Supplemental File NOT for Review

Understanding Organic Nonpoint-Source Pollution in Watersheds via Pollutant Indicators, Disinfection By-Product Precursor Predictors, and Composition of Dissolved Organic Matter

Yixiang Zhang [a], Xinqiang Liang [a,b,*]

[a] Department of Environmental Engineering, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China

[b] Zhejiang Provincial Key Laboratory of Water Pollution Control and Environmental Safety, Zhejiang University, Hangzhou 310058, China

* Corresponding author. Tel.: +86-571-88982018; E-mail address: liang410@zju.edu.cn.

Number of pages: 10

Number of Figures: 1

Number of Tables: 5

Table of Contents

Section S1. Information about the Web apps

1) The Web app “FRI” is about the fluorescence classification method of FRI (Chen et al., 2003).

2) The 4 boundary lines: Em = 330, Em = 380, Ex = 250, Ex = Em + 50 (the slope was calculated using SymPy).

3) The Web app “peak-picking” is about the fluorescence classification method of peak-picking (Coble, 1996).

4) The Web app “FRI” was created using PyCharm 2017.3.3, Notepad++ v7.5, Sublime Text, VIM 8.0.586, Flask, mysql-connector, Bokeh, Mako, jinja2, setuptools, Navicat for MySQL Version 12.0.23.

5) The Web app “peak-picking” was created using Anaconda Prompt, Notepad2 4.2.25, UltraEdit 24.20.0.51, conda, Django, psycopg2, Navicat for PostgreSQL 12.0.23.

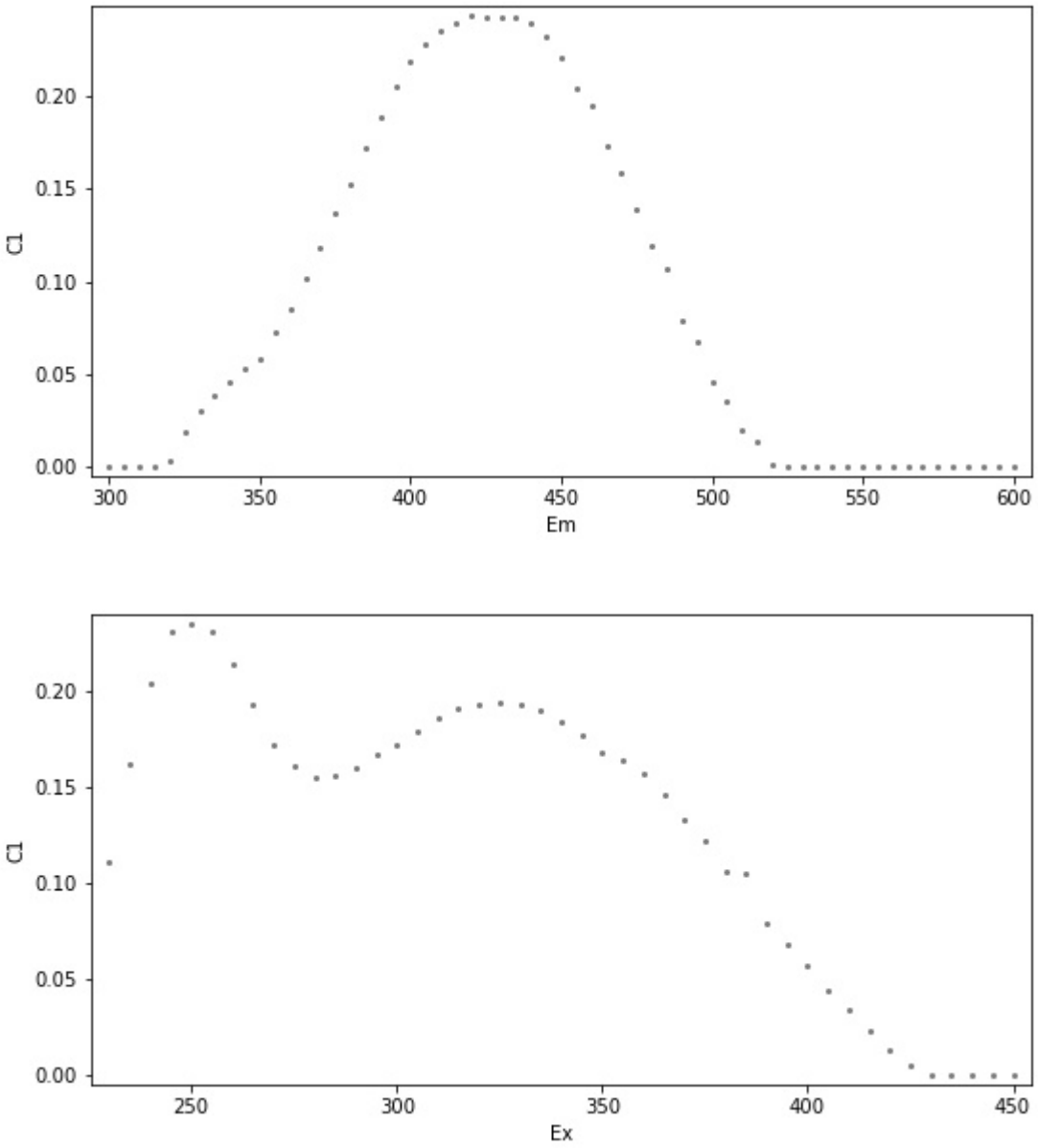

Fig. S1 Fluorescence emission and exciting loadings of PARAFAC C1. The figure was created using Glue, glueviz.

Table S1
Type of gradient analysis methods. The figure was created using WPS Presentation (10.2.0.5996).

	Gradient Analysis Methods		
Response Models	Indirect	Direct	Hybrid
Linear	PCA	RDA	hRDA
Unimodal	CA	CCA	hCCA
Unimodal (detrended)	DCA	DCCA	hDCCA

Table S2

The results of different gradient analysis methods and principal coordinates analysis.

Sum of all eigenvalues	Indirect	Direct	Hybrid
Linear	1.000	1.000	1.000
Unimodal	0.041	0.041	0.041
Unimodal (detrended)	0.041	0.041	0.041
Principal coordinates analysis		23 axes	

Table S3

The results of correspondence analyses of PARAFAC components, DBP FPs and pollution states.

	Principal eigenvalues	Value	Percentage
Simple correspondence analysis	1	0.123798	69.38%
	2	0.031076	17.42%
Multiple correspondence analysis	1	0.173	73%
	2	0.0192	8.12%
Joint correspondence analysis	1	0.267095	NA
	2	0.039613	NA

Simple correspondence analysis was performed using Apache OpenOffice 4.1.5, WPS Spreadsheets(10.2.0.5996), xlrd, xlutils, xlwt, tqdm, ca, conda. Multiple correspondence analysis was performed using openpyxl, xlsxwriter, pyprind, mca, Setuptools.

Table S4

The accuracy results of pollution state predictions using supervised and unsupervised learning.

Method	Accuracy
k-Means	92.86%
Naive Bayes	85.71%
Generalized linear model (xgboost, Python)	87.50%
Generalized linear model (xgboost, R)	87.50%
Logistic regression (CNTK)	100%
Logistic regression (glmnet, R)	98.21%
Generalized linear model (pyglmnet, Python)	87.50%
Rotation forest (rotationForest, R)	100%
Fuzzy c-means clustering (e1071, R)	92.86%
Probabilistic fuzzy c-means (scikit-cmeans, Python)	94.64%
Fuzzy c-means clustering (scikit-fuzzy, Python)	92.86%

Table S5
Linear regression analyses between PARAFAC components and DBP FPs.

	C1	C2	C3	C4
THM FP	w = 0.459, b = 1.028	w = 0.529, b = 0.970	w = 0.139, b = 1.063	w = 0.707, b = 0.964
HAA FP	w = 0.926, b = 0.705	w = 1.139, b = 0.570	w = 0.233, b = 0.639	w = 1.326, b = 0.597
HAN FP	w = 0.035, b = 0.070	w = 0.029, y = 0.069	w = 0.002, b = 0.077	w = 0.049, b = 0.067
CH FP	w = 0.039, b = 0.063	w = 0.040, b = 0.060	w = 0.003, b = 0.070	w = 0.050, b = 0.060

References

Chen, W., Westerhoff, P., Leenheer, J.A., Booksh, K., 2003. Fluorescence Excitation−Emission Matrix Regional Integration to Quantify Spectra for Dissolved Organic Matter. Environ. Sci. Technol. 37, 5701–5710. https://doi.org/10.1021/es034354c

Coble, P.G., 1996. Characterization of marine and terrestrial DOM in seawater using excitation-emission matrix spectroscopy. Mar. Chem. 51, 325–346. https://doi.org/10.1016/0304-4203(95)00062-3