Chemosphere 155 (2016) 283–291

Contents lists available at ScienceDirect

Chemosphere

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

Combining parallel factor analysis and machine learning for the classification of dissolved organic matter according to source using fluorescence signatures

C.W. Cuss [a, 1], S.M. McConnell [b], C. Guéguen [c, *]

[a] Environmental and Life Sciences Graduate Program, Trent University, ON, Canada
[b] Department of Computing and Information Systems, Trent University, ON, Canada
[c] Chemistry Department, Trent University, ON, Canada

HIGHLIGHTS

- Machine-learning applied to 1029 PARAFAC-modeled EEMs to classify 24 DOM sources.
- Classification accuracy: 97% river vs leachate; 93% leachate by species; 87% by river.
- Some machine learning algorithms achieved higher classification accuracies.
- Accuracy similar to NPLS-DA, but faster and with simultaneous multiclass comparison.
- Extending # components past cross-validated PARAFAC model improved accuracy.

ARTICLE INFO

Article history:
Received 3 January 2016
Received in revised form
17 April 2016
Accepted 18 April 2016
Available online 29 April 2016

Handling Editor: I. Cousins

Keywords:
Parallel factor analysis (PARAFAC)
Excitation-emission matrix (EEM)
Data mining/machine learning
Leaf leachate
K-nearest neighbours (kNN)
Dissolved organic matter (DOM)

ABSTRACT

Parallel factor (PARAFAC) analysis of dissolved organic matter (DOM) fluorescence has facilitated a surge of investigation into its biogeochemical cycling. However, rigorous, PARAFAC-based methods for holistically distinguishing DOM sources are lacking. This study classified 1029 PARAFAC-analyzed excitation-emission matrices (EEMs) measured using DOM isolated from 24 different leaf leachates, rivers, and organic matter standards using four machine learning methods (MLM). EEMs were also divided into subsets to assess the impact of experimental treatments (i.e. whole EEMs, size fractionation, mixtures, quenching) and dataset properties (i.e. different numbers of EEMs from each leachate/river) on classification. A split-half validated, 10-component PARAFAC model was extended to 12 components to remove consistent peaks evident in model residuals. The 12-component model performed better than the 10-component model, correctly classifying up to 80 additional EEMs, when the dataset included size-fractionated DOM or several different sources (i.e. many leaf species and rivers); however, the 10-component model performed better for whole-sample EEMs when comparing leaf leachates to rivers. The MLM correctly classified whole EEMs of riverine DOM by source with up to 87.0% accuracy, leachates with up to 92.5% accuracy, and distinguished leachates from rivers with 97.2% accuracy. A difference of up to 17.3% in classification accuracy was observed depending on the MLM method used with the following order: multilayer perceptron = support vector machine > *k*-nearest neighbours ≫ decision tree; however, performances differed widely depending on the data subset. Classification accuracy for whole and size-fractionated rivers compared to whole and size-fractionated leachates using N-way partial least-squares discriminant analysis (NPLS-DA; 97.7%) was similar to that achieved using MLM. Combining MLM with PARAFAC is an effective method for classifying DOM based on its fluorescence signature because PARAFAC can isolate meaningful fluorescent species and unlike PLSDA, MLM constructs a single model which simultaneously classifies EEMs as belonging to one of several categories. A complete

* Corresponding author. Department of Chemistry, Trent University, 1600 West Bank Drive, Peterborough, ON K9J 7B8 Canada.
E-mail addresses: cuss@ualberta.ca (C.W. Cuss), celinegueguen@trentu.ca (C. Guéguen).
[1] Present address: Department of Renewable Resources, University of Alberta, Canada.

http://dx.doi.org/10.1016/j.chemosphere.2016.04.061

accounting of carbon flows through ecosystems should include the processes and sources that contribute to the disparate fluorescence signatures of riverine and leached DOM.

1. Introduction

Dissolved organic matter (DOM) is a complex and dynamic mixture of molecules that arises from the death and exudation of organisms, and their subsequent decomposition. DOM plays various roles in natural and engineered ecosystems, including: serving as either a source or sink for carbon, through decomposition to produce carbon dioxide or methane, or sequestration in soils and sediments (Kalbitz and Kaiser, 2008; Jiao and Zheng, 2011; Lorenz, 2013); serving as a source of energy, nutrients, and protection from harmful UV light for microorganisms (Williamson et al., 2001; Zepp et al., 2007; Sleighter et al., 2014); binding with heavy metals and pollutants, thereby controlling their mobility and toxicity (Guéguen and Dominik, 2003; Aiken et al., 2011); forming carcinogenic disinfection byproducts by binding with chlorine during the disinfection of drinking water (Villanueva et al., 2004; Beggs and Summers, 2011). In these roles the effectiveness of DOM depends upon its composition and structure; however, these qualities are highly dynamic. Thus, cost effective and efficient methods for the frequent characterization of DOM are necessary to measure its changes.

Excitation-emission matrix (EEM) fluorescence spectroscopy is a rapid and cost-effective method for characterizing DOM. EEMs are generated by aligning the scans of multiple emission (Em) wavelengths, each measured at a different excitation (Ex) wavelength (Coble, 1996). Typically, each EEM contains thousands to tens of thousands of Ex/Em pairs, many of which are highly correlated. DOM fluorescence appears as correlated sets of Ex/Em pairs (or 'peaks') because it arises from a combination of individual molecules and both intra- and inter-molecular interactions (Lakowicz, 2006; Aiken, 2014; Sharpless and Blough, 2014). Thus, it is an indicator of both composition and structure.

Despite the overlap of fluorescent species in DOM EEMs, parallel factor (PARAFAC) analysis has proven successful for resolving EEMs into independent components that have shown a high degree of similarity across multiple studies (Stedmon et al., 2003; Ishii and Boyer, 2012; Murphy et al., 2014a; Parr et al., 2014). In addition to reducing the high dimensionality of EEMs (typically from thousands of Ex/Em pairs to < 10 components), PARAFAC has been used to distinguish DOM from the leachates of leaves from different tree species, from different rivers, and from different marine environments (Stedmon and Markager, 2005a; Jørgensen et al., 2010; Murphy et al., 2008; Cuss and Guéguen, 2013), and for tracing DOM isolated from different sources through ecosystems and biogeochemical processes (Stedmon and Markager, 2005b; Larsen et al., 2010; Osburn et al., 2012; Chen and Jaffé, 2014). The PARAFAC analysis of fluorescence has also been used to connect DOM to its effectiveness as a source of nutrients for microorganisms (Cuss and Guéguen, 2012a, 2015a), a binder of heavy metals (Yamashita and Jaffé, 2008; Chen et al., 2013; Cuss and Guéguen, 2014), and a precursor for disinfection byproducts (Beggs and Summers, 2011; Lyon et al., 2014). Thus, the discrimination of DOM isolated from different sources using fluorescence (e.g. Chen et al., 2010) enables the refined tracing of its biogeochemical dynamics, and can facilitate the identification of relative contributions in waters that contain DOM arising from multiple sources (Fellman et al., 2010), and in estuarine mixing (Stedmon and Markager, 2005a; Huguet et al., 2007). However, linking the contributions of DOM from different sources to the behaviour of DOM mixtures requires both the accurate distinction of the endmembers and the extraction of spectra that represent different fluorescent species. While mixing models have been applied to PARAFAC-modeled EEMs for estimating proportional contributions during the mixing of several endmembers (e.g. Larsen et al., 2015), source discrimination is useful for detecting the dominant contributor, or for determining the location of origin for a water sample. This type of pure endmember distinction has proven useful for distinguishing fresh and marine waters to test whether ballast water exchange has been completed using PARAFAC with N-way partial-least squares discriminant analysis (NPLS-DA) (Hall et al., 2005), and the intensity of fluorescence at two excitation-emission wavelength pairs (Murphy et al., 2006). Hence, distinguishing between pure endmembers can be useful in forensic applications that seek to distinguish between sourcewaters.

The discrimination of DOM from different sources using its fluorescence signature has been qualitatively achieved by comparing the relative levels of individual PARAFAC components (Stedmon et al., 2003; Carstea et al., 2014), by using ratios or indicators (Gabor et al., 2014; Huang et al., 2015), and by combining PARAFAC with multivariate analyses (e.g. principal component analysis; Cuss and Guéguen, 2013; Chen and Jaffé, 2014; Kothawala et al., 2014); however, there are a lack of methods for the quantitative estimation of differences that use the entire fluorescence signature (i.e. all PARAFAC components). Descriptions of the degree to which mixtures can be distinguished based on fluorescence composition, and of the degree of difference between end members and mixtures, also remain qualitative.

PARAFAC deconvolutes the fluorescence signatures of underlying fluorophores and thereby describe meaningful fluorescent species, whereas the components produced by NPLS-DA (Wittrup, 2000; Hall et al., 2005; Murphy et al., 2014b) and self-organizing maps with backpropagation artificial neural networks (SOM-BPNN; Bieroza et al., 2012) are not chemically meaningful. Thus, applying machine learning methods to classify DOM according to source using their PARAFAC-based signatures effectively combines the discriminatory power of methods like NPLS-DA and SOM-BPNN with the description of meaningful spectra achieved by PARAFAC. NPLS is also limited to binary comparisons in each model so that multiple analyses are required to group samples by source if more than two sources exist, and the classification of a sample as belonging to more than one source or no source is also possible (Hall et al., 2005). On the other hand, the machine learning methods used in this study are capable of simultaneous multi-source comparison, which saves valuable computation time and classifies samples as belonging to a single source.

In this study, 1073 EEMs were generated by measuring DOM from 24 different leaf leachates and rivers under similar physicochemical conditions, and decomposed using PARAFAC models with 10 and 12 components. For the first time, the resulting fluorescence compositions were classified by leachate/river using four established machine learning methods to quantify differentiability. The classification accuracy achieved by applying different machine learning methods and PARAFAC models were compared both with each other and NPLS-DA, and differences were related to the