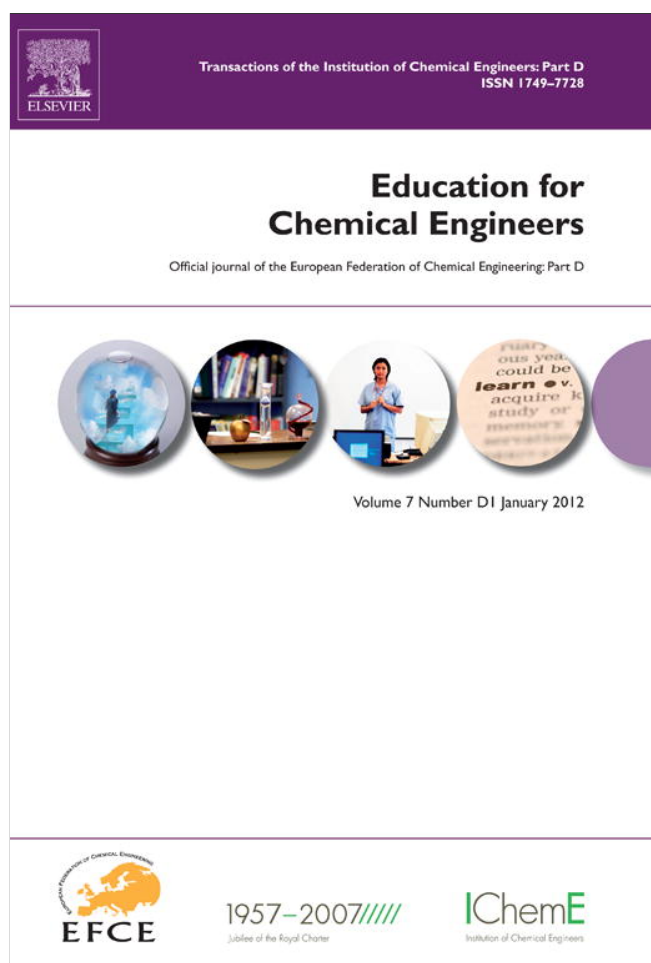


Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>

Contents lists available at [SciVerse ScienceDirect](http://SciVerse.ScienceDirect)

Education for Chemical Engineers

journal homepage: [www.elsevier.com/locate/ece](http://www.elsevier.com/locate/ece)

IChemE

## Exploratory analysis of excitation–emission matrix fluorescence spectra with self-organizing maps—A tutorial

Magdalena Bieroza<sup>a</sup>, Andy Baker<sup>b</sup>, John Bridgeman<sup>c,\*</sup>

<sup>a</sup> Centre for Sustainable Water Management, Lancaster Environment Centre, Lancaster LA1 4YQ, UK

<sup>b</sup> 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

<sup>c</sup> School of Civil Engineering, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

### A B S T R A C T

Large datasets are common in chemical and environmental engineering applications and tools for their analysis are in great demand. Here, the outputs of a series of fluorescence spectroscopy analyses are utilised to demonstrate the application of the self-organising map (SOM) technique for data analysis. Fluorescence spectroscopy is a well-established technique of organic matter fingerprinting in water. The technique can provide detailed information on the physico-chemical properties of water. However, analysis of fluorescence spectra requires the application of robust statistical and computational data pre-processing and analysis tools.

This paper presents a tutorial for training engineering postgraduate researchers in the use of SOM techniques using MATLAB<sup>®</sup>. Via a tutorial, the application of SOM to fluorescence spectra and, in particular, the characterisation of organic matter removal in water treatment, is presented. The tutorial presents a step-by-step example of the application of SOM to fluorescence data analysis and includes the source code for MATLAB<sup>®</sup>, together with presentation and discussion of the results. With this tutorial we hope to popularise this robust pattern recognition technique for fluorescence data analysis and large data sets in general, and also to provide educational practitioners with a novel tool with which to train engineering students in SOM.

© 2011 The Institution of Chemical Engineers. Published by Elsevier B.V. All rights reserved.

**Keywords:** Tutorial; Self-organising maps; Pattern recognition; Fluorescence; Organic matter

### 1. Introduction

The study of engineered processes often generates large, complex datasets and, whilst these datasets contain vast amounts of useful information, obtaining that information via data mining can often prove to be a significant hurdle, particularly for less experienced students and researchers. The University of Birmingham, UK, offers both generic and discipline-specific training in research methods to all its postgraduate researchers. As part of the development opportunities offered to engineering research students, a package of training in data analysis and statistical techniques has been implemented in the School of Civil Engineering. The school's cohort of research students is drawn from a broad base of initial disciplines, including chemical engineers, mathematicians and environmental scientists, in addition to civil

engineers. Senior researchers in the school recognised the need for training of new researchers in robust data analysis and pattern recognition techniques in order to elucidate key mechanisms from large, complex datasets. Included within this training is the use and application of several data mining techniques, including principal component analysis, parallel factor analysis, and artificial neural networks (ANN). To demonstrate the use of one form of ANN, the self-organizing map (SOM), a tutorial has been developed as a training tool. In the tutorial, the SOM is applied to the exploratory analysis of fluorescence data. The application of the SOM technique to the characterisation of organic matter removal in drinking water was first demonstrated by Bieroza et al. (2009a). However, a wide interest in the application of this technique to fluorescence data analysis created a demand for a step-by-step tutorial explaining the practical and computational

\* Corresponding author. Tel.: +44 121 414 5145; fax: +44 121 414 3675.

E-mail address: [j.bridgeman@bham.ac.uk](mailto:j.bridgeman@bham.ac.uk) (J. Bridgeman).

Received 18 March 2011; Accepted 29 October 2011

aspects of the algorithm in real life examples. To meet this demand, we developed this EEM-SOM tutorial for MATLAB®, which we now offer as a general teaching aid for SOM.

## 2. Fluorescence spectroscopy

Fluorescence excitation–emission matrix (EEM) spectroscopy has become a common method for the characterisation of organic matter in aqueous systems. The technique enables rapid, non-invasive and accurate characterisation of different organic matter fractions from various environments: terrestrial (Senesi et al., 1991), freshwater (Baker et al., 2008), estuarine (Stedmon et al., 2003), and marine (Coble, 1996). Recently, the EEM technique has been successfully utilised in drinking water treatment for comprehensive characterisation of organic matter removal across different treatment stages (Bierozza et al., 2009b, 2010).

Although acquisition of fluorescence spectra has become easier with recent advances in the spectrofluorometric technology, fluorescence data analysis still remains a computational and interpretive challenge. Standard techniques for fluorescence data mining, e.g. the peak-picking approach, regional integration technique and principal components analysis (PCA), operate solely on a portion of EEM data available or produce new, tentative variables that are difficult to interpret (Coble, 1996; Persson and Wedborg, 2001; Chen et al., 2003). Therefore, attempts have been made to facilitate other methods for fluorescence data analysis that overcome the drawbacks of the standard techniques and make use of the entire available EEM spectral information.

Parallel factor analysis (PARAFAC) is currently considered as a state-of-the-art tool for fluorescence EEM-organic matter characterisation (Hudson et al., 2007; Henderson et al., 2009). This multi-way technique has been shown to be useful in discriminating between aquatic samples of different origins (Stedmon et al., 2003; Murphy et al., 2008). Several PARAFAC tutorials and review papers are available (i.e. Andersen and Bro, 2003; Stedmon et al., 2003; Stedmon and Bro, 2008).

## 3. Artificial neural networks and self-organising maps

Artificial neural networks (ANNs) are powerful computational tools, frequently used in the modelling of water resources (Maier and Dandy, 2000; Dawson and Wilby, 2001; Bowden et al., 2005). ANNs can be described as mathematical models of a specific structure, consisting of a number of single processing elements (nodes, neurons), arranged in inter-connected layers. A typical artificial neural network is comprised of three layers, namely the input, hidden and output layers (Bos et al., 1993). The input layer is passive and presents the input data vector to the hidden layer through weighted connections. The overall output of the network is calculated as the sum of the outputs of the neurons in the final, output layer. The network is calibrated in the training stage, as weights connecting the layers are modified appropriately. The trained network needs to be validated on data not used in the training phase. If the trained network subsequently returns appropriate results for the independent dataset, it can be used as a calibration or classification model (Bos et al., 1993).

The self-organizing map (SOM, or Kohonen ANN) is a two-layered ANN that provides the conversion of nonlinear statistical relationships between high-dimensional data into

simple geometric relationships on a low-dimensional map, whilst keeping the most important topological and metric relationships of the input data (Kohonen, 2001). The SOM is an example of an unsupervised clustering algorithm in which any existing pattern is assigned to a category, not specified or not known a priori. It is often used in the exploratory data analysis to discern any reasonable relationships among the data, often without prior knowledge or assumptions on the given dataset. The SOM algorithm explores the input data to find and extract features, describing an elementary pattern of information that represents partial aspects or properties of an item (Kohonen, 2001). The SOM's feature in relation to fluorescence analysis can refer to the presence of a particular fluorophore (or group of fluorophores) or its specific spectral properties. This feature extraction facilitated by the SOM involves a nonlinear transformation of the input data onto a two-dimensional map.

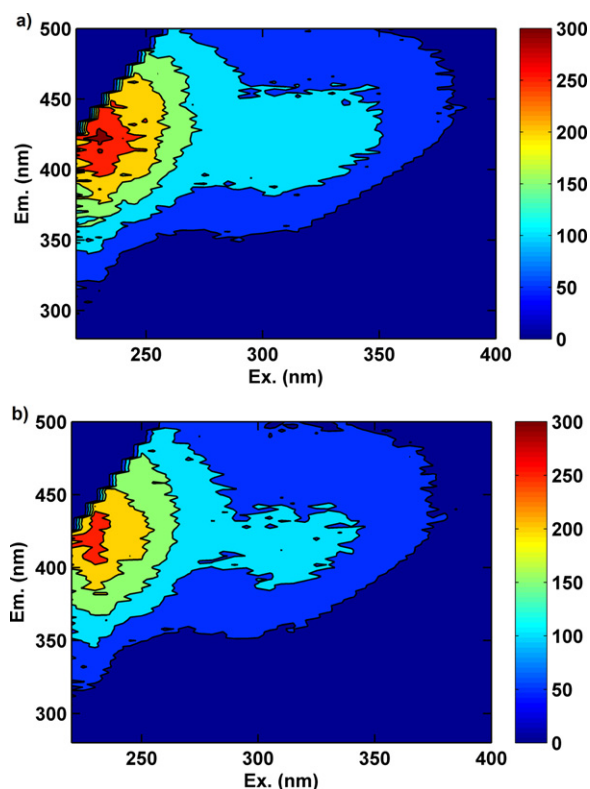
Examples of the application of SOMs to water management issues include classification of fluorescence data (Lee et al., 2005), modelling of water quality (Bowden et al., 2005), and characterisation of various hydrological processes, i.e. rainfall-runoff (Hsu et al., 2002; Kalteh et al., 2008). Recently, the authors have successfully adopted the procedure described in this tutorial for the evaluation of continuous fluorescence data and the detection of diesel pollution when monitoring river organic matter (Carstea et al., 2010). This approach was used successfully to discriminate between overlapping fluorophores. The authors concluded that an expert system incorporating SOM for the on-line evaluation of fluorescence data can be applied as an early warning system of failure of water quality in natural and engineered water systems, thus demonstrating the value of the SOM technique as a real-time water management tool, as well as a technique to be applied to laboratory data.

This tutorial is addressed to those in the wider fluorescence community who seek an alternative data mining tool, and also to everyone interested in the application of the SOM technique to a large environmental dataset. The tutorial consists of two parts; viz. data pre-processing and SOM application, and a brief analysis of results. In the first part, common problems regarding fluorescence data pre-processing and preparation for computational analysis are addressed. In particular, to demonstrate the possibility of combining different EEM analysis techniques, the application of the scatter removal tool from the *DOMFluor* toolbox for MATLAB® is presented (Stedmon and Bro, 2008). In the second part, the step-by-step application of the SOM algorithm to fluorescence data is presented and discussed. The use of the SOM visualization tools is presented and a brief discussion of the results is provided. This tutorial does not intend to present a thorough description of all available SOM tools, but provides a solid background to start the reader's own exploration of this powerful data mining technique.

## 4. Tutorial

### 4.1. Tutorial dataset

Organic matter removal is of the greatest importance to water treatment companies. Residual organic matter reacts with chlorine (when used as a disinfectant) leading to the formation of potentially harmful disinfection by-products. Fluorescence EEM spectroscopy offers rapid and non-invasive measurements with the possibility for incorporation into



**Fig. 1 – Excitation–emission matrices for raw (top) and corresponding clarified water (bottom). Location of the peak maxima: peak A (Ex. 230 nm/Em. 420 nm), peak C (Ex. 320 nm/Em. 420 nm), and peak T (Ex. 220 nm/Em. 330 nm). Fluorescence intensity in Raman units.**

on-line monitoring systems. It has been shown to provide an accurate assessment of organic matter removal efficiency during treatment processes (Bierzoza et al., 2009a). The dataset provided in this tutorial is taken from a study of fluorescence spectroscopy measurements on samples of raw and clarified water from 16 surface water treatment works (WTWs), collected monthly between August 2006 and February 2008. The WTWs treat a range of raw water types, each exhibiting different quantitative and qualitative organic matter properties and varying degrees of anthropogenic impact, (e.g. microbial pollution). A detailed interpretation of this dataset can be found in Bierzoza et al. (2009a).

Fluorescence data were collected by scanning excitation wavelengths in the range from 200 to 400 nm in 5 nm steps, and detecting the emitted fluorescence in 2 nm steps between 280 and 500 nm. For a detailed description of all laboratory analytical methods, refer to Bierzoza et al. (2009a). For the purpose of this tutorial, a subset of 72 EEMs of raw and clarified water derived from six WTWs was selected. During the preliminary analysis of the fluorescence data, the region with excitation wavelengths less than 240 nm was removed as it contained redundant and noisy signal. Thus, each of the resultant EEM ranged from 240 to 400 nm excitation and from 300 to 500 nm emission wavelengths respectively, producing an array  $37 \times 111$ .

Examples of typical EEMs for raw and clarified water are presented in Fig. 1. Fluorescence EEMs exhibit increased intensities in particular regions and these fluorescence regions can be attributed to both natural fluorescence (humic- and fulvic-like), defined as peaks A and C and microbial derived organic matter (tryptophan- and tyrosine-like fluorescence, defined as

**Table 1 – Excitation and emission wavelength pairs for principal peak fluorescence intensities.**

Peak	$\lambda_{Ex}$ (nm)	$\lambda_{Em}$ (nm)
Humic		
A	237–260	400–500
Humic (highly coloured)		
C	300–370	400–500
C <sub>1</sub>	320–340	410–430
C <sub>2</sub>	370–390	460–480
Tyrosine		
B <sub>1</sub>	225–237	309–321
B <sub>2</sub>	275	310
Tryptophan		
T <sub>1</sub>	275	340
T <sub>2</sub>	225–237	340–381
Humic (marine)		
M	290–310	370–410

peaks T and B) at shorter emission wavelengths (Coble, 1996) (Table 1). In drinking water treatment, organic matter removal results in the decrease in fluorescence intensity in all regions between raw and clarified water samples. From the differences in the fluorescence properties of raw and clarified water, Bierzoza et al. (2009a) inferred information regarding organic matter removal. In particular, the decrease in fluorescence intensity between raw and clarified water was correlated with the measured decrease in total organic carbon ( $R^2 = 0.91$ ).

## 4.2. EEM-SOM tutorial

### 4.2.1. DATA pre-processing

In preparing this tutorial, version 7.7.0.471 of MATLAB® has been used.

1. Before starting the tutorial, ensure that the folder *EEM-SOM tutorial* has been downloaded and saved to a known location.

Start a new MATLAB® session by double-clicking the MATLAB® shortcut on the Windows desktop. This opens the MATLAB® desktop application containing by default the Command Window, Workspace Window and Command History Window.

To use the SOM tutorial, your MATLAB® environment needs to locate all required files. Therefore select *File -> Set Path -> Add with subfolders...* and find the folder *EEM-SOM tutorial*. Then press *Save* and, after checking that the new path to your folder appeared at the top of combo box, close the editor. Now you are ready to start working with the *EEM-SOM tutorial*.

2. To load the tutorial workspace, type in the command window of MATLAB®:

```
load EEM.SOM.tutorial.mat
```

and press Enter. This will read the previously prepared binary files associated with fluorescence data into the tutorial workspace. The variable *OriginalData* is a three-way array of size  $72 \times 111 \times 37$  which contains 72 fluorescence EEMs (samples) each comprising 111 emission and 37 excitation wavelengths. The variables *Em* and *Ex* contain the emission and excitation wavelengths respectively, whereas in array *EmEx*, all 4107 ( $111 \times 37$ ) emission and excitation wavelength pairs are stored. Finally, the three variables *Samples*, *Sites* and *Variables* contain the number of samples, sample labels and fluorescence emission–excitation wavelength labels respectively. Double-clicking on any of the variables in the workspace



window opens an additional Variable Editor window that enables editing and analysis of the data. Alternatively, typing each of the variables in the command window shows the data and its structure.

3. The variable *OriginalData* contains raw fluorescence data normalized to Raman scatter peak value. To enhance the modelling of fluorescence data with SOM, the removal of Rayleigh and Raman scatter is recommended. There are several scripts and toolboxes facilitating scatter removal available for MATLAB®. Here, the *DOMFluor* toolbox for MATLAB® was used (Stedmon and Bro, 2008). The toolbox contains several useful functions for fluorescence data processing (i.e. identification of outlier EEMs) and modelling with parallel factor analysis (PARAFAC). The toolbox is available at (<http://www.models.life.ku.dk/source/DOMFluor>).

4. Install the *DOMFluor* toolbox, repeating the steps described in point 1. Prior to using the toolbox, you need to re-format the original data to prepare the structure containing all the relevant information on the fluorescence dataset. In MATLAB®, structures are multidimensional arrays that can contain different types of data, e.g. numerical, textual and logical in a single variable. Here, we can create a structure *EEM* that contains fluorescence intensity data for all samples, emission and excitation wavelengths, number of samples, and an additional back-up copy of the fluorescence data. In the command window, type:

```
EEM = struct('Ex', Ex, 'Em', Em, 'X', OriginalData, 'nEx', 37,
            'nEm', 111, 'nSample', Samples, 'XBackup', OriginalData)
```

and press Enter. This command generates a structure *EEM* with the following properties:

```
EEM =
Ex: [37 × 1 double]
Em: [111 × 1 double]
X: [72 × 111 × 37 double]
nEx: 37
nEm: 111
nSample: 72
XBackup: [72 × 111 × 37 double]
```

You can access any of the structure's elements by typing the structure's name followed by the appropriate field designator. For example typing:

```
EEM.nSample
```

returns the number of fluorescence samples stored in the structure *EEM* (72).

5. To remove scatter from the fluorescence data use the *EEMCut* function. For a detailed description of function arguments and output, refer to the MATLAB® command *help*. Type:

```
help EEMCut
```

In the input arguments of the *EEMCut* function the size of the first and second order scatter have to be defined. The function replaces the original values with a not-a-number constant (NaN) and zeros. Type:

```
EEM.cut = EEMCut(EEM, 40, 40, 15, 15, 'No');
```

For each sample, *EEMCut* plots the input and resultant EEM.

6. Prior to further modelling of fluorescence data with SOM, the three-way fluorescence data array has to be transformed (unfolded) to a two-dimensional array with a size of  $72 \times 4107$ . By entering the following commands:

```
DimX = size(EEM.cut.X);
EEM.som = reshape(EEM.cut.X, DimX(1), prod(DimX(2:end)));
```

first you create a variable *DimX* that contains information on the cut fluorescence data size, and then you can unfold the three-way array by using the command *reshape*. To check the size of the new variable *EEM.som* type:

```
size(EEM.som)
```

In the new variable, each row contains one sample. In the columns fluorescence intensity values are kept for each emission–excitation wavelength pair. The order of emission–excitation wavelength pairs is given in the array *EmEx*. All emission wavelengths are stored in the first row, whereas the corresponding excitation wavelengths are in the second row.

7. Before proceeding with the tutorial, the NaN constants inserted in the place of scatter values have to be replaced with zeros. This is achieved by typing:

```
index = find(isnan(EEM.som));
EEM.som(index) = zeros(size(index));
```

Firstly, the function *find* determines the indices of array elements that are NaNs and stores them in the variable *index*. Then, for the selected columns, zero values are inserted.

Columns containing zeros are redundant and if not removed can impede further modelling. Therefore typing:

```
empty = find(sum(EEM.som) == 0);
nonempty = find(sum(EEM.som) ~= 0);
EEM.som(:, empty) = [];
EmEx.som = EmEx;
EmEx.som(:, empty) = [];
```

first finds the columns of the *EEM.som* array that contain only zeros and then removes them from both the fluorescence intensity data and the array containing emission–excitation wavelengths pairs (*EmEx.som*). The final fluorescence dataset *EEM.som* comprises 72 samples and 2615 fluorescence emission–excitation pairs and this can be used in SOM modelling.

8. SOM toolbox for MATLAB® (Alhoniemi et al., 2002) can be downloaded from (<http://www.cis.hut.fi/projects/somtoolbox/download>). To get started, and familiarize yourself with the SOM terminology, type *help somtoolbox*. This entry contains a full list of available functions and references to useful, step-by-step tutorials:

```
som_demo1 SOM Toolbox demo 1: basic properties
som_demo2 SOM Toolbox demo 2: basic usage
som_demo3 SOM Toolbox demo 3: visualization
som_demo4 SOM Toolbox demo 4: data analysis
```

9. Similar to the *DOMFluor* toolbox, the *SOM* toolbox operates on data stored in a particular structure. Type:

```
SOM_data = som_data_struct(EEM_som, 'labels', Sites,
'comp_names', Variables);
```

to generate a *SOM\_data* structure containing fluorescence data (*EEM\_som*), sample labels (*Sites*), and variable (component) names (*Variables*).

Prior to SOM analysis, the data need to be normalized to improve the algorithm's numerical accuracy (Kohonen, 2001). In practice, the data can be normalized in the way that the variance of the dataset is equal to unity or the mean is subtracted from each variable (normalization in the range 0–1). The function *som\_normalize* also enables different types of normalization, i.e. logistic or histogram equalization (for a full list of the available normalizations refer to help *som\_normalize*).

To normalize the fluorescence data type:

```
SOM_data = som_normalize(SOM_data, 'var');
```

The parameters of the normalization are stored within the *SOM\_data* structure within the *comp\_norm* field designator.

#### 4.2.2. SOM analysis

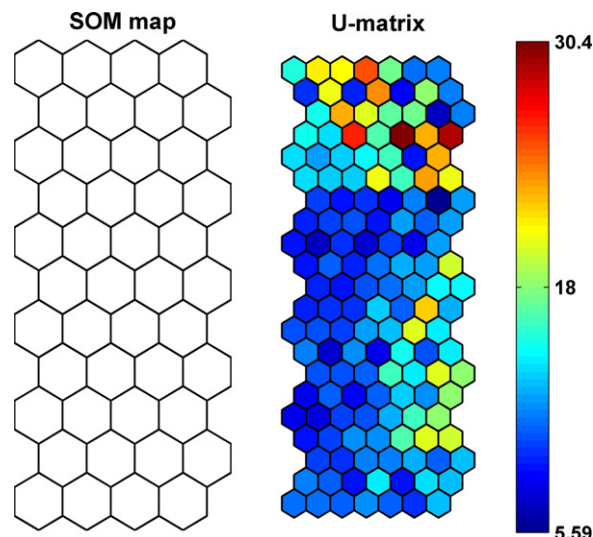
1. The self-organizing map is an example of a two-layered artificial neural network (ANN). The input layer constitutes the interface between the map and the input data. Each neuron from the input layer is fully connected with all input samples and the numerical representation of this connection is held in the associated reference vector. The reference vector contains weights that are adjusted in the training process to project the input data. The reference vector (also called the codebook or weight vector) can be defined as  $d_i = [d_{i1} \ d_{i2} \ \dots \ d_{im}]$ , where  $m$  is equal to the dimension of the input vectors (2615 fluorescence excitation–emission pairs). Each reference vector from the input layer is associated with one neuron from the output layer (two-dimensional map).

The SOM map building process consists of three principal steps. At first the size of the map is usually determined by finding the ratio of the two greatest eigenvalues of the input matrix. For the calculated size, a map is initialized using linear initialization along the two greatest eigenvectors of the input matrix. Then the SOM algorithm learns the pattern of the input data in the training process. Input data are simultaneously presented to the SOM network. For each sample, the output neuron with the reference vector most similar to the unfolded fluorescence spectra is selected (*the winning neuron*). Once the best-matching reference vector for each input vector is found, its weights and the weights of its neighbouring neurons are modified and moved towards the input vector (Eq. (1)):

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

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,  $h_p(i, k)$  describes the neighbourhood of the winning neuron,  $k$  is the number of epochs (a finite set of input patterns presented sequentially) and  $p$  is the index of the winning neuron.

During map training, both the learning rate (which describes speed of the training) and the radius of the neighbourhood decrease monotonically. Map training involves a two stage process; viz., the rough training phase (with a large neighbourhood radius) and the fine-tuning phase (with a small neighbourhood radius).



**Fig. 2 – The self-organizing map (left) and the U-matrix (right).**

To create a SOM map type:

```
SOM_map = som_make(SOM_data);
```

The whole process should take up to 10 min on a standard-speed computer.

The newly created map has the dimensions 11 neurons  $\times$  4 neurons. The quality of the map projection is estimated with two diagnostics: final quantization error (here 16.949) and final topographic error (here 0.056). The final quantization error is the average distance between each input vector and its best-matching vector and can be used as a measure of the map resolution. The final topographic error is defined as the proportion of all input vectors for which the first and second best-matching vectors are not adjacent (for measuring topology preservation) (Kohonen, 2001).

A new variable *SOM\_map* contains the specific characteristics of the map (i.e. topology, type of neighbourhood between map units, sample labels) and the results of the map training stored in the *codebook* field. The *codebook* contains a numerical representation of the map's reference vectors; thus it is a two-dimensional array of a size corresponding to the number of nodes (44) and the number of variables (2615 fluorescence pairs).

2. The SOM map evaluation begins with the analysis of the unified distance matrix (U-matrix) (Ultsch, 1993). Type:

```
som.show(SOM_map, 'empty', 'SOM map', 'umat', 'all');
```

The U-matrix demonstrates the distances between neighbouring map units that are calculated and visualized using grey or colour scale on the trained map (Park et al., 2003). Compared with the original map size (11  $\times$  4 neurons), the U-matrix comprises additional map units to visualize the distances between neurons (Fig. 2). High values on the U-matrix indicate large distances between neighbouring units and hence can be helpful in determining the cluster borders. Clusters typically form uniform areas of low values. Here the presence of a few clusters can be discerned, e.g. the cluster in the upper right side corner.

Type help *som.show* to learn more about this SOM visualization function.

To evaluate the sample distribution on the SOM map type:

```
SOM_map = som_autolabel(SOM_map, SOM_data, 'vote');
som_show(SOM_map, 'umat', 'all', 'empty', 'Samples');
som_show_add('label', SOM_map, 'Textsize', 12,
'TextColor', 'b',
'Subplot', 2);
```

First, the function *som\_autolabel* automatically labels the neurons. During training, each neuron can be the winning neuron for more than one sample, based on the similarities of the input data. Thus, each neuron stores the labels of all samples assigned to this neuron. In the *som\_autolabel* function with *vote* mode, only the label with the most instances is kept. Finally, the *som\_show\_add* function defines the properties of label visualization on the SOM map, i.e. text size, font colour. The command *som\_show\_add* can only be applied to *som\_show* visualization, therefore both commands should be pasted together in the command window.

As a result, both the U-matrix and sample distribution on the map are presented (Fig. 3). The samples grouped together have similar fluorescence properties, i.e. the same pattern of fluorophores, similar spectral location of the humic peaks. For instance, the upper right side corner on the U-matrix comprises clarified water samples from sites 1, 4 and 6, indicating similarities in the post-coagulation residual organic matter properties. Furthermore, it can be observed that the raw water samples are located in the bottom part of the SOM map, whereas the corresponding clarified water samples are to be found in the upper part. Thus, the vertical direction on the SOM map reflects the relative decrease in fluorescence intensity and organic matter quantity. Site 1 exhibits uniform and stable organic matter fluorescence properties of both raw and clarified water, whereas site 2 presents a more complex nature, with raw and clarified water samples distributed over the entire map. The horizontal map direction differentiates between sites 1 and 3, and hence the opposing organic matter character of those sites can be hypothesized.

3. Another basic visualization technique of the *som\_show* function is the hit histogram.

For each neuron, the hit characteristic is calculated on the basis of the map response to the input data. The hit characteristic shows how many times each neuron was the winning neuron for the dataset. Type:

```
hit1 = som_hits(SOM_map, SOM_data);
som_show(SOM_map, 'empty', 'Hits');
som_show_add('hit', [hit1], 'Marker', 'lattice', 'MarkerColor',
[1 0 0], 'Text', 'on', 'TextColor', [0 0 0], 'TextSize', 12);
```

The function *som\_hits* calculates the number of hits for the map over the given domain (here for all samples). The importance of each neuron (hit characteristic) can be presented graphically with different size markers or with a label showing the number of hits. It can be seen from Fig. 4a that data are uniformly distributed over the map, with a number of neurons located at the edges of the map being the most frequent neurons (neuron 7–4 hits, neuron 33–5 hits, neuron 34–6 hits, neuron 41–5 hits).

Similar to the single hit histogram, a multiple hit histogram can be defined. To compare the hit response for raw and corresponding clarified water for site 1 type:

## Samples

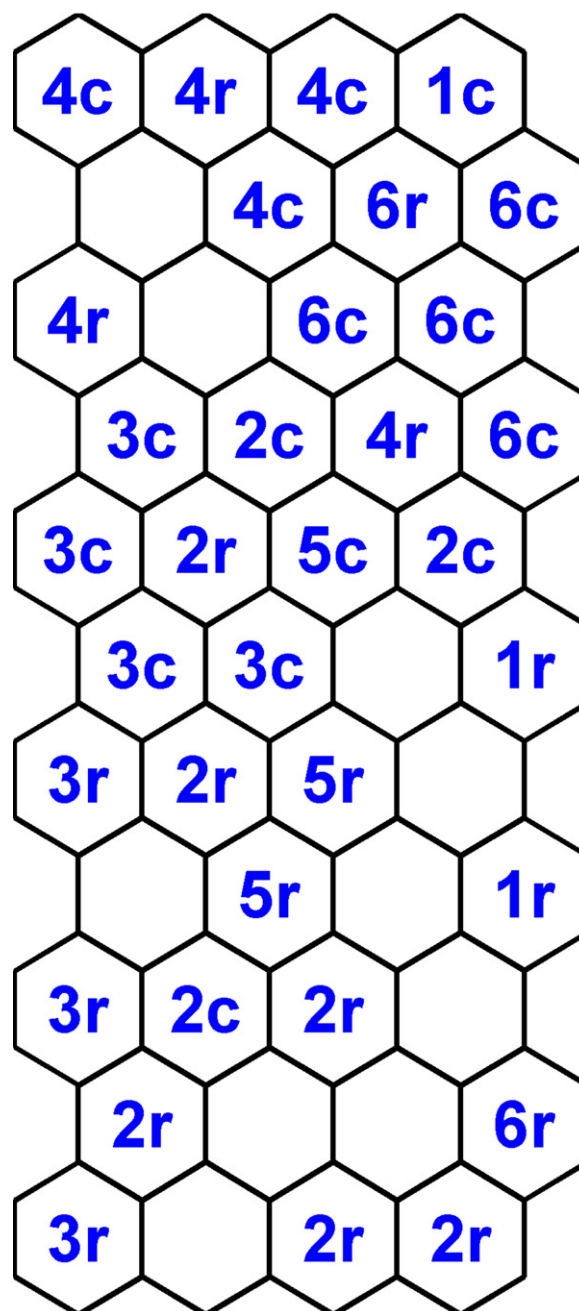


Fig. 3 – Sample distribution on the SOM map. Labelling used (e.g. 1r – site 1 raw water; 1c – site 1 clarified water).

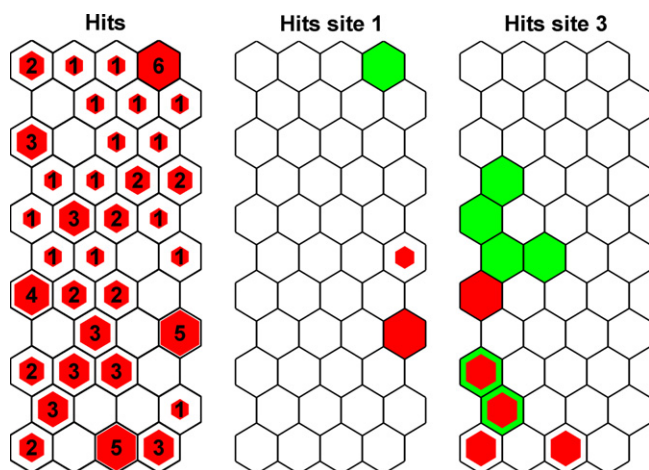
```
hit1 = som_hits(SOM_map, SOM_data.data(1:6,:));
hit2 = som_hits(SOM_map, SOM_data.data(7:12,:));
som_show(SOM_map, 'empty', 'Hits site 1');
som_show_add('hit', [hit1, hit2], 'Marker', 'lattice', 'MarkerColor',
[1 0 0; 0 1 0]);
```

Ignore the warnings regarding RGB colour. For site 1, two hit histograms are defined: *h1* for raw water (samples 1–6) and *h2* for the corresponding clarified water (samples 7–12).

Similarly for site 3 the above entry takes the following form:

```
hit1 = som_hits(SOM_map, SOM_data.data(25:30,:));
hit2 = som_hits(SOM_map, SOM_data.data(31:36,:));
som_show(SOM_map, 'empty', 'Hits site 3');
```





**Fig. 4 – Single hit histogram (a), multiple hit histograms site 1 (b) and site 3 (c). Red marker – raw water; green marker – clarified water.**

```
som_show_add('hit', [hit1, hit2], 'Marker', 'lattice', 'Marker-Color', [1 0 0; 0 1 0]);
```

A comparison of hit histograms for both sites (Fig. 4b and c), reveals the differences between both raw and clarified water organic matter properties. Site 3 shows a great variation in both raw and clarified water properties that may be indicative of differing organic matter removal efficiency. Conversely, site 1 demonstrates uniform fluorescence properties for both water types.

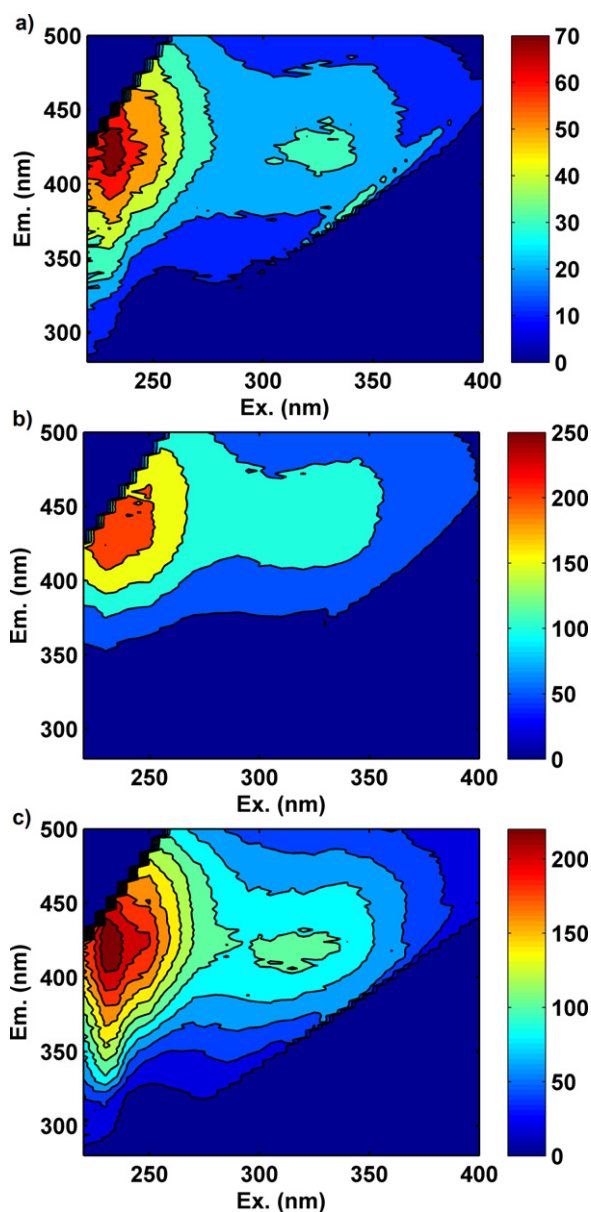
4. To correlate the location of a sample on the map with the fluorescence spectra, an analysis of the reference vectors should be performed. To retrieve the spectral information stored in the neuron of interest, (here neuron 34 pertinent to sample 1c with the highest number of hits), type:

```
EEM_den = som_denormalize(SOM_map);
REF_vec(1, nonempty) = EEM_den.codebook(34,:);
REF_eem = reshape(REF_vec, 111, 37);
contourf(Ex, Em, REF_eem(:,,:)), colorbar
xlabel('Ex. (nm)')
ylabel('Em. (nm)')
```

First, the de-normalized fluorescence data (function *som\_denormalize*) is transformed to include the empty columns removed prior to SOM modelling and reshaped to a three-way array (sample by emission by excitation wavelength). Then, a filled contour EEM is plotted (Fig. 5a).

The above procedure can be repeated for neuron 41 (5 hits, reflects spectral properties of sample 1r, Fig. 5b) and neuron 7 (4 hits, reflects spectral properties of sample 3r, Fig. 5c). A distinctive decrease in fluorescence intensity in all fluorescence regions can be observed between the raw and clarified water properties of site 1 (Fig. 5a and b). Furthermore, a shift towards lower emission wavelengths can be discerned for the corresponding clarified water sample, indicating selective removal of the more hydrophobic organic matter fraction. The raw water of site 3 (Fig. 5c) exhibits more hydrophilic and microbial character than its clarified water (Fig. 5b) as indicated by the presence of tryptophan-like fluorescence.

5. The importance of each fluorescence variable in determining the samples' distribution on the map can be depicted with the use of the component plane. Thus, for each excitation-emission wavelength pair, a corresponding



**Fig. 5 – Reference vectors of neuron 34 (a), neuron 41 (b), and neuron 7 (c).**

component plane can be obtained that enables a correlation between a sample's location on the map and fluorescence properties. In the current study there are 2615 variables, therefore only a subset can be displayed at one time. Type:

```
som_show(SOM_map, 'comp', [1:10]);
```

to display the first ten component planes, showing a very different response due to the presence of noise signal. The high values in the component plane denote a higher fluorescence intensity. To enhance the analysis, a few fluorescence variables have been selected from the entire fluorescence dataset to cover the most interesting fluorescence regions (Table 2).

After typing:

```
som_show(SOM_map, 'comp', [162, 187, 212, 234, 1112, 1137, 1162, 1187, 2021, 2031, 2056, 2081]);
```

it can be observed (Fig. 6) that for each excitation wavelength, with increasing emission wavelength, the centre of the highest



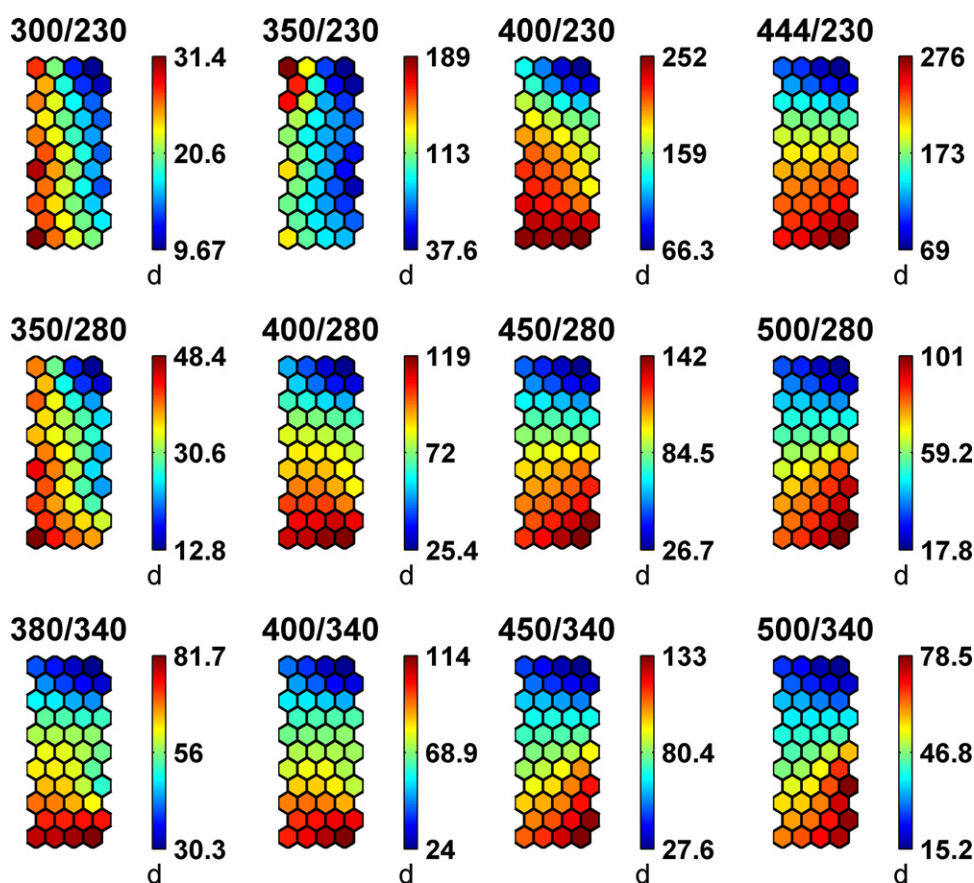


Fig. 6 – Selected component planes (excitation/emission wavelength nm).

**Table 2 – Emission and excitation wavelengths of the selected component planes.**

Ex/Em (nm)	300	350	400	450	500
230	162	187	212	234	–
280	–	1112	1137	1162	1187
340	–	2021	2031	2056	2081

values moves from the left, through the bottom of the map, to the right. Thus, the horizontal plane of the SOM map reflects the degree of aromaticity, whereas the vertical plane can be correlated with fluorescence intensity. Combining the observations derived from the component planes and samples' distribution provides useful information on the organic matter character. For example, the highest values on the component plane related to tryptophan-like fluorescence (excitation 280 nm, emission 350 nm), correspond with the location of site 3 raw water samples, indicating the predominance of microbial organic matter.

## 5. Discussion

The procedure described in this tutorial can be adopted directly for other environmental datasets when recognition of the underlying patterns is of interest but the large size of the dataset (large number of variables) prevents the application of common statistical methods, i.e. correlation between variables.

The first part of this tutorial described the pre-processing procedure pertinent to fluorescence excitation–emission data

including transformation of three-dimensional EEMs into two-way arrays. This step of the analysis is dependent on the particular dataset being considered and specific problems associated with data pre-processing. However for most environmental datasets, some general issues have to be addressed prior to further modelling; e.g. how to treat missing values. Here, NaNs were inserted in the place of scatter signal. As these data did not contain any important information, they were treated as a redundant data and simply removed from further SOM analysis. Nevertheless, for other datasets, missing values may hold vital process information which should be incorporated in the analysis after adequate treatment (e.g. interpolation). This decision is dependent upon the dataset in question.

Assuming that any environmental data can be presented as a two-way array of observations (e.g. sampling points) and variables describing the process (e.g. process parameters), the second part of the tutorial (SOM analysis) can be directly applied to any dataset. The speed of SOM modelling (map building) depends on both the number of observations and the number of variables. Simple tests were performed to assess the speed of SOM map formation using fluorescence datasets of varying sizes. It was found that the number of variables is of greater importance to the speed of the algorithm than the number of observations. For a constant number of observations (72), doubling the number of variables (from 2615 to 5230) resulted in a 6.5 times increase in map building time (2 and 13 min respectively). Modelling with varying number of observations did not generate significant changes to modelling time (for 36 samples, 1 min 55 s; for 72 samples, 2 min). It should be also noted that variables can demonstrate varying importance in explaining dataset variation; therefore the

more difficult the pattern which is analysed with the SOM, the longer the map building stage will be.

The SOM interpretation process involves evaluation of the sample distribution and importance of the particular variables. The latter can be analysed using component planes. However, as the number of process variables corresponds to the number of component planes to be evaluated, for large datasets this SOM interpretation tool can present certain limitations. For instance, in the example presented in this tutorial, there were 2615 fluorescence variables, for which the same number of component planes was formed. As the visual inspection of 2615 component planes could impede the interpretation process, a selection of interesting component planes was performed based on prior knowledge of the fluorescence properties of the dataset. Thus, component planes demonstrating fluorescence properties of certain fluorophores were analysed in more detail (Fig. 6). However, the availability of other SOM interpretation methods (i.e. analysis of hit histograms) can successfully overcome this potential limitation for large datasets.

The validation stage commonly used in calibration with neural networks is not performed with the SOM. In calibration tasks where an ANN is required to model the input–output relationship, the error of the prediction can be calculated for a validation dataset not used in the training process. In SOM, validation refers both to the comparison of network performance with the objectives and evaluation of its usefulness (i.e. finding the pattern, discriminating between sampling sites). The SOM explores the original data whilst preserving all topological and geometric properties and thus enables direct comparison between samples. In this way, the SOM output for a given dataset is always valid and its usefulness depends on the properties of the data themselves. The addition of new samples changes the distances between data vectors and therefore produces a new SOM map. Furthermore, in the SOM an important step is the validation of data to select a reliable dataset for training of the SOM algorithm (i.e. detection and removal of outliers and scatter). However, compared to other techniques used for fluorescence data analysis (e.g. PCA or PARAFAC), SOM offers high noise and fault tolerance that makes it an appropriate tool for analysis of fluorescence data without pre-processing, i.e. removing of scatter. Here the scatter was removed solely to enhance the speed of the analysis as the number of variables has a significant impact on the training times.

The main advantage of SOM over more commonly used techniques for fluorescence data analysis is that SOM is an entirely data-driven approach where the entire EEMs can be used to discriminate between samples, and consequently no assumptions have to be made regarding spectral location of fluorophores or final components. A further, not inconsiderable advantage of SOM over PARAFAC is its ability to analyse fluorescence data at short excitation wavelengths. The analysis of these interesting data is usually impeded due to limited experimental conditions and lack of instrumental capability. Here, the SOM modelling results were compared for both the full dataset (excitation wavelengths >200 nm) and the truncated EEMs (excitation wavelengths >240 nm). The size of the SOM map and the main topological and structural features remained the same for both datasets. There were only minor differences in sample distribution and number of hits. Thus, there is no limitation within the SOM to analyse the short UV spectral range.

## 6. Conclusions

This tutorial has presented a step-by-step application of a robust unsupervised SOM algorithm to fluorescence spectroscopy data. The technique was employed for the characterisation of fluorescence EEMs obtained for raw and partially treated waters at 16 WTWs in the Midlands region of the UK. The fluorescence data contain a substantial amount of information on the organic matter properties of the water samples. However the complexity of the data impedes any direct correlation between sample distribution and spectral properties of organic matter. Here, the SOM facilitated pattern recognition of the fluorescence data. With reference to the fluorescence differences between raw and partially treated water, the SOM enabled correlation of the fluorescence properties with organic matter removal efficiency and properties for particular WTWs. These results demonstrate that SOMs can be a robust tool in the analysis of fluorescence data characterising organic matter properties in aquatic samples.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.ece.2011.10.002](https://doi.org/10.1016/j.ece.2011.10.002).

## References

- Alhoniemi, E., Himberg, J., Parhankangas, J., Vesanto, J., 2002. SOM Toolbox., <http://www.cis.hut.fi/projects/somtoolbox/download>.
- Andersen, C.M., Bro, R., 2003. Practical aspects of PARAFAC modeling of fluorescence excitation–emission data. *Journal of Chemometrics* 17, 200–215.
- Baker, A., Tipping, E., Thacker, S.A., Gondar, D., 2008. Relating dissolved organic matter fluorescence and functional properties. *Chemosphere* 73, 1765–1772.
- Bierozza, M., Baker, A., Bridgeman, J., 2010. Assessing organic matter removal efficiency at water treatment works using fluorescence spectroscopy. *Drinking Water Engineering and Science* 3, 63–70.
- Bierozza, M., Baker, A., Bridgeman, J., 2009a. Exploratory analysis of excitation–emission matrix fluorescence spectra with self-organizing maps as a basis for determination of organic matter removal efficiency at water treatment works. *Journal of Geophysical Research-Biogeosciences* G00F07, 114, [doi:10.1029/2009JG000940](https://doi.org/10.1029/2009JG000940).
- Bierozza, M., Baker, A., Bridgeman, J., 2009b. Relating freshwater organic matter fluorescence to organic carbon removal efficiency in drinking water treatment. *Science of the Total Environment* 407, 1765–1774.
- Bos, M., Bos, A., van den Linden, W.E., 1993. Data processing by neural networks in quantitative chemical analysis. *The Analyst* 118 (4), 323–328.
- Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1—background and methodology. *Journal of Hydrology* 301, 75–92.
- Carstea, E., Baker, A., Bierozza, M., Reynolds, D., 2010. Continuous fluorescence excitation–emission matrix monitoring of river organic matter. *Water Research*, <http://dx.doi.org/10.1016/j.watres.2010.06.036>.
- Chen, W., Westerhoff, P., Leenheer, A., Booksh, K., 2003. Fluorescence excitation–emission matrix regional integration to quantify spectra for dissolved organic matter. *Environmental Science and Technology* 37, 5701–5710.
- Coble, P.G., 1996. Characterization of marine and terrestrial DOM in seawater using excitation–emission matrix spectroscopy. *Marine Chemistry* 51, 325–346.

- Dawson, C.W., Wilby, R.L., 2001. Hydrological modelling using artificial neural networks. *Progress in Physical Geography* 25 (1), 80–108.
- Henderson, R.K., Baker, A., Murphy, K.R., Hambly, A., et al., 2009. Fluorescence as a potential monitoring tool for recycled water systems: a review. *Water Research* 43 (4), 863–881.
- Hsu, K.L., Gupta, H.V., Gao, X.G., et al., 2002. Self-organizing linear output map (SOLO): an artificial neural network suitable for hydrologic modelling and analysis. *Water Resources Research* 38 (12).
- Hudson, N.J., 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.
- Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. *Environmental Modelling and Software* 23, 835–845.
- Kohonen, T., 2001. *Self-organizing Maps*, 3rd ed. Springer, Berlin.
- Lee, K.I., Yi, Y.S., Chung, S.W., et al., 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.
- Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling and Software* 15, 101–124.
- Murphy, K.R., Stedmon, C.A., Waite, T.D., Ruiz, G.M., 2008. Distinguishing between terrestrial and autochthonous organic matter sources in marine environments using fluorescence spectroscopy. *Marine Chemistry* 108, 40–58.
- Park, Y.S., 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.
- Persson, T., Wedborg, M., 2001. Multivariate evaluation of the fluorescence of aquatic organic matter. *Analytica Chimica Acta* 434, 179–192.
- Senesi, N., Miano, T.M., Provenzano, M.R., Brunetti, G., 1991. Characterization, differentiation and classification of humic substances by fluorescence spectroscopy. *Soil Science* 152, 259–271.
- Stedmon, C.S., 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.
- Stedmon, C.S., Bro, R., 2008. Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial. *Limnology and Oceanography Methods* 6, 572–579.
- Ultsch, A., 1993. Self-organizing neural networks for visualization and classification. In: Opitz, O., Lausen, B., Klar, R. (Eds.), *Information and Classification*. Springer-Verlag, Berlin, pp. 307–313.