Journal of Water Resource and Protection, 2019, 11, 244-279
http://www.scirp.org/journal/jwarp
ISSN Online: 1945-3108
ISSN Print: 1945-3094

Evaluation of Dissolved Organic Carbon Using Synchronized Fluorescence Emission Spectra and Unsupervised Method of Principal Component Analysis (PCA) and Independent Component Analysis (ICA)

Tais Cristina Filippe[1], Luana Mayumi Takahasi Marques[1], Heloise G. Knapik[2], Júlio César Rodrigues de Azevedo[1*], Jorge Costa Pereira[3]

[1]Department of Chemistry and Biology, Federal University of Technology—Paraná, Curitiba, Brazil
[2]Department of Hydraulic and Sanitation, Federal University of Technology—Paraná, Curitiba, Brazil
[3]Department of Chemistry, University of Coimbra, Coimbra, Portugal
Email: *jcrazevedo@utfpr.edu.br

How to cite this paper: Filippe, T.C., Marques, L.M.T., Knapik, H.G., de Azevedo, J.C.R. and Pereira, J.C. (2019) Evaluation of Dissolved Organic Carbon Using Synchronized Fluorescence Emission Spectra and Unsupervised Method of Principal Component Analysis (PCA) and Independent Component Analysis (ICA). *Journal of Water Resource and Protection*, **11**, 244-279.
https://doi.org/10.4236/jwarp.2019.113015

Received: January 21, 2019
Accepted: March 4, 2019
Published: March 7, 2019

Abstract

Dissolved organic matter (DOM) can be originated from autochthonous or allochthonous sources, where allochthonous DOM can be from pedogenic sources (humic substances—HSs) or anthropogenicsources (wastewater). The analysis of fluorescence emission, excitation, synchronous or excitation-emission matrix (EEM) have been used to identify the main source or probable contribution of dissolved compounds, such as humic acids (HA), fulvic acids (FA) and dissolved organic carbon (DOC) from sewage, but does not quantify. Fluorescence emission is a powerful technique to detect and qualify organic dissolved compounds but fails in quantitative aspects. In this work, we propose an in situ method for direct determination of DOC using synchronous fluorescence spectra with independent component analysis (ICA). Well known standard solutions were used for method development and validation. In this work, we show that it is possible to predict the number of independent contributions using an unsupervised method based on iterative Principal Component Analysis and Independent Component Analysis (PCA-ICA) approach over combined matrix results. Within these results it's also possible to see that with a very small amount of independent components it is possible to describe environmental samples of HA, FA and primary productivity (PP).

Keywords

Independent Component Analysis, Dissolved Organic Carbon, Spectra Deconvolution, Synchronized Fluorescence

1. Introduction

Disordered population growth in urban and rural areas has changed the water quality in different aquatic environments (rivers, lakes, ponds, streams). Water quality in rivers and lakes is related with the physical, chemical, and biological characteristics of an aquatic environment. Human occupancy has caused abiotic changes (nutrient cycling, metals and organic matter) and biotic. To evaluate and follow these changes, it is necessary to perform a distinct monitoring strategy in aquatic environments. However, Brazilian hydrographic basins are huge, where a more detailed monitoring is difficult or almost impossible because of the number of points to be sampled, the distances between points, the laboratory structure, and the necessary costs.

In the same context of water quality deterioration, eutrophication is a severe environmental damage and is essentially related with the entry of nutrients (nitrogen—N, and phosphorous—P) in water bodies. Consequently, it also becomes important to monitor the variations and origin of organic matter (OM) in aquatic environments, which can be among the main factors of biotic functioning of these ecosystems [1] [2] [3] [4].

It is important to highlight that the dissolved organic matter plays several important and beneficial roles in aquatic ecosystems: buffer with water low levels of alkalinity, nutrients transporting and cycling, controlling metals bioavailability and their speciation, organic pollutants solubility, and acting as light screen in the attenuation of ultraviolet radiation [5] [6], becoming an important parameter for the evaluation of possible contamination of these environments. Thus, organic matter (OM) is a relevant parameter for the determination and evaluation of different organic pollution in aquatic environments.

Biological oxygen demand (BOD), chemical oxygen demand (COD) and dissolved organic carbon (DOC) analyses supply quantitative information on the organic matter present in the aquatic ecosystems. BOD, a time-consuming analysis, focuses on the biodegradable fraction of the organic matter and may present subjective results due to inhibitors, biological condition, temperature and sampling storage. COD and DOC are both rapid techniques, but while COD can overestimate the organic content due to the chemicals used, DOC do not allow the distinction between the labile and refractory fractions. However, organic matter results from a complex mixture of substances with different structural compositions, indicating the need for a qualitative analysis to provide information to identify their sources, transformation, and degradation mechanisms [4] [5].

In recent years, there has been made considerable effort to characterize DOM, which is generally used to describe a broad group of organic compounds in all natural waters. Additionally, DOM in aquatic environments is subject to a series of events and origins [5], which implies a different approach for its properly determination. DOM can be originated from autochthonous or allochthonous sources. Autochthonous originates, for example, from the activity of phytoplankton or aquatic weeds. The allochthonous DOM can be from pedogenic sources (humic substances—HSs) or anthropogenic sources, such as industrial or domestic wastewater [2] [3] [5] [7].

The isolation of the dissolved substances, such as HSs is costly and time-consuming. After extracted, to obtain information about part of the structure of SHs (FA and HA), it is necessary to perform NMR-C13 analysis, NMR-H, FTIR, TGA, potentiometric titration, elemental analysis, and others.

Alternative techniques that utilize chemical differences such as molecular weight and molecular size are important for its characterization, because within the use of this technique it is possible to estimate the degree of complexity and its possible origin or source in the aquatic environment [1] [6].

Among the techniques available for the differentiation of fractions of the DOM in aquatic systems, it is important to highlight the spectroscopic absorption (in the ultraviolet to visible region) and the fluorescence emission. The main advantages of molecular fluorescence emission spectroscopy are its high sensitivity [8]. Another advantage is the variety of spectra that can be obtained: emission, excitation, synchronized, and emission-excitation matrix [1] [5] [6] [7] [8] [9] [10]. Complementarity, fluorescence spectrophometry is a relatively low cost technique and requires small amount of sample, small sample treatment such as filtration and addition of sample buffer, is a rapid analysis and is a non-destructive sample method [11].

Both visible ultraviolet and fluorescence spectrophotometry have been used by several researchers to identify DOM possible sources in the aquatic environment and to characterize the structural composition of dissolved organic carbon [1] [4] [6] [7] [8] [12].

Within the synchronized emission spectra it is possible to identify different peaks, such as humic substances with maximum intensity, usually, near 450 nm (fulvic acids) and between 465 and 500nm (humic acids) [8] [13]. In the region of 280 - 310 nm ($\Delta\lambda$ = 18 nm) the emission is attributed mainly to the aromatic amino acids and other volatile acids containing highly conjugated structures [13]. Studies conducted by Yu *et al.* [2] [4], Chen *et al.* [6], Ahmad and Reynolds [7] Pons *et al.* [9] emphasize the presence of peaks in this region (280 - 310 nm emission) to existing compounds in wastewater, like tryptophan and tyrosine. Rivers influenced by human activity may be characterized by high intensity region of similar proteins and fulvic acids [14]. In the same context, studies conducted by [15] showed that the fluorescence intensity related to fulvic acids and proteins is significantly higher in areas downstream sewage treatment than its upstream stations.

Studies have shown an efficient evaluation of DOM through fluorescence emission both in estuarine and ocean areas [16] such as rivers, lakes and sewage [1] [5] [6] [7] [9] [10] [17].

Some forms of data interpretation were developed, such as: the location distinct peaks [16], the integration of the excitation-emission matrix [6], the association of chemometric tools such as principal component analysis [18], by parallel analysis factors (PARAFAC) [19] and the use of parallel factors associated with other chemometric techniques [20], second derivative spectra of the synchronized [2] and analysis by excitation spectra and the emission spectra for ICA application standards for evaluating mixing solutions of aromatic (biphenyl, naphthalene and benzotriazol) compounds [21].

Considering these aspects, this paper summarizes some studied spectra of humic acid, fulvic acid, primary productivity, tyrosine and tryptophan, with the objective of developing a methodology to obtain a predictive model for determining DOC in field water samples based on synchronous fluorescence and independent component analysis (ICA).

2. Materials and Methods

2.1. Solutions and Sample Preparation

Considering the representativeness of environmental samples, humic Acid (HA) and fulvic Acid (FA) used for analysis were collected near the Patos Lagoon (Mato Grosso do Sul, Brazil) and extracted from the sediment (four samples) of the lagoon and also from the soil (two samples) near the pond. Soil and sediment samples were collected at the depth of 20 cm within a modified Petersen dredge. The samples were homogenized and transferred into sample bags (Whirl Pack) and preserved at −20˚C. Sediment and soil humic substances were extracted according to International Humic Substances Society (IHSS).

Leaves, branches and roots were removed from the collected sediment and soil. First, to obtain a pH equal to 1, HCL 1.0 $ml{\cdot}L^{-1}$ was added for each 100 g of sediment sample. Then, HCl 0.1 $ml{\cdot}L^{-1}$ was added while maintaining the ratio of 10 ml HCl per gram of sample. The suspension was stirred for 1 h and the supernatant separated by centrifugation. This supernatant was denominated "fulvic acid-I (FA-I)". The precipitate was neutralized with NaOH 1.0 $ml{\cdot}L^{-1}$ at pH 7. The following was added NaOH 0.1 $ml{\cdot}L^{-1}$, under N_2, until the ratio of 10 mL of NaOH/g of sample. The extraction was conducted with intermittent stirring overnight under N_2. The supernatant was separated by centrifugation, discarding the residue (humin). The supernatant was acidified to pH 1.0 with HCl 6.0 $ml{\cdot}L^{-1}$ with constant stirring and left to stand for one night. Again separation was performed by centrifugation, separating the precipitate (Humic Acid—HA) of the supernatant, which was called "Fulvic Acid II (FA-II)".

The fulvic acid was percolated on a column containing DAX-8 resin, at a flow rate of 15 times the volume of resin per hour. The effluent was discarded and the column was washed with deionized water, in the quantity of 0.65 column volumes.

The fulvic acid was removed by retro eluition, with NaOH 0.1 ml·L^{-1}, adding the amount equivalent to a volume of the column, followed by 2 to 3 column volumes of deionized water. Immediately acidified with HCl solution 6.0 ml·L^{-1} to pH 1.0. Then, HF was added to a final concentration of 0.30 ml·L^{-1}. The same procedure AF-I was applied to the supernatant called AF-II.

All eluates were reunited, remixed and reapplied through the column (DAX-8). Again the retro elution was carried out with the addition of NaOH 0.1 ml·L^{-1}. The final eluate was passed through a column containing a strong cationic resin in H^+ ions, using three times the number of moles of Na^+ ions contained in the solution. The eluate, containing fulvic acid was dried by lyophilization. These extracted FAs were prepared as described below.

The precipitate containing the humic acid (HA) was resuspended in KOH0.1 ml·L^{-1} (redissolved) under a nitrogen gas flow; solid KCl was added to obtain 0.3 ml·L^{-1} concentration in added K^+ ions. The mixture was centrifuged to remove suspended solids. HA was precipitated withthe addition of HCl 6.0 ml·L^{-1}, at pH 2, imposing constant stirring and left to stand for one night. The humic acid was separated by centrifugation (30,000 rpm, 30 min). The supernatant was discarded and the precipitate (HA) was transferred to a plastic container which was added HCl 0.1 ml·L^{-1} + HF 0.3 ml·L^{-1} (stirring for one night). This procedure was repeated and the supernatant was separated by centrifugation. The supernatant was discarded and the residue (AH) purified in a dialysis tube (12,000 Daltons) with deionized water until obtaining a negative test for chloride (three months). The humic acid sample was dried by lyophilization. These extracted HAs were prepared as described below.

The extraction of the aquatic fulvic and humic acids were was performed in 150 L of filtered water (0.45 mm), acidified (pH 2.0) with HCl. The 150 L of water was passed through a column containing DAX-8 resin, at a flow rate of 15 times the volume of resin per hour. The same procedure described above was employed.

Microalgae strain, *Scenedesmus subspicatus*, was provided by the laboratory from Instituto Ambiental do Paraná (IAP). Cultivation was carried out in Erlenmeyer flasks, controlling the temperature (25˚C) and light incidence. Nitrogen, phosphorus, and potassium (NPK) and micronutrients were added; cell culture was kept in movement (shaker type system). The samples were filtered and the filtrate, containing DOC derived from primary productivity, was used as solutions of primary productivity (PP).

Environmental samples (HA, FA, and PP) and commercial humic acid (Sigma-Aldrich) were prepared in buffer solution in order to obtain DOC concentrations ranging from 4.9 to 28.9 mg·L^{-1} for HAs, 4.5 to 38.2 mg·L^{-1} for the FAs and 13.8 to 48.5 mg·L^{-1} for the PP. Solutions for spectrophotometric study were prepared with 10 mM phosphate buffer (pH = 7.0 ± 0.1) using high quality water (ultrapure).

Standard solutions of tyrosine (#CAS 60-18-4; Mw 181.19 g mol^{-1}; pKa's 2.10, 9.21 and 10.46), tryptophan (#CAS 73-22-3; Mw 204.23 g mol^{-1}; pKa's 2.46 and

9.41), and phenanthroline (#CAS 66-71-7; Mw 180.21 g·mol^{-1}; pKa 4.27) were prepared using high purity (≥99%) analytical grade standards (Sigma-Aldrich) with no further purification process. Standard solutions were initially prepared with concentrations ranging from 0.4 to 40.0 mg C L^{-1}.Binary and ternary standard mixtures were prepared using tyrosine concentrations ranging from 0.284 to 2.13 mg C/L; tryptophan between 0.778 to 20.75 mg C L^{-1} and phenanthroline ranging 0.853 to 5.69 mg C L^{-1}.

2.2. Sample Characterization

The fluorescence spectra of standards (humic acid, fulvic acid or PP) were obtained in 1 cm quartz cuvettes (Cary Eclipse Fluorescence Spectrophotometer). Synchronous fluorescence spectra were determined using excitation (λ_{exc}) over the range of 250 - 600 nm, starting the emission scan at 268 nm ($\Delta\lambda = \lambda_{exc} + 18$ nm), often applied in the studies of humic substances, natural organic matter or wastewater [13] [17]. Thespectra were obtained applying the following conditions: slits 5 nm, scan speed of 240 nm/min, intensity of the Raman peak of ultrapure water (Ex/Em 275/303 nm) was used to examine changes in the fluorescence intensity signal and to normalize the data, all spectra were subtracted from the spectrum of water, spectrophotometric blank sample was obtained using phosphate buffer in the samples compartment. Spectral data was recorded under same experimental conditions and spectra were organized into specific matrices related to each case study.

As the features of fluorescence spectra may be changed according to the pH and ionic strength of the medium [13] [17] was added phosphate buffer (pH = 7.0 ± 0.1).

Each sample was characterized in terms of DOC determination and the respective spectrum recorded. The DOC for environmental samples (humic acid, fulvic acid and PP) was determined according to the methodology proposed by the manufacturer of Total Organic Carbon Analyzer (TOC 5000-A, Shimadzu). Values of the DOC present in standard solution were obtained by direct weighing.

Data treatment was made using two software packages: GNU Octave high-level interpreter language for calculations and plots and R-project "fast ICA" package for independent component analysis.

2.3. Data Treatment: Fundamentals

Under optimal conditions, solution fluorescence spectral information at each i wavelength can be viewed as an additive composition of independent light emissions (e_{ik}) of q fluorescent species present in solution at different concentration levels (c_k)

$$f_i = g\sum_{k=1}^{q} e_{ik}c_k \quad (1)$$

where g is a nonspecific constant depending upon experimental conditions and for that reason will be discarded for further response modelling purposes.

Combining all spectra information for $i=\{1,\cdots,n\}$ recorded spectra wavelengths of $j=\{1,\cdots,m\}$ solutions in a unique data matrix we obtain

$$F = EC \tag{2}$$

were F represents the overall information matrix assembled as m column fluorescence spectra vectors.

Independent component analysis is a numerical iterative method for finding underlying factors or components from multivariate (multidimensional) statistical data [22].

The origins of Independent Component Analysis are related with studies in the context of neural network modelling in early 1980s [22] [23] but it was probably firstly reported by [24].

As indicated, Independent Component Analysis try to decompose a complex signal system (F) into a linear combination of signal sources (S) convoluted with a mixing matrix (A)

$$F = SA \tag{3}$$

Equations (2) and (3) are notably similar and thus ICA may be helpful in retrieving information from unknown complex light emitted mixture spectra.

In order to perform this deconvolution ICA uses orthogonal negentropy approximation to maximize independent component contributions.

As many other algorithms that works with information variability, ICA algorithm starts with a previous variable centring step in order to enter into the variability dimension.

Defining a specific number of independent components to be retrieved ($c \le m$), column centred data matrix is then "whitened" by projecting the data onto its c principal component directions

$$F_c = FK \tag{4}$$

where F_c matrix is a compressed version of centred data matrix into a c sub-space.

Using the negentropy approximation ICA estimates the "un-mixing" orthogonal matrix (W) in order to estimate source signals ($\hat{S}$)

$$\hat{S} = F_c W \approx S \tag{5}$$

This estimated sources correspond to a representation of original signal sources, Equation (3), in "c" subspace.

If the number of imposed components (c) matches the number of emitting species (k), theoretically ICA will be able to find out specific component spectral emitting contributions (E), Equation (1).

In a previous work [25], we described the methodology to use ICA in order to retrieve spectral contributions ($\hat{S}$) and with this spectral information estimate respective concentration profiles in solution solving eq.1 for mixtures composition (Eq. 6)

$$\hat{C} = \left[\hat{S}^{\mathrm{T}}\hat{S}\right]^{-1}\hat{S}^{\mathrm{T}}F \tag{6}$$

After this estimation process, we developed polynomial models able to describe DOC results in several standards and sample solutions.

In this work we are interested in directly use ICA estimated mixing matrix (A), Equation (3), as a component contribution information, Equation (1), in order to estimate DOC via polynomial models.

Since in real situations the number of emitting species (k) and their specific emission profiles are unknown, different criteria was used in order to verify if ICA retrieved information was able to describe each considered system.

ICA ability to recover initial spectral information was evaluated with residual relative error (%RE):

$$\%\mathrm{RE} = \left(\overline{f}\right)^{-1} \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\hat{f}_{ij} - f_{ij}\right)^2} \tag{7}$$

where $\hat{f}_{ij}$ and f_{ij} represents the estimated and original spectra and $\overline{f}$ represents the mean overall fluorescence intensity.

In our experience with ICA, we found out that estimated mixing matrix (A) can lead to erroneous conclusions since increasing the number of imposed components (c) we began to have correlated information, revealing that we are using redundant information.

Another strategy is to use Principal Component Analysis (PCA); PCA [26] is a very basic and important tool in chemometrics—it is able to compress all information into essential relevant factors, discarding redundant information.

Imposing PCA analysis to estimate the mixing matrix (A)

$$A = \alpha\Lambda\theta^{\mathrm{T}} \tag{8}$$

we can obtain the respective scores (α) and loadings (θ). Eigenvalue matrix (Λ) was used to access and evaluate retrieved information index from p most relevant eigenvalues

$$\%\mathrm{Rec}(p) = 100\left(\frac{\sum_{l=1}^{p}\lambda_l}{\sum_{l=1}^{q}\lambda_l}\right) \tag{9}$$

After solving the problem of estimating the best number of components necessary to describe the fluorescence spectra information, it is necessary to consistently define the best model able to describe the predicted DOC values of each mixture.

Assuming least squares standard approach (nonstochastic dependent variable with additive normal independent error on dependent variable) [27], polynomial equations (η) were used to describe dependent variable (y_i) and best parameter estimates were obtained in order to minimize model error (SS_{res})

$$SS_{res} = \sum_{i=1}^{n}\left(y_i - \eta_i\right)^2 \tag{10}$$

Two basic polynomial approaches were used. The simpler approach was to consider a first degree polynomial approach ($\eta(1)$) were

$$\eta(1)_i = b_0 + \sum_{i=1}^{c} b_i x_i \tag{11}$$

using $p = c + 1$ parameter model accounting for c predictors (x_i) and including a constant parameter (b_0).

Second approach was based on a full second degree polynomial model ($\eta(2)$)

$$\eta(2)_i = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i=1}^{k} \sum_{j=1}^{k} b_{ij} x_i x_j \tag{12}$$

with a total of $p = \left(c^2 + 3c + 2\right)/2$ parameters, which contains a constant coefficient (b_0), c individual first degree contributions (b_i), c individual second degree contributions (b_{ii}) and $\binom{c}{2}$ combined (b_{ij}) responses.

Modelling is related with defining a good function that is able to follow experimental data and also may be used in order to predict results. For this purpose different indicators were used.

Model fitting ability was evaluated by relative residual error (%RE), R-squared index (R^2), parameter statistical significance and Akaike information criterion, while model predicting ability was evaluated with cross-validation techniques related with Jack knife resampling and data k-folding.

Mean residual error (σ_{fit})

$$\sigma_{fit} = \sqrt{\frac{SS_{res}}{n - p}} \tag{13}$$

is the most common estimative to evaluate model bias in the fitting process.

Since there are different ranges for the dependent variable, the decision was to compute this estimate as a relative residual error, in respect to dependent variable central estimate ($\overline{y}$)

$$\%\text{RSE} = 100\left(\frac{\sigma_{fit}}{\overline{y}}\right) \tag{14}$$

avoiding thus scale effects and corresponding misinterpretations.

R-squared statistics (R^2) is based on ANOVA regression analysis assumptions [26] and reveals the total amount of relative information described by the model in fitting a given data set

$$R^2 = 1 - \left(SS_{res} / SS_{tot}\right) \tag{15}$$

where SS_{res} and SS_{tot} correspond to the residual and total sum of squares.

In general, increasing the number of model parameters (p) we also increase the model ability to fit data and thus it is frequent to end with an overparameterized model.

For this reason, when comparing different models, it is crucial to use R-squared adjusted (R^2_{adj}) in order to compensate the impact of extra p parameters

$$R^2_{adj} = 1 - \left(\frac{SS_{res}/(n-p)}{SS_{tot}/(n-1)}\right) = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-p-1}\right) \tag{16}$$

Since we are using linear models, least squares approach provides the best parameter estimates and full statistical support and thus it is possible to evaluate

parameter significance based on respective estimates of position (b_i) and dispersion, $\sigma(b_i)$, considering the null hypothesis of $H_0 : b_i = 0$

$$TV = \frac{|b_i|}{\sigma(b_i)} \le t^b_{\alpha(n-p)} \tag{17}$$

where $t^b_{\alpha(n-p)}$ refers to t-student bilateral critical value at $100(1-\alpha)\%$ confidence level.

Akaike information criterion (AIC) [28] evaluates the relative quality of a statistical model for describing a given dataset. In the case of homogeneous variance experimental error, maximum-likelihood estimation is replaced by unweighed least squares approach and thus AIC assumes a very convenient simple expression (Equation (18)):

$$\mathrm{AIC} = n\log\left(\frac{SS_{res}}{n}\right) + 2p \tag{18}$$

accounting on both—fitting quality (SS_{res}/n) and the number of used parameters. AIC does not provide a statistical test for a given model but it may be used in order to compare their ability in fitting data with minimal parameters.

When working with finite small datasets Equation (18) may originate biased information estimates and thus a sample size correction should be made

$$\mathrm{AIC}_c = \mathrm{AIC} + \left(\frac{2p(p+1)}{n-p-1}\right) \tag{19}$$

Two cross-validation strategies were used in order to evaluate model predicting ability in describing a specific dataset—the conservative "jack knife" and the drastic "unfolding" approach.

Jackknife resampling strategy was used in order to estimate Root Mean Square Prediction Error (RMSPE) here evaluated as a relative estimate to mean response

$$\%\mathrm{RMSPE} = \left(\frac{100}{\bar{y}}\right)\sqrt{\frac{SS_{res(.)}}{n-1}} \tag{20}$$

were $SS_{res(.)}$ represents residual error estimated for "leave-one-out" jackknife resampling strategy.

In order to evaluate maximal prediction error we proceed with a 1/4 "unfolding" resampling strategy—original data set was divided in 4 similar subsets (a, b, c and d) and thus combined in order to obtain 6 different combinations of 50% data each (ab, ac, ad, bc, bd, and cd). Then, each of these subsets were used to calibrate DOC response and then the obtained model was used to estimate DOC in the correspondent unused 50% data set. Prediction Error ($\%\mathrm{PE}_{1/2}$) was thus estimated as

$$\%\mathrm{PE}_{1/2} = \left(\frac{100}{\bar{y}}\right)\sqrt{\frac{\sum_{i=1}^{3n}(\eta_i - y_i)^2}{3n}} \tag{21}$$

When using second degree polynomial models, Equation (12), the number of parameters (p) increases significantly with the number of considered independ-

ent contributions (c), but some of these parameters have a small impact on model fitting ability in describing the dataset. Incremental F-test may be used to test the relevance of a given parameter testing iteratively least significant parameter with

$$TV = \frac{\Delta\sigma_{fit}^2}{\sigma_{pe}^2} = \frac{SS_{res(n-p)} - SS_{res(n-p-1)}}{\sigma_{fit(n-p-1)}^2} \quad (22)$$

If TV exceeds the predicted unilateral Fisher critical value for 1 and (n-p-1) degrees of freedom at 0.01 significance level, null hypothesis ($H_0 : \Delta\sigma_{fit}^2 \leq \sigma_{pe}^2$) is no longer considered as valid, indicating that tested parameter is crucial to ensure a good data fit.

3. Results and Discussion

In a previous study with ICA [25], we discussed the application of spectral signal deconvolution in order to obtain "fluorescence emission specific profiles" and with this adjusted signals estimate respective contribution in terms of concentration and with this information estimate sample charge in terms of dissolved organic carbon.

Now we are concerned in directly using the mixing matrix information (A) in order to estimate accurate component contribution predictors for DOC. In order to better understand ICA results, we studied simple standard solutions and then applied the developed strategy to environmental samples.

3.1. Standard Solutions

First results are related with simple standard solutions in order to develop a procedure for the applicability of the mixing matrix information (A(c × m)) as component information and thus estimate DOC in respective solutions.

For that purpose, small molecular weight conjugated molecules of pure compounds were used as standards in the study of the applicability of ICA processing ability.

In the process of developing an analytical strategy different situations were considered with these standards solutions. A first approach was considering the individual ICA treatment of single solution standards.

Figure 1 presents the synchronous fluorescence spectra. There are 24 spectra at 8 concentration levels (in triplicates) for tyrosine (**Figure 1(a)**—five concentration levels), tryptophan (**Figure 1(b)**—six concentration levels), and phenanthroline (**Figure 1(c)**—five concentration levels).

After this individual standard solutions approach, another test was to combine all single standard spectra solutions in a global data Ensemble (72 spectra of single standard solutions) and compare this results with All SRM sample solutions (246 spectra containing single standard solutions, binary mixtures and ternary mixtures).

Figure 2 presents the spectra related with the Ensemble of 72 single solution

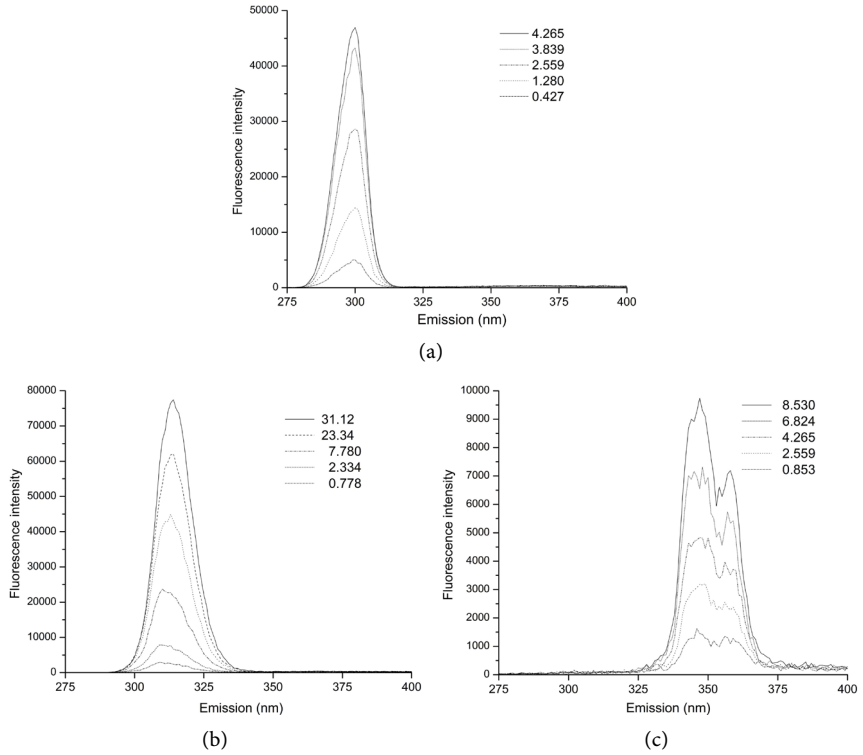

Figure 1. Synchronous fluorescence spectra ($\Delta\lambda = (\lambda ex - \lambda em) = 18$ nm) for single standard solutions each at 8 different concentration levels, in mM, dissolved in 10 mM phosphate buffer (pH 7.0), of (a) tyrosine (Tyr); (b) tryptophan (Tryp); and (c) phenanthroline (Phen).

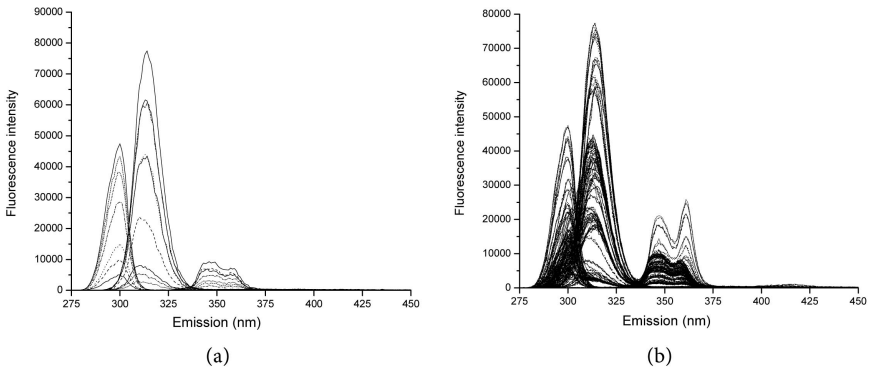

Figure 2. Synchronous fluorescence spectra ($\Delta\lambda = 18$ nm) for the Ensemble of spectra (a) single components and (b) All SRMs (single, binary and ternary mixtures of tyrosine, tryptophan, and phenanthroline).

spectra and All SRMs spectra (72 single components, 44 binary, and 130 ternary mixtures in a total of 246 spectra).

ICA is a very powerful analytical tool in deconvoluting mixed information [22], but it requires the previous definition of the number of components (c) that may be accounted in order to correctly describe relevant spectral information, which, in real situations is not an obvious task.

Working with the simpler situation (single standard solutions), a first approach

was to start ICA analysis imposing a small number of components (c = 1) and then successively increasing this number until achieve an excessive number of components (c = 5). By inspecting respective spectral component contributions profiles (S) we could estimate the number of independent components (**Figure 3**).

From **Figure 3** it is clear that imposing an excessive number of components significant noise increases and spectral profiles becomes progressively deformed when comparing with original spectral signals, cf. **Figure 3**.

From these figures it is also evident that deconvoluted signal sources are estimated regardless its actual orientation—for example, residual signal contribution identified in **Figure 3(b)** as c = 2 is the same as c = 3 in **Figure 3(c)** but is inverted on **Figure 3(d)** (source c4) and in **Figure 3(e)** (source c = 5).

In this simple single standard solutions, it can be observed that for tyrosine (**Figures 3(a)-(e)**) and phenanthroline (**Figures 3(k)-(o)**), a single component is sufficient to recover all relevant spectral information—the second retrieved contribution presents essentially residual spectral information.

In the case of tryptophan (**Figures 3(f)-(j)**), there is evidence for a need of two spectral contributions in order to justify the fluorescence band—the third component reveals residual signal contribution.

Figure 4 presents the information contained in mixing matrix (A) for the cases presented in **Figure 3**. Since this information is related with each component contribution we represented them in respect to the concentration of dissolved organic carbon (DOC, mM). From **Figure 4** it is possible to observed different scenarios related between mixing matrix information and actual solute concentration—for tyrosine case, positive slope linear contributions, nearly invariant and negative slope linear dependencies.

Comparing with respective signal sources, **Figure 3**, in tyrosine's case, v1 is always related with main signal at 300 nm. In **Figures 3(a)-(d)**, deconvoluted signal is typically upwards oriented while in **Figure 4(e)** it was downward oriented; consequently, in **Figure 4** A to 4D mixing information is positive while in **Figure 4(e)** is now negative.

Like other deconvolution techniques as PCA, with ICA there is this orientation difficulty—deconvoluted components are correct in modulus but not necessarily well defined in terms of signal. Since original data (spectra) were previously centred, it is obvious the tendency for some negative values in mixing matrix and, in real and more complex systems it is obvious a huge difficulty in finding the correct orientation of signal sources and respective contributions.

In order to increase system complexity and test ICA for its ability in deconvoluting signals we have considered an Ensemble case (were all previous single standard spectra were assembled in a single data matrix) and new mixtures were prepared (binary and ternary mixtures), composing a final data matrix by Ensemble matrix and these new actual mixtures spectra and named as All SRM case.

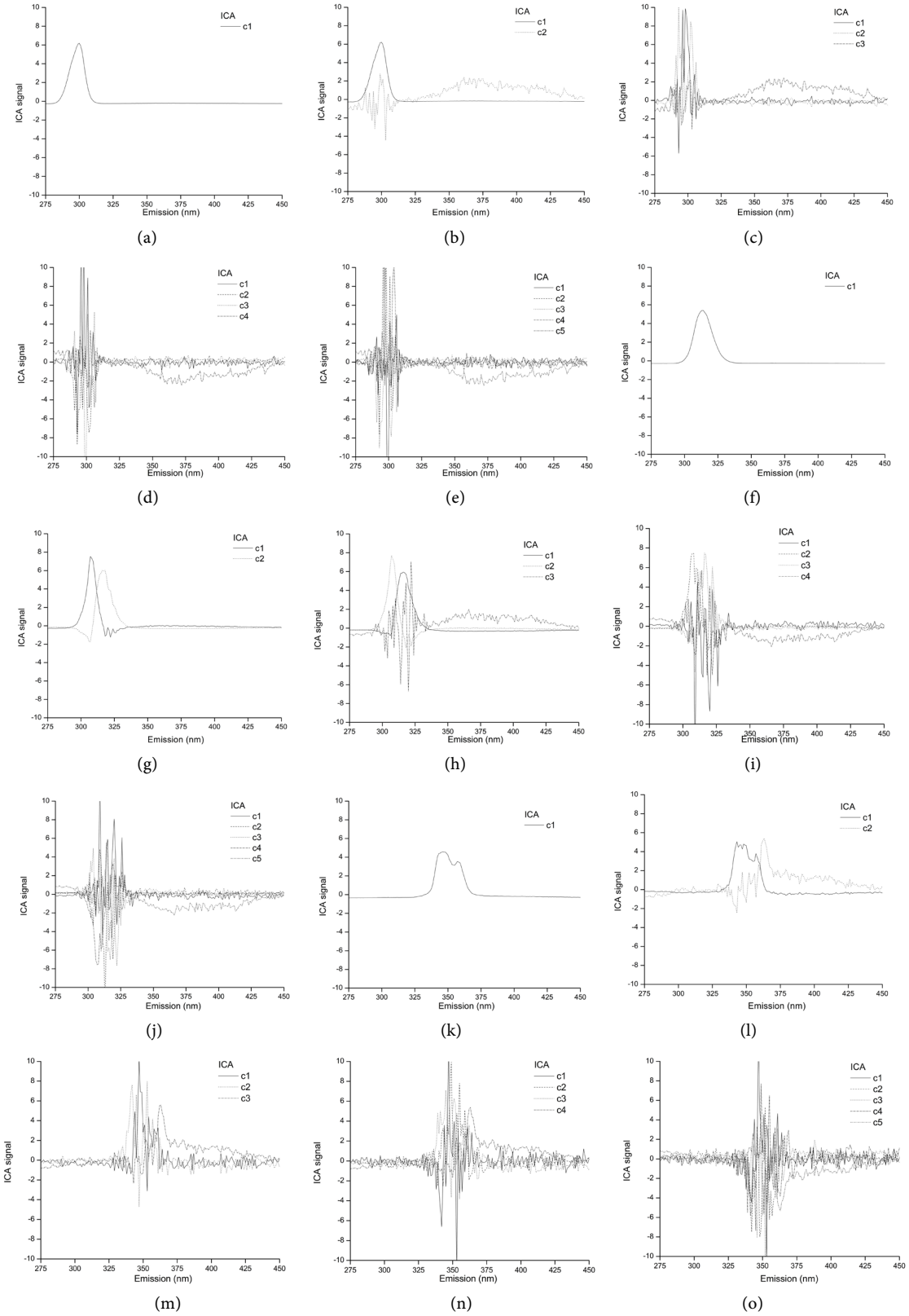

Figure 3. Evolution of spectral component contributions with the number of imposed components (c) in ICA analysis, from first component (c = 1, on top) to fifth component (c = 5, bottom), of single standard solutions of tyrosine ((a)-(e)), tryptophan ((f)-(j)) and phenanthroline ((k)-(o)).

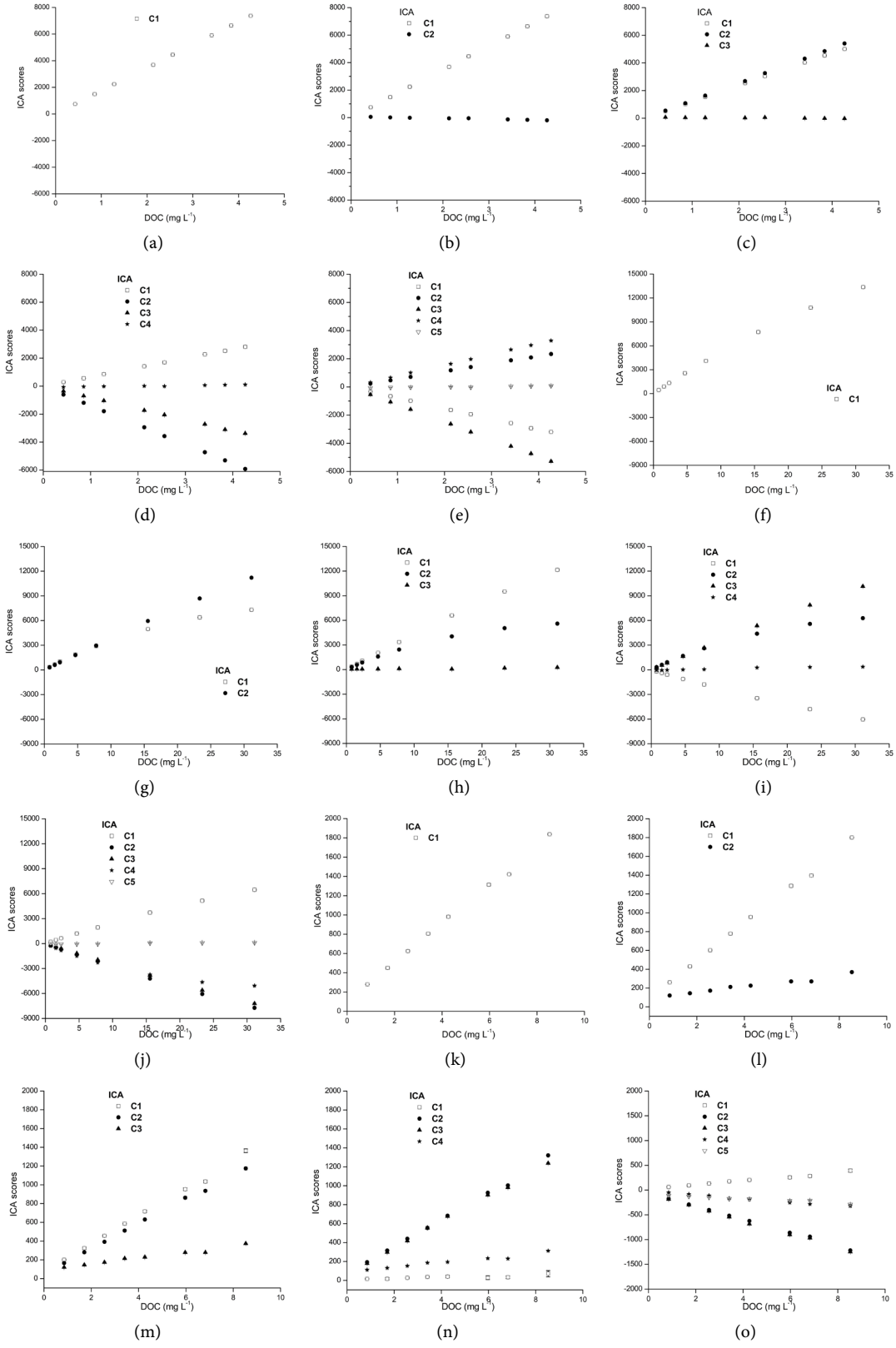

Figure 4. Evolution of spectral component contributions with the number of imposed components in ICA analysis (c), from first component (c = 1, on top) to fifth component (c = 5, bottom), of single standard solutions of tyrosine ((a)-(e)), tryptophan ((f)-(j)) and phenanthroline ((k)-(o)).

Figure 5 and Figure 6 present spectral contribution deconvolution and respective component contributions obtained by imposing c = 2 till c = 6 components in ICA deconvolution of the ensemble case and all SRM's solutions.

From Figure 5 it is possible to observe that these simple situations become more difficult to analyse when including in simultaneous different solutes and a large number of samples. However ICA is able to correctly extract each component contribution and thus be able to identify individual contributions—comparing spectral contributions in Ensemble case, c = 1 in Figures 5(a)-(d) was estimated in Figure 5(e) in downward mode; looking into All SRM case, same signal is now associated with c = 2 (Figures 5(f)-(h)) and c = 4 (Figure 5(i) and Figure 5(j)) are both related with tyrosine contribution given by c = 1 in Figure 3(a) and Figure 3(b).

As previously stated, increasing system complexity ICA maintains its blind source deconvoluting ability but source signal recognition and its correct orientation becomes harder to do with the inconvenient of dealing with incorrect mixing information dependency. In addition, as it can be observed, increasing the number of imposed independent contributions in ICA noisier signal sources are obtained and sometimes artificially fragmented into complementary signals. When dealing with real environmental sample solutions this fact will be difficult to circumvent since we do not know previously the number of components to set on ICA processing step in order to correctly retrieve independent components and respective contributions.

For this reason different evaluating situations were tested in order to define the correct number of independent components.

Combining spectral components (S) with component contributions (A), Equation (3), and calculating the relative residual error, Equation (14) imposing p = 0, between centred spectra and reproduced spectra we are able to compute a mean bias estimative for each deconvolution case. Obtained results are presented on Figure 7.

From Figure 7, if we consider the simpler case (single standard solutions in Figure 7(a)), it is possible to observe that there is a basal evolution of residual error in respect to the number of imposed ICA components—decreasing c there is a smooth increased tendency for increasing relative residual error. Like in PCA scree plots, a drastic positive shift in relative error evidences important information suppression and thus defining the number of required components to describe data matrix as the previous in-line value. With this diagnose in mind, from Figure 7(a) it is easy to establish c = 1 for tyrosine and c = 2 for tryptophan. However, there is some evidence for using c= 1 for phenanthroline.

With this fact in consideration, we guess for Ensemble and All SRM's cases a total of c = 4 independent components. Looking now at Figure 7(b), in the Ensemble case c = 4 is the evidence and for All SRM's case the plausibility of c = 4 or c = 5 is not well defined.

From Figure 4 we can also see that imposing an excessive number of components in ICA deconvoluting process, component contribution matrix starts to

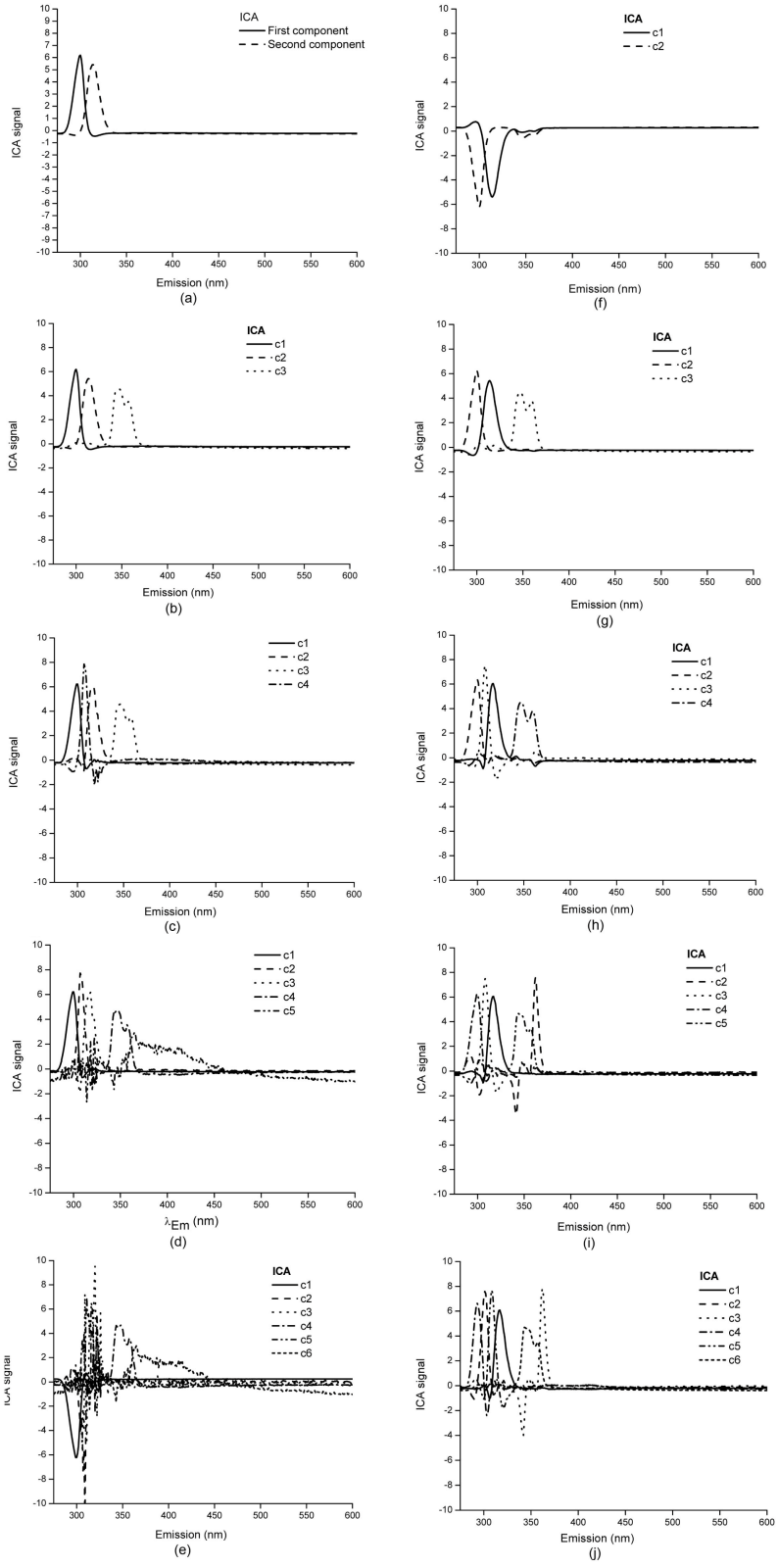

Figure 5. Evolution of spectral component contributions with the number of imposed components in ICA analysis (c), from second component (c = 2, on top) to sixth component (c = 6, bottom), of the Ensemble of single standard solutions ((a)-(e)) and in All SRM case ((f)-(j)).

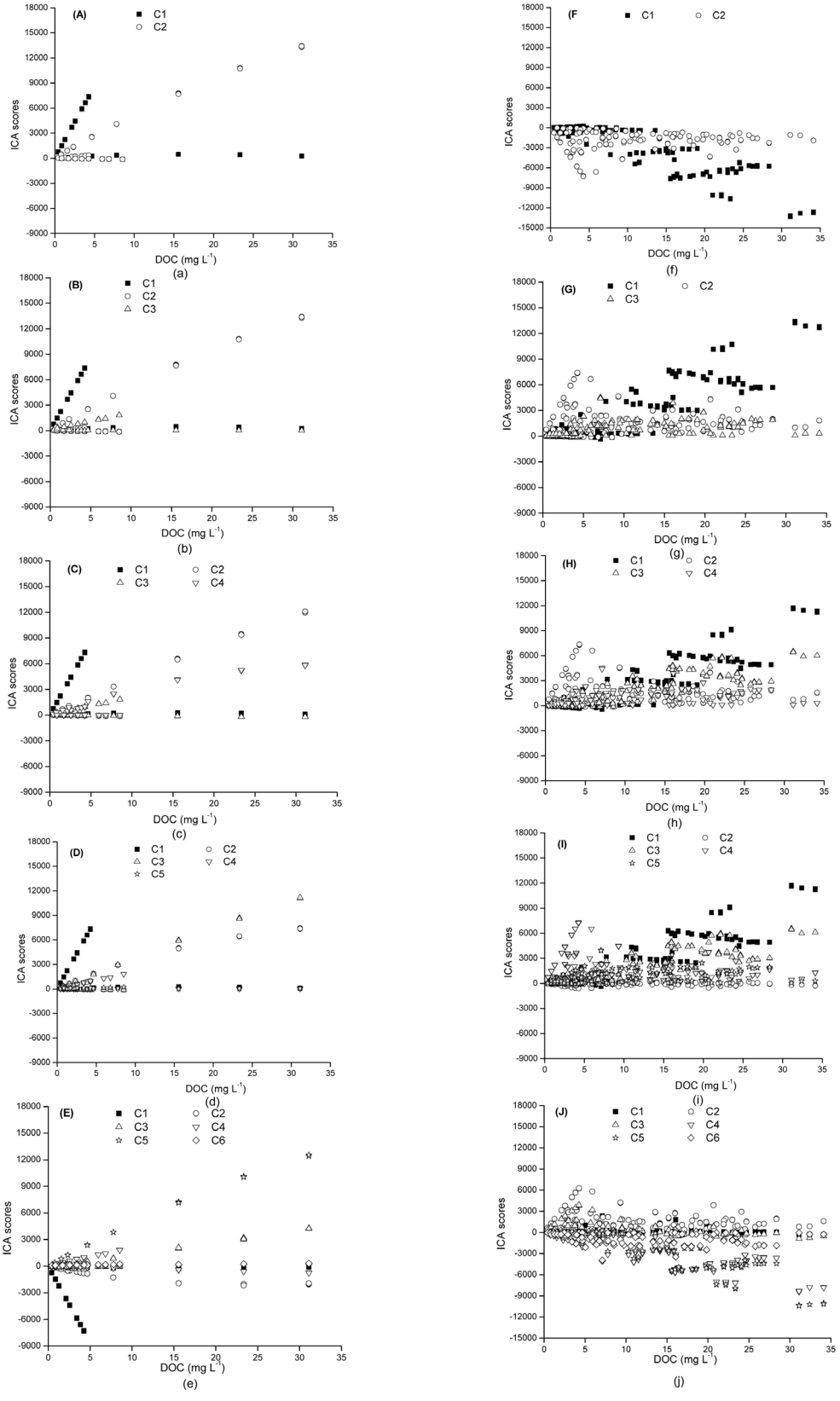

Figure 6. Evolution of component contributions with the number of imposed components in ICA analysis (c), from second component (c = 2, on top) to sixth component (c = 6, bottom), of the ensemble of single standard solutions ((a)-(e)) and in all SRM case ((f)-(j)).

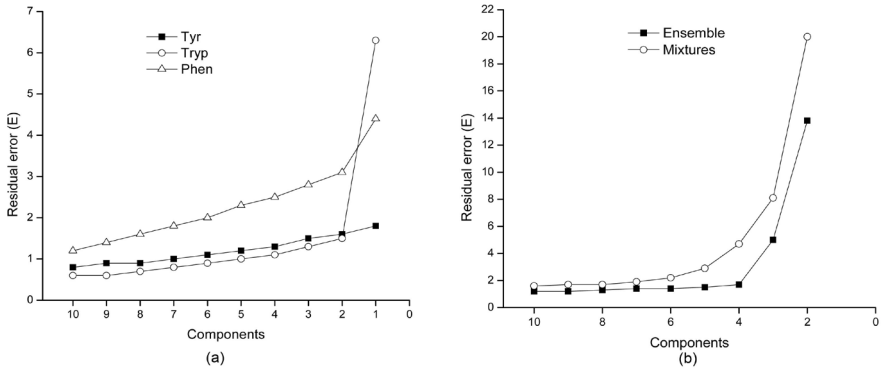

Figure 7. Residual error (%RE) evolution in respect to original spectra recovery for (a) single SRM solutions and (b) for Ensemble and All SRMs.

reveal significant correlation between estimated components. With this in mind it was crucial to develop and establish an unsupervised method for information condensation and thus refine the number of effective relevant contributions.

For that reason we believe that principal component analysis may be useful in dealing with this situation, predicting the correct number of independent components for each case. In Table 1, we present PCA retrieved information index, Equation (9), of ICA's mixing matrix for single standard solutions (Tyr, Tryp, and Phen) in terms of considered PCA components (columns) at each imposed c for ICA analysis (in rows).

From Table 1 it is clear that this is a very simple system (a single standard solution case), since over 80% information recovery may be obtained with a single component for all tested cases.

Since we are interested in describing all relevant information in mixing matrix we should stipulate a very high recovery performance, different from 100%, in order to obtain a certain constant value, related with the effective number of independent components.

In this perspective, imposing a minimum of 99.5% information recovery it is possible to find a more or less consistent pattern—tyrosine and tryptophan are consistently pointing out to c = 2 independent components when previous ICA analysis imposes c < 15. For phenanthroline the suggestion varies between 2 and 3. In the same way, imposing an excessive number of components (c > 15) different suggestions are also obtained.

This phenomenon is related with the increase of basal noisy contribution in respect to the increase of imposed c components in ICA analysis.

In order to clarify the correct choice of the number of independent components to be used in ICA, we suggest an iterative process where convergence of c may be obtained. We recommend to start imposing a high value in c (e.g.: c = 20) and perform ICA analysis; perform PCA analysis on ICA obtained mixing matrix and estimate the number of relevant contributions (f1) able to recover at least 99.5% information (Table 2).

On the next iteration, start imposing c2 = f1 in ICA analysis and proceed again with PCA analysis over mixing matrix obtaining f2.

Table 1. PCA information recovery of ICA mixing matrix for single standard solutions (Tyr, Tryp, and Phen): in lines we present each ICA case (imposing c = 20 to c = 2).

	c	1	2	3	4	5	6	7	8	9	10
Tyrosine	20	95.1	98.0	99.1	99.3	**99.5**	99.7	99.8	99.9	99.9	99.9
	15	92.2	98.8	99.3	**99.7**	99.9	99.9	99.9	100		
	10	98.8	**99.7**	99.8	99.9	100					
	8	87.7	**99.9**	99.9	100						
	7	90.6	**99.9**	100							
	6	92.2	**99.9**	100							
	5	98.4	**99.9**	100							
	4	98.6	**100**	100							
	3	94.6	**100**	100							
	2	99.1	**100**								
Tryptophan	20	96.5	99.3	**99.7**	99.8	99.9	99.9	99.9	99.9	100	100
	15	92.8	**99.8**	99.9	99.9	99.9	100				
	10	93.2	**99.7**	99.9	99.9	100					
	8	98.2	**99.5**	99.9	100						
	7	98.7	**99.5**	100							
	6	98.6	**99.5**	100							
	5	98.5	**99.5**	100							
	4	99.0	**99.7**	100							
	3	92.2	**99.7**	100							
	2	99.4	**100**								
Phenanroline	20	86.9	92.2	94.9	96.6	97.3	98.5	98.9	99.4	**99.6**	99.7
	15	87.5	93.2	97.1	98.0	98.7	99.3	**99.6**	99.8	99.9	100
	10	93.2	**99.7**	99.9	99.9	100	100	100			
	8	98.6	99.2	**99.5**	99.7	99.9	100				
	7	98.1	98.8	**99.4**	99.8	99.9	100				
	6	98.8	99.4	**99.8**	99.9	100					
	5	98.4	99.4	**99.7**	100						
	4	86.2	**99.5**	100							
	3	99.3	**99.9**	100							
	2	99.4	**100**								

Table 2. Estimative of the number of independent contributions in ICA mixing information matrix at various imposed contributions.

	f	1	2	3	4	5	6	7	8	9	10	11
	30	47.8	85.0	97.5	99.0	**99.6**	99.8	99.9	99.9	99.9	100	
Ensemble	25	53.7	89.6	99.4	**99.7**	99.9	99.9	99.9	99.9	100		
(72)	20	54.1	93.6	99.2	**99.8**	99.9	99.9	99.9	100			
	15	44.7	84.6	94.8	99.3	**99.9**	99.9	100				
	10	59.7	92.7	**99.6**	99.8	100						
	8	67.6	92.8	99.3	**99.7**	100						
	6	61.9	91.7	98.6	**99.7**	100						
	5	52.8	90.0	99.2	**99.9**	100						
	4	56.7	89.7	**99.8**	100							
	3	47.6	87.0	**100**								
	2	59.6	100									
All SRM	30	62.4	86.4	96.3	97.8	99.0	**99.7**	99.8	99.9	99.9	99.9	99.9
(246)	29	61.1	86.2	94.9	97.9	99.0	**99.5**	99.7	99.8	99.8	99.9	99.9
	28	62.1	83.7	93.7	97.2	98.6	**99.5**	99.8	99.9	99.9	99.9	99.9
	27	64.5	86.0	95.9	97.6	98.8	**99.6**	99.7	99.8	99.9	99.9	100
	26	68.7	85.6	95.7	97.5	98.7	**99.7**	99.8	99.9	99.9	99.9	99.9
	25	57.8	86.8	96.1	98.0	98.9	**99.6**	99.8	99.8	99.9	99.9	100
	24	59.9	85.3	95.6	97.6	98.9	**99.6**	99.8	99.8	99.9	99.9	100
	23	61.8	85.9	95.9	97.9	99.0	**99.8**	99.9	99.9	99.9	100	
	22	64.6	85.8	96.0	98.0	99.1	**99.8**	99.9	99.9	100		
	21	62.6	82.3	91.6	95.7	99.2	**99.8**	99.9	99.9	100		
	20	63.0	84.2	94.8	97.3	98.8	**99.8**	99.9	99.9	100		
	19	66.9	84.1	92.5	97.6	99.1	**99.8**	99.9	100			
	18	66.3	83.8	92.1	97.4	99.0	**99.8**	99.9	99.9	100		
	17	58.2	81.3	91.1	96.8	98.6	99.3	**99.8**	99.9	100		
	16	59.7	81.8	95.3	97.6	98.9	**99.7**	99.8	99.9	99.9	100	
	15	59.2	80.3	92.5	95.8	98.1	99.3	**99.7**	99.8	99.9	100	
	14	67.5	86.1	95.4	97.5	98.9	**99.8**	99.9	100			
	13	59.7	81.2	94.1	97.7	98.9	**99.7**	99.8	99.9	100		
	12	47.1	76.2	92.5	95.8	98.1	**99.6**	99.8	99.9	100		
	11	60.2	82.6	95.5	98.0	99.3	**99.8**	99.9	100			
	10	41.6	73.7	92.2	96.3	98.9	**99.7**	99.9	100			
	9	40.0	74.8	91.1	96.0	98.6	**99.6**	99.9	100			
	8	40.3	71.2	94.2	98.0	**99.5**	99.9	100				
	7	38.7	69.7	87.2	95.9	**99.6**	99.9	100				
	6	42.7	79.3	95.7	**99.5**	99.9	100					
	5	47.0	82.5	96.1	**99.7**	100						
	4	50.4	82.6	99.4	**100**							
	3	43.8	77.5	100								
	2	60.2	100									

The correct number of independent components is defined when convergence is obtained (ci = fi).

To demonstrate this strategy, let's assume we are studying tyrosine solutions. If we impose c1 = 20 in ICA we obtain f1 = 5.On next iteration we impose c2 = 5 and obtain f = 2.

On the third iteration we converged to c3 = f3 = 2 revealing that tyrosine spectral information is composed by two independent contributions. Looking into **Figures 3(a)-(c)** and **Figures 4(a)-(c)**, we clearly see that there are in fact 2 independent components with specific source signals and mixing information—one of them have a linear dependency on tyrosine concentration and the other is almost independent.

Let's see if this process is also useful in other standard cases. In **Table 2** we present the results obtained for the Ensemble and for All SRM's cases.

From **Table 2** it is clear that the Ensemble case points out towards c = 3 fundamental independent components while in the All SRM's case it points out to c = 4 independent contributions.

Neglecting background independent contribution, previously we concluded that Tyr and Phen may be described with a single component while Tryp seems to require two components to describe spectral information and mixing matrix ICA contribution. When treated in simultaneous they behave like a c = 3 component case for the Ensemble and c = 4 component for All SRM's case. From **Figure 5(h)** it is easy to relate Tyr contribution (c = 2) to **Figure 3(a)** (c = 1), Tryp contributions (c = 1 and c = 3) to **Figure 3(g)** (c = 1 and c = 2) and Phen (c = 4) to **Figure 3(k)**.

Now we are going to evaluate the ability to fit and predict DOC values using mixing matrix values.

Table 3 presents the modelling results of DOC in terms of quality of fit and predictive ability. From **Table 3** it is clear that in single standard solutions (Tyr, Tryp, and Phen) there is a very good data description of DOC values with mixing matrix information via simple first degree models—these linear models are able to describe more than 99.5% of response with c ≥ 1 for Tyr and c ≥ 2 for Tryp and Phen cases, with low residual error (<2% DOC $mg{\cdot}L^{-1}$).

Considering AIC and AICc information criteria best fitting models are ICA1 for Tyr, ICA4 for Tryp and ICA2 for Phen.

Considering the conservative cross-validation %RMSEP, Equation (20), best predicting models are obtained with ICA1 for Tyr and Phen and ICA2 for Tryp; these results are in accordance to the results obtained with the robust predicting ability %PE1/2, Equation (21).

From these results it is possible to state that the analysis of these single standard solutions, which is necessary to impose c = 1 for Tyr and Phen and c = 2 for Tryp, DOC may accurately be estimated and predicted with a first degree polynomial curve (p = 2).

Looking at the Ensemble case, fitting results are also good but the overall number of independent components is comprehensibly higher, c ranging from 4

Table 3. Modelling performance of independent component contribution to describe DOC values of standard solutions in simpler cases (Tyr, Tryp, and Phen), in the Ensemble (72 spectra) and in All SRM solutions (246 spectra).

System	Model	p	StdFit	%RE	R^2	R^2adj	AIC	AICc	%RMSPE	%PE1/2
Tyr	ICA1	2	0.01	0.5	0.9999	0.9999	−208.4	−207.8	0.1	0.5
(24)	ICA2	3	0.01	0.5	0.9999	0.9999	−206.4	−205.2	0.1	36.0
	ICA3	4	0.01	0.5	0.9999	0.9999	−205.3	−203.2	0.1	140.0
	ICA4	5	0.01	0.5	0.9999	0.9999	−202.1	−198.8	0.1	93.4
	ICA5	6	0.01	0.5	0.9999	0.9999	−200.8	−195.9	0.1	141.5
Tryp	ICA1	2	0.82	7.6	0.9945	0.9939	−6.9	−6.3	1.7	16.6
(24)	ICA2	3	0.16	1.5	0.9998	0.9998	−82.6	−81.4	0.3	3.6
	ICA3	4	0.15	1.4	0.9998	0.9998	−84.3	−82.2	0.3	34.1
	ICA4	5	0.13	1.2	0.9999	0.9999	−91.4	−88.1	0.3	45.3
	ICA5	6	0.12	1.1	0.9999	0.9999	−90.2	−85.3	0.3	175.7
Phen	ICA1	2	0.12	2.8	0.9979	0.9977	−100.1	−99.5	0.6	10.5
(24)	ICA2	3	0.08	1.9	0.9991	0.9990	−117.3	−116.1	0.4	39.5
	ICA3	4	0.08	1.9	0.9991	0.9989	−114.9	−112.8	0.4	39.0
	ICA4	5	0.08	1.9	0.9991	0.9989	−112.3	−109.0	0.4	43.4
	ICA5	6	0.08	1.8	0.9993	0.9990	−113.0	−108.0	0.4	45.2
Ensemble	ICA2	3	2.47	42.3	0.8907	0.8858	133.4	133.8	5.1	54.7
(72)	ICA3	4	0.59	10.2	0.9938	0.9934	−70.5	−69.9	1.3	83.6
	ICA4	5	0.16	2.7	0.9996	0.9995	−260.1	−259.2	0.3	45.6
	ICA5	6	0.16	2.7	0.9996	0.9995	−258.8	−257.5	0.3	103.2
	ICA6	7	0.16	2.7	0.9996	0.9995	−256.7	−254.9	0.4	96.8
	ICA4²(15)	15	0.09	1.6	0.9999	0.9999	−323.7	−315.1	0.3	68.5
	ICA4²p(10)	10	0.09	1.6	0.9999	0.9998	−328.9	−325.3	0.3	68.5
	ICA4²p(6)	6	0.12	2.1	0.9997	0.9997	−295.5	−294.2	0.3	54.6
	ICA5²(21)	21	0.07	1.3	0.9999	0.9999	−341.1	−322.6	0.3	154.1
	ICA5²p(12)	12	0.07	1.3	0.9999	0.9999	−359.3	−354.0	0.2	154.1
	ICA6²(28)	28	0.08	1.3	0.9999	0.9999	−312.6	−274.8	0.9	64.8
	ICA6²p(12)	12	0.07	1.2	0.9999	0.9999	−368.3	−363.0	0.2	104.4
All SRM's	ICA2	3	4.40	36.0	0.8231	0.8209	731.8	731.9	2.3	65.2
(246)	ICA3	4	2.96	24.2	0.9204	0.9190	537.6	537.8	1.6	84.8
	ICA4	5	2.12	17.3	0.9594	0.9585	374.1	374.4	1.1	84.5
	ICA5	6	0.73	6.0	0.9951	0.9950	−145.6	−145.2	0.4	85.8
	ICA6	7	0.72	5.9	0.9954	0.9952	−155.7	−155.2	0.4	84.2
	ICA3²(10)	10	1.73	14.2	0.9734	0.9723	280.5	281.4	1.0	53.0
	ICA3²p(8)	8	1.73	14.2	0.9731	0.9722	279.2	279.8	1.0	53.0
	ICA4²(15)	15	0.92	7.5	0.9926	0.9922	−24.0	−21.9	0.5	41.3
	ICA4²p(10)	10	0.92	7.5	0.9926	0.9922	−24.0	−23.0	0.5	41.3

Continued

$ICA5^2(21)$	21	0.39	3.2	0.9987	0.9986	−436.2	−432.1	0.2	69.5
$ICA5^2p(16)$	16	0.39	3.2	0.9987	0.9986	−441.9	−439.5	0.2	38.4
$ICA5^2p(10)$	10	0.42	3.4	0.9984	0.9984	−415.0	−414.1	0.2	31.6
$ICA6^2(28)$	28	0.38	3.1	0.9988	0.9986	−436.3	−428.9	0.2	62.2
$ICA6^2p(17)$	17	0.39	3.2	0.9987	0.9986	−449.8	−447.1	0.2	62.2
$ICA6^2p(7)$	7	0.76	6.2	0.9948	0.9946	−126.5	−126.0	0.4	78.7

ICA—first degree polynomial model, Equation (11); ICA^2—second degree polynomial models, Equation (12); p—number of model parameters; StdFit—model residual error, Equation (13); %RE—model bias, Equation (14); R^2—r squared value, Equation (15); R^2adj—adjusted r squared value, Equation (16); AIC—Akaike information criteria, Equation (18), AICc—readjusted Aikaike information criteria, Equation (19); %RMSPE—jackknife prediction error, Equation (20); %PE1/2—50% unfolding prediction error, Equation (21).

to 6. Akaike AIC and AICc are in accordance with cross-validation results, stating that c = 4 is the correct number of independent components to be used in describing DOC solutions using ICA mixing information matrix via a single degree linear polynomial (p = 5).

Using second degree polynomial model no further significant performance increase is observed and model flexibility is penalizing its predicting ability—cross-validation results points to higher prediction errors.

In respect to All SRM case, first degree polynomial model requires c ≥ 5 to give satisfactory results in terms of fitting error (~6%), describing results (~99.5% of DOC description) with a relatively small predicting error. Better results are obtained considering c = 5 and using respective parsimonious models in order to ensure a good fitting without extra deterioration of its prediction ability.

From this preliminary study with simple standard solutions, some conclusions are drown:

1) It is possible to predict the number of independent contributions using a unsupervised method based on iterative PCA-ICA approach over mixing matrix results;

2) Predicted components are consistent with spectra signal recovery;

3) These estimated components are able to accurately describe DOC via linear first degree polynomial models;

4) Some difficulties arrive when dealing with real mixtures—it may be present some interfering effect and thus a relative lack on signal linearity in respect to solute concentration;

5) Fitting results a predicting ability may be increased using quadratic polynomial models;

6) Comparing spectral deconvoluted contributions it is possible to identify bands related to each component and thus identify its contribution in solution.

3.2. Environmental Samples

Humic (HA), fulvic acids (FA), and plankton (PP) isolated samples were also analysed in separate and in Ensemble. **Figure 8** presents the respective spectra.

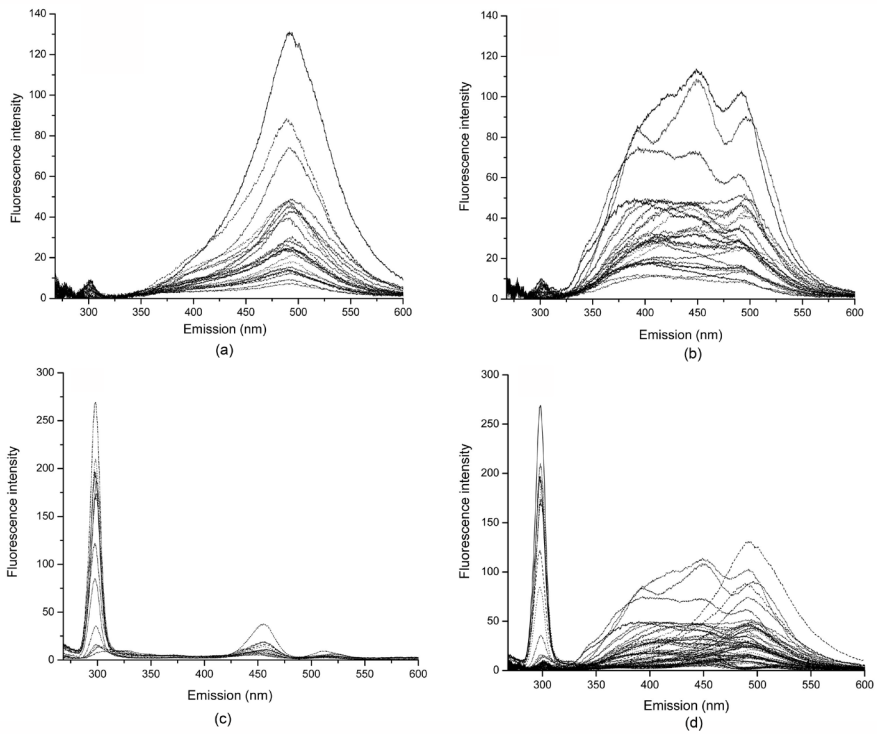

Figure 8. Synchronous fluorescence spectra ($\Delta\lambda$ = 18 nm) for (a) humic acid (HA, 24 spectra); (b) fulvic acid (FA, 25 spectra); (c) primary productivity (PP, 14 spectra) and (d) the Ensemble of this environmental samples (HFP = HA + FA + PP, 63 spectra).

From **Figure 8** it is possible to observe that each representative environmental sample have typical major emission contributions but some minor emission contributions are also present and seem very similar in respect to maxima emitting light—for instance, humic acid presents typically a maxima about 490 nm in a broad conical emitting profile between 350 - 620 nm, but also have a small maxima near 300 nm.; fulvic acids have a broad band composed by at least 3 important signals in the range 330 - 600 nm and present also a small maxima near 300 nm; plankton have a maxima emission at near 300 nm and small maxima contributions in 420 - 550 nm range.

These band superimpositions may be resolved by ICA as common basic bands that the total number of required independent sources will be less than expected sum of individual components.

Figure 9 and **Figure 10** present the evolution of deconvoluted signal sources from c = 3 to c = 10 in HA, FA, PP, and HFP ensemble. From **Figure 9** it makes some spectral sense to use c = 5 for HA, c = 7 for FA, c =4 for PP. In **Figure 10** it makes some spectral sense to use c = 7 for HFP ensemble (HA + FA + PP).

Figure 11 presents the evolution of relative fitting error with the number of imposed components in ICA treatment for this environmental samples (HA, FA, PP, and HFP ensemble).

Figure 11 strongly suggests the need of a single source signal to describe PP, 2 signals for HA, 3 for FA and an overall 3 major contributions for HFP Ensemble.

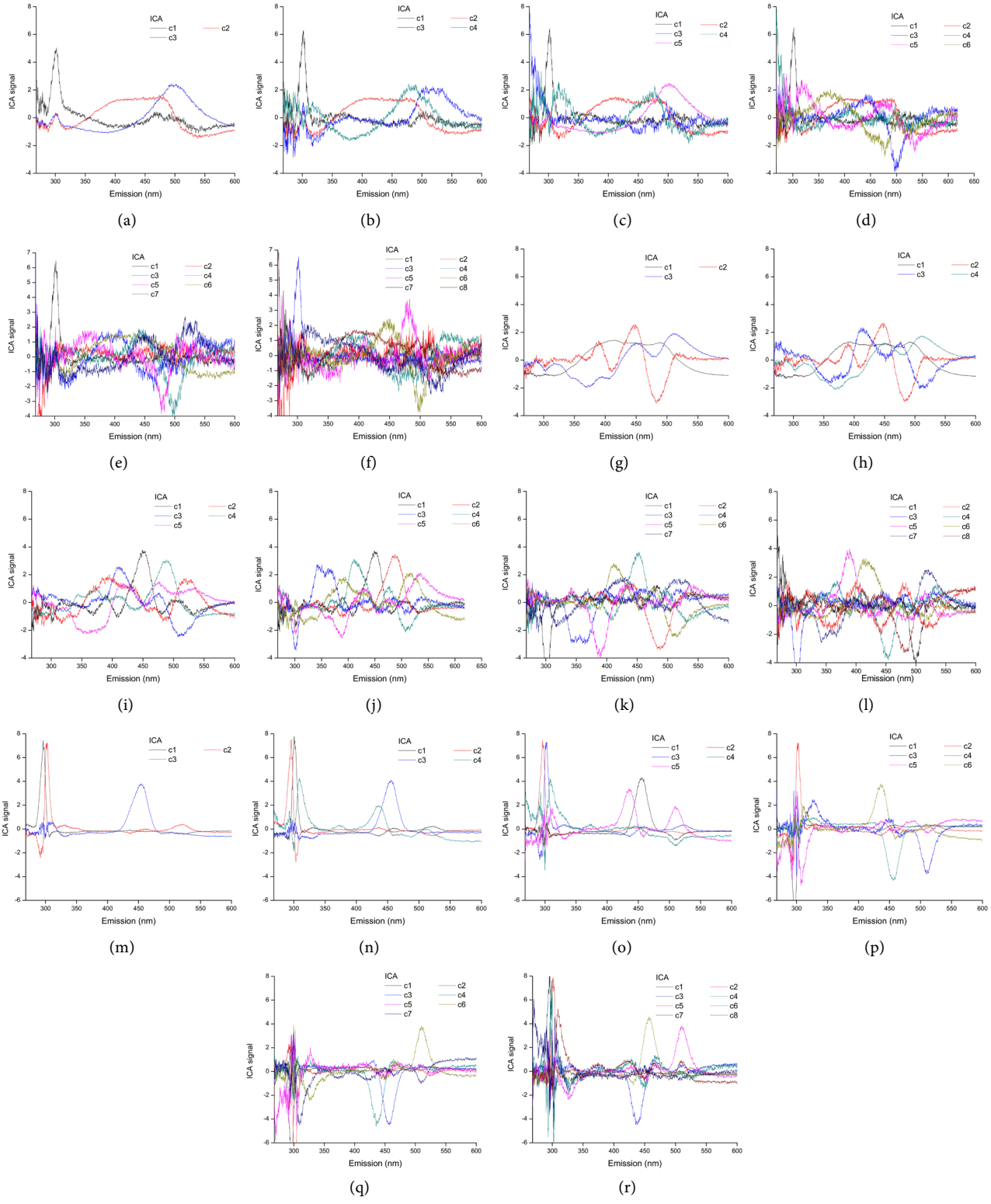

Figure 9. Estimated source signal components given by ICA from c = 3, 4, 5, 7 and8, for HA ((a)-(f)), FA ((g)-(l)) and PP ((m)-(r)).

If a line is drawn over **Figure 11** plots a more complete description may be obtained suggesting c = 5 for HA and PP, c = 6 for FA and c = 7 for HFP Ensemble.

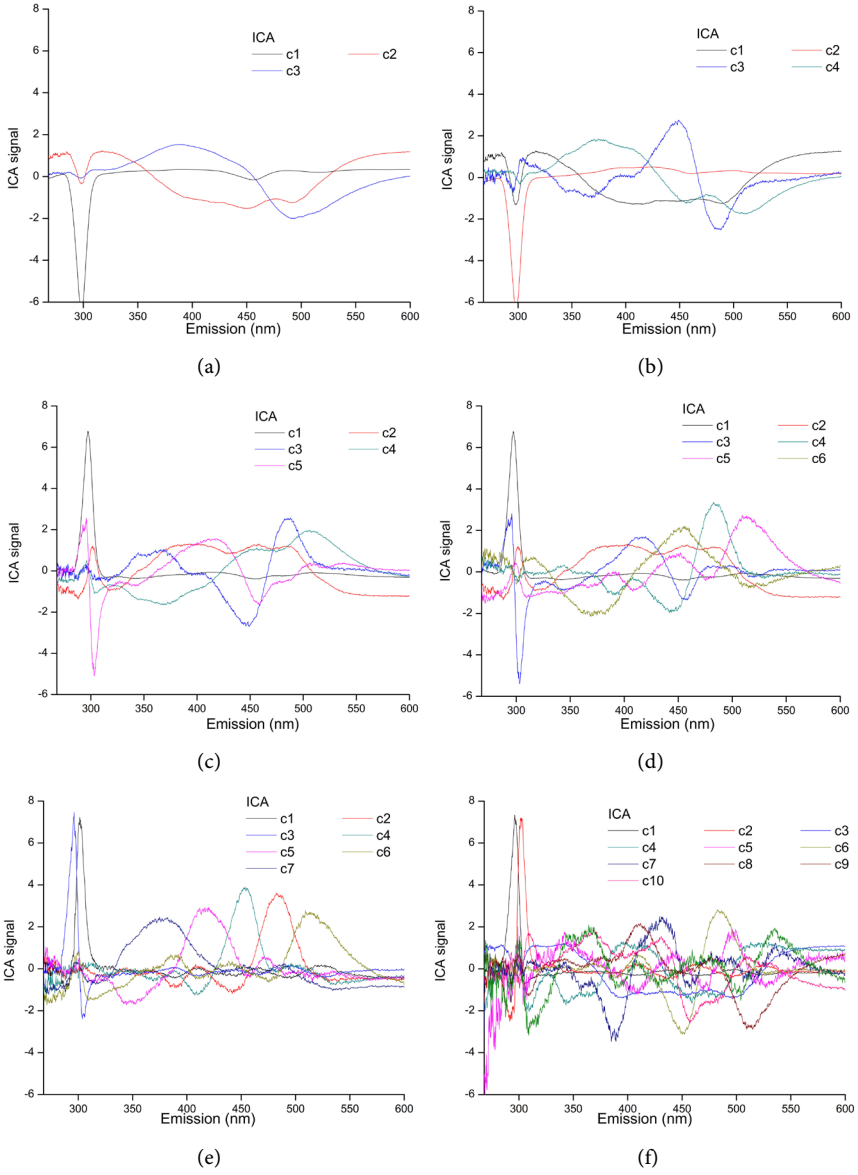

Figure 10. Estimated source signal components given by ICA from third component (c = 3) to seventh and tenth component (c = 10) for the Ensemble (HFP).

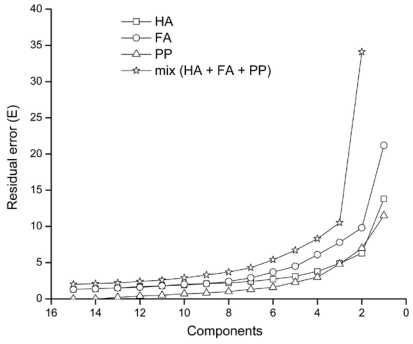

Figure 11. Relative residual error in fluorescence signal restoration with deconvoluted ICA components for humic acids (HA), fulvic acids (FA), plankton (PP), and in the Ensemble (HFP) of signals.

Table 4 presents the results of the PCA analysis applied to several attempts of ICA deconvolutions in order to define the minimum number of independent contributions in mixing matrix for HA, FA, PP, and HFP Ensemble (Table 4).

From Table 5 it is possible to obtain information for c = 3 component contributions in HA and PP, c = 4 contributions for FA and in Table 6, information for c = 5 contributions in HFP ensemble case. Conciliating all sources of information, we guess the need: for 3 to 5 signal sources in HA, 4 to 7 in FA, 3 to 6 signals in PP (Table 5) and for 5 to 7 signals in the ensemble HFP (Table 6).

Using this information we start to model retrieved ICA mixing matrix information for the DOC experimental results (Table 5 and Table 6).

The direct descriptions of DOC via first degree linear models fail in general in describing experimental results. For HA imposing c = 5 we obtain a model bias of about 30% with 68.7% description of DOC results and a prediction error of about 8% (in an optimistic point of view). These results do not improve significantly till c = 10.

In FA case, using c = 6 it is possible to obtain a model bias of about 25%, however it is only able to describe about 75% of response information with an optimistic prediction error of 6%.

In PP case, using c = 5, model bias is about 23% and can only explain over 62% of response with an optimistic prediction error of about 8%.

Since the interest is only on evaluating all samples DOC response, no further modelling effort will be explored here.

Considering the ensemble case with c = 7 and using a first degree polynomial (ICA7), it is obtained an overall model bias about 36%, able to describe about 58% of DOC results and with an optimistic predicting error of about 5%.

Extending to c = 15 and using first degree polynomial model results do not become significantly better. However, if we explore polynomial second degree model, Equation (12), these results are significantly better.

With only c = 5 and second degree full model ($ICA^2(21)$), it is possible to reduce average model bias to 28% and describe over 74% of DOC response, but with an increased prediction error estimative of near 10%. Prediction error can be successfully reduced to about 4% with parsimonious $ICA^2p(10)$ without significant change in model bias neither in response description.

Choosing c = 6 and with second degree full model ($ICA6^2(28)$), it is possible to achieve 24% average model bias, with an ability to describe 80% of DOC response with an optimistic prediction error of 8%. Opting for parsimonious model ($ICA6^2p(20)$), it is possible to decrease the prediction error to 5%.

Imposing c = 7 and using $ICA7^2(36)$ model, it still obtain a mean model bias about 24%, an ability to describe about 81% of DOC information with a large predicting error of 34% (in an optimistic point of view). If the model is refined to obtain parsimonious $ICA7^2p(19)$ model, the predict error on d can be reduced significantly to 5%, maintaining the overall model performance.

Choosing c = 8 and $ICA8^2(45)$, it is possible to obtain significantly better results—overall mean model bias of 14% with a 94% DOC response description,

Table 4. Estimative of independent components required in order to describe HA (24 spectra), FA (25 spectra), PP (14 spectra) and HFP ensemble (63 spectra) estimated with PCA analysis over ICA's mixing matrix.

System	c	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
HA	20	89.0	95.0	97.0	97.8	98.3	98.7	99.1	99.4	**99.6**	99.7	99.8	99.9	99.9	100	100
24	19	90.0	95.7	97.5	98.7	99.1	99.4	**99.6**	99.7	99.8	99.9	99.9	99.9	100		
	18	77.4	89.7	95.2	98.0	98.6	99.0	99.4	**99.6**	99.7	99.8	99.9	99.9	100		
	17	87.6	96.3	97.8	98.9	99.3	**99.5**	99.7	99.8	99.9	99.9	100				
	16	90.8	96.9	98.0	98.8	99.4	**99.7**	99.8	99.9	99.9	99.9	100				
	15	87.5	95.7	97.8	98.6	99.3	**99.6**	99.8	99.9	99.9	99.9	100				
	14	91.2	97.5	98.5	99.1	**99.5**	99.6	99.8	99.9	99.9	99.9	100				
	13	94.0	98.2	99.1	**99.5**	99.7	99.8	99.9	99.9	100						
	12	87.1	96.5	98.8	**99.5**	99.7	99.9	99.9	100							
	11	93.5	97.8	99.0	**99.5**	99.7	99.9	99.9	100							
	10	81.9	96.0	97.8	99.1	**99.7**	99.9	99.9	100							
	9	92.8	97.4	98.9	**99.6**	99.8	99.9	100								
	8	93.6	98.3	99.2	**99.8**	99.9	100									
	7	93.5	98.3	99.4	**99.9**	100										
	6	92.6	97.6	99.2	**99.9**	100										
	5	86.9	98.6	99.3	**99.9**	100										
	4	95.5	99.1	**100**												
	3	92.3	99.2	**100**												
	2	92.0	100													
FA	20	77.6	94.8	96.8	97.9	98.7	99.0	99.3	**99.5**	99.6	99.7	99.8	99.9	99.9	99.9	100
25	19	79.0	88.7	95.1	97.6	98.4	98.8	99.1	99.4	**99.6**	99.7	99.8	99.9	99.9	99.9	100
	18	70.2	88.8	93.3	95.5	96.9	98.1	98.8	99.2	**99.5**	99.7	99.8	99.8	99.9	99.9	100
	17	62.6	82.7	90.3	95.7	97.6	98.6	99.1	99.4	**99.6**	99.7	99.8	99.9	99.9	100	
	16	73.4	90.0	94.9	96.9	98.0	98.7	99.2	99.4	**99.7**	99.8	99.9	99.9	100		
	15	81.9	96.9	98.2	99.0	99.3	**99.5**	99.7	99.8	99.9	99.9	99.9	100			
	14	53.2	87.5	93.6	95.9	97.5	98.6	99.4	**99.6**	99.8	99.9	99.9	100			
	13	84.9	97.6	98.4	99.0	99.3	**99.6**	99.8	99.9	99.9	100					
	12	74.0	88.6	92.3	95.0	96.9	98.4	99.1	**99.5**	99.8	100					
	11	74.1	94.5	96.8	98.0	98.8	99.4	**99.6**	99.8	99.9	100					
	10	82.0	97.5	98.7	99.3	**99.6**	99.8	99.9	100							
	9	75.1	94.6	97.6	98.9	**99.5**	99.7	99.9	100							
	8	81.5	96.9	98.6	99.3	**99.6**	99.9	100								
	7	75.6	96.5	98.1	99.4	**99.7**	100									

Continued

	6	66.1	96.6	98.5	99.3	**99.9**	100								
	5	72.9	97.7	99.2	**99.8**	100									
	4	57.4	88.2	96.9	**100**										
	3	50.0	93.7	100											
	2	64.9	**100**												
PP	12	64.6	78.1	87.2	93.4	97.5	98.8	**99.5**	99.8	99.9	100				
14	11	66.6	82.5	92.5	96.4	98.0	98.9	99.4	**99.8**	100					
	10	63.4	80.4	91.3	96.7	98.3	99.3	**99.7**	99.9	100					
	9	64.0	82.5	93.7	98.1	99.3	**99.7**	99.9	100						
	8	61.1	80.6	90.6	96.5	98.8	**99.9**	100							
	7	52.2	76.4	89.4	96.6	98.7	**99.9**	100							
	6	61.9	84.8	96.5	99.0	**99.6**	100								
	5	64.5	92.2	97.3	**99.5**	100									
	4	74.6	99.0	**100**											
	3	74.6	99.0	**100**											
	2	81.4	100												
HFP	30	57.2	84.3	94.2	95.8	96.9	97.8	98.3	98.8	99.0	99.3	**99.5**	99.6	99.7	99.8
Ensemble	25	68.7	85.0	95.4	97.6	98.4	98.8	99.1	99.4	**99.5**	99.6	99.7	99.8	99.8	99.9
63	20	60.6	85.5	94.7	97.0	98.1	98.8	99.2	99.4	**99.6**	99.7	99.8	99.9	99.9	100
	19	49.9	85.8	95.2	97.8	98.4	98.9	99.2	99.4	**99.6**	99.8	99.8	99.9	99.9	100
	18	65.5	88.4	97.7	98.4	98.9	99.2	**99.5**	99.7	99.8	99.8	99.9	99.9	100	
	17	67.5	86.6	97.6	98.3	98.8	99.2	**99.5**	99.7	99.8	99.8	99.9	99.9	100	
	16	73.1	88.1	97.5	98.5	98.9	99.3	**99.6**	99.7	99.8	99.9	99.9	99.9	100	
	15	67.8	88.1	97.8	98.6	99.0	99.4	**99.7**	99.8	99.8	99.9	99.9	100		
	14	61.3	82.3	95.9	97.7	98.6	99.1	**99.5**	99.7	99.8	99.9	99.9	100		
	13	63.1	87.6	97.6	98.6	99.2	**99.6**	99.7	99.8	99.9	100				
	12	61.2	85.6	97.0	98.3	99.2	**99.6**	99.7	99.8	99.9	100				
	11	55.7	81.2	97.1	98.6	99.2	**99.6**	99.8	99.9	99.9	100				
	10	68.5	86.1	96.7	98.7	99.3	**99.7**	99.8	99.9	100	100				
	9	62.6	84.8	97.9	98.9	99.4	**99.7**	99.8	99.9	100					
	8	63.1	83.7	97.6	98.9	99.4	**99.7**	99.9	100						
	7	62.3	82.9	98.4	99.3	**99.6**	99.9	100							
	6	58.3	83.7	97.3	99.0	**99.6**	100								
	5	55.4	85.8	97.8	99.4	**100**									
	4	53.9	84.4	98.0	100										
	3	57.4	85.6	100											

Table 5. Modelling performance of independent component contribution to describe DOC values of standard solutions in simpler cases (HA, FA, PP).

System	Model	p	StdFit	%RE	R^2	R^2adj	AIC	AICc	%RMSPE	%MEP
HA	ICA2	3	6.25	45.5	0.3953	0.3046	92.0	93.2	11.0	61.1
24	ICA3	4	4.08	29.7	0.7546	0.7029	73.2	75.3	6.9	61.9
	ICA4	5	4.18	30.4	0.7550	0.6869	76.4	79.7	7.7	64.8
	ICA5	6	4.18	30.4	0.7682	0.6863	78.7	83.6	7.6	63.0
	ICA6	7	3.75	27.3	0.8235	0.7463	76.2	83.2	7.0	62.9
	ICA7	8	3.87	28.1	0.8235	0.7294	80.8	90.4	7.6	60.4
	ICA8	9	3.77	27.4	0.8429	0.7420	83.3	96.1	7.4	38.2
	ICA9	10	3.76	27.4	0.8538	0.7413	87.6	104.5	7.5	54.9
	ICA10	11	3.57	26.0	0.8780	0.7663	90.3	112.3	7.1	83.2
FA	ICA2	3	9.93	45.6	0.3216	0.2246	118.7	119.9	9.8	80.8
24	ICA3	4	9.67	44.4	0.3862	0.2634	119.1	121.1	10.7	47.1
	ICA4	5	8.65	39.8	0.5320	0.4088	115.4	118.6	11.2	44.6
	ICA5	6	5.54	25.5	0.8174	0.7565	95.4	100.1	5.7	66.4
	ICA6	7	5.03	23.1	0.8576	0.7990	93.1	99.7	5.6	37.7
	ICA7	8	4.71	21.6	0.8822	0.8233	92.8	101.8	5.4	36.0
	ICA8	9	4.79	22.0	0.8853	0.8165	97.1	109.1	5.3	43.8
	ICA9	10	4.37	20.1	0.9105	0.8465	96.6	112.4	4.8	33.9
	ICA10	11	4.47	20.6	0.9123	0.8381	102.7	123.0	5.5	32.6
PP	ICA2	3	7.09	25.7	0.6338	0.5239	59.9	62.3	7.7	47.3
14	ICA3	4	6.94	25.2	0.6810	0.5393	62.0	66.4	7.2	60.8
	ICA4	5	5.90	21.4	0.7930	0.6636	61.0	68.5	6.4	17.5
	ICA5	6	6.24	22.6	0.7937	0.6168	67.4	79.4	7.5	90.5
	ICA6	7	6.67	24.2	0.7939	0.5535	76.1	94.8	9.1	96.6
	ICA7	8	5.46	19.8	0.8816	0.6921	80.5	109.3	12.6	40.9
	ICA8	9	5.93	21.5	0.8838	0.6224	98.4	143.4	14.3	37.2
	ICA9	10	6.01	21.8	0.9043	0.5853	126.0	199.4	24.3	23.5
	ICA10	11	1.75	6.3	0.9939	0.9606	148.1	280.1	5.2	28.6

ICA—first degree polynomial model, Equation (11); ICA^2—second degree polynomial models, Equation (12); p—number of model parameters; StdFit—model residual error, Equation (13); %RE—model bias, Equation (14); R^2—r squared value, Equation (15); R^2adj—adjusted r squared value, Equation (16); AIC—Akaike information criteria, Equation (18), AICc—readjusted Aikaike information criteria, Equation (19); %RMSPE—jackknife prediction error, Equation (20); %PE1/2—50% unfolding prediction error, Equation (21).

Table 6. Modelling performance of independent component contribution to describe DOC values of HFP ensemble (63 spectra).

System	Model	p	StdFit	%RE	R^2	R^2adj	AIC	AICc	%RMSPE	%MEP
HFP	ICA3	4	8.67	43.3	0.4382	0.3995	276.6	277.3	5.6	107.4
Ensemble	ICA4	5	8.49	42.4	0.4699	0.4234	275.4	276.4	5.7	103.5
	ICA5	6	7.31	36.6	0.6137	0.5723	257.9	259.4	4.9	109.6
	ICA6	7	7.37	36.8	0.6145	0.5654	260.3	262.3	5.2	83.3
	ICA7	8	7.26	36.3	0.6324	0.5779	259.9	262.6	5.1	102.8
	ICA8	9	7.32	36.6	0.6327	0.5703	262.6	266.0	5.3	101.7
	ICA9	10	7.39	37.0	0.6328	0.5622	265.4	269.6	5.7	94.0
	ICA10	11	7.25	36.2	0.6538	0.5791	264.6	269.8	5.4	93.8
	ICA11	12	6.45	32.2	0.7313	0.6668	251.7	258.0	5.1	95.6
	ICA12	13	6.42	32.1	0.7385	0.6691	253.2	260.6	5.2	83.7
	ICA13	14	6.48	32.4	0.7389	0.6628	256.4	265.2	5.4	84.3
	ICA14	15	6.55	32.7	0.7393	0.6561	259.8	270.0	5.5	139.9
	ICA15	16	6.19	31.0	0.7714	0.6919	255.1	267.0	5.1	94.4
	$ICA5^2$(21)	21	5.66	28.3	0.8293	0.7419	257.4	280.0	9.8	80.9
	$ICA5^2$p(10)	10	5.59	27.9	0.7901	0.7498	230.2	234.4	4.2	99.4
	$ICA5^2$p(5)	5	7.56	37.8	0.5811	0.5444	260.7	261.8	5.0	56.2
	$ICA5^2$(4)	4	7.76	38.8	0.5504	0.5194	262.7	263.4	5.1	56.2
	$ICA6^2$(28)	28	4.88	24.4	0.8944	0.8075	266.4	314.2	8.0	101.4
	$ICA6^2$p(20)	20	4.67	23.4	0.8809	0.8241	230.2	250.2	4.9	101.4
	$ICA6^2$p(12)	12	5.69	28.4	0.7906	0.7404	236.0	242.3	4.3	101.4
	$ICA6^2$p(8)	8	6.67	33.4	0.6895	0.6435	249.3	251.9	4.4	101.4
	$ICA6^2$p(5)	5	7.53	37.6	0.5848	0.5484	260.2	261.3	5.0	101.4
	$ICA6^2$(4)	4	7.81	39.0	0.5457	0.5144	263.5	264.2	5.1	80.6
	$ICA7^2$(36)	36	4.82	24.1	0.9206	0.8106	319.2	421.6	34.0	97.9
	$ICA7^2$p(19)	19	4.47	22.3	0.8887	0.8395	221.7	239.3	5.2	97.9
	$ICA7^2$p(16)	16	4.75	23.7	0.8658	0.8191	221.6	233.4	6.2	97.9
	$ICA7^2$p(11)	11	5.41	27.1	0.8114	0.7707	227.9	233.1	4.7	97.9
	$ICA7^2$(10)	10	6.34	31.7	0.7304	0.6786	246.0	250.2	5.6	97.9
	$ICA8^2$(45)	45	2.76	13.8	0.9826	0.9366	382.5	626.0	50.6	101.4
	$ICA8^2$p(29)	29	2.92	14.6	0.9634	0.9311	206.7	259.4	5.8	103.5
	$ICA8^2$p(18)	18	4.76	23.8	0.8709	0.8181	226.9	242.4	4.5	103.5
	$ICA8^2$p(11)	11	5.65	28.3	0.7893	0.7439	233.3	238.5	4.4	95.8
	$ICA8^2$p(5)	5	8.24	41.2	0.5081	0.4649	271.5	272.6	6.0	95.8
	$ICA8^2$p(2)	2	9.92	49.6	0.2690	0.2446	291.3	291.5	6.4	45.2

ICA—first degree polynomial model, Equation (11); ICA^2—second degree polynomial models, Equation (12); p—number of model parameters; StdFit—model residual error, Equation (13); %RE—model bias, Equation (14); R^2—r squared value, Equation (15); R^2adj—adjusted r squared value, Equation (16); AIC—Akaike information criteria, Equation (18), AICc—readjusted Aikaike information criteria, Equation (19); %RMSPE—jackknife prediction error, Equation (20); %PE1/2—50% unfolding prediction error, Equation (21).

but with a bad predicting error of over 50%. Model refinement in order to obtain the parsimonious model ICA8^2(29) can allows to maintain other performance indicators and reduces significantly predicting error to about 6%.

From Table 3 it is possible to observe that further model refinement in order to remove additional parameters and thus obtain also parsimonious models do not benefit on better results—models become more rigid and unable to describe experimental results.

With these results, it's also possible to observe that with a very small amount of independent components (c = 5) it is possible to describe all Ensembles of environmental samples if parsimonious second degree polynomial models is used.

4. Conclusions

In this work, we explored ICA mixing matrix values in order to directly use that numerical information in DOC estimation of standard and environmental sample solutions based upon synchronous fluorescence spectra ($\Delta\lambda$ = 18 nm).

When dealing with standard solutions, the developed strategy seems to work very well.

Linear models were accurately used to estimate and predict DOC with a first degree polynomial model with very few independent components (c = 1 or 2) and parameters (p = 2 or 3).

Same approach was still valid for single solute standard spectra ensemble, where accurately can still determine and estimate DOC with only c = 4 spectral sources and first degree polynomial model (ICA4).

When standard mixtures were analysed, we noticed some interference that oblige to use a second degree polynomial models in order to maintain accurate determinations.

From this study some conclusions are drown:

1) It is possible to predict the number of independent contributions using a unsupervised method based on iterative PCA-ICA approach over mixing matrix results;

2) Predicted components are consistent with spectra signal recovery;

3) These estimated components are able to accurately describe DOC via linear first degree polynomial models;

4) Some difficulties arrive when dealing with mixtures—it may be present some interfering effect;

5) Using quadratic polynomial models fitting results are better–lower deviations and grater predicting ability;

6) Comparing spectral deconvoluted contributions it is possible to identify bands related to each component and thus identify its contribution in solution.

When dealing with actual environmental representative samples results were not so satisfactory for several reasons.

Firstly, we have some amount of experimental error in DOC determinations, highly greater than theoretical values of standard solutions.

Secondly, we have proved that mixtures present significant interference—its behaviour do not exactly matches the behaviour of independent component solution ensemble. This lack of spectral consistency may causes difficulties in calibration and quantification process.

With this work we have proved to be possible to perform DOC estimation based on fluorescence spectra.

However this optimization process is not still fully optimized.

Acknowledgements

Authors are thankful for FCT Fundação para a Ciência e Tecnologia do Ministério da Educação e Ciência de Portugal (PEst-OE/QUI/UI0313/2014). The financial support from CAPES (BEX 12102/13-0), the Coimbra Chemistry Centre and Department of Chemistry and Biology/UTFPR.CNPq by Bolsa Produtividade (proc. 302736/2016-6) and call MCTIC/CNPq N˚ 28/2018 (proc. 407157/2018-2).

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

[1] Oliveira, J.L., Boroski, M., Azevedo, J.C.R. and Nozaki, J. (2006) Spectroscopic Investigation of Humic Substances in a Tropical Lake during a Complete Hydrological Cycle. *Acta Hydrochimica et Hydrobiologica*, **34**, 608-617. https://doi.org/10.1002/aheh.200400659

[2] Yu, H., Song, Y., Tu, X., Du, E., Liu, R. and Peng, J. (2013) Assessing Removal Efficiency of Dissolved Organic Matter in Wastewater Treatment Using Fluorescence Excitation Emission Matrices with Parallel Factor Analysis and Second Derivative Synchronous Fluorescence. *Bioresource Technology*, **144**, 595-601. https://doi.org/10.1016/j.biortech.2013.07.025

[3] Knapik, H.G., Fernandes, C.V.S., Azevedo, J.C.R. and Porto, M.F.A. (2014) Applicability of Fluorescence and Absorbance Spectroscopy to Estimate Organic Pollution in Rivers. *Environmental Engineering Science*, **31**, 653-663. https://doi.org/10.1089/ees.2014.0064

[4] Yu, H., Song, Y., Liu, R., Hongwei Pan, H., Xiang, L. and Qian, F. (2014) Identifying Changes in Dissolved Organic Matter Content and Characteristics by Fluorescence Spectroscopy Coupled with Self-Organizing Map and Classification and Regression Tree Analysis during Wastewater Treatment. *Chemosphere*, **113**, 79-86. https://doi.org/10.1016/j.chemosphere.2014.04.020

[5] Westerhoff, P. and Anning, D. (2000) Concentrations and Characteristics of Organic Carbon in Surface Water in Arizona: Influence of Urbanization. *Journal of Hydrology*, **236**, 202-222. https://doi.org/10.1016/S0022-1694(00)00292-4

[6] Chen, J., Gu, B. Leboeuf, E.J., Pan, H. and Dai, S. (2002) Spectroscopic Characterization of the Structural and Functional Properties of Natural Organic Matter Fractions. *Chemosphere*, **48**, 59-68. https://doi.org/10.1016/S0045-6535(02)00041-3

[7] Ahmad, S.R. and Reynolds, D.M. (1995) Synchronous Fluorescence Spectroscopy of Wastewater and Some Potential Constituents. *Water Research*, **29**, 1599-1602.

https://doi.org/10.1016/0043-1354(94)00266-A

[8] Senesi, N. (1990) Molecular and Quantitative Aspects of the Chemistry of Fulvic Acid and Its Interactions with Metal Ions and Organic Chemicals. Part II. The Fluorescence Spectroscopy Approach. *Analytica Chimica Acta*, **232**, 77-106. https://doi.org/10.1016/S0003-2670(00)81226-X

[9] Pons, M., Bonté, S.L. and Potier, O. (2004) Spectral Analysis and Fingerprinting for Biomedia Characterization. *Journal of Biotechnology*, **113**, 211-230. https://doi.org/10.1016/j.jbiotec.2004.03.028

[10] Carstea, E.M., Baker, A., Bieroza, M. and Reynolds, D. (2010) Continuous Fluorescence Excitation-Emission Matrix Monitoring of River Organic Matter. *Water Research*, **44**, 5356-5366. https://doi.org/10.1016/j.watres.2010.06.036

[11] Baghoth, S.A., Sharma, S.K. and Amy, G.L. (2011) Tracking Natural Organic Matter (NOM) in a Drinking Water Treatment Plant Using Fluorescence Excitation-Emission Matrices and PARAFAC. *Water Research*, **45**, 797-809. https://doi.org/10.1016/j.watres.2010.09.005

[12] Spencer, R.G.M., Baker, A., Ahad, J.M.E, Cowie, G.L., Ganeshram, R., Upstill-Goddard, R.C. and Uher, G. (2007) Discriminatory Classification of Natural and Anthropogenic Waters in Two U. K. Estuaries. *Science of the Total Environment*, **373**, 305-323. https://doi.org/10.1016/j.scitotenv.2006.10.052

[13] Peuravuori, J., Koivikko, R. and Pihlaja, K. (2002) Characterization, Differentiation and Classification of Aquatic Humic Matter Separated with Different Sorbents: Synchronous Scanning Fluorescence Spectroscopy. *Water Research*, **36**, 4552-4562. https://doi.org/10.1016/S0043-1354(02)00172-0

[14] Baker, A. (2002) Spectrophotometric Discrimination of River Dissolved Organic Matter. *Hydrological Process*, **16**, 3203-3213. https://doi.org/10.1002/hyp.1097

[15] Baker, A. (2001) Fluorescence Excitation-Emission Matrix Characterization of Some Sewage-Impacted Rivers. *Environmental Science and Technology*, **35**, 948-953. https://doi.org/10.1021/es000177t

[16] Coble, P.G., Del Castillo, C.E. and Avril, B. (1998) Distribution and Optical Properties of CDOM in the Arabian Sea during the 1995 Southwest Monsoon. *Deep-Sea Research Part II*, **45**, 2195-2223. https://doi.org/10.1016/S0967-0645(98)00068-X

[17] Senesi, N., Miano T.M., Provenzano, M.R. and Brunetti, G. (1989) Spectroscopy and Compositional Comparative Characterization of I.H.S.S. Reference and Standard Fulvic and Humic Acids of Various Origins. *Science of the Total Environment*, **81-82**, 143-156. https://doi.org/10.1016/0048-9697(89)90120-4

[18] Boehme, J., Cobles, P., Conmy, R. and Stovall-Leonard, A. (2004) Examining CDOM Fluorescence Variability Using Principal Component Analysis: Seasonal and Regional Modeling of Three-Dimensional Fluorescence in the Gulf of Mexico. *Marine Chemistry*, **89**, 3-14. https://doi.org/10.1016/j.marchem.2004.03.019

[19] Yamashita, Y. and Jaffé, R. (2008) Characterizing the Interactions between Trace Metals and Dissolved Organic Matter Using Excitation-Emission Matrix and Parallel Factor Analysis. *Environmental Science & Technology*, **42**, 7374-7379. https://doi.org/10.1021/es801357h

[20] Hall, G. and Kenny, J.E. (2007) Estuarine Water Classification Using EEM Spectroscopy and PARAFAC-SIMCA. *Analytica Chimica Acta*, **581**, 118-124. https://doi.org/10.1016/j.aca.2006.08.034

[21] Gao, L. and Ren, S. (2012) Integrating Independent Component Analysis with Artificial Neural Network to Analyze Overlapping Fluorescence Spectra of Organic Pollutants. *Journal of Fluorescence*, **22**, 1595-1602.

https://doi.org/10.1007/s10895-012-1100-y

[22] Hyvarinen, A, Karhunen, J. and Oja. E. (2001) Independent Component Analysis. John Wiley & Sons, New York. https://doi.org/10.1002/0471221317

[23] Tharwat, A. (2018) Independent Component Analysis: An Introduction. *Applied Computing and Informatics*, In Press, Corrected Proof. https://doi.org/10.1016/j.aci.2018.08.006

[24] Abed-Meraim, K., Loubaton, P. and Moulines, E. (1997) A Subspace Algorithm for Certain Blind Identification Problems. *IEEE Transactions on Information Theory*, **43**, 499-511. https://doi.org/10.1109/18.556108

[25] Brehm, F.A., Azevedo, J.C.R., Pereira, J.C. and Burrows, H.D. (2015) Direct Estimation of Dissolved Organic Carbon Using Synchronous Fluorescence and Independent Component Analysis (ICA): Advantages of a Multivariate Calibration. *Environmental Monitoring and Assessment*, **187**, 703. https://doi.org/10.1007/s10661-015-4857-z

[26] Jolliffe, I.T. (2002) Principal Component Analysis. Springer Series in Statistics, 2nd Edition, New York.

[27] Rao, C.R., Toutenburg, H. and Heumann, C. (2008). Linear Models and Generalizations: Least Squares and Alternatives. Springer Series in Statistics, 2nd Edition, New York.

[28] Dobson, A.J. (2002) An Introduction to Generalized Linear Models. 2nd Edition, Chapman & Hall/CRC, Boca Raton.