Research Article

Environmetrics

Received: 10 April 2009, Accepted: 25 January 2010, Published online in Wiley Online Library: 17 July 2010

(wileyonlinelibrary.com) DOI:10.1002/env.1045

Classification and calibration of organic matter fluorescence data with multiway analysis methods and artificial neural networks: an operational tool for improved drinking water treatment

Magdalena Bieroza[a]*, Andy Baker[b] and John Bridgeman[c]

Fluorescence spectroscopy enables fast and sensitive analysis of environmental samples containing various organic matter constituents. However, to retrieve valuable information from fluorescence spectra, robust techniques for data analysis should be employed. Here, different multivariate analysis methods and artificial neural networks (ANNs) were applied for decomposition and calibration of fluorescence excitation–emission matrices (EEMs). This is the first paper summarizing the application of different data mining methods, from multiway analysis to ANNs, for fluorescence EEMs technique employed to characterize organic matter properties and removal in the field of drinking water treatment. Fluorescence analysis was carried out on municipal water treatment samples of raw and partially-treated water. Parallel factor analysis (PARAFAC) method and self-organizing maps were used to analyse EEMs, extract information on the organic matter constituents and reduce the dimensionality of the data to enhance the efficiency of calibration methods. Partial least squares (PLS), multiple linear regression (MLR) and neural network with back-propagation were employed for calibration of fluorescence data with actual total organic carbon (TOC) concentrations. All models except PARAFAC-MLR produced consistent results with correlation coefficient $R^2 = 0.93$ for validation dataset. This is the first such comparative analysis of fluorescence data modelling that clarifies fundamental fluorescence data analysis questions regarding the suitability of different decomposition and calibration methods.

Keywords: data mining; multivariate statistics; neural networks; fluorescence spectroscopy; organic matter removal

1. INTRODUCTION

Fluorescence spectroscopy has been successfully utilized in environmental studies, providing a robust, holistic and non-destructive source of information on organic matter function and structure in a variety of environments, from marine (Mopper and Schultz, 1993; Coble, 1996; de Souza Sierra *et al*., 1997; Clark *et al*., 2002; Stedmon *et al*., 2003; Boehme *et al*., 2004) to freshwater ecosystems (Battin, 1998; Mounier *et al*., 1999; McKnight *et al*., 2001; Patel-Sorrentino *et al*., 2002; Cammack *et al*., 2004; Hudson *et al*., 2007). Recently, attempts have been made to facilitate the use of fluorescence analysis in drinking and wastewater treatment studies for deriving surrogates for organic matter presence and removal and total organic carbon (TOC; Baker, 2001; Cumberland and Baker, 2007; Hudson *et al*., 2008), and for organic matter fractions characterization (molecular weight, degree of hydrophobicity; Bengraïne and Marhaba, 2003; Wu *et al*., 2003; Belzile and Guo, 2006). Organic matter assessment is a growing concern for water companies responsible for providing reliable and clean drinking water supplies. During the treatment processes organic matter is reduced effectively; however, insufficient removal can result in formation of carcinogenic disinfection by-products (DBPs; mainly trihalomethanes) as an effect of side reaction with disinfectant. Commonly available techniques of organic matter quantification are time- and labour-consuming, therefore alternative and novel approaches, like fluorescence spectroscopy should be explored and utilized to enable faster and more robust solution to DBPs detection. The increased interest of drinking

* *Correspondence to: M. Bieroza, School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K. E-mail: magdalena.bieroza@gmail.com*

a *Department of Civil Engineering, University of Bristol, Queen's Building, University Walk, Bristol BS8 1TR, U.K.*

b *Connected Waters Initiative, University of New South Wales, 110 King Street, Manly Vale NSW 2093, Australia*

c *School of Civil Engineering, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.*

water companies in the potential application of fluorescence spectroscopy to operational monitoring of water treatment works (WTWs) results from the method's simplicity, sensitivity, speed of analysis and low cost, combined with potential for incorporation into an online monitoring system. In the current study, fluorescence spectroscopy has been used for the determination of organic matter at 16 WTWs across the Midlands region of the UK. The work has concentrated on characterizing organic matter removal across different stages of municipal water treatment and providing qualitative information on organic matter properties relevant to treatment processes efficiency. Specifically, the use of fluorescence spectroscopy as a rapid screening tool for assessing the presence of carcinogenic DBPs, formed when organic matter combines with disinfectant (e.g. chlorine), has been evaluated.

Fluorescence spectroscopy is a flexible, rapid and portable organic matter characterization tool, with the excitation–emission matrix (EEM) being the state-of-the art technique used (Hudson *et al*., 2007). An excitation–emission matrix portrays the inherent fluorescence properties of organic matter constituents. By scanning fluorescence intensity over a range of different excitation and emission wavelengths, a three-dimensional matrix is produced, with increased fluorescence intensities in particular regions, dependent on the organic matter origin (natural versus anthropogenic) and fractions (humic or fulvic like organic matter). Typical EEMs of raw and partially-treated water are presented in Figure 1. Increased fluorescence intensity in particular regions of the matrix indicates the presence of specific fluorophores, of which the most common in freshwater are humic-like, fulvic-like and protein-like (tryptophan- or tyrosine-like). The relative change in fluorescence intensity between raw and partially-treated water provides useful information on the degree of removal of organic matter and thus WTW performance.

Although EEMs can be a substantial source of environmental information on organic matter composition and variability, the high-dimensionality (e.g. spatial variability by temporal variability by fluorescence emission by fluorescence excitation) and nonlinearity of the data generate difficulties in data interpretation and extraction of useful information. Standard techniques for EEM analysis include identification of particular fluorophores independently, by assessment of the position of the fluorescence peak and intensity ('peak-picking'), and spectra subtraction for the evaluation of the relative changes in the fluorescence indicative of changes in organic matter quantity and quality (e.g. organic matter removal during the treatment processes; Stedmon *et al*., 2003; Boehme *et al*., 2004). Therefore, to handle a large amount of fluorescence data, a more comprehensive approach is warranted, which both preserves the important topological and metric relationships of the fluorescence data (high-dimensionality and nonlinearity) and establishes valid correlations between various process parameters.

Advanced EEMs analysis techniques can be classified by the algorithms used and purpose of the analysis. The most common techniques include different multivariate analysis tools from chemometric analysis (multiway analysis: principal components analysis (PCA), partial least squares (PLS) analysis, parallel factor analysis (PARAFAC); multiple linear regression (MLR), multivariate curve resolution (MCR)) and other computation and modelling techniques such as artificial neural networks (ANNs) and fuzzy logic. Examples of the methods used for EEMs analysis found in literature are presented in Table 1.

In fluorescence analysis, multivariate techniques are commonly used in the exploratory analysis, pattern recognition and multivariate calibration of spectra. Exploratory analysis aims to gather information on the dataset and formulate hypotheses worth testing without prior knowledge of the regularities and patterns in the data. This approach often employs unsupervised decomposition algorithms (e.g. SOM, PCA). The multivariate calibration involves development of the mathematical model relating the fluorescence intensity of a component (fluorophore) with the concentration from a set of reference samples (calibration) and prediction of the component concentration from the unknown sample fluorescence spectra (validation; Martens and Naes, 1989; Bos *et al*., 1993).

The differences between the calibration methods lie in the way the parameters of those algorithms are optimized (Despagne and Massart, 1998). ANNs represent the models without any constraints on the parameters, which are fitted to minimize the calibration samples squared

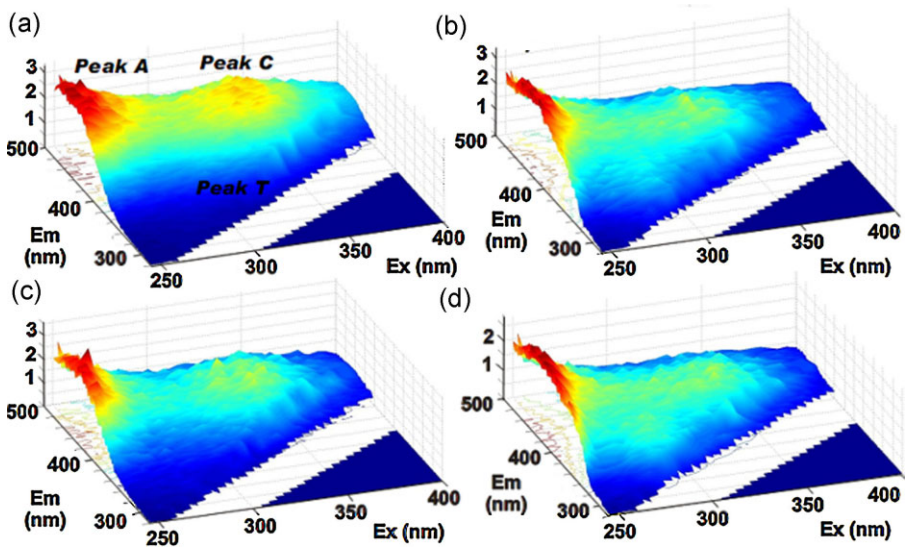

Figure 1. Three-dimensional EEMs of two sites of contrasting organic matter properties (a—raw water Site 1, b—raw water Site 7, c—partially-treated water Site 1, d—partially-treated water Site 7). EEM features: peak A—humic-like fluorescence, peak C—fulvic-like fluorescence, peak T—tryptophan-like fluorescence. This figure is available in colour online at wileyonlinelibrary.com./journal/environmetrics

Table 1. Examples of application of multivariate techniques to EEMs analysis (CAL—calibration, CLAS—classification)

Reference	Topic	Approach	Method	Aim of analysis
McAvoy *et al.*, 1992	DOM	ANNs/multiway analysis	ANNs (BP) vs. PLS	CAL
Henrion *et al.*, 1997	Algae	Multiway analysis	3-way PCA	CLAS
Marhaba *et al.*, 2000	DOM	Multiway analysis	PCA	CLAS
Marhaba *et al.*, 2000	DOM	EEMs exploratory analysis	spectra analysis	CAL
Saurina *et al.*, 2000	Water quality	Chemometrics	MCR-ALS	CLAS
Persson and Wedborg, 2001	DOM	Multiway analysis	PCA vs. PLS	CLAS
Wolf *et al.*, 2001	Fermentation processes monitoring	ANNs	ANNs (BP)	CLAS/CAL
Bengraïne and Marhaba, 2003	DOM	Multiway analysis	PCA vs. PLS	CAL
Chen *et al.*, 2003	DOM	EEMs exploratory analysis	Regional integration	CLAS
Guimet *et al.*, 2005 sdss	Olive oils	Multiway analysis	3-way PCA vs. PARAFAC	CLAS
Marhaba *et al.*, 2003	DOM	Multiway analysis	PLS	CAL
Scott *et al.*, 2003	olive oils	ANNs	ANNs (BP, SFAM, RBF) vs. PCA	CLAS
Stedmon *et al.*, 2003	DOM	Multiway analysis	PARAFAC	CLAS/CAL
Boehme *et al.*, 2004	DOM	Multiway analysis	PCA	CLAS
Antunes *et al.*, 2005	Humic-like fluorescence	Chemometrics	MCR-ALS	CAL
Guimet *et al.*, 2005	olive oils	Multiway analysis	3-way PCA vs. PARAFAC vs. N-PLS	CLAS
Hall *et al.*, 2005	ballast waters	Multiway analysis	PARAFAC vs. N-PLS	CLAS
Lee *et al.*, 2005	Fermentation processes monitoring	ANNs/multiway analysis	ANNs (SOM, BP) vs. PCA, PLS	CLAS/CAL
Rhee *et al.*, 2005	Fermentation processes monitoring	ANNs	ANNs (SOM)	CLAS
Divya and Mishra, 2007	fuels	Multiway analysis	PARAFAC vs. N-PLS vs. 3-way PCA	CLAS
Spencer *et al.*, 2007	DOM	Multiway analysis	PCA vs. DA	CLAS
Ohno *et al.*, 2008	DOM	Multiway analysis	PARAFAC	CLAS

residuals. In the decomposition models, such as PCA, PLS and PARAFAC, the parameters are obtained with constraints such as scores of orthogonality, maximization of the variance (PCA) or covariance of the independent and dependent datasets (PLS), and with a non-negativity constraint in PARAFAC models (Bro, 1998; Despagne and Massart, 1998).

The multivariate methods are often utilized in combination, where the preliminary analysis aims to reduce the amount of data and extract the most important features of the fluorescence data and the second step involves the calibration with known standards and concentrations. In particular, the outcome of data decomposition methods, e.g. PCA and PARAFAC, is often used as input to other models, e.g. regression models (MLR, PLS) or ANNs (Bro, 1998; Scott *et al.*, 2003; Lee *et al.*, 2005).

The main objective of this study was to apply selected advanced data analysis techniques to the fluorescence data characterizing WTW performance. Two fluorescence spectra decomposition methods for determination of number and type of fluorophores were compared; i.e. PARAFAC (the most-commonly used in EEMs analysis) and self-organizing map (SOM), a type of ANN. Additionally, different calibration algorithms (multiway partial least squares (N-PLS), MLR, ANNs) were tested for relating fluorescence data with TOC concentrations. In particular, the following questions were addressed:

- Can PARAFAC model outperform other decomposition techniques of fluorescence data?
- Can PARAFAC model outperform standard peak-picking technique in identification of additional spectral components?
- Which calibration technique (N-PLS, MLR, ANNs) best explains the fluorescence—TOC removal relationship?

2. MATERIALS AND METHODS

2.1. Fluorescence data

Fluorescence spectroscopy measurements and TOC analyses were carried out on samples of raw and partially-treated (clarified) water from 16 surface WTWs, collected monthly between August 2006 and February 2008 (Bieroza *et al.*, 2009). The treatment works are located in the Midlands region, central UK and are owned and operated by Severn Trent Water Ltd.

Organic matter fluorescence was measured using a Cary Eclipse Fluorescence Spectrophotometer (Varian, Surrey, UK), by scanning excitation wavelengths from 200 to 400 nm in 5 nm steps, and detecting the emitted fluorescence in 2 nm steps between 280 and 500 nm. Excitation and emission slit widths were set to 5 nm and photomultiplier tube voltage to 725 V. In order to maintain the consistency of measurement conditions, blank scans with a sealed cell containing deionized water and the measurement of the intensity of Raman line of water at 348 nm excitation wavelength were run systematically following the method described by Baker (2002). The mean Raman value during the study period was 22.3 intensity units, 1 S.D. = 0.5). All the fluorescence intensities were corrected and calibrated to a Raman peak intensity of 20 units at 396 (392–400) nm emission wavelength. Organic matter fluorescence was measured on unfiltered samples in 4 ml cuvettes. Arising from the fluorescence measurements, EEMs were obtained for each water sample, displaying the intensity of fluorescence within the sample against the wavelengths at which excited fluorophores emitted the light.

TOC was measured using a Shimadzu TOC-V-CSH analyser with auto-sampler TOC-ASI-V. The non-purgable organic carbon (NPOC) determination method was employed and the result NPOC was calculated as a mean of the three valid measurements. The typical error of the analyses was less than 10%, indicating sufficient precision of the TOC measurements.

Each month, water samples were collected from the 16 WTWs over a period of 2 days; the samples being stored cool and in the dark until analysis, between 3 and 7 days from collection. Storage test experiments were undertaken to demonstrate that degradation of water samples was insignificant under these storage conditions (typical variations, including both increase and decrease in fulvic-like (peak C) and tryptophan-like fluorescence intensities was 4.4 and 5.3% respectively; typical change in TOC was less than 5%). The typical error of the fluorescence analyses was less than 5%. Sample stability assessment was carried out on the basis of peak-picking approach, and so the results should be treated as an estimation of real changes in each fluorescence region.

Examples of EEMs of raw and clarified waters are given in Figure 1. The decrease in fluorescence intensity in fulvic-like region between raw and clarified water can be correlated with organic matter removal, measured as a decrease in TOC. Both fluorescence and TOC measurements were carried out independently on paired samples, providing an overall, significant relationship between observed decrease in fulvic-like fluorescence intensity and TOC removal ($R^2 = 0.90$).

2.2. Data pre-processing

Data were processed with scripts written in Matlab® 7.0. In particular, the N-way toolbox for Matlab (Andersson and Bro, 2000) and SOM toolbox version 2 (Kohonen, 1998) were used for the implementation and validation of the PARAFAC, N-PLS and SOM models. All data analyses were carried out on a 512 MB Dual Pentium III PC computer.

Prior to modelling, the fluorescence data were pre-processed. Firstly, fluorescence spectra were normalized to the Raman scatter peak (at 348 nm excitation wavelength) of deionized water by subtracting the Raman signal from the raw data (Nieke *et al*., 1997; Determann *et al*., 1998; Stedmon *et al*., 2003).

The Rayleigh (both first and second order) and Raman scattering were handled by removing and replacing with interpolated values (Bahram *et al*., 2006). Scatter should be removed before further data analysis to enhance the EEM modelling efficiency of decomposition models such as PCA and PARAFAC (Bahram *et al*., 2006). The position of the Rayleigh scatter was defined as the diagonal, where the excitation wavelength equals that of the emission (first-order) or double excitation (second-order). The Raman position was at a constant energy shift with respect to the first-order Rayleigh scatter. The widths of Rayleigh (first- and second-order) and Raman scatter were set at [25, 25], [10, 10] and [10, 10] nm, respectively.

Finally, data scaling was performed to reduce the concentration effects exhibited by intensity (Boehme *et al*., 2004). Both single centring and scaling of the dataset was applied (centring across first mode and scaling within first mode to standard deviation; Bro, 1997).

To enhance the speed of analyses, regions of EEMs containing limited (low fluorescence regions) and redundant information (Rayleigh and Raman, region of excitation wavelengths larger than emission wavelengths) were removed (strips of removed data can be observed in Figure 1). As a result, the output EEMs ranged from 220 to 400 nm excitation and from 300 to 500 nm emission wavelengths, respectively.

The fluorescence data in regions where excitation exceeded emission were replaced with missing data (NaN, inserted at emission data at wavelength < Ex wavelength + 20 nm) and zeros (inserted at emission wavelengths that are < Ex wavelength – 20 nm) (Bro, 1998; Stedmon *et al*., 2003).

2.3. Methods

2.3.1. Multiway analysis

EEMs are examples of three-way arrays, where for every sample the fluorescence emission intensity is measured at several wavelengths for different excitation wavelengths (Bro, 1998). As a result, fluorescence data can be represented as a three-way matrix, $I \times J \times K$, where the index I refers to the sample number, J to emission wavelength and K to the excitation wavelength. For the analysis of the multiway data (data arranged in three or more dimensions), the most appropriate approach is the multiway analysis, which is the extension of multivariate analysis. Similarly to bidimensional models of data decomposition like PCA and PLS, in the multiway analysis the fluorescence data are decomposed into sets of scores and loadings, describing the most important features of the original dataset (Bro, 1997).

Multiway methods used in the fluorescence data analysis include the PARAFAC model producing unique, chemically meaningful solutions, multiway PCA models (also described as Tucker models) and N-PLS, which is the two-block model with both sets of independent and dependent data being decomposed simultaneously (Bro, 1998).

The advantage of the three-way models results from the adequacy of the fluorescence data format (sample × emission × excitation) and the model structure, which reflects the true underlying parameters (Bro, 1998). Essentially, the PARAFAC model provides an estimation of the fluorophore number and relative concentration in each sample, as well as the spectral properties of the components (emission and excitation

wavelengths), which can be used to interpret the chemical composition of the components (Bro, 1998; Stedmon *et al*., 2003; Ohno *et al*., 2008). Consequently, the components in the three-way models are easier to interpret than in the corresponding two-way models (e.g. PARAFAC vs. two-way PCA). Nevertheless, the robustness of three-way models is highly dependent on trilinearity of the data and selection of the correct number of components.

The theory and applications of the multiway methods have been thoroughly presented in the literature (Bro, 1997, 1998; Henrion *et al*., 1997; Andersson and Bro, 2000; Guimet *et al*., 2005; Hall *et al*., 2005). Examples of the fluorescence data analysis using PARAFAC and N-PLS include characterization of the organic matter–metal binding process (Ohno *et al*., 2008), quantitative determination of the kerosene fraction present in diesel (Divya and Mishra, 2007), classification of ballast water (Hall *et al*., 2005), estuarine water (Stedmon *et al*., 2003) and edible oils (Guimet *et al*., 2005).

2.3.2. Artificial neural networks

ANNs are powerful computational tools, frequently used in modelling studies (Andrews and Lieberman, 1994; Häck and Kohne, 1996; Daliakopoulos *et al*., 2005; Alunkaynak, 2007). ANN can be described as a mathematical model of a specific structure, consisting of the number of the single processing elements (nodes, neurons), arranged in inter-connected layers. Active neuron multiplies each input vector by its weight, sums the products and passes the sum through a transfer function to produce the output.

There is a substantial amount of published work which employed multifarious ANNs in modelling fluorescence and water quality data. Scott *et al.* (2003) used ANNs for pattern recognition of olive oil fluorescence spectra. Li *et al.* (2000) applied ANNs for classification of the fluorescence properties of multicomponent mixtures of fluorescent dyes, whereas Wolf *et al.* (2001) associated the fluorescence measurements of biofilm reactors with process performance parameters. A similar approach was presented by Häck and Kohne (1996) to estimate wastewater process parameter concentrations (e.g. NH_4) using ANNs (input data included pH value, conductivity, redox potential, turbidity).

The self-organizing map (SOM, Kohonen ANN) approach provides the conversion of nonlinear statistical relationships between high-dimensional data into simple geometric relationships on a low-dimensional map, while keeping the most important topological and metric relationships of input data (Kohonen, 1998; Rhee *et al*., 2005; Garcia *et al*., 2007). A complete description of the SOM algorithm can be found in (Kohonen, 2001).

SOMs are often used in pattern recognition and feature extraction, data compression and as a pre-processing tool for other networks (Hammerstrom, 1993). Examples of SOMs use include classification of fluorescence data to monitor the fermentation processes (Lee *et al*., 2005; Rhee *et al*., 2005), and exploratory analysis of metalloproteins based on X-ray fluorescence analysis of the metal ions (Garcia *et al*., 2007).

3. RESULTS

3.1. Multiway analysis

3.1.1. PARAFAC

A PARAFAC model with non-negativity constraint on all modes (samples, emission and excitation) was implemented in Matlab® (Bro, 1997). For each water type (raw and clarified) and for the whole dataset, the PARAFAC models were calibrated independently.

Initially, the series of models with varying number of components were run to discern the outlying data in each mode (samples, emission and excitation wavelengths) by analysing the leverages and shape of the emission and excitation loadings of each model created. Due to distinctively higher deviations from the average for all the data, the excitation wavelengths lower than 240 nm were excluded from the further analysis (low signal to noise ratio).

Core consistency analysis (CORCONDIA) was performed to determine the optimum number of components (Bro, 1997). The appropriate models with the highest number of components, highest variance explained and the valid CORCONDIA value (raw 84.6%, clarified 92.1%, all data 90.7%) were selected (Table 2). Additionally, the residual sum of squares was analysed and plotted against the number of components. The curves flatten out for three components indicating the optimum number of PARAFAC components, which is in accordance with the results obtained from CORCONDIA.

However, to derive a valid number of PARAFAC components, not only the statistical diagnostics were evaluated but also the emission and excitation spectra were analysed (Figure 2).

The emission loadings demonstrate well-defined, single peaks, whereas in the excitation loadings, double-peaks for first two components, and increased signal in loading of third component between 245 and 260 nm, can be discerned. This inferior appearance of the excitation loadings could be attributed either to the presence of some unexplained by trilinear model variance related to another component (visible as shoulders in the loadings) and/or model impediment due to occurrence of non-systematic noise in the fluorescence signal at lower excitation wavelengths. Therefore, the validity of additional PARAFAC models was evaluated, adopting approach presented by Andersen and Bro (2003). The original PARAFAC solution was found with the application of solely non-negativity constraint on all three modes (samples, emission and excitation wavelengths). As suggested by Andersen and Bro (2003), the unimodality constraint can significantly reduce the interference from the minor artefacts in the data and thus can enhance the interpretation and selection of the valid components. For raw, clarified and all-dataset fluorescence spectra, three- and four-component PARAFAC models with the unimodality constraint in the emission and excitation modes were tested. The application of the unimodality constraint in case of three-component model did not improve the percentage of the explained variance and core consistency diagnostic, however, decreased the shoulder observed in the loading of the third component. For the four-component constrained model the variance explained remained unchanged compared with the results of unconstrained models, with minor increase in core consistency value for the raw water model (from 4.5 to 8.5%). For the clarified water and

Table 2. Variance explained, residual sum of squares and CORCONDIA result of three PARAFAC models for first seven PARAFAC components (C)

C	Raw			Clarified			All		
	Variance explained (%)	Residual sum of squares	CORCONDIA (%)	Variance explained (%)	Residual sum of squares	CORCONDIA (%)	Variance explained %	Residual sum of squares	CORCONDIA (%)
1	97.0	34 005.5	100.0	97.8	29 913.9	100.0	97.7	70 243.9	100.0
2	99.2	13 982.5	95.4	98.7	18 230.5	98.9	98.9	33 747.4	96.9
3	99.4	9546.2	84.6	99.1	13 175.8	92.1	99.2	24 001.5	90.1
4	99.5	7645.6	4.5	99.2	11 518.2	−6.7	99.3	20 623.0	14.1
5	99.6	6937.0	5.0	99.2	11 112.7	−5.4	99.4	19 533.2	3.3
6	99.6	6699.3	7.4	99.2	10 950.7	8.7	99.4	18 085.9	−1.9
7	99.6	6320.9	1.0	99.3	10 489.7	1.2	99.4	17 267.9	1.1

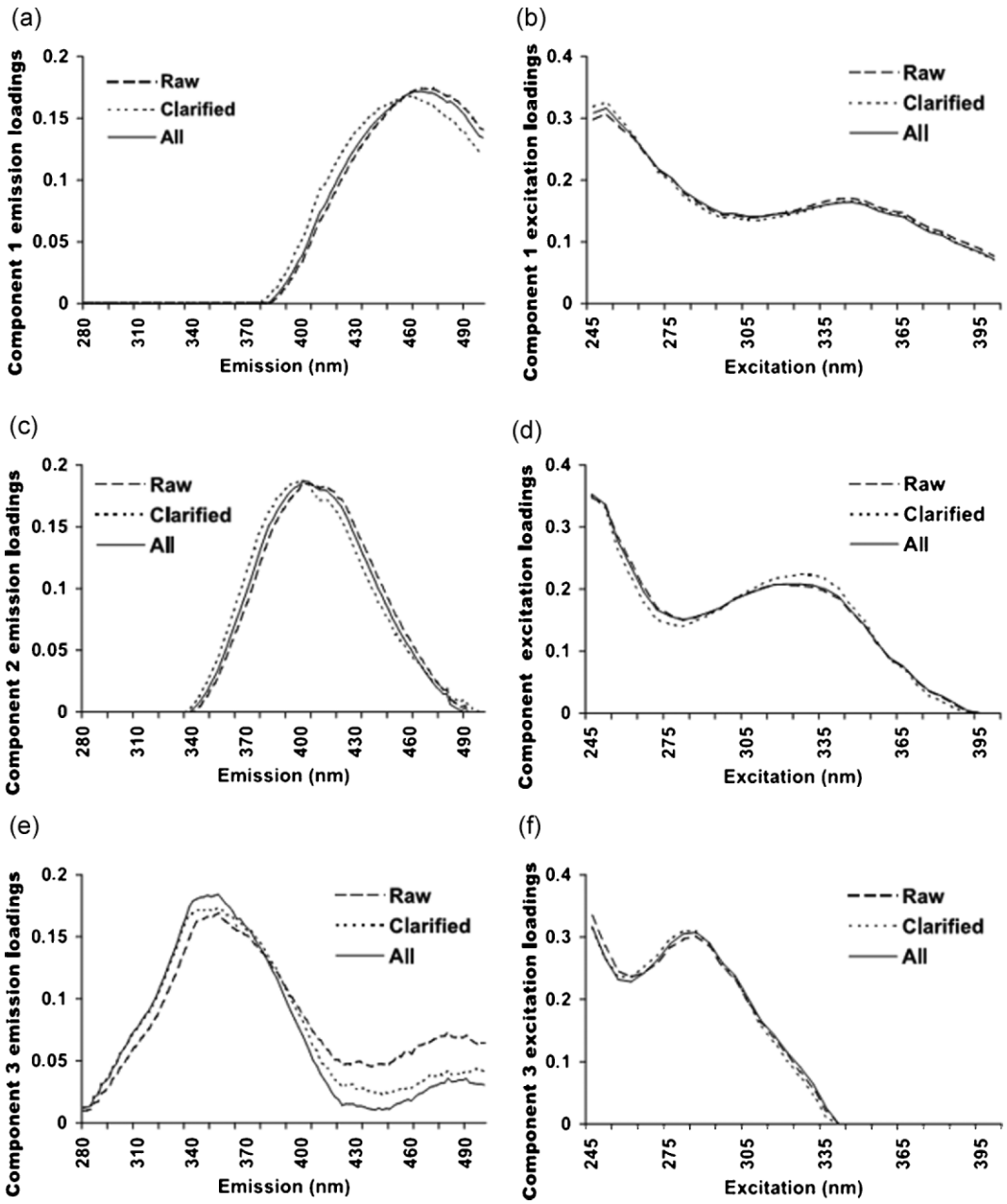

Figure 2. Three-component PARAFAC model excitation and emission loadings of raw, clarified and all data. Component 1: emission (a), excitation (b); Component 2: emission (c), excitation (d); Component 3: emission (e), excitation (f)

all-dataset models a corresponding decrease in CORCONDIA percentage was observed. More explicit were differences in emission and excitation loadings in four-component PARAFAC models (Figure 3).

In the four-component model, the second and fourth components (Figure 3c, d, and g, h) exhibit similar spectral properties to the corresponding second and third components from the unconstrained models (Figure 2c–f). However, the initial first component from the three-component model was decomposed in two separate components, the first and third, with similar emission loadings but varying excitation loadings (maxima below 245 and at 350 nm). In the clarified water four-component model, the emission and excitation spectra of the first and second components demonstrate different signatures, with lower emission in the first component loading and a distinctively unique spectral position in the excitation profile of the second component. Hence, the PARAFAC analysis reveals a variation in fluorescence character over the range of samples, pertinent to various organic matter sources, and water types, which reflects removal of specific organic matter fractions during the treatment process.

From the visual appearance of the four-component clarified water model, it appears that the first three components might be a linear combination of two real components, and thus this solution seems to be not valid. The instability of the four-component can be inferred from the CORCONDIA analysis as indicated with negative value of this diagnostic (−6.5%) for clarified water and low value (8.5%) for raw water.

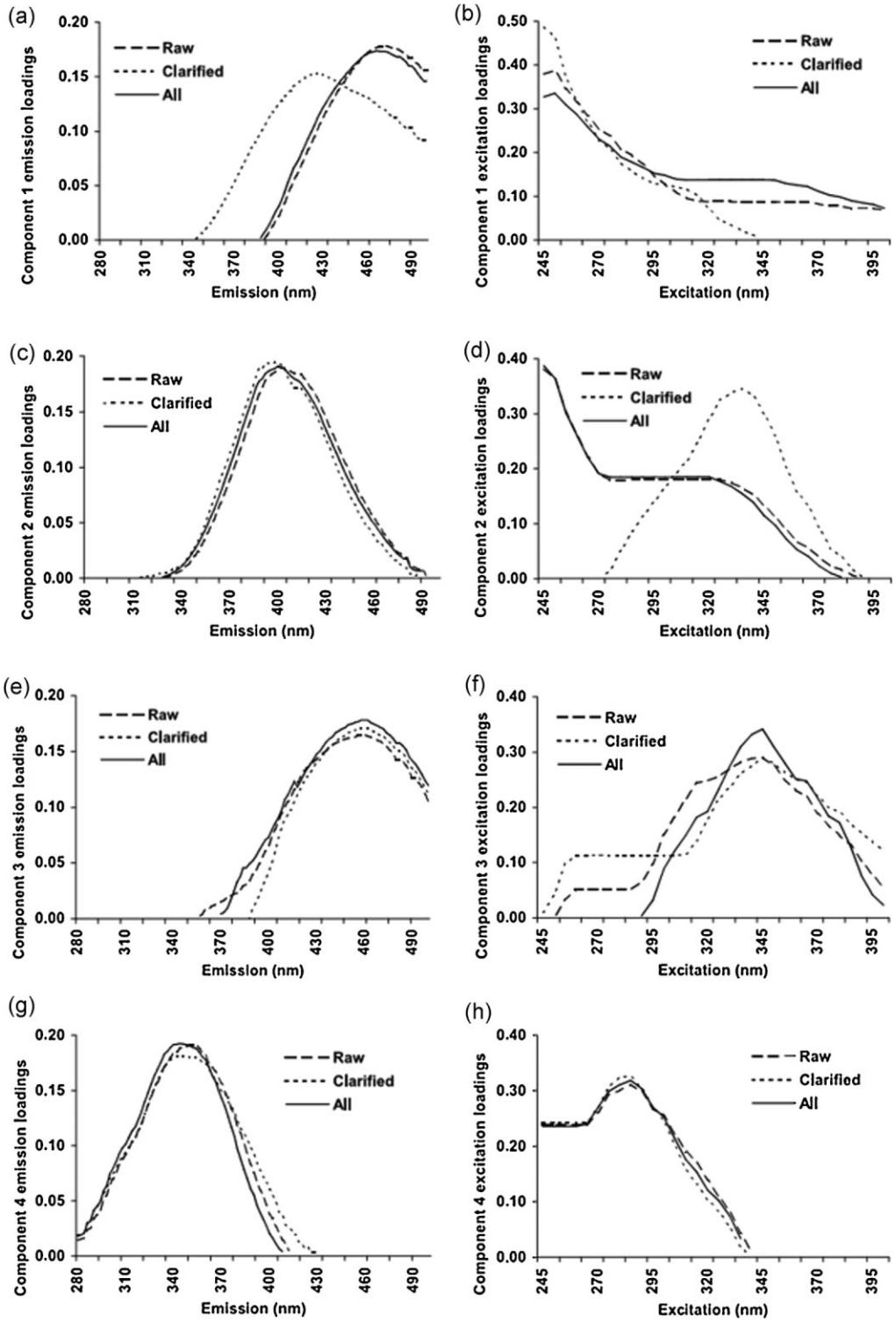

Figure 3. Four-component PARAFAC model excitation and emission loadings of raw, clarified and all data. Component 1: emission (a), excitation (b); Component 2: emission (c), excitation (d); Component 3: emission (e), excitation (f); Component 4: emission (g), excitation (h)

A valid model in terms of statistical diagnostics, with consistent emission and excitation loadings for raw, clarified and all fluorescence data was obtained for three components. However, from the analysis of the loadings it can be seen that two first components produced double-excitation maxima indicative of the presence of more components. An evaluation of more complex models (four and five-components) confirmed this hypothesis. Both the components with emission wavelength around 400 and 460 nm can be partitioned into two more components of different excitation spectra (less than 245 and at 350). Despite this inherent complexity of the components, the

higher-component models were unstable as derived from CORCONDIA analysis (low and negative values) and from the repeated PARAFAC analyses. For the same datasets, different PARAFAC models were found, which is in contradiction with the PARAFAC model assumption of producing a unique solution.

Therefore, in the further analysis, only the three-component PARAFAC model was validated with the split-half analysis and the residuals evaluation. For three components, the models were found to be valid with the variance explained for raw, clarified and all data being 99.5, 99.2 and 99.3%, respectively.

The number and order (variance explained) of the three components derived independently from raw, clarified and all data models exhibit the same pattern with three components, where emission and excitation loadings resemble each other (Figure 2).

Component 1 exhibits a wide emission spectra with maxima at 460 and 470 nm for clarified and raw water, respectively (Figure 2a), with the two excitation maxima at 250 and 345 nm (Figure 2b). The emission and excitation maxima of component 2 occur at shorter wavelengths compared to component 1 (402 nm and below 245 nm; Figures 2c and d). The emission spectrum of component 3 is maximum at 350 nm wavelength with excitation spectra consisting of two peaks (below 245 and 285 nm; Figures 2e and f). Results of earlier studies (Coble, 1996; Bro, 1997; Blough and del Vecchio, 2002; Stedmon *et al*., 2003; Hall *et al*., 2005) indicate that the three PARAFAC components can be identified as a visible humic-like fluorescence associated with peak C (component 1), UV humic-like fluorescence corresponding to peak A (component 2) and protein-like fluorescence (peak T, Tryptophan, component 3). When compared with four-component constrained PARAFAC model, the fourth component (Figure 3g and h) corresponds to Tryptophan-like fluorescence and its emission and excitation spectra are similar to the third component in three-component model (Figure 2e and f). However, the application of the unimodality constraint reduces the 'shoulder' in Tryptophan excitation loading between 245 and 260 nm. The difference between water types can be discerned from the analysis of the first three components. The raw water and all data four-component models exhibit the presence of two UV humic-like fluorescence peaks with different emission maxima (the first and second component; Figure 3a–d), and one visible humic-like fluorescence as a third component (peak C; Figure 3e and f). For the clarified model, the first component corresponds to UV humic fluorescence with lower emission wavelength (430 nm; Figure 3a and b), and the second and third to visible humic fluorescence with emission maximum at 400 nm (second; Figure 3c and d) and at 470 nm (third; Figure 3e and f). From this comparison appears that in the clarification process the UV humic-like fluorescence of higher emission wavelength is removed more than other fractions.

The score plots reveal interesting correlations between humic-like and tryptophan-like fluorescence. The raw water characteristics of reservoir sites 1, 6 and 14 exhibit minor variations in the microbial fraction compared to sites 3 and 7, which are prone to significant changes in tryptophan-like fluorescence intensity related to the distinctive contribution of algae (data not presented here). This division into two groups of sites, first with stable tryptophan-like (variation in humic-like fluorescence which can be indicative of the presence of different fractions) and second with stable fulvic-like fluorescence (pre-dominance of the one fraction of humic organic matter with variable microbial inputs) is even more discernible for clarified water scores. Furthermore, for clarified water a significant relationship exists ($R^2 = 0.92$) between scores of components 1 and 2.

For the majority of sites, a similar pattern of temporal variation in PARAFAC scores can be inferred. For both sites the score variation is greater for clarified water. Components 1 and 2, corresponding to fulvic and humic-like fluorescence (respectively peak C and peak A fluorescence), exhibit increased scores in the winter with maximum in January, whereas the lower values are pre-dominant over the summer months. The opposite applies to component 3, which is indicative of microbial fraction of organic matter content. The winter and late summer months (July) can be characterized with lower scores on component 3 and elevated values can be observed in the April and May.

These PARAFAC results on the composition of organic matter are in accordance with visual inspection of EEMs and peak-picking approach suggesting that there are three main fluorophores present in the water samples. Moreover, the analysis of PARAFAC models with higher number of components suggests the presence of more fluorophores, which cannot be validated for the entire dataset or even water types (raw and clarified). Therefore, to derive a valid model containing all variation in fluorescence spectra, a solution with fewer components has to be chosen. A lack of good, overall diagnostic for the selection of the number of valid components impedes an interpretation of PARAFAC model and makes an analysis a time-consuming process.

3.1.2. N-PLS

The scores of the three PARAFAC components for raw and clarified water models, which can be treated as a chemical representation of the fluorescence data, were correlated with the organic matter removal parameter (actual TOC removal) via the two-way PLS method (PLS-PARAFAC model). Additionally, two N-PLS models of raw and clarified samples were correlated with TOC removal independently (N-PLS-RAW and N-PLS-CLA models) and the MLR model containing PARAFAC scores as the independent variables was developed.

The X-block in PLS-PARAFAC model comprised of 290 samples (rows) × 6 PARAFAC scores (columns), whereas for both N-PLS-RAW and N-PLS-CLA models the whole EEMs were exploited, producing the three-way X matrices of size 290 × 111 × 32, where the second and third indices describe the emission and excitation wavelength number. The datasets were divided into calibration and validation sets by selecting 30% of validation samples covering the whole range of data variation according to the discrimination obtained from the projection of the first PARAFAC components of raw and clarified water.

The datasets were mean-centred across the first mode prior to all N-PLS analyses. The optimum number of components was selected according to leave-one-out cross-validation. For four latent variables (LVs) in PLS-PARAFAC model, the root mean squared error of prediction (RMSEP) is the lowest, indicating the appropriate number of components. For N-PLS models of raw and clarified samples, the best fit of the model was obtained for six LVs.

The quality of predictions of the developed models was evaluated using the total variance explained (independent X-block and dependent data Y-block) and the root mean squared error of cross-validation (RMSECV), which accounts for the ability of a model to predict new sample. The PLS-PARAFAC model demonstrated a better fit with the variance explained by the four-component model $y = 89.4\%$, $x = 97.3\%$ and cross-validation error RMSECV = 5.3% when compared with 6-components N-PLS-RAW ($y = 86.8\%$, $x = 81.1\%$,

Table 3. Variance explained and root mean squared error of validation (RMSECV) of the three PLS models for first seven latent variables (LV)

LV	PLS-PARAFAC			N-PLS-RAW			N-PLS-CLA		
	Variance explained (%)		RMSECV (%)	Variance explained (%)		RMSECV (%)	Variance explained (%)		RMSECV (%)
	X-block	Y-block		X-block	Y-block		X-block	Y-block	
1	45.4	75.7	9.0	34.9	74.3	10.2	62.8	37.9	14.7
2	86.3	81.4	7.5	72.5	79.7	8.5	71.4	75.9	9.6
3	91.5	87.5	6.1	75.1	81.6	9.4	76.1	77.2	9.9
4	97.3	89.4	5.3	78.8	82.2	9.5	78.5	79.9	9.4
5	99.7	89.7	5.4	79.8	85.2	9.3	79.0	83.9	9.1
6	100.0	89.7	5.4	81.1	86.8	8.6	79.3	87.6	8.0
7	100.0	89.7	5.4	81.5	88.5	9.3	79.5	90.1	8.4

RMSECV = 8.6%) and N-PLS-CLA ($y = 87.6\%$, $x = 79.3\%$, RMSECV = 8.0%) models (Table 3). These results confirm that the correlation between TOC removal and fluorescence properties is stronger when raw and clarified waters are modelled simultaneously (as a combination of PARAFAC scores) than when EEMs of raw and clarified water are modelled independently (N-PLS-RAW and N-PLS-CLA). Thus, the PARAFAC method provides a powerful decomposition tool and enables extraction of the most important spectral features from the original datasets. Data pre-processing with PARAFAC prior to multivariate calibration makes complex fluorescence data analysis feasible (where the independent dataset consists of two water types). Additionally, a multilinear regression model was applied to the fluorescence data, as this method is commonly used in multivariate calibration.

Figure 4a and b presents the relationship between actual and modelled TOC removal data for the PLS-PARAFAC and MLR analyses. PLS method provides better determination of the TOC removal (correlation coefficient $R^2 = 0.93$ (PLS), $R^2 = 0.82$ (MLR)).

Scores of the first two LVs in the PLS-PARAFAC model of both the training and test set facilitate discrimination between sites. Sites 1 and 6 are clearly separated from the rest of the sites as indicated by the positive correlation with the first LV. These sites exhibit distinctive organic matter properties; namely prevailing hydrophobic character and higher molecular weight, which can give rise to improved TOC removal efficiency. Therefore, the first LV of the PLS-PARAFAC model can be related to the degree of hydrophobicity, as indicated by the negative correlation with more hydrophilic sites (16, 10 and 5).

The N-PLS prediction of the TOC removal separately from the fluorescence data of raw and clarified water results in similar correlation coefficients (0.81 for raw and 0.84 for clarified). The emission loadings in N-PLS-RAW model indicate that factors 2 and 4 are indicative of fulvic-like fluorescence and microbial fraction (tryptophan-like fluorescence). The blue-shifted emission wavelength of the corresponding factor 2 for clarified water confirms the improved removal of more hydrophobic organic matter fraction. An interesting correlation between scores of the first two LVs (identified as peak C emission wavelength and peak C intensity) of the N-PLS-CLA model can be discerned. Whilst the overall correlation coefficient value is 0.64, the correlations for particular sites appear to be stronger (e.g. site 1, $R^2 = 0.97$). This would appear to suggest that the sites are optimized individually to TOC concentration. For the whole dataset, the peak C intensity vs. TOC removal correlation demonstrates significant variability in both TOC concentration and organic matter composition between the individual sites.

3.2. Artificial neural networks

3.2.1. SOM

Prior to construction of the map, the pre-processed fluorescence data (see Section 2.2) was unfolded to a one-dimensional vector and autoscaled along all spectral variables, with mean variance of each variable set to zero and a standard deviation of one. This resulted in the generation of a 2515×625 matrix, where the first dimension denotes the number of unfolded fluorescence emission–excitation pairs and the second is the number of raw and clarified samples. The complete input matrix was presented to the nodes of SOM input layer simultaneously (batch mode) and neuron weights were initialized using linear initialization along the two greatest eigenvectors. The size of the output layer was determined by training different maps and comparing the quality measures, final quantization error (the average distance between each input vector and its best-matching unit (BMU) for measuring map resolution) and final topographic error, defined as the proportion of all input vectors for which first and second BMUs are not adjacent (for measuring topology preservation; Kohonen, 2001). The optimum map contained 258 units (size 43×6), calculated as a ratio of two greatest eigenvalues of the training data. The final quantization and topographic errors were low; 0.72 and 0.04 respectively.

The unified distance matrix algorithm (U-matrix; Ultsch, 1993) and the k-means algorithm (Jain and Dubes, 1998) were used to distinguish the subsets of fluorescence data (clusters). The U-matrix is the most common graphical representation of the SOM structure, in which distances between neighbouring map units are calculated and visualized using grey or colour scale on the trained map (Park *et al*., 2003). High values on the U-matrix indicate large distance between neighbouring units and thus can be helpful in determining the cluster borders. Clusters typically form uniform areas of low values.

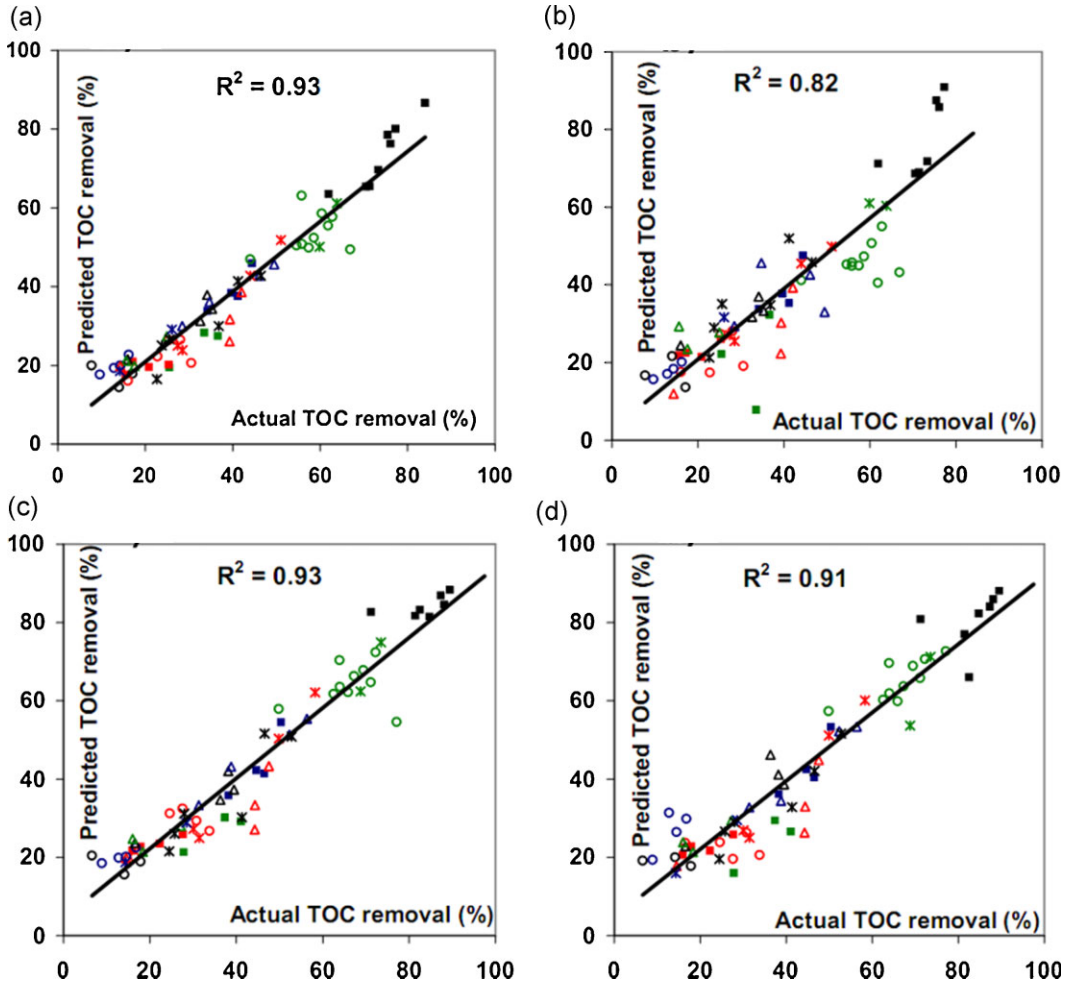

Figure 4. Predicted versus actual TOC removal for PLS-PARAFAC (a), MLR model (b), PARAFAC-BP (c) and SOM-BP (d). This figure is available in colour online at wileyonlinelibrary.com./journal/environmetrics

The number of clusters determined with the k-means algorithm was chosen according to the correlation between mean squared error and number of clusters and the Davies–Bouldin index (DBI; Davies and Bouldin, 1979). The difference of the mean squared error between three and four clusters was less than 5% and the DBI was the lowest for three clusters. Therefore, the fluorescence data can be classified into three classes. This result is in accordance with the PARAFAC analysis, where the valid number of components was also three.

From the analysis of U-matrix and component planes for two sites of contrasting organic matter properties (site 1 and site 7), important characteristics of selected clusters can be inferred (Figure 5).

The component planes depict the values of prototype vectors for different variables (components). It can be seen that for each sample component planes project the EEMs onto a lower-dimensional map. Horizontal and vertical axes correspond to fluorescence emission and excitation wavelengths, with increasing values from the bottom to the top and from the left to the right, respectively. The raw water fluorescence signature of Site 1 exhibits increased intensities in peak A and C regions (UV humic- and fulvic-like fluorescence), whereas the intensities at site 7 are blue-shifted (lower emission wavelengths) indicating more hydrophilic character of organic matter. Moreover, the presence of increased tryptophan-like fluorescence can be discerned in the component plane of site 7 for both raw and clarified waters. The partitioning of the U-matrix into three classes is based on differences in excitation wavelength, as the transition from cluster 1 to cluster 3 corresponds to the increasing excitation wavelengths. Therefore, cluster 1 can be identified as UV humic-like fluorescence (peak A), and both cluster 2 and 3 contain the fulvic and protein-like fluorescence of higher (cluster 2) and lower (cluster 3) excitation wavelengths (Figure 5b). A similar decomposition of fluorescence spectra was obtained with PARAFAC analysis, where three factors were discerned regarding peak C, peak A and tryptophan-like fluorescence. These results indicate that SOM can be considered as a surrogate technique of fluorescence spectra decomposition to PARAFAC modelling, with advantages including easier interpretation of the results.

For each sample, normalized weights of three SOM classes were obtained and used as an input matrix for training and validating back-propagation neural network.

3.2.2. BPNN

A three layer feed-forward type of neural network with back-propagation learning rule, the most popular type of ANN, was used to calibrate fluorescence and TOC removal data.

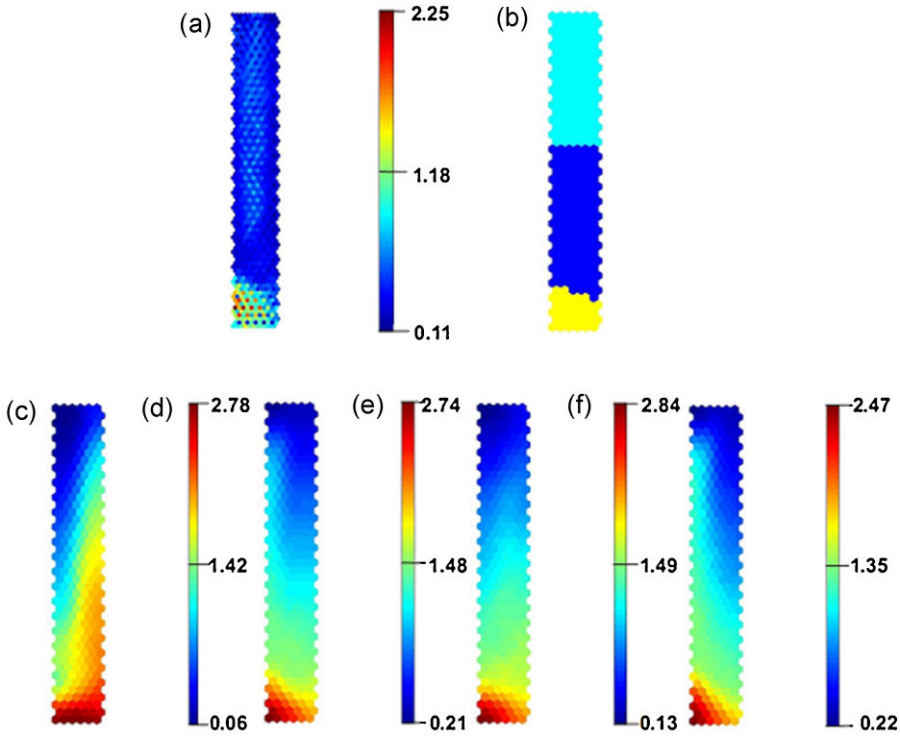

Figure 5. Visualization of the SOM map for fluorescence data: U-matrix (a), clusters (b), component planes: sample 12 (c, site 1, September 2007, raw), sample 144 (d, site 7, September 2007, raw), sample 144 (e, site 1, September 2007, clarified), sample 466 (f, site 7, September 2007, clarified). This figure is available in colour online at wileyonlinelibrary.com./journal/environmetrics

Firstly, the network topology (number of layers and number of nodes in hidden layer) was assessed, to produce a network large enough to train the fluorescence data but small enough to generalize well the regularities within the data. To obtain a feasible network with acceptable training times (typically less than 1 h), the total size of the input fluorescence intensity data (111 emission × 32 excitation wavelengths × 2 water types) was reduced. In the first case, the PARAFAC scores of raw and clarified water were used as an input vector (six nodes, PARAFAC-BP model). The second approach employed the information on useful emission–excitation combinations derived from the SOM classification (SOM-BP model).

The input layer comprised six nodes, while the output had only one node, denoting the TOC removal. One hidden layer with sigmoid transfer function was used in the model, as recommended for the purpose of multivariate calibration by others (Smits *et al*., 1994; Despagne and Massart, 1998). Initially, the cross-validation algorithm was used as a convergence criterion to optimize the learning epoch size and avoid overtraining and to select the optimum number of hidden nodes (Li *et al*., 2000; Wolf *et al*., 2001). Additionally, the calibration error for different settings of networks was compared. The topology corresponding to the lowest error determined in five trials with different sets of initial weights was retained for the validation.

The best test network architecture for both models (six input nodes, four hidden nodes, one output node) was set to train with the same training set used for N-PLS analysis until convergence was achieved. The Levenberg–Marquardt algorithm (Marquardt, 1963) with early stopping for improving generalization and preventing data overfitting was employed for training the networks. The Levenberg–Marquardt technique is faster and less easily trapped in local minima than other optimization methods, as the back-propagation (BP) momentum coefficient is decreased during the training in relation to error gradient information (Marquardt, 1963; Basheer and Hajmeer, 2000). The early stopping algorithm monitors the validation error and the training stops when data begin to overfit as indicated by the increasing validation error.

The average squared error between network outputs (predicted by network) and targets (actual TOC removal; mean square error (MSE)) was used as a network performance function. Calibration, validation errors and the training times were in accordance for both back-propagation models (PARAFAC-BP performance: calibration (0.005), validation (0.004) and SOM-BP performance: calibration (0.006), validation (0.005)).

4. DISCUSSION

The results of validation of the neural networks are given in Figure 4 (c, d), where outputs of the networks (PARAFAC-BP and SOM-BP) were plotted against target values. The correlation coefficients for both models were similar (PARAFAC-BP $R^2 = 0.93$ and SOM-BP $R^2 = 0.91$) and consistent with PLS models (PLS-PARAFAC and MLR). However, three models (PARAFAC-PLS, PARAFAC-BP and SOM-BP) would appear to outperform the MLR model in respect of the strength of the relationship. The similar results obtained for the

validation dataset suggest that the combined approaches of multiway analysis and ANNs can be successfully utilized in modelling of fluorescence data. Here, prior to calibration of fluorescence data with organic matter removal derived from actual TOC concentrations, two decomposition methods were applied. This enabled the reduction of initial dataset dimensionality from more than 2500 emission–excitation wavelengths to few PARAFAC components or SOM clusters, and therefore to extract most important features of the dataset in form of PARAFAC scores and loadings and SOM normalized weights. The similar results obtained in the calibration, confirm that both methods adequately retrieved information on the most important fluorescence features of the datasets.

The PARAFAC solution has an advantage over other decomposition methods including PCA or SOM, of generating components that can be related to the real fluorophores on the basis of emission and excitation loadings, provided that an appropriate number of components are selected. However, it was found in our study that the determination of the number of PARAFAC components is a non-trivial task, as the different model diagnostics provided equivocal results. Therefore, this approach requires a close supervision of a domain-expert in the validation process. On the other hand, the SOM data decomposition is achieved by the positioning of samples on two-dimensional map according to their metric and topological properties. As the SOM algorithm operates unsupervised, the produced output needs to be evaluated with the use of different statistical and visualization techniques, which applicability is data-dependent. Therefore, similar to the PARAFAC approach, model evaluation necessitates the use of different validation tools. Moreover, both methods require careful data pre-processing including outlier removal in all three modes (samples, emission and excitation wavelengths). The SOM solution compared with PARAFAC derived for drinking water data appears to be more stable as the ANNs are more noise-tolerant models. As discussed before, for the validated three-component PARAFAC model, appears that there is some unexplained variance remaining that pertains to another components or noise in the data. However models with higher number of components were unstable and not valid.

The regression models used in this study, PLS and ANN with back-propagation algorithm (back-propagation neural networks (BPNN)), produced consistent results as indicated with correlation coefficients (Figure 4). The BPNN approach is more flexible than the standard optimization methods and therefore more difficult to implement as its flexibility can pose a danger of overfitting the calibration data and producing unreliable results. Moreover, prior to modelling, the BPNN network requires defining a topology (number of nodes in a hidden layer) and parameters that describe the speed of the training process and maintain the optimum search stability (the learning rate and the momentum factor). There are several rules of thumb that facilitate BPNN network design but a trial-and-error procedure is commonly used to adjust back-propagation algorithm parameters and obtain a feasible network topology. However, once appropriately designed, trained and validated, the BP network can be a robust predictive tool provided that a substantial amount of training data was available. The BPNN solution advantages include also fault and noise tolerance (ability of processing noisy, uncertain data), self-modelling, self-learning (by example) and generalization capabilities.

The PLS algorithm appears to be simpler as it does not require defining a set of parameters in the training phase. The validation procedure involves selection of the appropriate number of LVs, which is the crucial step in generating a valid and robust model. However, there are techniques and statistical diagnostic available that enhance this process, like the leave-one-out cross-validation and the RMSEP. The PLS algorithm explicitly incorporates dimension reduction and generates components explaining the most important features of the given dataset. Therefore this regression tool can be successfully used in exploratory data analysis. Here, the analysis of PLS scores provided a useful information relating the organic matter properties (degree of hydrophobicity) with efficiency of organic matter removal. Thus, for the purpose of prediction of organic matter removal in drinking water systems with fluorescence analysis the PLS method appears to be more suitable and outperforms the BPNN technique.

5. CONCLUSIONS

The application of different decomposition and calibration models used for fluorescence data has been presented. In particular, fluorescence data characterizing organic matter removal across different stages of the treatment process was first reduced with PARAFAC and SOM models prior to calibration with actual TOC removal data. It was found that fluorescence data decomposition can be successfully performed using the PARAFAC or SOM approaches. Although the number of fluorophores present in a water sample can be determined with visual inspection of EEM, with the substantial number of EEMs involved in analysis this method is not efficient. Applied decomposition methods, e.g. PARAFAC and SOM, can enhance the speed of analysis and provide quantitative information on fluorescence data in the form of PARAFAC scores or SOM normalized weights, together with qualitative assessment of fluorophores (number and type). Both methods require time-consuming data preparation (typically several hours presuming user programming proficiency) including removal of regions of EEMs containing redundant information (i.e. Raman and Rayleigh scatter). The most popular PARAFAC model requires supervised evaluation of the obtained components, while for the unsupervised SOM approach the characteristic features of the data are selected automatically. This is a distinct advantage when analysing the samples from similar sources and with a uniform pattern of few fluorophores.

The decomposed fluorescence data into PARAFAC scores or SOM weights can be further exploited and correlated with calibration data. From the results obtained in this study, the best results of fluorescence—TOC removal calibration were obtained with the use of PLS and BP techniques, whereas the MLR model was less accurate. Combined approaches, namely PARAFAC with PLS and SOM with BP, utilized for modelling of fluorescence data, produced similar results when assessed in terms of the final correlation coefficient ($R^2 = 0.93$ and 0.91). However, the analysis of PLS component scores and loadings provided additional information on the dataset and enabled correlation of the fluorescence properties of organic matter with its removal efficiency. Further improvements in the models' accuracy are limited by the error of fluorescence and TOC measurements which is variable within a range of 1–10%.

The acquisition of fluorescence data is becoming easier and faster due to the increasing availability of more sensitive and faster spectrofluorometric measurement techniques. However, the substantial amount of data produced requires robust techniques for data analysis and extraction of useful information. Here, the most common techniques utilized in fluorescence data analysis were briefly introduced and evaluated. This is the first paper summarizing the application of different data mining methods to fluorescence EEMs to characterize organic

matter properties and removal in the field of drinking water treatment. A novel comparative analysis of commonly used methods, including multiway and ANNs approaches, has enabled us to compare the suitability of different data mining methods. The similar results obtained using different decomposition and calibration models should be an incentive for the fluorescence community to employ not only standard solutions, like PARAFAC and MLR, but also other statistical tools. This paper provides the first insight into their methodology and application. However, the results also show that there is little difference between advanced methods of data decomposition and conventional peak-picking approach in both the selection of valid fluorophores and their quantification. Therefore, those conventional methods can in some cases be easier and more efficient to implement and should be considered first prior to implementing more complex techniques, like PARAFAC or SOM.

Acknowledgements

The authors are grateful for the financial and logistical support provided by Severn Trent Water Ltd and the University of Birmingham. The authors also acknowledge the laboratory support provided by Dr Ian Boomer and Mr Andy Moss.

REFERENCES

Alunkaynak A. 2007. Forecasting surface water level fluctuations of lake Van by artificial neural networks. *Water Resources Management* **21**(2): 399–408.
Andersen CM, Bro R. 2003. Practical aspects of PARAFAC modeling of fluorescence excitation–emission data. *Journal of Chemometrics* **17**: 200–215.
Andersson CA, Bro R. 2000. The N-way toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems* **52**(1): 1–4.
Andrews JM, Lieberman SH. 1994. Neural network approach to qualitative identification of fuels and oils from laser induced fluorescence spectra. *Analytica Chimica Acta* **285**(1–2): 237–246.
Antunes MCG, Esteves da Silva JCG. 2005. Multivariate curve resolution analysis excitation-emission matrices of fluorescence of humic substances. *Analytica Chimica Acta* 546: 52–59.
Bahram M, Bro R, Stedmon C, Afkhami S. 2006. Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. *Journal of Chemometrics* **20**(3–4): 99–105.
Baker A. 2001. Fluorescence excitation–emission matrix characterization of some sewage-impacted rivers. *Environmental Science and Technology* **35**(5): 948–953.
Baker A. 2002. Fluorescence excitation–emission matrix characterization of river waters impacted by a tissue mill effluent. *Environmental Science and Technology* **36**(7): 1377–1382.
Basheer IA, Hajmeer M. 2000. Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods* **43**(1): 3–31.
Battin TJ. 1998. Dissolved organic matter and its optical properties in a blackwater tributary of the upper Orinoco river, Venezuela. *Organic Geochemistry* **28**(9–10): 561–569.
Belzile C, Guo L. 2006. Optical properties of low molecular weight and colloidal organic matter: application of the ultrafiltration permeation model to DOM absorption and fluorescence. *Marine Chemistry* **98**(2–4): 183–196.
Bengraïne K, Marhaba TF. 2003. Comparison of spectral fluorescent signatures-based models to characterize DOM in treated water samples. *Journal of Hazardous Materials B* **100**(1–3): 117–130.
Bieroza M, Baker A, Bridgeman J. 2009. Relating freshwater organic matter fluorescence to organic carbon removal efficiency in drinking water treatment. *Science of Total Environment* **407**: 1765–1774.
Boehme J, Coble P, Conmy R, Stovall-Leonard A. 2004. Examining CDOM fluorescence variability using principal component analysis: seasonal and regional modelling of three-dimensional fluorescence in the Gulf of Mexico. *Marine Chemistry* **89**(1–4): 3–14.
Bos M, Bos A, van der Linden WE. 1993. Data processing by neural networks in quantitative chemical analysis. *The Analyst* **118**(4): 323–328.
Bro R. 1997. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* **38**(2): 149–171.
Bro R. 1998. Multi-way analysis in the food industry. Models, algorithms, and applications. *PhD Dissertation*. Department of Dairy and Food Science, Royal Veterinary and Agricultural University, Denmark.
Cammack WKL, Kalff J, Prairie YT, Smith EM. 2004. Fluorescent dissolved organic matter in lakes: relationship with heterotrophic metabolism. *Limnology and Oceanography* **49**(6): 2034–2045.
Chen W, Westerhoff P, Leenheer JA, Booksh K. 2003. Fluorescence excitation-emission matrix regional integration to quantify spectra for dissolve organic matter. *Environmental Science and Technology* 37: 5701–5710.
Clark CD, Jimenez-Morais J, Jones G, Zanardi-Lamardo E, Moore CA, Zika RG. 2002. A time-resolved fluorescence study of dissolved organic matter in a riverine to marine transition zone. *Marine Chemistry* **78**(2–3): 121–135.
Coble PG. 1996. Characterization of marine and terrestrial DOM in seawater using excitation–emission spectroscopy. *Marine Chemistry* **51**(4): 325–346.
Cumberland SA, Baker A. 2007. The freshwater dissolved organic matter fluorescence-total organic carbon relationship. *Hydrological Processes* **21**(16): 2093–2099.
Daliakopoulos IN, Coulibaly P, Tsanis IK. 2005. Groundwater level forecasting using artificial neural networks. *Journal of Hydrology* **309**(1–4): 229–240.
Davies DL, Bouldin DW. 1979. Cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**(2): 224–227.
Blough NV, Del Vecchio R. 2002. Photobleaching of chromophoric dissolved organic matter in natural waters: kinetics and modeling. *Marine Chemistry* **78**(4): 231–253.
de Souza Sierra MM, Donard OF, Lamotte M. 1997. Spectral identification and behavior of dissolved organic fluorescent material during estuarine mixing processes. *Marine Chemistry* **58**(1–2): 51–58.
Despagne F, Massart DL. 1998. Neural networks in multivariate calibration. *The Analyst* **123**(11): 157R–178R.
Determann S, Lobbes JM, Reuter R, Rullkötter J. 1998. Ultraviolet fluorescence excitation and emission spectroscopy of marine algae and bacteria. *Marine Chemistry* **62**(1–2): 137–156.
Divya O, Mishra AK. 2007. Multivariate methods on the excitation emission matrix fluorescence spectroscopic data of diesel–kerosene mixtures: a comparative study. *Analytica Chimica Acta* **592**(1): 82–90.
Garcia JS, da Silva GA, Arruda MA, Poppi RJ. 2007. Application of Kohonen neural network to exploratory analyses of synchroton radiation X-ray fluorescence measurements of sunflower metalloproteins. *X-ray Spectrometry* **36**(2): 122–129.
Gontarski CA, Rodrigues PR, Mori M, Prenem LF. 2000. Simulation of an industrial wastewater treatment plant using artificial neural networks. *Computers and Chemical Engineering* **24**(2–7): 1719–1723.
Guimet F, Ferré J, Boqué R. 2005. Rapid detection of olive-pomace oil adulteration in extra vrgin olive oils from the protected denomination of origin 'Siurana' using excitation–emission fluorescence spectroscopy and three-way methods of analysis. *Analytica Chimica Acta* **544**(1–2): 143–152.

Häck M, Kohne M. 1996. Estimation of wastewater process parameters using neural networks. *Water Science and Technology* **33**(1): 101–115.
Hall GJ, Clow KE, Kenny JE. 2005. Estuarial fingerprinting through multidimensional fluorescence and multivariate analysis. *Environmental Science and Technology* **39**(19): 7560–7567.
Hammerstrom D. 1993. Working with neural networks. *IEEE Spectrum* **30**(7): 46–53.
Henrion R, Henrion G, Böhme M, Behrendt H. 1997. Three-way principal components analysis for fluorescence spectroscopic classification of algae species. *Fresenius' Journal of Analytical Chemistry* **357**(5): 522–526.
Hudson NJ, Baker A, Reynolds D. 2007. Fluorescence analysis of dissolved organic matter in natural, waste and polluted waters—a review. *River Research and Applications* **23**(6): 631–649.
Hudson NJ, Baker A, Ward D, Brunsdon C, Reynolds D, Carliell-Marquet C, Browning S. 2008. Fluorescence spectrometry as a surrogate for the BOD_5 test in water quality assessment: an example from South West England. *Science of the Total Environment* 391: 149–158.
Jain A, Dubes R. 1998. Algorithms for Clustering Data. Prentice-Hall: New Jersey.
Kiers H. 1991. Hierarchical relations among 3-way methods. *Psychometrika* **56**(3): 449–470.
Kohonen T. 1998. The self-organizing map. *Neurocomputing* **21**(1): 1–6.
Kohonen T. 2001. Self-organizing Maps (3rd edn), Springer: Berlin.
Lee KI, Yim YS, Chung SW, Wei J, Rhee JI. 2005. Application of artificial neural networks to the analysis of two-dimensional fluorescence spectra in recombinant E. coli fermentation processes. *Journal of Chemical Technology and Biotechnology* **80**(9): 1036–1045.
Li Q, Yao X, Chen X, Liu M, Zhang R, Zhang X, Hu Z. 2000. Application of artificial neural networks for the simultaneous determination of a mixture of fluorescent dyes by synchronous fluorescence. *The Analyst* **125**(11): 2049–2053.
Marhaba TF, Bengraine K, Pu Y, Arago J. 2003. Spectral fluorescence signatures and partial least squares regression: model to predict dissolved organic carbon in water. *Journal of Hazardous Materials* B97: 83–97.
Marhaba TF, Van D, Lee Lippincott R. 2000. Rapid identification of dissolved organic matter fractions in water by spectral fluorescent signatures. *Water Research* 34(14): 3543-3550.
Marquardt D. 1963. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics* **11**(2): 431–441.
Martens H, Naes T. 1989. Multivariate Calibration. John Wiley & Sons: Chichester.
McAvoy TJ, Su HT, Wang NS, He M, Horwath J, Semerjian H. 1992. A comparison of neural networks and partial least squares for deconvoluting fluorescence spectra. *Biotechnology and bioengineering* 40: 53–62.
McKnight DM, Boyer EW, Westerhoff PK, Doran PT, Kulbe T, Andersen DT. 2001. Spectrofluorometric characterization of dissolved organic matter for indication of precursor organic material and aromaticity. *Limnology and Oceanography* **46**(1): 38–48.
Mopper K, Schultz CA. 1993. Fluorescence as a possible tool for studying the nature and water column distribution of DOC components. *Marine Chemistry* **41**(1–3): 229–238.
Mounier S, Braucher R, Benaïm JY. 1999. Differentiation of organic matter's properties of the Rio Negro basin by cross-flow ultra-filtration and UV-spectrofluorescence. *Water Research* **33**(10): 2363–2373.
Nieke B, Reuter R, Heuermann R, Wang H, Babin M, Therriault JC. 1997. Light absorption and fluorescence properties of chromophoric dissolved organic matter (CDOM), in the St. Lawrence Estuary (Case 2 waters). *Continental Shelf Research* **17**(3): 235–252.
Ohno T, Amirbahman A, Bro R. 2008. Parallel factor analysis of excitation–emission matrix fluorescence spectra of water soluble soil organic matter as basis for the determination of conditional metal binding parameters. *Environmental Science and Technology* **42**: 186–192.
Park YS, Céréghino R, Compin A, Lek S. 2003. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecological Modelling* **160**(3): 265–280.
Patel-Sorrentino N, Mounier S, Benaim JY. 2002. Excitation–emission fluorescence matrix to study pH influence on organic matter fluorescence in the Amazon basin rivers. *Water Research* **36**(10): 2571–2581.
Persson T, Wedborg M. 2001. Multivariate evolution of the fluorescence of aquatic organic matter. Analytica Chimica Acta 434: 179–192.
Rhee JI, Lee KI, Kim CK, Yim YS, Chung SW, Wei J, Bellgardt KH. 2005. Classification of two-dimensional fluorescence spectra using self-organizing maps. *Biochemical Engineering Journal* **22**(2): 135–144.
Saurina J, Leal C, Compano R, Granados M, Tauler R, Prat MD. 2000. Determination of triphenyltin in sea-water by excitation-emission matrix fluorescence and multivariate curve resolution. *Analytica Chimica Acta* 409: 237–245.
Scott SM, James D, Ali Z, O'Hare WT, Rowell FJ. 2003. Total luminescence spectroscopy with pattern recognition for classification of edible oils. *The Analyst* **128**(7): 966–973.
Smits JRM, Melssen WJ, Buydens LMC, Kateman G. 1994. Using artificial neural networks for solving chemical problems. Part I. Multi-layer feed-forward networks. *Chemometrics and Intelligent Laboratory Systems* **22**(2): 165–189.
Spencer RGM, Baker A, Ahad JME, Cowie GL, Ganeshram R, Upstill-Goddard RC, Yjer G. 2007. Discriminatory classification of natural and anthropogenic waters in two UK estuaries. *Science of the Total Environment* 373: 305–323.
Stedmon CS, Markager S, Bro R. 2003. Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy. *Marine Chemistry* **82**(3–4): 239–254.
Ultsch A. 1993. Self-organizing neural networks for visualization and classification. In Information and Classification, Opitz O, Lausen B, Klar R (eds). Springer-Verlag: Berlin; 307–313.
Wolf G, Almeida JS, Pinheiro C, Correia V, Rodrigues C, Reis MA, Crespo JG. 2001. Two-dimensional fluorometry coupled with artificial neural networks: a novel method for on-line monitoring of complex biological processes. *Biotechnology and Bioengineering* **72**(3): 297–306.
Wu FC, Evans RD, Dillon PJ. 2003. Separation and characterization of NOM by high-performance liquid chromatography and on-line three-dimensional excitation emission matrix fluorescence detection. *Environmental Science and Technology* **37**(16): 3687–3693.
Zupan J, Gasteiger J. 1991. Neural networks: a new method for solving chemical problems or just a passing phase? *Analytica Chimica Acta* **248**(1): 1–30.

