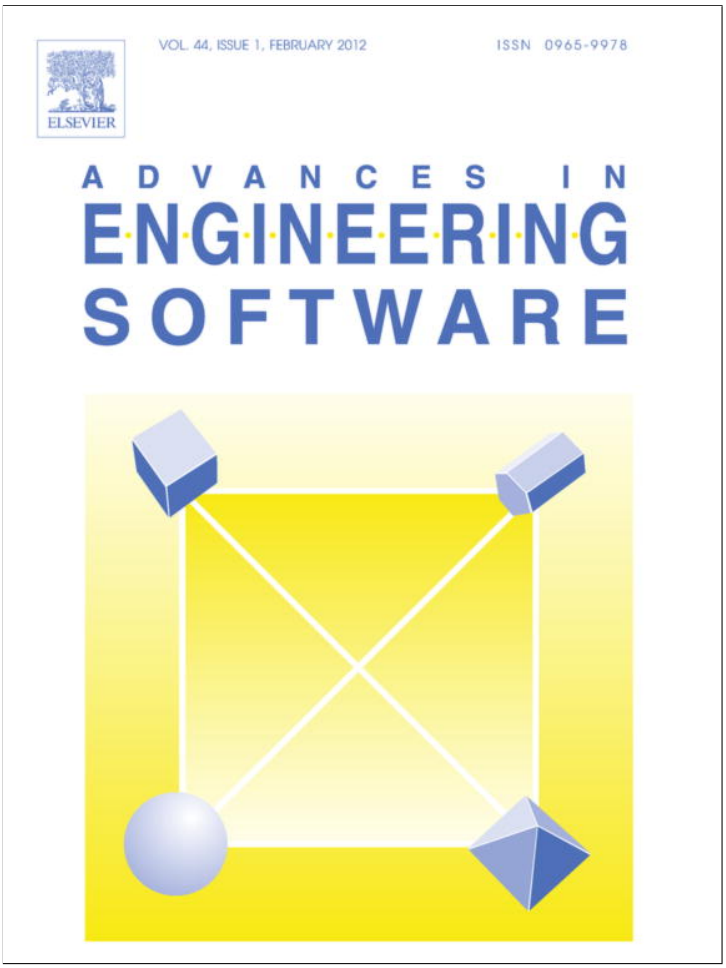

Advances in Engineering Software 44 (2012) 126–135

Contents lists available at ScienceDirect

Advances in Engineering Software

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

New data mining and calibration approaches to the assessment of water treatment efficiency

M. Bieroza [a], A. Baker [b], J. Bridgeman [c,*]

[a] Centre for Sustainable Water Management, Lancaster Environment Centre, Lancaster, LA1 4YQ, United Kingdom
[b] School of Civil and Environmental Engineering and School of Biology, Earth and Environmental Sciences, The University of New South Wales, 110 King St., Manly Vale, NSW 2093, Australia
[c] School of Civil Engineering, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

ARTICLE INFO

Article history:
Available online 24 June 2011

Keywords:
Data mining
Multivariate analysis
Pattern recognition
Artificial neural networks
Fluorescence spectroscopy
Organic matter removal

ABSTRACT

For the first time, the application of different robust data mining techniques to the assessment of water treatment performance is considered. Principal components analysis (PCA), parallel factor analysis (PARAFAC), and a self-organizing map (SOM) were used in the analysis of multivariate data characterising organic matter (OM) removal at 16 water treatment works. Decomposed fluorescence data from PCA, PARAFAC and SOM were used as input to calibrate fluorescence data with OM concentrations using stepwise regression (SR), partial least squares (PLS), multiple linear regression (MLR), and neural network with back-propagation algorithm (BPNN). The best results were obtained with combined PARAFAC/PLS and SOM/BPNN. Both the numerical accuracy and feasibility of the adopted solutions were compared and recommendations on the use of the above techniques for fluorescence data analysis are presented.

1. Introduction

Trihalomethanes (THMs) are the most common disinfection by-products formed during the disinfection of drinking water with chlorine [1–3]. The potential adverse health effects of THMs have been reported by many authors and have resulted in the tightening of regulatory standards for THMs to 100 μg/l in the UK (absolute standard) and 80 μg/l in the US (based on running annual average) [4–7].

The formation of THMs is a result of chlorine reacting with organic matter present in water. Organic matter removal in water treatment is achieved by physico-chemical processes (coagulation, flocculation and clarification, with some additional removal in filtration and granular activated carbon processes), prior to disinfection with chlorine. However, the inherent physical and chemical complexity of organic matter determines its selective removal and hence the presence of recalcitrant residual organic matter that can lead to the formation of THMs.

Thus, water companies attempt to improve overall organic matter removal efficiency and also to develop more accurate analytical techniques for the identification of THMs in treated water. Standard methods of THM formation prediction involve time-consuming laboratory analyses. Thus, surrogate parameters assessing organic matter removal and THM formation in water have been investigated, e.g. total organic carbon (TOC) removal, ultraviolet absorbance (UV), fluorescence spectroscopy [8–12]. Fluorescence has distinct advantages over other surrogate organic matter removal parameters; specifically, the speed of analysis, accuracy of the measurements and comprehensive organic matter characterisation [13,14]. Fluorescent organic matter compounds are excited with light of different wavelengths, and emission light is detected producing a three-dimensional output (an excitation–emission matrix, EEM) comprising more than 4000 fluorescence data points. The EEM exhibits increased fluorescence intensities in particular spectral regions (fluorescence peaks), depending on the organic matter constituents present in a sample and their relative concentration (Fig. 1). In an earlier study, Bieroza et al. [15] related organic matter properties and removal to fluorescence properties derived from EEMs of raw and partially treated (clarified) water from 16 water treatment works (WTWs) in the Midlands region of the UK.

From the fluorescence EEMs, the presence and relative concentration of particular fluorophores (fluorescent organic matter fractions) can be derived. For potable water samples, three fluorescence peak regions are of particular interest (Fig. 1): fulvic-like fluorescence (peak C, fluorescence excited between 300 and 340 nm, and emitted between 400 and 460 nm), humic-like fluorescence (peak A, fluorescence excited between 220 and 250 nm, and emitted between 400 and 460 nm), and tryptophan-like fluorescence (fluorescence excited between 270 and 280 nm and emitted between 330 and 370 nm). Fluorescence fulvic- and

* Corresponding author. Tel.: +44 121 414 5145; fax: +44 121 414 3675.
E-mail address: j.bridgeman@bham.ac.uk (J. Bridgeman).

0965-9978/$ - see front matter © 2011 Civil-Comp Ltd and Elsevier Ltd. All rights reserved.
doi:10.1016/j.advengsoft.2011.05.031

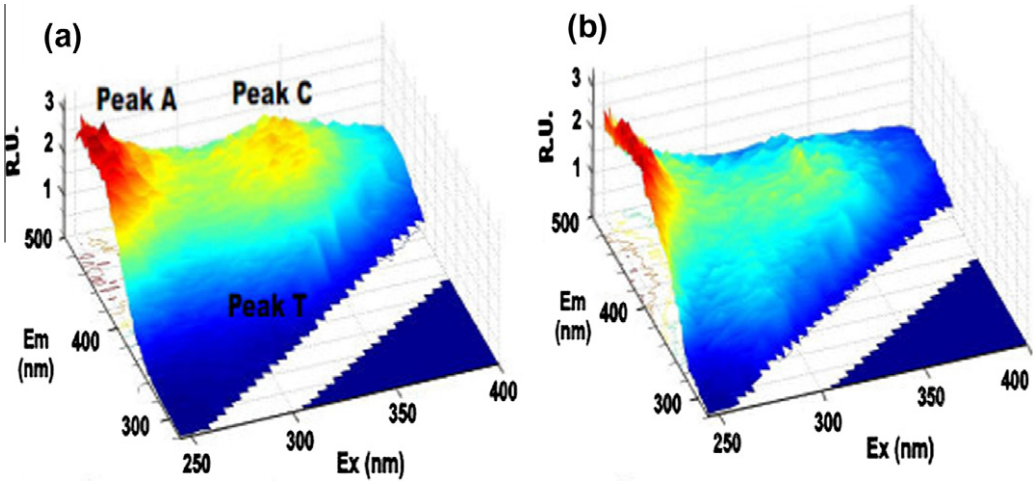

Fig. 1. Fluorescence EEMs of raw and partially-treated (clarified water) after scatter removal. Em – emission wavelength, Ex – excitation wavelength, R.U. – fluorescence intensity in 10^2 Raman units.

humic-like regions can be attributed to natural sources of organic matter (e.g. derived from the decomposition of plant tissues), whereas tryptophan-like fluorescence indicates presence of labile, algal and microbial derived organic matter including anthropogenic pollution [16,17]. Fulvic-like fluorescence intensity has been demonstrated to correlate with TOC concentration [13,18]. Moreover, previous studies have indicated that the fulvic-like fluorescence emission wavelength may be used as surrogate parameter for organic matter aromaticity and hydrophobicity [14,19].

In the standard EEM examination process, the fluorescence intensities and the spectral location of the fluorescence peaks are derived (known as the "peak-picking" approach). However, to discern regularities in a dataset containing a substantial number of EEMs, robust data mining techniques are required. The aim of this paper is to evaluate the use of selected data mining and calibration techniques for the comprehensive characterisation of organic matter removal at WTWs using fluorescence spectroscopy. In the work presented here, different pattern recognition and calibration algorithms are tested to retrieve fluorescence information on organic matter variability and to correlate fluorescence properties with organic matter removal in drinking water treatment. A range of pattern recognition techniques are used to reduce and decompose the most important features of the multivariate input fluorescence dataset. Specifically, the outcomes of principal components analysis (PCA), parallel factor analysis (PARAFAC), and self-organizing map (SOM) analysis are evaluated and compared with the standard peak-picking method. Furthermore, different calibration algorithms (multiple linear regression (MLR), stepwise regression (SR), partial least squares analysis (PLS), and artificial neural network with back-propagation learning (BPNN)) are used to correlate organic matter properties derived from the above fluorescence analysis with measured TOC removal at 16 WTWs.

2. Analytical methods

2.1. Fluorescence

Fluorescence and TOC analyses were carried out on samples of raw and partially-treated (clarified) water collected monthly between August 2006 and February 2008 from 16 WTWs owned and operated by Severn Trent Water Ltd.

Fluorescence EEMs were collected using a Cary Eclipse Fluorescence Spectrophotometer (Varian, Surrey, UK) equipped with Peltier temperature controller. For each unfiltered water sample, the fluorescence was measured in duplicate by scanning the excitation wavelengths from 200 to 400 nm in 5 nm steps, and detecting the emission intensity in 2 nm steps between 280 and 500 nm. Excitation and emission slit widths were 5 nm. To maintain the consistency of measurements and standardise the fluorescence data, all fluorescence intensities were corrected to Raman peak intensity of 20 units measured for deionised water at 348 nm excitation and 396 nm emission wavelengths.

2.2. TOC

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.

3. Data analysis

3.1. Data pre-processing

Prior to all statistical and computational analyses, fluorescence data, normalized to the intensity of Raman line of water at 348 nm excitation wavelength, were pre-processed to remove redundant fluorescence regions of EEMs. The Rayleigh and Raman scatter features (i.e. EEM regions with excitation wavelengths exceeding emission wavelengths and with excitation wavelengths less than 240 nm) were all removed as containing limited fluorescence information and low signal to noise ratio [20,21]. The resultant EEMs ranged from 240 to 400 nm excitation and from 300 to 500 nm emission wavelengths respectively. Finally, fluorescence data scaling (data variance normalized to one) and mean-centring (subtracting of variable means) was performed to reduce the concentration effects exhibited by intensity [22,23].

The final dataset used in the data mining analyses comprised 290 raw and 290 clarified water samples and 2515 fluorescence excitation–emission wavelengths. For the purpose of the computational analyses, the input fluorescence data are represented as a vector of length, m (number of fluorescence intensities measured at m excitation–emission wavelength pairs), and number of observations (raw and clarified water samples), n.

All advanced fluorescence data analyses were carried out in Matlab® 7.7 with the Statistics Toolbox 7.0 and Neural Network

Toolbox 6.0.1, on a 512 MB Dual Pentium III PC computer. For PARAFAC and PLS algorithms, the N-way toolbox for Matlab® was used [24], and the SOM implementation was obtained from SOM toolbox version 2 [25].

3.2. Decomposition methods

Fluorescence EEMs contain abundant information on organic matter quantity and quality. However, the analysis of fluorescence data containing more than 4000 data points per EEM presents serious computational difficulties. Therefore, a common approach in fluorescence data analysis is the decomposition of the original EEMs into a set of fewer fluorescence parameters (of size k, and $k \leqslant m$) measured at certain fluorescence regions, e.g. maximum fulvic-like fluorescence intensity. This "peak-picking" method is useful when the spectral properties of organic matter constituents are known or assumed *a priori*, and the analysis is restricted to the fluorescence regions of particular interest (supervised analysis). However, in many cases no preceding assumptions or knowledge on the fluorescence data variability are given, and the aim of the analysis is to extract the most characteristic fluorescence features of the dataset.

To retrieve the most important information on organic matter composition and to reduce the data dimensionality, the raw and clarified water EEMs were processed with three decomposition algorithms: Principal Components Analysis (PCA), Parallel Factor Analysis (PARAFAC) and Self-Organising Maps (SOM); methods which have previously been used for the exploratory analysis of fluorescence data [20,23,26,27].

PCA is a multivariate method of high-dimensional data simplification, where the original data matrix, X, is reduced to a number of principal components calculated as the directions of maximum variance of combined variables. The calculated principal components are in order of decreasing variance, where the first principal component describes the greatest variance within the dataset and all successive components account for the variance in decreasing order of magnitude. Depending on the overall variability in data explained by the primary components, in PCA often just the first few components are analysed to investigate any valid correlations and relate them to input data structure and functions. The linear PCA projection is defined as:

$$Y = XT, \tag{1}$$

where Y is the pattern matrix (scores matrix) of size $n \times k$, X is the input data matrix of size $n \times m$, and T is the transformation matrix (loadings matrix) of size $m \times k$. The loadings matrix expresses the importance of each variable (excitation–emission pair) in the original data matrix, whereas the scores are coordinates of samples in PCA projection. A detailed description of PCA transformation can be found in the literature [26,28–30]. For examples of the application of PCA for fluorescence data, the reader is referred to Persson and Wedborg [26], Boehme et al. [23], Spencer et al. [31].

PARAFAC is a three-way decomposition model commonly used in fluorescence data analysis to extract the most important fluorescence components (models of fluorophores) along with their spectral properties (emission and excitation wavelengths) and relative concentration in a sample. Although PARAFAC is commonly described as three-way version of PCA with data decomposition into scores and loadings, there are significant differences between models regarding the input data structure and model constrains [20,22]. Unlike PCA, where an infinite number of models with equal fit exist, the PARAFAC model produces a unique, chemically meaningful solution [22].

The PARAFAC algorithm decomposes the three-way data array $I \times J \times K$ (sample by excitation wavelength by emission wavelength) into a trilinear model that minimizes the sum of squares of the residuals ε_{ijk}:

$$X_{ijk} = \sum_{f=1}^{F} A_{fi} B_{fj} C_{fk} + \varepsilon_{ijk} \tag{2}$$

where X_{ijk} is the fluorescence intensity measured at emission wavelength j and excitation wavelength k for sample i, A, B and C are the loading matrices and F is the number of components [22].

Previous PARAFAC implementations include characterization of the organic matter-metal binding process [32], quantitative determination of the kerosene fraction present in diesel [33], classification of ballast water [34], estuarine water [20] and edible oils [35].

The SOM is a powerful pattern recognition algorithm based on a two-layered Artificial Neural Network (ANN), consisting of a number of interconnected single processing units called neurons or nodes. ANNs have the ability to learn the pattern from the input features (pattern recognition) or model input–output relationship (calibration) based on training algorithms in which weight vectors stored in connections between neurons are adjusted to minimise the overall error of network prediction [36–37].

Like PCA, SOM is an example of an unsupervised clustering algorithm in which any existing pattern is assigned to one of the categories, not specified or not known *a priori*. In PCA and SOM, the feature extraction from the input domain is performed via linear (PCA) or nonlinear (SOM) transformations of the input data on the lower dimensionality k principal components (PCA) or a k-dimensional map (SOM). In a SOM network, the connection weights of size of the input data m are stored in input neurons and during training are projected onto k-dimensional output space [38]. The analysis of the network output provides the basis for extraction of relationships and regularities from the original data.

The pattern recognition with SOM involves an iterative process of neurons weights adjustment, in which for each input sample the neuron with weights most similar to the input vector is first identified (best-matching unit, BMU). The weights of the BMU and its neighbouring neurons are modified according to Eq. (3):

$$w_i(k+1) = w_i(k) + \varepsilon(k) h_p(i,k)\{x_j(k) - w_i(k)\} \tag{3}$$

where $w_i(k)$ is the previous weight of neuron, $w_i(k+1)$ is the new weight of neuron, $\varepsilon(k)$ is the learning rate (describes the speed of the training process), $h_p(i,\ k)$ describes the neighbourhood of the BMU, k is the number of epochs (a finite set of input patterns presented sequentially) and p is the index of the BMU neuron.

For the implementation details and examples of data analysis with the SOM, the reader is referred to Kohonen [25,38].

In data decomposition, the original dataset containing n samples and m variables is projected onto k-dimensional space and new coordinates (numerical transformation of m variables) are calculated for each sample. Of the three data decomposition methods evaluated in this study, PCA and PARAFAC represent multi-way algorithms, where the original dataset is projected onto a k components (principal components in PCA) coordinate plane with the components scores being the numerical representations of variables. In SOM, the original data are projected on k neurons arranged in a two-dimensional map, where the original variables are transformed into neurons weights. Therefore, pattern recognition involves analysis of the samples' distribution in the new coordinate planes calculated with each algorithm (Fig. 2).

The relationship between the original space and the PCA, PARAFAC component space or SOM map is expressed by loadings in multi-way analysis and SOM component planes. In fluorescence data decomposition, the PCA and PARAFAC loadings and SOM component planes demonstrate the importance of particular excitation–emission combinations (fluorophores) for the model and

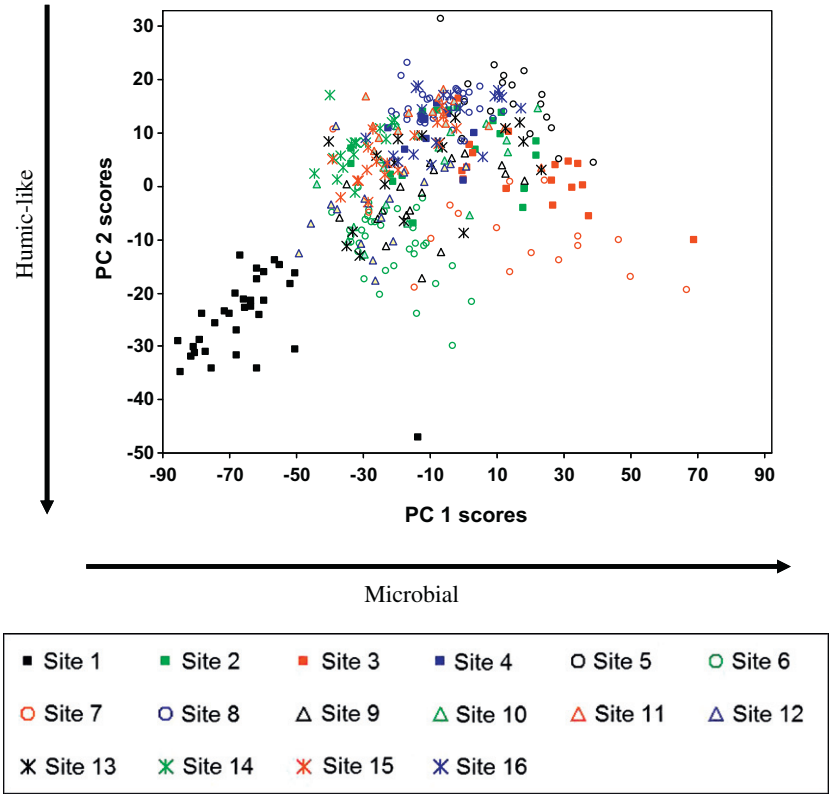

Fig. 2. Fluorescence data distribution with PCA.

therefore provide a basis for interpretation of the variation in organic matter properties between samples.

To provide meaningful information on the fluorescence data patterns using decomposition methods, the calculated models have to be valid in terms of the optimum number of components (PCA), optimum number and chemical interpretation of components (PARAFAC), and optimum number of neurons comprising the map (SOM). In the SOM approach the map size is determined by the size of input data, and the ratio of two greatest eigenvalues is commonly used to calculate the map size [38]. In this drinking water fluorescence study the final map contained 120 nodes (size 15 × 8).

3.3. Calibration methods

In the multivariate calibration of fluorescence spectra, a mathematical model relating the fluorescence properties of fluorophores (e.g. intensity) to an actual, measured quantity (e.g. TOC concentration) is developed. A set of reference samples containing both independent variables (explanatory variables) and target values is firstly used to calibrate the model (pattern learning), and then the model's prediction accuracy is tested (validation) for an unknown set of samples. Here, the fluorescence spectra were calibrated with measured TOC removal for 16 WTWs. The initial fluorescence dataset, decomposed into PCA, PARAFAC scores and SOM normalized weights, was used as an input for different calibration models. To evaluate the efficacy of the selected decomposition methods in differentiating the most important features of fluorescence spectra, the complete EEMs and peak-picking results were also used in the calibration. The fluorescence dataset was divided into calibration and validation sets by selecting 25% (70 samples) as validation samples covering the whole range of data variation in projection of the first two principal components. The details of the input datasets for the calibration models can be found in Table 1.

Table 1
Details of the fluorescence dataset used in the calibration.

Model	Data	Calibration matrix size	Validation matrix size
EEM	Complete EEMs of raw and clarified water	220 × 5030	70 × 5030
PEAK	Fulvic-like fluorescence intensity and emission wavelength, tryptophan-like fluorescence intensity of raw and clarified water	220 × 6	70 × 6
PARAFAC	Scores of three PARAFAC components of raw and clarified water	220 × 6	70 × 6
PCA	Scores of first three PCA components of raw and clarified water	220 × 6	70 × 6
SOM	Normalized weights of three SOM clusters of raw and clarified water	220 × 6	70 × 6

Four different regression algorithms were tested: multiple linear regression (MLR), stepwise regression (SR), partial least squares analysis (PLS), and artificial neural network with back-propagation learning (BPNN). These methods are frequently used in optimisation analyses [33,39–40].

In a standard MLR model, all dependent variables are simultaneously regressed onto all the independent variables to minimise the squared error of the predictions, according to Eq. (4):

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \quad (4)$$

where Y is the predicted dependent variable, b_0–b_n are partial regression coefficients, and X_1–X_n are independent variables.

Both stepwise regression and partial least squares analysis can be considered as extensions of the standard MLR model. The SR analysis aims to select a statistically significant subset

of independent variables in decreasing order of importance in predicting the dependent variable. The analysis involves sequentially testing the independent variables and excluding from the regression variables those with p-values greater than 0.05. Although the stepwise regression model has clear advantages of selecting and grading of independent variables, it is important to be aware of some of the method limitations. Some authors [41,42] underline three possible misconceptions with regard to the application of the stepwise regression method: lack of theory or model to support the stepwise-derived equation and the possibility of incorporating accidental variables; the conceivable inclusion of falsely-best variables; and finally, the possible violation of the model specification due to the addition of extra variables to the analysis.

The PLS algorithm enables sequential decomposition of an array of independent variables into a multi-linear model in which the scores have the maximal covariance with the yet unexplained variation of the dependent variable. In the PLS model, two sets of variables, X and Y, are simultaneously decomposed into a sum of k variables:

$$X = TP^T + E = \sum t_k p_k^T + E \tag{5}$$

$$Y = UQ^T + F = \sum u_k q_k^T + F \tag{6}$$

where T and U are the score matrices, P and Q are the loading matrices, E and F are the residual matrices. The matrix Y, containing for example concentration data of a particular fluorophore, can be derived from independent data (matrix X, containing for example the fluorescence spectra) according to Eq. (7):

$$Y = TBQ^T + F \tag{7}$$

where B is the regression coefficient matrix for scores T and U.

The artificial neural network with back propagation learning (BPNN) is the most common optimisation algorithm of neural networks. For the given input, the actual output is compared with the desired output, and weights are adjusted iteratively to minimize the error of the entire network [36,43,44]. The weight adjustment and error calculation is propagated backwards from the output layer until the input layer is reached. In the BPNN, three parameters are defined: the learning rate, the momentum factor and the range in which the initial weights are randomized [36]. The learning rate describes the speed of the training process, while the momentum coefficient is used in weight updating to maintain the optimum search stability [37]. The determination of both parameters is a trade-off between the speed of the training and the likelihood of finding the global, rather than local, minimum. Therefore a trial-and-error procedure is commonly used to adjust back-propagation algorithm parameters [43].

Here, the Levenberg–Marquardt algorithm with early stopping was employed for preventing the data from overfitting [46]. The optimal network architecture was found to comprise six input nodes (denoting 6 PCA, PARAFAC scores and SOM normalised weights, Table 1), four hidden nodes and one output node (denoting 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). The number of hidden neurons was chosen based on analyses of the performance of different networks. In the EEM model, for each sample a complete matrix containing 2515 fluorescence excitation–emission wavelengths was simultaneously presented to the network.

4. Results

4.1. Fluorescence data decomposition

The importance of a given component in multi-way analysis is calculated as a variance explained by this component. The first component can be attributed to the most important spectral features of the dataset and successive components are less important as indicated by the decreasing explained variance. In this study, the first three PCA components explained 69.3% of the total variance (Fig. 2, Table 2). Thus, the PCA analysis revealed the presence of many distinctly different fluorescence features of similar importance (explained variance).

In this study, determination of the optimum number of PARAFAC components was a complex and challenging operation which involved analysis of both statistical diagnostics (variance explained by the model, core consistency analysis (CORCONDIA) [40,45]) and component loadings. The apparently most appropriate PARAFAC model (i.e. the one with the highest number of components, highest explained variance and valid CORCONDIA value (closest to 100%)) was selected (three-component model, Table 2). However, visual inspection of the emission and excitation loadings for the three-component PARAFAC model revealed the likely presence of more components. From Fig. 3 it can be seen that emission loadings of component 1 and 2 demonstrate well-defined, single peaks, whereas in the excitation loadings, double-peaks can be discerned for the first two components, together with an increased signal in the emission loading of the third component between 470 and 500 nm. Double-peaks in the loadings can be attributed to the presence of variance which is related to another component (visible as shoulders in the loadings) and which is unexplained by the trilinear model. An evaluation of more complex models (four and five-components) confirmed this hypothesis, as the first two components were partitioned into two more components of a different excitation spectra. However, the higher-component models were unstable as indicated by both the low CORCONDIA values and also from the results of repeated PARAFAC analyses. For the same dataset, different PARAFAC models were found, which is in contradiction with the PARAFAC model assumption of producing a unique solution.

PCA and PARAFAC components and SOM component planes are models of groups of fluorophores with similar fluorescence characteristics. On the basis of the spectral properties of the PCA and PARAFAC components derived from the component loadings, the three components were recognized as protein-like, humic-like and fulvic-like fluorescence, a result which is in agreement with visual inspection of EEMs and the fluorescence peaks discerned (Table 3). However, the order and specific spectral properties of each component vary between the PCA and PARAFAC models.

The tryptophan-like component in the PCA results exhibits a broader peak with excitation wavelengths between 250 and 300 nm and emission between 300 and 350 nm, obscured by the region of the removed fluorescence of the Raman line. The spectral

Table 2
PCA and PARAFAC model. VAR_e – variance explained (%).

Component	PCA			PARAFAC	
	VAR_e	Sum of VAR_e	Residual sum of squares	Sum of VAR_e	CORCONDIA %
1	47.3	47.3	70243.9	97.7	100.0
2	14.4	61.7	33747.4	98.9	96.9
3	7.6	69.3	24001.5	99.2	90.1
4	1.7	71.0	20623.0	99.3	14.1
5	1.0	72.0	19533.2	99.4	3.3
6	0.8	72.8	18085.9	99.4	-1.9
7	0.6	73.4	17267.9	99.4	1.1

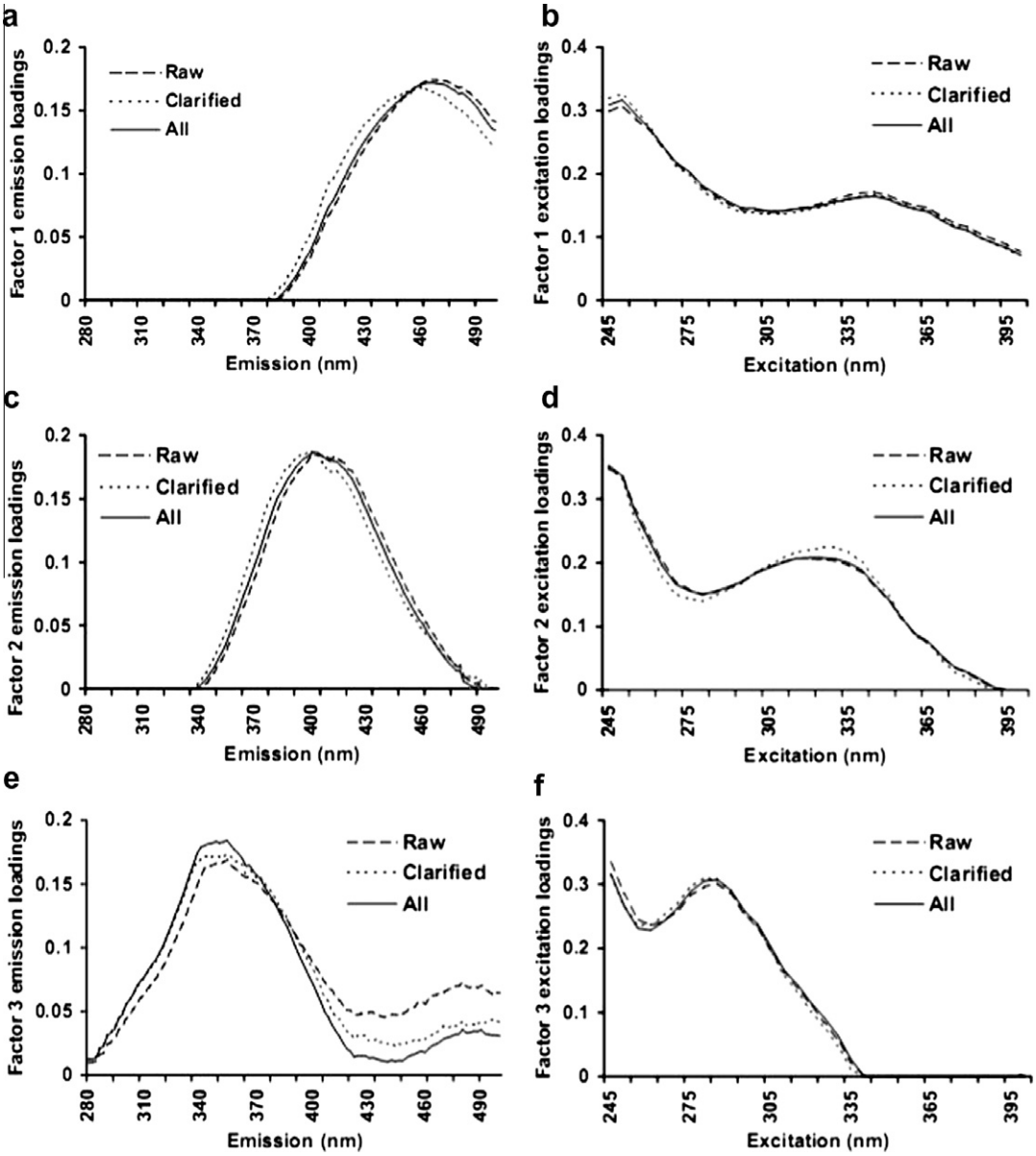

Fig. 3. Excitation and emission loadings of raw, clarified and all data derived from PARAFAC modelling. Factor 1: (a) emission loading, (b) excitation loading; Factor 2: (c) emission loading, (d) excitation loading; Factor 3: (e) emission loading, (f) excitation loading.

position of this fluorescence peak demonstrates more complex properties compared with a well-defined tryptophan-like PARAFAC component and thus it can be attributed to protein-like fluorescence rather than simply tryptophan-like. The humic-like component in both models has a similar emission spectra with distinctly different excitation wavelengths. The shift towards higher excitation wavelengths in the loadings of PC2 suggests the presence of a secondary fluorophore, i.e. the visible fraction of fulvic-like fluorescence. The fulvic-like PCA component indicates the presence of an organic matter fraction with a lower degree of hydrophobicity, rendering it more difficult to remove during the treatment processes. The relative change in this fluorescence between raw and clarified water appears to correlate with TOC removal and thus is indicative of organic matter removal efficiency. Conversely, the PARAFAC fulvic-like component exhibits higher emission wavelengths pertinent to more a hydrophobic OM fraction, which is easier to remove in conventional water treatment.

In the SOM approach, the component planes depict the values of weights vectors for different fluorescence variables and allow correlation between the samples' distribution and excitation–emission wavelengths. Three major directions of fluorescence properties change were observed. (Fig. 4). The horizontal and vertical axes of the map were found to correspond to fluorescence emission and excitation wavelengths, with increasing values from the top to the bottom and from the left to the right respectively. Moreover, the diagonal joining the upper left and lower right corners of the map was the line of the greatest changes in variance within the dataset and discriminated the sites of radically different organic matter spectral properties.

Evaluation of the results of the decomposition models provides significant information on the differences between WTWs in organic matter character and removal, differences in organic matter composition between raw and clarified water, and the relationship between organic matter character and the efficiency of its removal.

All methods revealed the presence of two groups of WTWs of distinctly unique spectral properties; the upland reservoir sites 1 and 6, and anthropogenically influenced sites 2, 3, 5, 7. The SOM shows that there exists good discrimination between raw and

Table 3
Characteristics of three PCA and PARAFAC components identified for drinking water fluorescence dataset. Excitation and emission wavelengths maxima and identified fluorophores. Wavelength in brackets denotes secondary maximum.

PCA				PARAFAC			
Comp.	Exc. max (nm)	Em. max (nm)	Fluorophore	Comp.	Exc. max (nm)	Em. max (nm)	Fluorophore
1	250–300	300–350	Tryptophan -like	1	(<250) 345	460	Fulvic-like
2	260–300	400	Humic-like	2	<245	400	Humic-like
3	350	400	Fulvic-like	3	285	350	Tryptophan-like

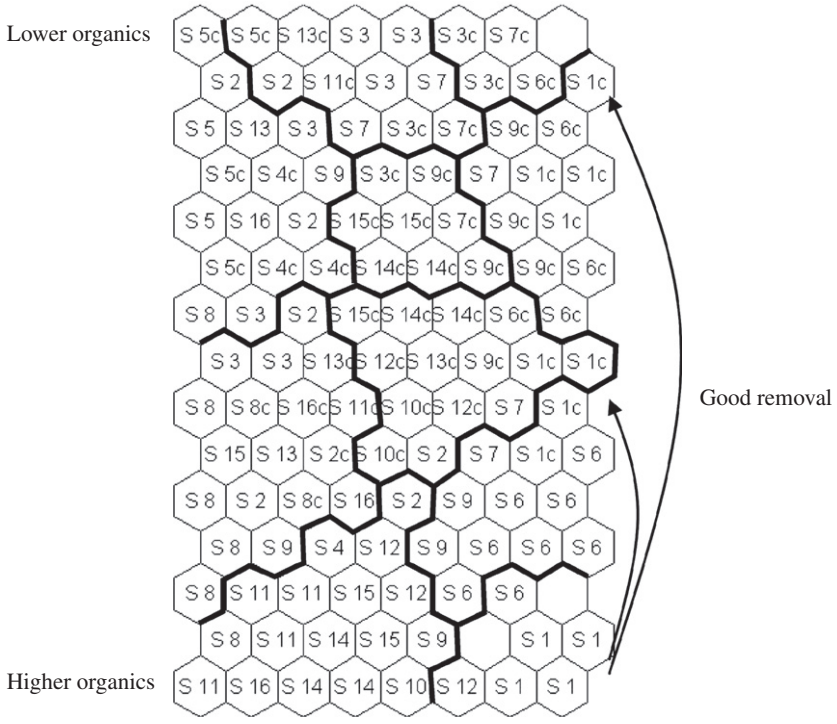

Fig. 4. Fluorescence data distribution with SOM. '*c*' = clarified water, no suffix = raw water.

Table 4
Prediction accuracy and prediction error for selected decomposition and calibration methods.

Decomposition model	Prediction accuracy (R^2)				Prediction error (RMSEP)			
	MLR	SR	PLS	BPNN	MLR	SR	PLS	BPNN
EEM	0.75	0.94	0.94	0.92	1.90	0.52	0.52	0.56
PEAK	0.87	0.82	0.92	0.95	0.73	0.78	0.78	0.45
PARAFAC	0.63	0.93	0.93	0.94	1.29	0.55	0.54	0.51
PCA	0.78	0.85	0.85	0.90	0.94	0.77	0.77	0.64
SOM	0.74	0.93	0.93	0.93	1.21	0.56	0.64	0.60

clarified water fluorescence properties for the first group, whereas the opposite can be observed for the second cluster. For reservoir sites 1 and 6, the raw water organic matter exhibits a higher degree of hydrophobicity and lower microbial fraction (tryptophan-like fluorescence). Therefore, the raw water characteristics enhance organic matter removal efficiency (better removal of more aromatic organic matter) as indicated by the greatest changes in the PC1 scores and the distances on the SOM map between raw and clarified water. Conversely, sites 3 and 7 demonstrate poor organic matter removal which is related to the predominance of highly-variable, riverine organic matter with a significant contribution of highly-microbial and hydrophilic fractions.

4.2. Fluorescence data calibration

In the previous section, the application of PCA, PARAFAC and SOM analyses for pattern recognition of fluorescence data was presented. Here, the quantitative outcomes of the decomposition methods were used for the calibration of fluorescence data with TOC removal. Different calibration methods were evaluated: multiple linear regression (MLR), stepwise regression (SR), partial least squares analysis (PLS), and artificial neural network with back-propagation learning (BPNN). To evaluate the ability of the decomposition methods to reduce the initial dataset to the most important fluorescence features, complete EEMs (EEM model) and peak fluorescence parameters (tryptophan- and fulvic-like fluorescence of raw and clarified water) (PEAK model) were also used in calibration models (Table 1).

The models' performances are compared in Table 4 based on prediction accuracy (how well the calibrated model performs when run on unknown samples) and prediction error (i.e. the error of prediction for unknown samples). First, the abilities of different data mining approaches to select the most important features of fluorescence EEMs were compared. Generally, the higher the

prediction accuracy, the greater percentage of variance is explained by the combined data mining-calibration model. For a given calibration technique, a higher prediction accuracy is obtained for a data mining approach which provides a better numerical representation of the fluorescence data. It can be seen that the PARAFAC and SOM approaches generally provided better decomposition of fluorescence data than PCA (Table 4). The accuracy and residual error of prediction was similar for EEM, PARAFAC and SOM models, indicating that those decomposition techniques can successfully facilitate fluorescence feature extraction from the whole EEM. However, the peak-picking approach (PEAK model) was equally good in BPNN modelling and only slightly poorer for SR and PLS models. Thus, for the fluorescence datasets with known or similar pattern of fluorophores, the peak-picking approach can produce similar numerical results to the PARAFAC and SOM models without time-consuming data pre- and post-processing. However, not only do PARAFAC and SOM provide good numerical representation of fluorescence data, but they also offer additional tools for advanced fluorescence data analysis. Thus, the PARAFAC and SOM approaches are better in providing direct correlations between quantitative and qualitative properties of fluorescence data. In all cases except SR analysis, PCA performance was poorer than the PEAK model, hence the use of the standard peak-picking approach is recommended prior to PCA analysis.

The regression models used in this study produced consistent results, as indicated by the correlation coefficients and prediction errors (Table 4). For independent validation fluorescence data, the poorest results were obtained for the MLR model which accounted for 63–87 % of the total variance explained. The SR, PLS, and BPNN regression models produced similar results, however for all decomposition methods the latter was slightly more efficient (higher correlation coefficients and lower prediction errors).

5. Discussion

5.1. Fluorescence data decomposition

From the above analysis it can be seen that the evaluated decomposition models provided both quantitative and qualitative information on organic matter characteristics and removal at WTWs. The computation times of all decomposition algorithms are similar and do not exceed 20 min on an average speed computer. There are however differences in the ease of applicability of the given method pertinent to interpretation of the results, post-processing etc.

Both the PCA and PARAFAC models decomposed the fluorescence data into a set of components related to the fluorophores. In the PARAFAC model, supervised evaluation of the obtained components based on the analysis of performance statistics, emission and excitation loadings is required. However, as presented here, the ambiguous results of different diagnostic tools impeded the interpretation of the PARAFAC components. PCA and SOM provide an unsupervised data decomposition which can also pose difficulties in the interpretation process. PARAFAC components can be related to the real fluorophores, whilst PCA components are more complex and represent groups of fluorophores or particular OM properties.

In the SOM approach the characteristic features of the data are selected automatically. This is a distinct advantage when analyzing samples from similar sources and with a uniform pattern of fluorophores. Unlike PCA and PARAFAC, the SOM enables easier interpretation of the samples' distribution and fluorescence variables due to several visualization techniques available (e.g. component planes, hit histograms; for details see Kohonen [38]). The geometric distances on the SOM between raw and clarified water samples indicated a degree of similarity in spectral properties of organic matter which correlated with removal efficiency. The greater the spread of water samples of a particular type on the map, the more variation in spectral properties was observed.

5.2. Fluorescence data calibration

In practical applications, it is not only the numerical accuracy of a calibration technique that should be considered, but also the feasibility of the approach is of great importance. The advantages and disadvantages of the two best calibration techniques are discussed below.

The BPNN algorithm is considered to be more flexible than the standard regression methods and therefore more challenging to implement as its flexibility can pose a danger of overfitting the calibration data and producing unreliable results. However, the advantages of this approach include fault and noise tolerance (the ability to process noisy, uncertain data), self-modelling, self-learning (by example) and generalization capabilities. Prior to modelling, the BPNN network requires the topology (number of nodes in a hidden layer) to be defined, together with 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 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 BPNN can be a robust predictive tool provided that a substantial amount of training data is available.

Compared to the BPNN, the PLS algorithm is simpler as it does not require parameter definition in the training phase. However, the validation procedure involves the selection of an appropriate number of components (latent variables), which is the crucial step in generating a valid and robust model. The leave-one-out cross-validation and the root mean squared error of prediction (RMSEP) are the most common techniques and statistical diagnostic used in components selection. 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 also successfully used in exploratory data analysis. Here, four PLS components were selected. The analysis of PLS scores provided useful information relating the organic matter properties (degree of hydrophobicity) with the efficiency of organic matter removal.

6. Summary

Fluorescence excitation–emission spectroscopy is commonly used in organic matter characterisation. A common problem in practical applications of fluorescence analysis is the selection of effective data mining techniques for the decomposition of the large fluorescence datasets which are produced. The selection of appropriate pattern recognition tools is crucial to developing an understanding of the analysed processes or to reduce the computation time and complexity of various calibration problems utilizing fluorescence data. Here, fluorescence analysis was used to characterise organic matter removal efficiency in water treatment. To facilitate fluorescence data reduction to the most important fluorescence properties and to analyse the relationship between fluorescence predictors and TOC removal for 16 WTWs, three different, commonly used decomposition algorithms (principal components analysis, PCA; parallel factor analysis, PARAFAC; and a self-organizing map, SOM)) were evaluated and compared with the standard peak-picking approach in calibration tests. The application of fluorescence analysis for the assessment of water

treatment performance provided a basis for comparison of advantages and limitations of applied data decomposition techniques. Visual inspection of EEMs and peak-picking suggested the presence of three main fluorophores in raw and partially-treated water samples: fulvic-, humic-, and tryptophan-like fluorescence. The PARAFAC model revealed the presence of more potential fluorophores that could not have been validated for the entire dataset or specific water type (raw and clarified). Therefore, to derive a valid PARAFAC model containing all variations in fluorescence spectra, a solution with fewer components had to be chosen. A lack of good, overall diagnostic for the selection of the number of valid components impeded robust interpretation of the PARAFAC model and made analysis a time-consuming process. Compared to the PARAFAC approach, components derived from the PCA analysis were more difficult to identify as particular fluorophores on the basis of loadings interpretation. Thus, it is concluded that the PCA algorithm can be successfully used in the initial fluorescence data analysis to provide an insight into data variation and distribution. However, more advanced analysis requiring the prediction of fluorophores, their importance and relative concentration can be facilitated by PARAFAC. When the fluorophores' composition is uniform between samples and sites, standard peak-picking and the SOM analysis can successfully outperform lengthy PARAFAC modelling. While the peak-picking approach facilitates basic fluorescence data analysis, the SOM model enables advanced interpretation of fluorescence data, samples' distribution between sites and water types, and TOC removal-organic matter relationship. The calibration of TOC removal with fluorescence data decomposed with PARAFAC and SOM confirm the similar abilities of those models in identifying the most important fluorescence features.

From the calibration methods examined in this study, three produced consistent results in TOC removal prediction modelling from fluorescence data (stepwise regression, partial least squares analysis, and artificial neural network with back-propagation algorithm). However, the lack of theory or statistical model to support the stepwise-derived model means that, for the purpose of the prediction of organic matter removal in drinking water systems, the PLS and BPNN models are considered to be the most useful.

Acknowledgements

The authors are grateful to Severn Trent Water Ltd. and the University of Birmingham for financial and logistical support.

References

[1] Rook JJ. Formation of haloforms during chlorination of natural waters. Water Treat Exam 1974;23(2):234–43.
[2] Young JS, Singer PC. Chloroform formation in public water supplies – case study. J Am Water Works Assoc 1979;71(2):7–95.
[3] Carlson M, Hardy D. Controlling DBPs with monochloramine. J Am Water Works Assoc 1998;90(2):95–106.
[4] Dunnick JK, Melnick RL. Assessment of the carcinogenic potential of chlorinated water: experimental studies of chlorine, chloramines, and trihalomethanes. J National Cancer Institute 1993;85(10):817–22.
[5] US EPA, Research plan for microbial pathogens and disinfection by-products in drinking water; 1997.
[6] Moudgal CJ, Lipscomb JC, Bruce RM. Potential health effects of drinking water disinfection by-products using quantitative structure toxicity relationship. Toxicology 2000;147:109–31.
[7] Jackson P, Hall T, Young W, Rumsby P. A review of different national approaches to the regulation of THMs in drinking water, DEFRA/DWI; 2008.
[8] Amy GL, Chadik PA, Chowdhury ZK. Developing models for predicting trihalomethane formation potential and kinetics. J Am Water Works Assoc 1987;79(7):89–97.
[9] Li C-W, Benjamin MM, Korshin GV. Use of UV spectroscopy to characterize the reaction between NOM and free chlorine. Environ Sci Technol 2000;34(12):2370–575.
[10] Parsons SA, Jefferson B, Goslan EH, et al. Natural organic matter – the relationship between character and treatability. Water Sci Technol – Wat Supp 2004;4(5–6):43–8.
[11] Ates N, Yetis U, Kitis M. Effects of bromide ion and natural organic matter fractions on the formation and speciation of chlorination by-products. J Environ Eng 2007;133(10):947–54.
[12] Yang X, Shang C, Lee W, et al. Correlations between organic matter properties and DBP formation during chloramination. Water Res 2008;42:2329–39.
[13] Hudson NJ, Baker A, Reynolds D. Fluorescence analysis of dissolved organic matter in natural, waste and polluted waters – a review. River Res Appl 2007;23(6):631–49.
[14] Baker A, Tipping E, Thacker SA, Gondar D. Relating dissolved organic matter fluorescence and functional properties. Chemosphere 2008;73:1765–72.
[15] Bieroza M, Baker A, Bridgeman J. Relating freshwater organic matter fluorescence to organic carbon removal efficiency in drinking water treatment. Environ Sci Technol 2009;407:1765–74.
[16] Nguyen M-L, Westerhoff P, Baker L, et al. Characteristics and reactivity of algae-produced dissolved organic carbon. J Environ Eng 2005;131(11):1574–82.
[17] Hudson NJ, Baker A, Ward D, et al. Fluorescence spectrometry as a surrogate for the BOD_5 test in water quality assessment: an example from South West England. Sci Total Environ 2008;391(1):149–58.
[18] Cumberland SA, Baker A. The freshwater dissolved organic matter fluorescence-total organic carbon relationship. Hydrol Process 2007;21(16):2093–9.
[19] Kalbitz K, Geyer W, Geyer S. Spectroscopic properties of dissolved humic substances – a reflection of land use history in a fen area. Biogeochemistry 1999;47:219–38.
[20] Stedmon CS, Markager S, Bro R. Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy. Marine Chem 2003;82(3–4):239–54.
[21] Bahram M, Bro R, Stedmon C, Afkhami A. Handling of Rayleigh and Raman scatter for PARAFAC modelling of fluorescence data using interpolation. J Chemometrics 2006;20:99–105.
[22] Bro R, 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; 1998.
[23] Boehme J, Coble P, Commy R, Stovall-Leonard A. Examining CDOM fluorescence variability using principal component analysis: seasonal and regional modelling of three-dimensional fluorescence in the Gulf of Mexico. Marine Chem 2004;89(1–4):3–14.
[24] Andersson CA, Bro R. The N-way toolbox for MATLAB. Chemom Intellig Lab Syst 2000;52(1):1–4.
[25] Kohonen T. The self-organizing map. Neurocomputing 1998;21(1):1–6.
[26] Persson T, Wedborg M. Multivariate evaluation of the fluorescence of aquatic organic matter. Anal Chim Acta 2001;434:179–92.
[27] Lee KI, Yim YS, Chung SW, et al. Application of artificial neural networks to the analysis of two-dimensional fluorescence spectra in recombinant *E. coli* fermentation processes. J Chem Technol Biotechnol 2005;80(9):1036–45.
[28] Despagne F, Massart DL. Neural networks in multivariate calibration. The Analyst 1997;123(11):157R–78R.
[29] Scott SM, James D, Ali Z, et al. Total luminescence spectroscopy with pattern recognition for classification of edible oils. The Analyst 2003;128:966–73.
[30] Wolf G, Almeida J, Crespo JG, Reis MA. An improved method for two-dimensional fluorescence monitoring of complex bioreactors; 2007.
[31] Spencer RGM, Baker A, Ahad JME, et al. Discriminatory classification of natural and anthropogenic waters in two UK estuaries. Sci Total Environ 2007;373:305–23.
[32] Ohno T, Amirbahman A, Bro R. 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. Environ Sci Technol 2008;42:186–92.
[33] Divya O, Mishra AK. Multivariate methods on the excitation emission matrix fluorescence spectroscopy data of diesel-kerosene mixtures: A comparative study. Anal Chim Acta 2007;592(1):82–90.
[34] Henrion R, Henrion G, Böhme M, Behrendt H. Three-way principal components analysis for fluorescence spectroscopic classification of algae species. Fresenius J Anal Chem 1997;357(5):522–6.
[35] Guimet F, Ferré J, Boqué R. Rapid detection of olive-pomace oil adulteration in extra virgin olive oils from the protected denomination of origin "Siurana" using excitation–emission fluorescence spectroscopy and three-way methods of analysis. Anal Chim Acta 2005;544(1–2):143–52.
[36] Bos M, Bos A, van den Linden WE. Data processing by neural networks in quantitative chemical analysis. The Analyst 1993;118(4):323–8.
[37] Basheer IA, Hajmeer M. Artificial neural networks: fundamentals computing design and application. J Microbiol Methods 2000;43(1):3–31.
[38] Kohonen T. Self-organizing maps. 3rd ed. Berlin: Springer; 2001.
[39] Wolf G, Almeida JS, Pinheiro C, et al. Two-dimensional fluorometry coupled with artificial neural networks: A novel method for on-line monitoring of complex biological processes. Biotechnol Bioeng 2001;72(3):297–306.
[40] Hall GJ, Clow KE, Kenny JE. Estuarial fingerprinting through multidimensional fluorescence and multivariate analysis. Environ Sci Technol 2005;39(19):7560–7.

[41] Thompson B. Stepwise regression and stepwise discriminant-analysis need not apply here – a guidelines. Educ Psychol Measure 1995;55(4).
[42] Hopkins OS. Careful consideration necessary when using stepwise regression. J Am Water Works Assoc 2005;97(7):144.
[43] Zupan J, Gasteiger J. Neural networks: a new method for solving chemical problems or just a passing phase? Anal Chim Acta 1991;248(1):1–30.
[44] Hammerstrom D. Working with neural networks. IEEE Spectrum 1993;30(7):46–53.
[45] Bro R. PARAFAC. Tutorial and applications. Chemom Intellig Lab Syst 1997;38(2):149–71.
[46] Marquardt D. An algorithm for least-squares estimation of nonlinear parameters. SIAM J Appl Math 1963;11(2):431–41.