ELSEVIER

Journal of Hazardous Materials B97 (2003) 83–97

Journal of Hazardous Materials

www.elsevier.com/locate/jhazmat

Spectral fluorescence signatures and partial least squares regression: model to predict dissolved organic carbon in water

Taha F. Marhaba*, Karim Bengraïne, Yong Pu, Jaime Aragó

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

Received 10 June 2002; received in revised form 15 August 2002; accepted 16 August 2002

Abstract

Spectro-fluorescence signature (SFS) of water samples contains information that may be used to quantify dissolved organic carbon (DOC) if combined with multivariate analyses. A model was built through SFS and partial least squared (PLS) regression. The SFSs of 219 samples of natural water along the Raritan River and Millstone River watersheds located in central New Jersey, and their corresponding DOC concentrations were used to build the model. Calibration, full cross-validation, and prediction performances of various models were statistically compared before optimal model selection. The final selected model, tested on the Passaic River watershed in northern New Jersey, provided a bias of 0.028 mg/l and a root mean squared error of prediction (RMSEP) of 0.35 mg/l. Linked to PLS, SFS can be a quality and cost effective method to perform on-line rapid DOC measurements.

Keywords: Spectrofluorescence signature (SFS); Dissolved organic carbon (DOC); Partial least squared regression (PLS); Watershed; New Jersey

Abbreviations: DOC, dissolved organic carbon; DOM, dissolved organic matter; EEM, emission–excitation matrix; PC, principal components; PCA, principal component analysis; PLS, partial least-squared regression; PCR, principal component regression; PVWC, Passaic valley water commission; mg/l, milligrams/liter; ml, milliliter; MLR, multi-linear regression; NOM, natural organic matter; RMSEC, root mean squared error of calibration; RMSEP, root mean squared error of prediction; SFS, spectrofluorescence signature; μl, micro-liter; USEPA, United States Environmental and Protection Agency; UV254, ultraviolet radiation at 254 nm; SWTR, surface water treatment rule; λ_m, emission wavelength; λ_x, excitation wavelength

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

0304-3894/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0304-3894(02)00246-7

1. Introduction

Natural organic matter (NOM) present in drinking water sources is a complex matrix of organic materials. NOM has autochthonous and allochthonous origins, and is composed of humic and non-humic materials. NOM can be evaluated through surrogate measurements such as total organic carbon (TOC), dissolved organic carbon (DOC), and spectroscopic methods. UV absorbance spectroscopy is a technique widely used to characterize the dissolved organic matter (DOM) in water. Yet, compared to fluorescence spectroscopy, it shows lower selectivity and applicability to a wide range of organic character [1,2].

Spectrofluorescence occurs after excitation of organic materials' fluorophores by a high-energy light source that raises the energy levels of the electrons within the materials. As in other spectrometric methods, Beer's law describes the absorbed energy of any given fluorophore as:

$$A = \varepsilon pc \tag{1}$$

where A is the absorbance, ε the molar absorptivity, p the path length and c is the concentration of the present fluorophores. However, the SFS of a DOM water sample, contains more than one fluorophore, so Beer's law is expanded to a form accounting for all the fluorophores present in organic mixtures:

$$A = p \sum \varepsilon_i c_i \tag{2}$$

where ε_i and c_i are the molar absorptivity and the concentration of the ith fluorophore present in the water sample, respectively. While the energy level returns to ground, fluorescent light is emitted. The quantum yield (Φ) of a given fluorophore is the ratio of the number of quanta emitted as fluorescence by the total number of quanta absorbed. If F is the fluorescence intensity of a fluorophore at a λ_x–λ_m (excitation–emission) wavelength combination, F is defined as:

$$F = \ln(10)\, c\, \mathrm{Io}(\lambda_i)\, p\varepsilon\Phi\, \gamma(\lambda_j) \tag{3}$$

where $\mathrm{Io}(\lambda_i)$ represents the intensity of the excitation light, and $\gamma(\lambda_j)$ the fraction of F at λ_j wavelength. For a specific combination of γ_x–λ_m, $x(\lambda_i)$ and $y(\lambda_j)$, Roch [3] expresses F as:

$$F = cx(\lambda_i)y(\lambda_j) \tag{4}$$

By defining the two vectors, $x = x(\lambda_i)$, and $y = y(\lambda_j)$, F can be expressed in a matrix form as:

$$F = cxy^{\mathrm{T}} \tag{5}$$

For more than one fluorophore,

$$F = \sum F_i = \sum c_i x y_i^{\mathrm{T}} \tag{6}$$

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 coordinates of excitation and emission wavelengths. SFS represents a significant amount of data through a fingerprint of the sample of water. Therefore, multivariate analysis can be used to find patterns, structures and correlations.

There are numerous publications using chemometrics in combination with UV-Vis spectrometry and near infrared (NIR) spectroscopy, but far less with nuclear magnetic resonance (NMR), and spectrofluorescence. Combining UV and principal component regression (PCR), Egan et al. [4] measured carboxyhemoglobin in forensic blood samples. Dahlén et al. [5], correlated nitrate and organic carbon in a PLS2 model with UV spectra to assess ground water quality.

Haaland et al. [6] applied PLS and PCR to NIR spectral data. However, they raised concern over the presence of interfering molecular species that have spectral variance in the calibration data. Swierenga et al. [7] suggested a way to enhance the robustness of NIR calibrated models by proceeding to a robust variable selection. Therefore, instead of modeling the external variation, robust variable selection excludes external spectral variation before modeling. Harmer et al. [8] used ^{1}H NMR spectroscopy and PLS to establish the mathematical relationship between 14 fitted NMR parameters and the properties of a large set of bituminous Australian coals.

SFS and PLS have been recently used by Baunsgaard et al. [9] to evaluate the quality of solid sugar samples, by Goicoechea and Olivieri [10] to determine tetracycline in blood serum, and by Persson and Wedborg [11] to predict the relative percentage of different water masses present in surface water samples from the Baltic Sea or the Skagerrak deep water.

The use of SFS to characterize the DOC has been so far mainly based on a visual comparison of contour and landscape plots. Marhaba et al. [12] used PCA to process and analyze an entire spectral region applying a post-processing technique to identify specific characteristics of the sub-fraction of DOM in water. The sub-fractions are the six different fractions categorized by Leenheer [13] as hydrophobic acid, hydrophobic base, hydrophobic neutral, hydrophilic acid, hydrophilic base, and hydrophilic neutral, isolated and fractionated using resin adsorption chromatography. Marhaba et al. [14] identified regions within a SFS that were characteristic of the six fractions. Furthermore, Marhaba and Pu [15] correlated the SFS to the DOC of each fraction by multilinear regression (MLR). The post-processing technique included intensity of fluorescence, slope and the area of each of the six major peaks. However, when tested on a larger number of samples collected along different watersheds and periods, results were not conclusive and an increase of the bias was observed. The origin of the bias can be explained by, first, the importance of the seasonal pattern that was ignored since the calibration set contained samples from 1 month only (May 1998), and secondly by a possible lack of accuracy and rigor in choosing the major peak representing a given fraction. The risk that the selected combination of wavelengths, and not the entire SFS, used to characterize each fraction does not reflect all relevant DOM fluorophores was a concern.

Consequently, this paper exploits the entire SFS of natural water samples to build a model for direct prediction of the DOC (mg/l) and to finally test it. Calibrations were done using monthly samples from two watersheds in central New Jersey collected during a 1-year period in order to capture the variation in the organic quality. Models were then tested on samples of a third watershed in northern New Jersey. Models' performances were compared at the calibration, validation and prediction steps of the modeling process.

2. Materials and methods

2.1. Samples collection and treatment

Water samples were collected between October 2000 and October 2001 in three major watersheds serving central and northern New Jersey (Fig. 1), following a pre-defined program (Table 1). The sampling collection totalized 377 samples from 41 locations.

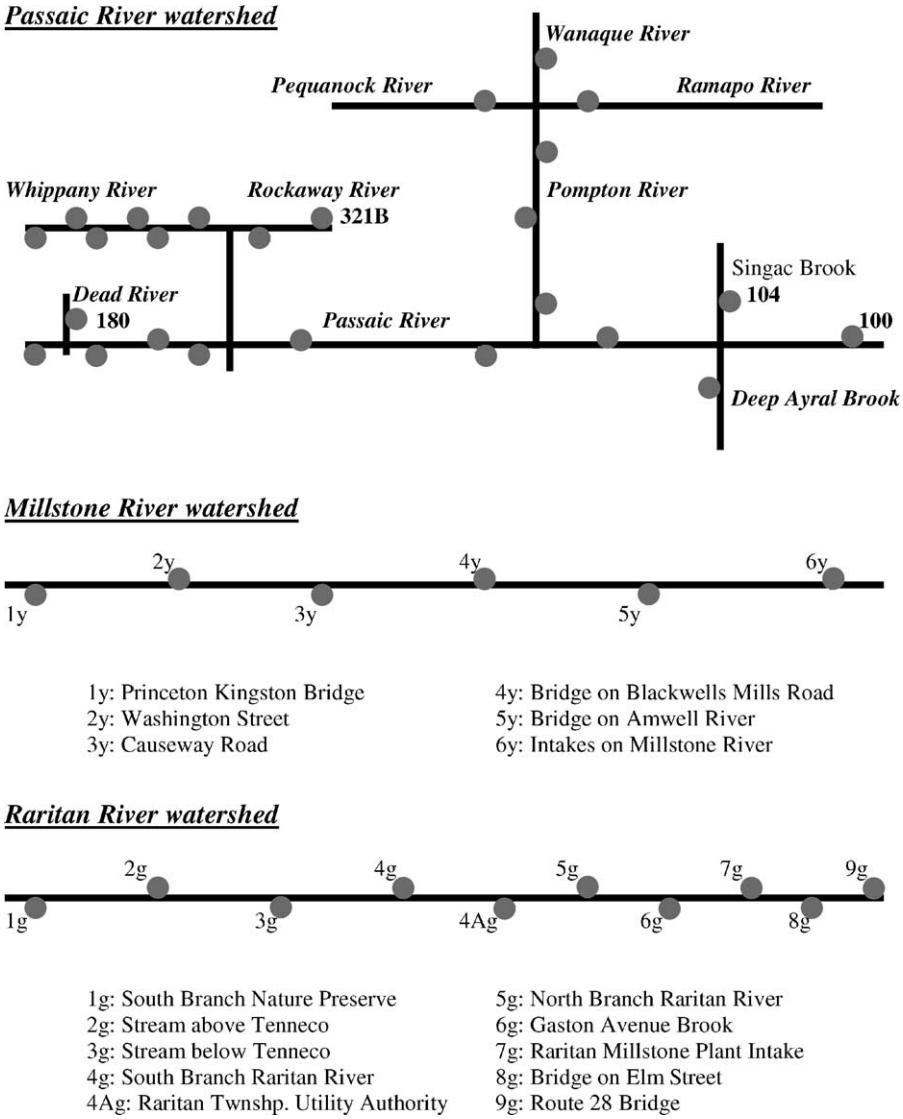

Fig. 1. The sampling locations along the studied watersheds.

Table 1
Data collection schedule on three New Jersey's watershed

	Watersheds		
	Passaic River (red (R))	Millstone River (green (G))	Raritan River (yellow (Y))
9/2000	0	6	9
10/2000	0	6	10
11/2000	0	6	10
12/2000	0	6	10
01/2001	22	6	9
02/2001	0	5	10
03/2001	12	6	10
04/2001	25	6	10
05/2001	10	4	10
06/2001	10	6	10
07/2001	25	6	10
08/2001	12	6	10
09/2001	12	6	10
10/2001	25	6	10
Total	153	81	138

Red, green and yellow are the different color codes.

Fig. 2 shows a schematic of the methodology adopted. Sample collection, transfer of custody, transportation, and preservation were strictly 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 Applied Water Research Center at 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) within 24 h after sample collection to remove suspended particles that might interfere in both the SFS acquisition and the DOC analyses.

2.2. Analytical methods

The DOC analyses were performed using a Phoenix 9000 carbon analyzer using the method of sodium persulfate oxidation (Standard Methods [16]).

The 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. They are 30 cases for each individual SFS; these cases correspond to the emission frequencies that range from 249 to 633 nm with 6 nm sample spacing. 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.

Finally, the working database was made of SFSs, which correspond to 1950 combinations of λ_x–λ_m for each sample. When exported to the Unscrambler software (Camo A/S, Trondheim, Norway [17]) the matrix was transposed in order to have each sample defined as

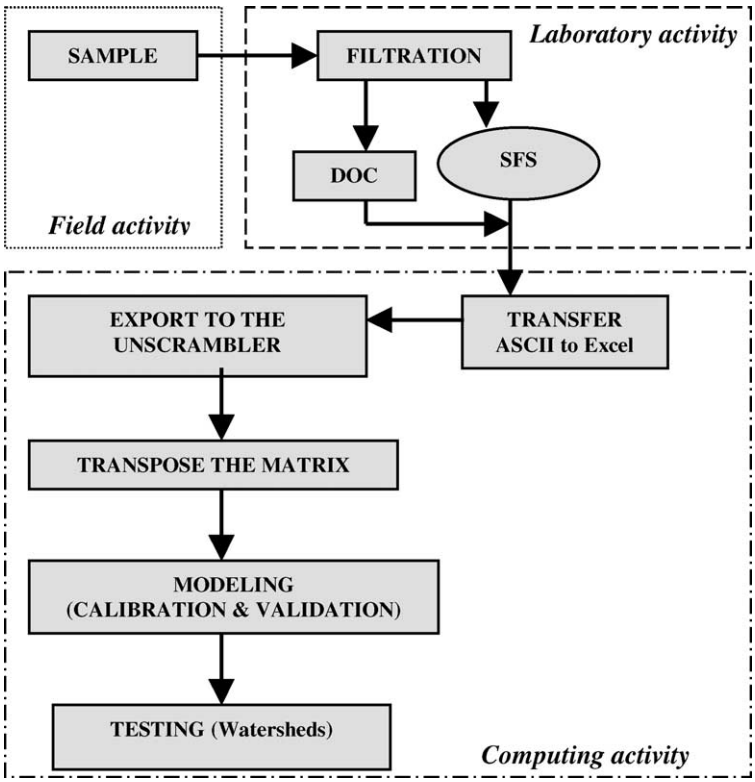

Fig. 2. Chart of the methodology adopted.

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 final matrix used has the dimensions (372 samples $\times$ 1951), which represent 725,772 elements.

The Unscrambler software package version 7.6 (Camo ASA, Trondheim, Norway) was used for all computing analyses.

2.3. Modeling, evaluation and interpretation

PCA belongs to the class of projection methods. This technique reduces the dimensionality of a large dataset of interrelated variables, while retaining as much as possible of the variation present in the data. Geometrically, PCA finds directions in space along which the distance between data points is the largest, which leads to the linear combinations of the initial variables that contribute most to making the samples different from each other. Mathematically, PCA is based on the decomposition of an original data matrix $\boldsymbol{X}$ in series of linear terms and a residual matrix. In a matrix form we have, $\boldsymbol{X} = \boldsymbol{SP}'$, where $\boldsymbol{X}$ is the analyzed data matrix with the dimensions $(i \times j)$, the product $\boldsymbol{SP}'$ is the new linear combinations of the raw data $\boldsymbol{X}$, $\boldsymbol{S}$ the score matrix with dimensions $[i \times \min(i, j)]$, and

$\boldsymbol{P}'$ is the loading matrix with dimensions $[j \times \min(i, j)]$. There are a significant number of principal components (PCs), designed by $\boldsymbol{f}$, equal to the rank of the $\boldsymbol{X}$ matrix that is relevant in describing the information in $\boldsymbol{X}$. The final model built in PCA has the form $\boldsymbol{X} = \boldsymbol{S}_f \boldsymbol{P}_f + \boldsymbol{E}$, where $\boldsymbol{S}_f$ is the score matrix with dimensions $i \times f$, $\boldsymbol{P}_f$ is the loading matrix with dimensions $j \times f$, and E is the residual matrix with X dimensions.

PLS is a multivariate iterative projection method for modeling a relationship between a dependent variable ($\boldsymbol{Y}$), and an independent variable ($\boldsymbol{X}$). PLS models 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. PLS1 deals with only one response at a time, while PLS2 can handle several responses ($\boldsymbol{Y}_i$) simultaneously. A PLS algorithm presents similarities with PCA. PLS uses the variance in the $\boldsymbol{Y}$ matrix to decompose the SFSs and calculate a model within the error limits. PCA and PLS are widespread algorithms for calibration of spectrometer data values and evaluation of unknown measurements spectra. However, the increasing number of data due to techniques such as hyperspectral imaging or SFS, induced the attempt to develop algorithms for faster PCA analysis and PLS regression [18,19]. In the classic algorithm used in the Unscrambler software described in [20], all data are centered as for a full size model. Both variables and samples weights are 1.0 all times. Literature on theory and applications of PCA and PLS is abundant, and should be consulted for thorough explanations [21–23].

After calibration, the software quickly computes a full cross validation. In this iterative process, the same samples are used for both model estimation and testing. One sample is left out from the calibration dataset and the model calibrates on the remaining ones. Then, the value for the left-out sample is predicted and the prediction residual is computed. The process is repeated until every object has been left out and tested. The model was built, calibrated and full-cross validated, using 219 samples from Raritan River and Millstone River watersheds in Central New Jersey, while the prediction, the ultimate stage in multivariate analysis, was done using 153 samples of Passaic River watershed in North New Jersey. None of the samples used at the prediction step are frozen. This precaution is set to incorporate the external influence of environmental parameters such as local point and non-point source of pollution, pH, and temperature, from a location to another as part of the built model. Additionally, all the SFS tri-dimensional points (λ_x–λ_m–I) have the same weight.

The statistical evaluation of the model relies on parameters such as slope, offset, bias, or the average value of the difference between predicted and measured values, given as:

$$\text{bias} = \frac{1}{I}\sum_{i=1}^{l}(\hat{y}_i - y_i) \tag{7}$$

where I represent the number of calibration samples, $\hat{y}_i$ the predicted response and y_i the measured DOC.

The uncertainty of future predictions was estimated with the root mean square error of prediction (RMSEP).

$$\text{RMSEP} = \sqrt{\frac{1}{I_{\text{tot}}}\sum_{s=1}^{N_{\text{seg}}}\frac{1}{y\ \text{weights}^2}\sum_{i=1}^{\text{Is}}\text{Fiys}(i, j)^2} \tag{8}$$

where N_{seg} represents the total number of segment at the cross-validation step. The variance (Eq. (9)) in DOC that is explained by the model was given as:

$$r_{k_1k_2} = \frac{\sum_{i \in s_k} (x_{jk_i} - \bar{x}_{k_i})(x_{jk_2} - \bar{x}_{k_2})}{(I-1)S_x(k_1)S_x(k_2)} \tag{9}$$

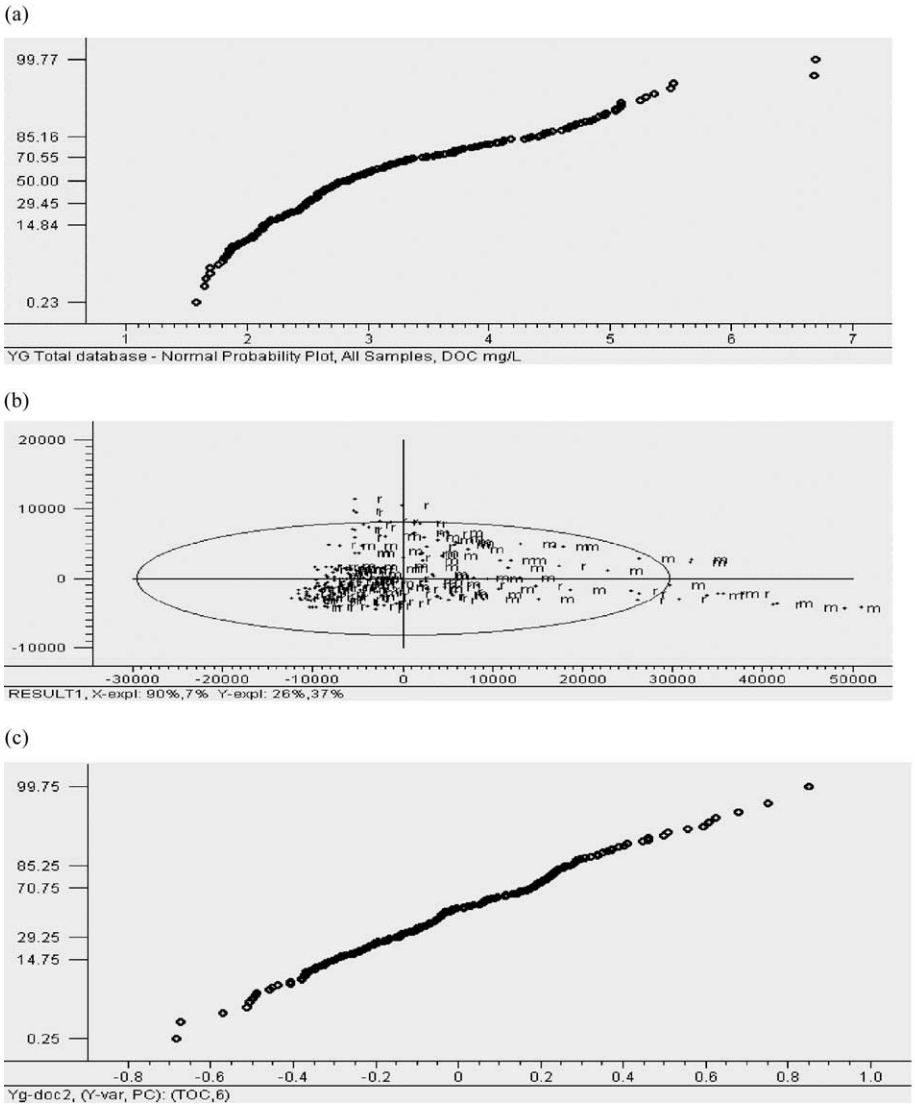

Fig. 3. Distribution of the raw data represent DOC values of Raritan River watershed and Millstone River watershed with and without outliers. (a) All 219 samples: effects (x-axis) vs. normal distribution (y-axis); (b) Hotelling T^2 score plot of the 219 samples: T^2 statistics vs. critical limit ($\alpha = 0.05$ and $F = 3$) (c) Effects (x-axis) vs. normal distribution (y-axis) for Yg-5 model.

All the parameters and algorithms used herein are mathematically expressed in Camo ASA [17].

3. Results and discussion

It is expected from all structured datasets to encounter unrepresentative data in the form of outliers. Therefore, the first step taken herein was to closely look at the data in order to point

Table 2
The conditions of each model tested using Raritan River and Millstone River watersheds samples

	Models								
	Yg-1	Yg-2	Yg-3	Yg-4	Yg-5	Yg-6	Yg-7	Yg-8	Yg-9
Calibration set	218	218	217	212	194	193	193	191	191
PCs calculated	6	6	6	6	6	6	6	6	6
PCs suggested	6	6	6	6	6	4	6	4	6
Total dataset	219	219	219	219	219	219	219	219	219
Samples kept out	1	1	2	7	25	26	26	28	28
Kept out samples list	5g0101	4Ag0801	4Ag0801	8g1200	4Ag1100	4Ag1100	4Ag1100	4Ag1100	4Ag1100
			5g0101	5g0101	4Ag0900	4Ag0900	4Ag0900	4Ag0900	4Ag0900
				4g0301	5g1000	5g1000	5g1000	5g1000	5g1000
				5g0301	2g1200	2g1200	2g1200	2g1200	2g1200
				4g0401	4g1200	4g1200	4g1200	4g1200	4g1200
				5g0401	4Ag1200	4Ag1200	4Ag1200	4Ag1200	4Ag1200
				4Ag0801	5g1200	5g1200	5g1200	5g1200	5g1200
					6g1200	6g1200	6g1200	6g1200	6g1200
					8g1200	8g1200	8g1200	8g1200	8g1200
					9g1200	9g1200	9g1200	9g1200	9g1200
					5g0101	5g0101	5g0101	5g0101	5g0101
					3g0301	3g0301	3g0301	3g0301	3g0301
					3g0401	3g0401	3g0401	5g0301	5g0301
					4Ag0401	4Ag0401	4Ag0401	3g0401	3g0401
					5g0701	5g0701	5g0701	4Ag0401	4Ag0401
					4g0801	4g0801	4g0801	4Ag0501	4Ag0501
					4g0901	4g0901	4g0901	5g0701	5g0701
					1y1200	1y1200	1y1200	4g0801	4g0801
					3y0301	3y0301	3y0301	4g0901	4g0901
					4y0401	4y0401	4y0401	1y1200	1y1200
					5y0401	5y0401	5y0401	3y0301	3y0301
					6y0401	6y0401	6y0401	4y0401	4y0401
					6y0501	1y0601	1y0601	5y0401	5y0401
					4y1001	6y0501	6y0501	6y0401	6y0401
					6y1001	4y1001	4y1001	1y0601	1y0601
						6y1001	6y1001	6y0501	6y0501
								4y1001	4y1001
								6y1001	6y1001

Table 3
Calibration and full cross-validation performances

	Models								
	Yg-1	Yg-2	Yg-3	Yg-4	Yg-5	Yg-6	Yg-7	Yg-8	Yg-9
Calibration									
Slope	0.8174	0.7976	0.8233	0.8985	0.9039	0.8879	0.874	0.8629	0.8625
Offset	0.5634	0.6247	0.5424	0.3087	0.2801	0.331	0.365	0.3946	0.3965
R^2	0.9041	0.8931	0.9073	0.9479	0.9507	0.9159	0.9349	0.9289	0.9287
RMSEC	0.4169	0.439	0.398	0.2939	0.2615	0.324	0.2913	0.2962	0.2968
SEC	0.4178	0.44	0.3989	0.2946	0.2621	0.3247	0.2921	0.297	0.2976
Bias	-8.4×10^{-8}	1.3×10^{-8}	7.6×10^{-8}	5.11×10^{-8}	3.62×10^{-8}	0.0082	7×10^{-8}	3.74×10^{-9}	-7.7×10^{-8}
Validation									
Slope	0.7797	0.7552	0.7815	0.87	0.8698	0.8629	0.8548	0.8879	0.8875
Offset	0.6812	0.7574	0.6715	0.3939	0.3768	0.3946	0.4186	0.331	0.3327
Q^2	0.8647	0.847	0.8679	0.9281	0.9258	0.9289	0.9209	0.9159	0.9156
RMSEP	0.4914	0.5205	0.4712	0.3435	0.319	0.2962	0.32	0.324	0.3247
SEP	0.4926	0.5217	0.4723	0.3443	0.3199	0.297	0.3208	0.3247	0.3254
Bias	0.0012	0.0019	0.0006	−0.0015	−0.0026	3.74×10^{-9}	−0.0018	0.0082	0.0083

Table 4
Results of the prediction DOC mg/l using the 138 samples from Passaic River watershed dataset

	Yg-1	Yg-2	Yg-3	Yg-4	Yg-5	Yg-6	Yg-7	Yg-8	Yg-9
Slope	0.7621	0.7126	0.7415	0.9899	0.9777	0.8728	0.9777	0.9628	0.9621
Offset	0.5788	0.7498	0.6862	0.0171	0.1126	0.08967	0.0994	0.08967	0.0916
P^2	0.6213	0.5713	0.6786	0.8965	0.948	0.9162	0.8949	0.9162	0.9164
RMSEP	1.0949	1.1735	0.9376	0.5181	0.349	0.4502	0.5169	0.4502	0.4493
SEP	1.0512	1.1289	0.8951	0.5194	0.349	0.4488	0.5183	0.4488	0.4478
Bias	−0.3175	−0.333	−0.2879	−0.0207	0.0286	−0.0505	0.01562	−0.0505	−0.0508

those outliers. The normal probability plot in Fig. 3(a) shows that the data could be normally distributed without some outliers. Two samples at the upper right are clearly isolated thus not correlated. Another group of samples affecting the distribution is located at the bottom left part of the graph. Another way to point out non-correlated data is the plot of score vector 1 versus score vector 2 of a PLS applied on all 219 SFSs as presented in Fig. 3(b). This plot confirms the presence of a large group of outliers and the hotelling T^2 ellipse ($F = 3.00$) shows it clearly. Taking a closer look at the identity of the outliers revealed differences between the two watersheds. Spatially the samples are mainly from Millstone River watershed, and show a higher DOC than the average value of 2.1 mg/l. Sampling location 4Ag, the effluent of a sewage treatment plant (STP), has a strong influence on station 5 g which appears to carry the gradient of DOC released in 4Ag. Thus, sampling station 5 g is frequently ousted throughout the sampling season: October and December

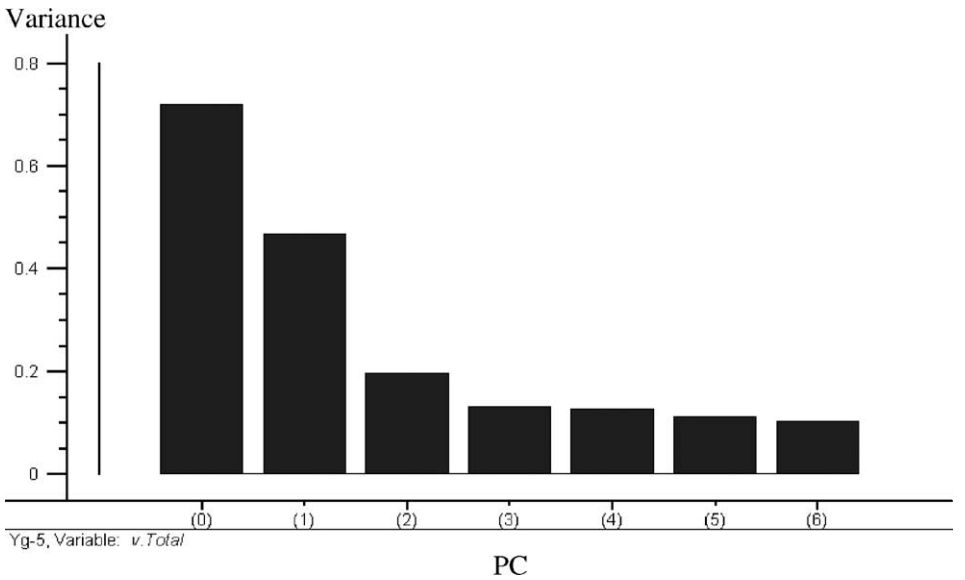

Fig. 4. Evolution of the residual variance with the number of principle components extracted for Yg-5 model.

2000, January, March and July 2001. Temporally, only 4 samples out of 28 are from the dry period of the year—June to September—the majority are related to high level runoff associated with either precipitations or deicing which affect the DOC measurements through higher chloride concentrations. In conclusion, it is important to mention that this first step allowed the isolation of the outliers.

Once the outliers were known, the second step consisted in building different models in which one, two and more outlier samples were removed from the calibration set. This step might concern dozens of scenarios, so it is important to keep track of the sample identity removed as shown in Table 2. Generally, the dataset stays unchanged while part of the spectra that is not related to the measured response is removed. It is the opposite herein as the spectra are unchanged but outliers are progressively removed. Another important difference is the size of the dataset of calibration-validation. Swierenga et al. [7] suggested a technique based on variable selection in order to enhance the robustness of a calibration model using NIR spectra. This technique uses only a subset of spectral values insensitive to the independent variable instead of using the whole spectra.

For each of the nine models it is still important to verify that the distribution of the data is normal as showed in Fig. 3(c) for model Yg-5. Table 3 presents 9 models out the 17 did that showed a variability $R^2 > 0.90$. Since both samples and variables have the same weight, the choice of the best model relies on the comparison of the correlation coefficient R^2, Q^2 (R for the calibration, Q for the validation, and P for the prediction), the root mean squared error of calibration RMSEC, the root mean squared error of prediction RMSEP and the bias. The choice is obviously a compromise between the number of samples kept

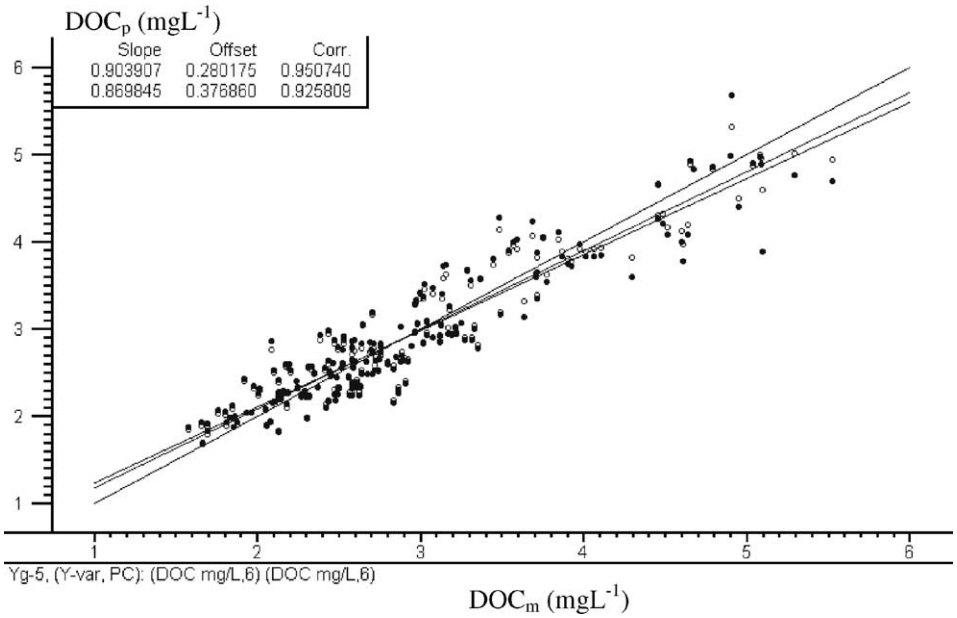

Fig. 5. Measured (DOCm) vs. predicted (DOCp) DOC in mg/l by full cross-validation of Yg-5 model on Millstone River and Raritan River watersheds dataset.

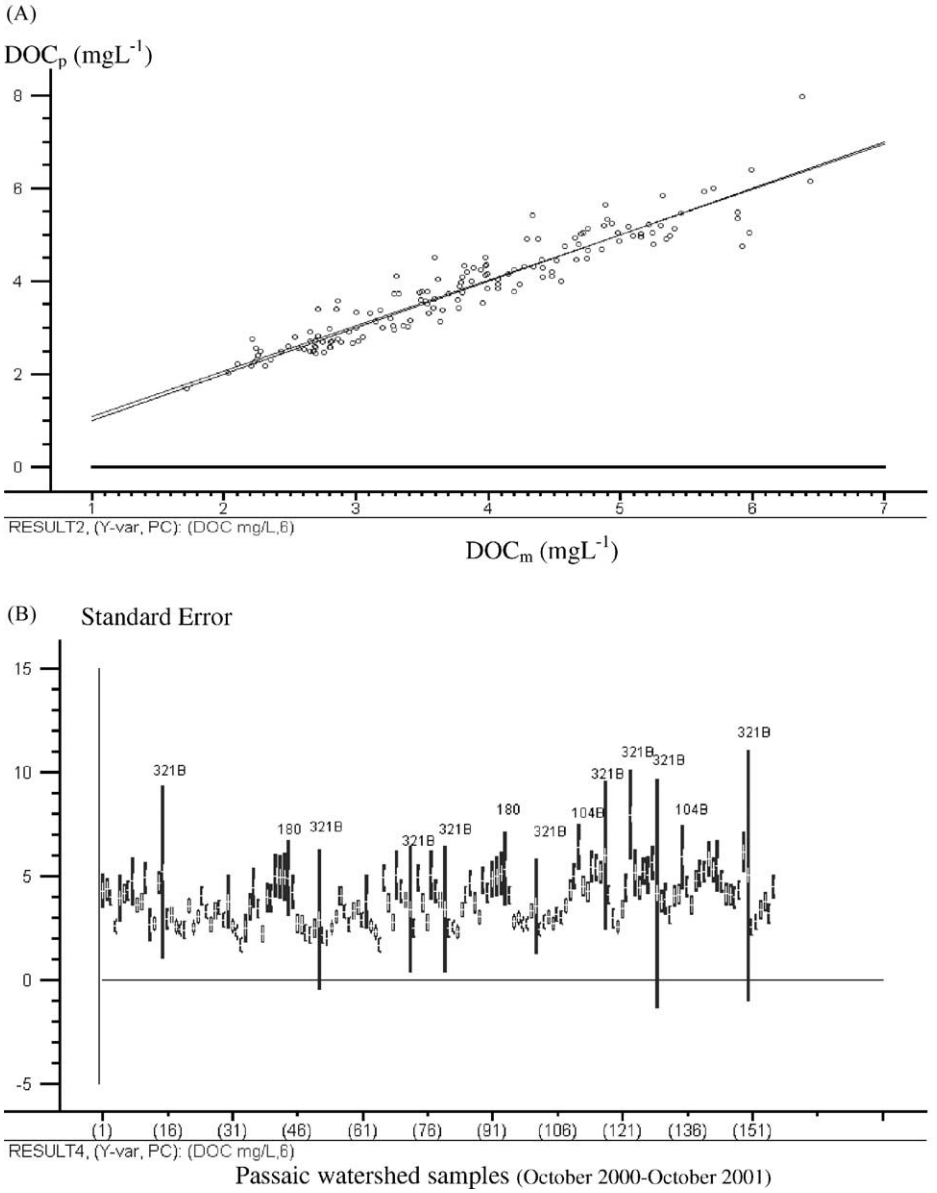

Fig. 6. Prediction of DOC in Mg/l on Passaic River watershed samples. (A) Predicted DOC vs. measured DOC in (mg/l) on Passaic River Watershed using model Yg-5; (B) evolution of the standard error of prediction along Passaic watershed dataset.

out, the goodness of the correlation on the calibration dataset and the fitness and accuracy of the prediction on the validation dataset. This strategy allows minimization of the number of samples to remove from the calibration set while attempting to obtain an image as real as possible of the studied watersheds.

From Table 4, it appears that keeping more samples out or extracting less principal components than suggested, would decrease the goodness of the correlation. Furthermore, minimizing the number of outliers would have an impact on the standard errors of calibration and validation as well as the bias, which has to be the lowest possible.

The models Yg-4 (7 samples out), Yg-5 (25 samples out) and Yg-6 (26 samples out) seem to offer a good compromise especially at the validation with Q^2 ranging between 0.9281 and 0.9289, bias between -0.0015 and 3.74×10^{-9}, and RMSEP between 0.2962 and 0.2962, and SEP ranging between 0.297 and 0.3444. In Fig. 4 the evolution of the residual variance for Model Yg-5, is effective between PC1 and PC4 and tends to be constant from PC4 to PC6. Therefore, and as shown in Table 2, two models (Yg-6 and 8) with less PCs than suggested were tested. It was expected to have a range scale "minima–maxima" style at the final prediction level on Passaic River watershed samples with models Yg-4, 5, and 6.

The testing on the Passaic River watershed samples confirmed that Yg-5 (Fig. 5) provides the best statistics, goodness and fit, while compared to all the models tested as seen in Table 4; a high P^2, a lowest SEP, an intermediate RMSEP, and a bias of 0.0286. The comparison between measured and predicted values of DOC concentrations showed that 31.6, 48.1, and 20% of the predicted values fell within ± 3, ± 7, and ± 12% of the measured DOC values, respectively.

Once the model is built and tested, one must look at the origin of the bias to point out sampling locations that behave differently. Both the plots of the predicted versus the measured DOC in Fig. 6(A) and of the residual variance in Fig. 6(B) for Yg-5 are used to this end. The sampling station 321B on Upper Whippany River is frequently out of the range, and to a certain extent sampling stations 104b on Singac Brook, 180 on Dead River as well. The origin of the bias (2.8%) probably is due to the composition of the DOM in these locations, and the high concentration of chloride ions that may interfere with measured results used to calibrate the model. A mean *t*-test comparison of the measured and predicted DOC, to determine whether the averages of the two sets are significantly different given a 95% confidence interval, confirmed the probability of this result, assuming the null hypothesis, is 0.32 ($t_{\text{ref}} = 0.985$ and a standard deviation of 1.10). Another important parameter to estimate the performance of Yg-5 model was accuracy, which revealed a value of 2%. This result is certainly due to the large amount of sample collected from diversified sources.

4. Conclusion

- This paper presented two of the multivariate regression problems generally encountered, outliers and evaluation of data due to complex datasets. Furthermore, the use of PLS circumvented the selection of emission/excitation wavelengths, making SFS a reliable tool for determining DOC.

- SFS correlated the DOC of surface water using PLS. Prediction of DOC showed a bias (0.286 mg/l), which is lower than the standard deviation of the DOC measurements (1.027 mg/l).
- The SFS-PLS correlation methodology can be adapted to other parameters linked to the organic content in water such as DBPs, chlorophyll-a, and chlorine demand, which may be of high interest to water purveyors.
- SFS-PLS is time (<2 min per sample) and cost effective.
- The methodology based on a realistic compromise between the number of samples kept out of the calibration set, and the statistical performances obtained for different models, guaranteed no loss of reliable spectral information while optimizing the model's goodness and fitness.

Acknowledgements

This work was funded in part by New Jersey Department of Environmental Protection (NJDEP). The authors thank Dr. R. Lee Lippincott of NJDEP Division of Science, Research and Technology for his valuable support, Philip Roosa of Passaic Valley Water Commission and Oleg Kostin of Elizabethtown water company.

References

[1] Y.V. Orlov, I.G. Persiantsev, S.P. Rebrick, S.M. Babichenko, J. Soc. Photo-Opt. Instrum. Eng. 2503 (1995) 150.
[2] T.F. Marhaba, J. Environ. Eng. (2000) 145.
[3] T. Roch T, Anal. Chim. Acta 356 (1997) 61.
[4] W.J. Egan, W.E. Brewer, S.L. Morgan, Appl. Spectrosc. 53 (3) (1999) 218.
[5] J. Dahlén, S. Karlsson, J. Bäckström, J. Hagberg, H. Petterssson, Chemosphere 40 (2000) 71.
[6] D.M. Haaland, L. Han, T.M. Niemczyk, Appl. Spectrosc. 53 (4) (1999) 390.
[7] H. Swierenga, F. Wulfert, O.E. de Noord, A.P. Weijer, A.K. Smilde, L.M.C. Buydens, Anal. Chim. Acta 411 (2000) 121.
[8] J.R. Harmer, T.G. Callcott, M. Maeder, B.E. Smith, Fuel 80 (2000) 1341.
[9] D. Baunsgaard, L. Muck, L. Nørgaard, Appl. Spectrosc. 54 (3) (2000) 438.
[10] H.C. Goicoechea, A.C. Olivieri, Anal. Chem. 71 (1999) 4361.
[11] T. Persson, M. Wedborg, Anal. Chim. Acta 434 (2001) 179.
[12] T.F. Marhaba, R.L. Lippincott, D. Van, Fresenius J. Anal Chem. 366 (2000) 22.
[13] J.A. Leenheer, Environ. Sci. Technol. 15 (5) (1981) 578.
[14] T.F. Marhaba, D. Van, R.L. Lippincott, Wat. Res. 34 (14) (2000) 3543.
[15] T.F. Marhaba, Y. Pu, J. Hazard. Mater. 73 (2000) 221.
[16] Standard methods for the examination of water and wastewater, 19th ed., American Public Health Association, American Water Works Association, Water Pollution Control Federation, 1995.
[17] Camo ASA, The unscrambler 7.6 user manual, Trondheim, 1998, Norway.
[18] F. Vogt, M. Tacke, Chemom. Intell. Lab. Syst. 59 (2001) 1.
[19] B. Walczak, B. van den Bogaert, D.L. Massart, Anal. Chem. 68 (1996) 1742.
[20] H. Martens, T. Næs, Multivariate Calibration, 2nd ed., Wiley, New York, 1991.
[21] K. Esbensen, T. Midtgaard, S. Schönkopf, in: A.S. Camo (Ed.), Multivariate Analysis in Practice, Trondheim, 1994.
[22] P. Geladi, B.R. Kowaslki, Anal. Chim. Acta 185 (1986) 1.
[23] A. Höskuldsson A, J. Chemom. 2 (1988) 211.