Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Hazardous Materials B108 (2004) 207–211
Journal of Hazardous Materials
www.elsevier.com/locate/jhazmat

Short communication

Predicting organic loading in natural water using spectral fluorescent signatures

Karim Bengraïne, Taha F. Marhaba*

Department of Civil and Environmental Engineering, New Jersey Institute of Technology University Heights, Newark, NJ 07102, USA

Received 27 February 2003; received in revised form 10 December 2003; accepted 12 December 2003

Abstract

Spectral fluorescent signature (SFS) is a rapid, reagent free and inexpensive technique, which has great potential for environmental monitoring of aqueous systems, especially for predicting dissolved organic carbon (DOC) along natural waters. This technical note aimed to examine the possibility to use SFS associated with partial least squares regression (PLS) to assess the organic loading in natural water. A model was built using samples of water collected between October 1999 and February 2002 on the Passaic River at Little Falls, NJ, USA. A correlation was established between measured DOC, SFS, and the corresponding daily registered flow from United States Geological Survey (USGS) New Jersey's streamflow database. The methodology presented herein looks promising in making use of the significant organic characteristics information contained in a SFS for application and use in spatial and temporal water quality management and treatment.

Keywords: Spectrofluorescence signature (SFS); Partial least square regression (PLS); Dissolved organic carbon (DOC); Streamflow; Loading; New Jersey

1. Introduction

Water treatment systems in recent years have had to contend with ever more stringent regulations such as those concerning disinfection by-products (DBPs). Thus, to have a method that could determine target levels quickly would be of great advantage to operating treatment facilities. From that perspective, this work aims to fill the need for an accurate and cost effective technique for surface water monitoring of organic loading using SFS and PLS.

The SFS, also called emission-excitation matrix (EEM), is the total sum of emission spectra of a sample at different excitation wavelengths, recorded as a matrix of fluorescent intensity in coordinate of excitation and emission wavelengths. Multivariate analyses, such as PLS, can be used on this fingerprint of the water sample to find patterns, structures and correlations. Indeed, compared to classical wet chemistry, SFS and PLS diverse applications have been proven effective in rapidity, cost, and performance [1–4]. The combination SFS-PLS was suggested by Marhaba et al. [4] as a surrogate parameter to predict organic loading, chlorine residual, chlorine demand, and DBPs.

In this note, the entire and raw SFS of natural water samples were used to predict organic loading, which is the product of the flow (m^3/s) and the DOC (mg/l). Calibration, full cross-validation, and testing were done on SFSs of monthly samples collected from the Passaic River at Little Falls, NJ and the daily corresponding flow obtained from the USGS–New Jersey's database [5].

2. Materials and methods

2.1. Origin of data and sample treatment

The USGS operates a network of sites throughout the State of New Jersey where streamflow is measured in cubic feet per second. The flow data (m^3/s) used in this study are taken from the USGS-online maintained database [5], which provides both real-time and compulsory data. In addition, between 1998 and 2002 the authors have maintained a database of different parameters for more than 25 stations within the Passaic watershed, as shown in Fig. 1. Different parameters were taken into account—chemical, biological, and physical—that described the watershed's hydrochem-

* Corresponding author. Tel.: +1-973-642-4599; fax: +1-973-596-5970.
E-mail address: Marhaba@njit.edu (T.F. Marhaba).

0304-3894/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2003.12.002

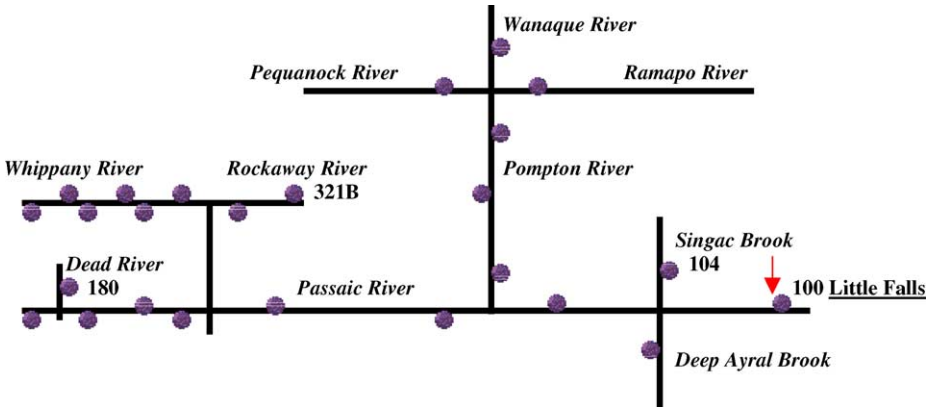

Fig. 1. Localization of the studied station on Passaic River at Little Falls (New Jersey).

istry. Herein, only UV254, Specific ultraviolet absorbance (SUVA), DOC, and the SFSs were used. Station 100, Passaic River at Little Falls, corresponding to the USGS 01388500 gaging-station, was the only one that could be used because SFS and Flow measurements were made exactly at the same location.

Water samples were collected between October 1999 and February 2002. The dataset totaled 17 samples as shown in Table 1. Sample collection, transfer of custody, transportation, and preservation procedures were strictly followed in accordance with the project's data quality objectives. The samples were collected in lot certified quality-assured 250-ml amber glass bottles, labeled with appropriate color and code, and transported the same day to the New Jersey Institute of Technology (NJIT). Samples were stored in a dark cooler room at 4 °C. Prior to any analytical measurements, the samples were filtered through nylon 0.45 μm membranes (Advantec MFS Inc., Pleasanton, CA, USA) within 24 h after sample collection to remove suspended particles that might interfere in both the SFS acquisition and the DOC analyses.

Table 1
Data measured or calculated for NJDEP's Passaic River at Little Falls (sampling station 100 corresponding to USGS 01389500 streamflow gaging station)

Date of sampling	UV254 (nm)	DOC (mg/l)	SUVA (nm l/mg)	Flow (m^3/s)[a]	Load (gs^{-1})
10/6/99	0.085	2.53	0.033	27.92	70.63
12/9/99	0.064	2.74	0.023	20.24	47.36
3/8/00	0.087	2.49	0.035	25.26	61.38
4/5/00	0.088	1.64	0.032	40.78	103.58
12/5/00	0.114	3.79	0.030	6.00	22.74
1/9/01	0.124	3.90	0.031	12.12	35.36
2/6/01	0.115	2.65	0.043	9.65	
3/13/01	0.067	1.50	0.044	46.44	102.63
4/10/01	0.078	2.00	0.057	77.88	194.70
5/8/01	0.077	7.26	0.015	9.51	23.58
7/3/01	0.093	3.65	0.025	13.25	
8/8/01	0.106	7.98	0.021	2.26	8.04
9/11/01	0.075	4.26	0.017	3.56	11.15
10/5/01	0.091	4.71	0.019	N/A	N/A
11/9/01	0.078	4.85	0.016	N/A	N/A
1/5/02	0.109	4.01	0.027	1.19	
2/5/02	0.105	3.66	0.028	0.72	0.72
Mean	0.095	3.74	0.029	21.63	48.60
Minimum	0.064	1.5	0.015	1.04	3.80
Maximum	0.124	7.98	0.057	75.89	78.04

N/A: not available.
[a] USGS–New Jersey online database http://nj.usgs.gov/gen_tbl_pg.

2.2. Analytical methods

The DOC analyses were performed using a Phoenix 9000 carbon analyzer using the method of sodium persulfate oxidation (Standard Methods 5310-D, 1995). The UV measurements at 254 nm were made with a Perkin-Elmer Spectrophotometer.

A Hitachi F4500 fluorescence spectrophotometer (Tokyo, Japan) equipped with 150-W ozone free Xenon lamp was used for the fluorescence measurements. The samples were recorded in a 1-cm quartz cuvette of 4-ml volume sample size and excited from 225 to 399 nm wavelengths. The scan speed was set at 30,000 nm/min and the slit (λ_x–λ_m) at 10/10 nm with a voltage of 700 V.

The working matrix was made of SFSs corresponding to 1950 combinations of λ_x–λ_m per sample. The matrices were exported to the Unscrambler software version 7.6 (Camo A/S, Trondheim, Norway) [6] and transposed in order to have each sample (i.e., SFS) defined as an object (row) and each of the 1950 wavelength combinations λ_x–λ_m defined as a variable (column). By adding the measured DOC, the flow, the loading and its log value, the final matrix used had 17 rows for 1954 columns.

2.3. Modeling

In the SFS-PLS methodology that was fully described elsewhere [4], PLS calibrates a relationship between a dependent variable ($\boldsymbol{Y}$), and an independent variable ($\boldsymbol{X}$) by modeling both $\boldsymbol{X}$ and $\boldsymbol{Y}$ simultaneously to find the latent variables in $\boldsymbol{X}$ that will predict the latent variables in $\boldsymbol{Y}$ the best. This is done using the variance in the $\boldsymbol{Y}$ matrix to de-

compose the SFSs and calculate a model within the error limits.

3. Results and discussion

The daily streamflow data accumulated at the USGS-Little Falls gage between January 1998 and March 2002 are presented in Fig. 2. The measured DOC values in mg/l, the corresponding streamflow in m^3/s, loading in g/s, UV254 in nm, and specific ultraviolet absorbance (SUVA) measured or calculated for samples collected between October 1999 and February 2002 are presented in Table 1. Indeed each sample was associated to an SFS. In Table 2, results—slope, offset, correlation, root mean squared error, squared error, and bias—of linear ($y = ax+b$) regressions between parameters are presented. Flow and loading followed the same evolution with a correlation coefficient of 0.9245. When flow increased loading increased as well. The loadings were quite similar but oppositely correlated to SUVA (0.7649) and DOC (-0.7763). Acknowledging that all correlations involving organic loadings were lacking fitness and accuracy, DOC, which is an aggregate organic parameter that does not provide information of the characteristics of organic matter, was better related to the loading than UV254, which gives information at a specific point of the spectrum (254 nm) where it is unlikely that all the organic substances are sensitive.

If simple correlations with DOC, UV254 and SUVA were lacking robustness and accuracy, it was expected that the large amount of information present in SFSs would be of a great advantage to develop a much more robust model relating to dissolved water organic loadings. Using the matrix of SFSs collected for each sample of the database, it was possible to build a model through PLS, following the methodology described elsewhere [4], to predict the organic loadings. Since the dataset consisted of seventeen measurements taken at different periods between 1999 and 2002, it was difficult to divide it into two subsets that span the same variation. Therefore, full cross-validation was used to validate the calibration model built on the fluorescence data. However, preliminary precautions were taken. The scores of $\boldsymbol{X}$, which was the raw fluorescence matrix for each PLS factor, were used in the initial calibrations to find $\boldsymbol{X}$–$\boldsymbol{Y}$ relational outliers, but none of the samples were discarded. In addition, a Hotelling T-square test was performed to guarantee the absence of outliers. The accuracy or errors in the model were expressed as

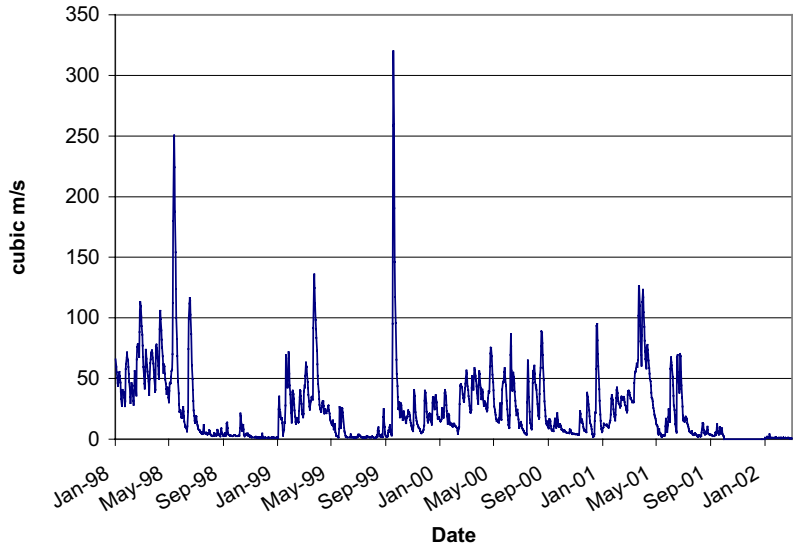

Fig. 2. Flow evolution in cubic meter per second on Passaic River at Little Falls (station 100) corresponding to USGS 01389500 streamflow gaging station.

Table 2
Results of linear correlations ($y = ax + b$) between UV254 (nm), SUVA (nm l/mg), flow (m^3/s), and loading (g/s) sampled on Passaic River at Little Falls (Station 100—USGS 01389500)

	UV254 vs. loading	DOC vs. loading	SUVA vs. loading	Flow vs. loading	Flow vs. UV254	DOC vs. flow	DOC vs. UV254	SUVA vs. FLOW
Elements	15	15	15	15	15	15	15	15
Slope	−593.728	−20.8579	2231.95	1.4342	−0.0003	−15.1982	0.0039	1601.64
Correlation	−0.3478	−0.7763	0.7649	0.9245	−0.4226	−0.8776	0.2737	0.8515
R.M.S.E.	59.2737	57.1269	59.3188	32.5283	29.3160	27.8557	3.5094	29.3499
S.E.	31.9293	32.8538	31.9144	15.0967	20.5845	21.6270	1.2234	20.5673
Bias	50.6148	47.49.80	50.6764	29.0753	−21.5395	18.4227	−3.3026	21.6010

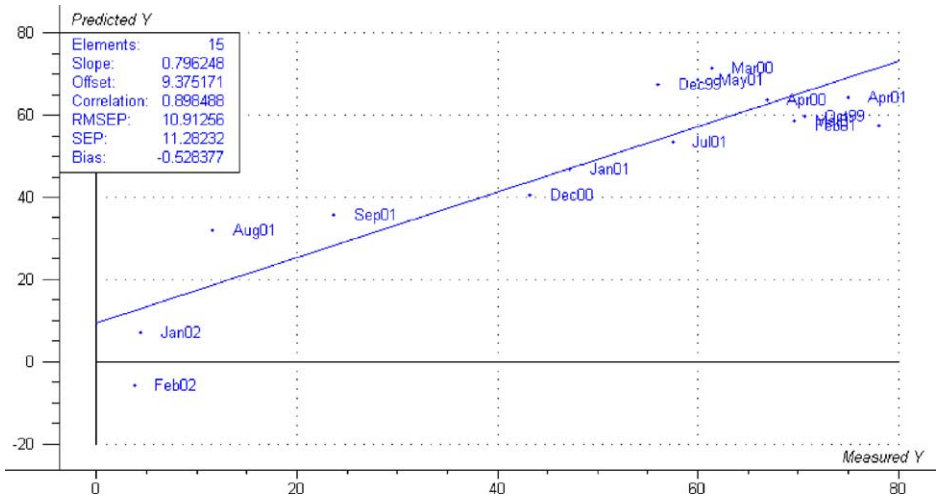

Fig. 3. Prediction of the organic loadings (gs^{-1}) using the samples collected on Passaic River at Little Falls (Station 100—USGS 01389500).

root mean square error of prediction (R.M.S.E.P.), and the bias, which is the systematic difference between predicted and measured values or the average value of the residuals. Robustness was appreciated through the value of the correlation. The PLS modeling results of dissolved organic loading is summarized in Table 2, while the testing curve—measured versus predicted values—is displayed in Fig. 3. Compared to the simple correlations, robustness was ameliorated (from $R = 0.77$ to 0.964) but the main improvement concerned the bias, which was low at the calibration (-1.065×10^{-6}), and -0.528 at the full cross-validation. Further, the comparison between measured and predicted values of loading showed that 45, 39, and 16% of the predicted values fell within 7, 9 and 17% of the measured loading values, respectively. This model had an error of 6.597 at calibration and 10.912 at validation (see Table 3), which is high compared to the range (3.8–78.04) of the reference values, yet five times smaller than the one of the DOC versus loadings (57.13) correlation. The model failed especially for laminar flow corresponding to drought periods (February and January 2002). Given the size and the range of the test set used herein, it is expected that more samples from different locations along Passaic River would improve the SFS-based model performances for predicting organic loading, as long as the sampling is made at the USGS gage station.

The method described in this paper ultimately has practical applications potential for water utilities and other water resources related organizations. Such applications involve rapid prediction of organic matter upstream of a water treatment plant intake as well as through water treatment. This aids in the optimization of undesirable organic matter removal during treatment without time consuming analyses that become impractical to perform in continuously changing water quality conditions. In addition, the method has the potential in being applied for rapid spatial and temporal analyses of watersheds to determine organic loadings from point and non-point sources.

Table 3
Calibration, full cross-validation and testing statistical performances of the PLS model established between spectral fluorescent signatures (SFS) and loading on Passaic River at Little Falls (Station 100—USGS 01389500)

	Calibration	Full cross-validation
Elements	15	15
Slope	0.9223	0.7962
Offset	3.4341	9.3751
Correlation	0.9640	0.8984
R.M.S.E.	6.5970	10.9125
S.E.	6.8285	11.2823
Bias	-1.06×10^{-6}	-0.5283

4. Conclusion

In this short note, it has been shown that spectral fluorescent signatures obtained from water samples at Little Falls New Jersey, associated to USGS streamflow measurements from the same location, could be used to calibrate a model predicting dissolved organic loading. Compared to DOC, UV254, and SUVA, SFS was found to be better correlated to the loading of surface water through a PLS (0.964), with a very low bias (-1.06×10^{-6}). This note showed that the large amount of information contained in an SFS makes it a useful tool to utilize multivariate analysis to correlate data that are organically related such as DOC, loading, and disinfection by-products. Indeed, these parameters are of great interest in improving drinking water quality, and having a unique method able to inform the practitioner on the quality of the product before, during and after treatment, would offer a time and cost saving solution.

Acknowledgements

This work has been funded by New Jersey Department of Environmental Protection under the A-280 Act, and by

New Jersey Institute of Technology. The authors thank Dr. R. Lee Lippincott for his significant contributions and Jaime Arago for maintaining the databases and routine data for calibration analyses.

References

[1] H.C. Goicoechea, A.C. Olivieri, Enhanced synchronus spectrofluorimetric determination of tetracycline in blood serum by chemometric analysis. Comparison of partial least-squares and hybrid linear analysis calibrations, Anal. Chem. 71 (1999) 4361–4368.

[2] D. Baunsgaard, L. Muck, L. Nørgaard, Evaluation of the quality of solid sugar samples by fluorescence spectroscopy and chemometrics, Appl. Spectrosc. 54 (3) (2000) 438–444.

[3] T. Persson, M. Wedborg, Multivariate evaluation of the fluorescence of aquatic organic matter, Anal. Chim. Acta 434 (2001) 179–192.

[4] T.F. Marhaba, K. Bengraine, Y. Pu, J. Arago, Spectral fluorescent signature and partial least squares regression: model to predict dissolved organic carbon in water, J. Hazard. Mater. B97 (2003) 83–97.

[5] United States Geological Survey–New Jersey online database (http://nj.usgs.gov/gen_tbl_pg).

[6] Camo ASA (1998) The Unscrambler 7.6 User Manual. Trondheim, Norway.