 (2010) 1414–1421.
[46] M. Wiggli, A. Smallcombe, R. Bachofen, Reflectance spectroscopy and laser confocal microscopy as tools in an ecophysiological study of microbial mats in an alpine bog pond, J. Microbiol. Methods 34 (1999) 173–182.
[47] D.A. Bryant, N.U. Frigaard, Prokaryotic photosynthesis and phototrophy illuminated, Trends Microbiol. 14 (2006) 488–496.
[48] R. Tauler, I. Marques, E. Casassas, Multivariate curve resolution applied to three-way trilinear data: study of a spectrofluorimetric acid–base titration of salicylic acid at three excitation wavelengths, J. Chemom. 12 (1998) 55–75.
[49] A. de Juan, R. Tauler, Comparison of three-way resolution methods for non-trilinear chemical data sets, J. Chemom. 15 (2001) 749–771.
[50] https://www.smn.gob.ar/.