Comparison of discrete Fourier transform (DFT) and principal component analysis/DFT as forecasting tools for absorbance time series received by UV-visible probes installed in urban sewer systems

Leonardo Plazas-Nossa and Andrés Torres

ABSTRACT

The objective of this work is to introduce a forecasting method for UV-Vis spectrometry time series that combines principal component analysis (PCA) and discrete Fourier transform (DFT), and to compare the results obtained with those obtained by using DFT. Three time series for three different study sites were used: (i) Salitre wastewater treatment plant (WWTP) in Bogotá; (ii) Gibraltar pumping station in Bogotá; and (iii) San Fernando WWTP in Itagüí (in the south part of Medellín). Each of these time series had an equal number of samples (1051). In general terms, the results obtained are hardly generalizable, as they seem to be highly dependent on specific water system dynamics; however, some trends can be outlined: (i) for UV range, DFT and PCA/DFT forecasting accuracy were almost the same; (ii) for visible range, the PCA/DFT forecasting procedure proposed gives systematically lower forecasting errors and variability than those obtained with the DFT procedure; and (iii) for short forecasting times the PCA/DFT procedure proposed is more suitable than the DFT procedure, according to processing times obtained.

Key words | absorbance, forecast, Fourier transform, principal component analysis, UV-Vis sensor

Leonardo Plazas-Nossa (corresponding author)
Universidad Distrital Francisco José de Caldas,
Facultad de Ingeniería,
Bogotá,
Colombia
E-mail: *lplazasn@udistrital.edu.co*;
plazas-l@javeriana.edu.co

Leonardo Plazas-Nossa
Andrés Torres
Grupo de Investigación Ciencia e Ingeniería del Agua y el Ambiente,
Facultad de Ingeniería,
Pontificia Universidad Javeriana,
Bogotá,
Colombia

INTRODUCTION

Water quality in sewer systems can quickly change due to climatic variables or to wastewater discharges in sewer systems. In recent years, developments in optics and electronics have allowed the mixing or joining of full UV-Vis spectrometry with small scale robust sensors to monitor water quality online and *in situ*. On-line UV-Vis spectrometry can be used to monitor and control water systems (van den Broeke 2007). One of the main applications concerns the monitoring of wastewater treatment plants (WWTPs), especially for the different treatment stages, in order to assess pollutants' loads and the efficiency of organic materials treatments (such as chemical oxygen demand (COD) and biochemical oxygen demand (BOD_5)), nitrates and nitrites, as well as total suspended solids (TSS) (Rieger *et al.* 2004).

In addition to the development of methods needed to calibrate these sensors (Torres & Bertrand-Krajewski 2008), time series analysis is necessary in order to infer pollutants' related absorbance spectra like periodicities and their relationships with wavelengths, for both UV (200 to 400 nm) and visible (400 to 750 nm) ranges (Gruber *et al.* 2005). Time series analysis can be used for forecasting UV-Vis absorbance spectra. There are many methods for forecasting that can be used to predict future events, and which can be divided into two basic types: qualitative and quantitative methods. Quantitative forecast methods can be grouped into two classes: univariate and causal models. Forecasting techniques include regression analysis, time series regressions and Box–Jenkins, ARMA or ARIMA models (Box *et al.* 1994). Even though some experiences related with forecasting of water quality time series exist in the literature (see e.g. Faruk 2010; Yan *et al.* 2010; Halliday *et al.* 2012), to the authors' knowledge fewer examples (see e.g. Plazas-Nossa & Torres 2013) have been reported for forecasting of UV-Vis time series with short time steps (acquisition time in the order of 1 min).

This paper aims to compare forecast results obtained using the discrete Fourier transform (DFT) (Proakis & Manolakis 2007) (Plazas-Nossa & Torres 2013), with those

doi: 10.2166/wst.2014.011

obtained by combining principal components analysis (PCA) (Shlens 2009) with DFT (*i.e.* PCA/DFT).

MATERIALS AND METHODS

spectro::lyser$_{TM}$ UV-Vis sensors are submersible probes of approximately 65 cm length and 44 mm diameter, used to register light attenuation (absorbance) on-line in relative continuous time (one signal per minute), with a light source provided by a Xenon lamp, for wavelengths of 200 to 750 nm, with intervals of 2.5 nm (Langergraber *et al.* 2004). For the present study, time series are composed of 1051 absorbance spectra at each sampling site: (i) for Salitre WWTP effluent (Bogotá D.C.) from 30th June 2011 at 7:16 h to 1st July 2011 at 2:43 h; (ii) for Gibraltar pumping station (GPS), (Bogotá D.C.) from 18th October 2011 at 16:17 h to 19th October 2011 at 9:47 h; and (iii) for San Fernando WWTP effluent, Itagüí (Medellín metropolitan area) from 16th October 2011 at 6:04 h to 17th October 2011 at 17:10 h. The data analysis was done with the general mathematical software R (R Development Core Team 2013).

DFT procedure

Spectral analysis is used to find periodicities included in time series. For this purpose DFT was chosen in order to move from time domain to frequency domain. This technique converts a finite number of equally spaced samples (discrete points) into a number of coefficients of a finite combination of complex sinusoids (components), ordered by their frequencies that have the same number of sample values (Proakis & Manolakis 2007). The proposed methodology consists of sorting the components based on their importance, assessed according to their amplitude. Afterwards, an elimination of components is undertaken from lower to higher importance, to finally return to time domain using IFFT (Inverse Fast Fourier Transform) that converts complex sinusoids (components) into a finite number of discrete points, returning back from the frequency domain to the time domain (Proakis & Manolakis 2007). This is done for all the wavelengths and all the components from the zero-component, which is the signal average in time domain and is the only component systematically taken into account for all of the analysis. The procedure begins removing, for each wavelength time series, values from 1 to 525 (out of 1051 values contained in the original data set) from the original time series to make the forecast procedure with the component values obtained according to the DFT calculation. After this step, only the 10 most important components are used. These are then used as the information to apply the IFFT to return back from the frequency domain to the time domain. The comparison between values of the original time series and the values obtained for the forecasting time series is made by using the normalized root-mean-square deviation (NRMSD) (see Equation (1)). NRMSD is a measurement of the differences between values predicted by the proposed methodology and the values observed, divided by the range of observed values. Figure 1 shows a flowchart explaining the DFT methodology.

In Equation (1), Val_{STi} is the original time series absorbance value for time index i, Val_{IFFTi} is the processed absorbance value for time index i and $Obs_{\max}$–$Obs_{\min}$ is the original time series amplitude range:

$$\mathrm{NRMSD} = \frac{\sqrt{\frac{\sum_{i=1}^{n}\left(Val_{IFFT_i} - Val_{ST_i}\right)^2}{n}}}{Obs_{\max} - Obs_{\min}} * 100 \qquad (1)$$

PCA procedure

PCA methodology aims to construct a linear transformation by choosing a new coordinate system from the original data set, in which the variance of greater size of the data set is captured in the first axis (called the first principal component), the second largest variance is the second axis, and so on. In order to build this linear transformation, a covariance matrix or correlation coefficient matrix must be built first. The goal is to transform an X data set given with dimensions $n \times m$, to another data set of lower dimension $n \times l$, with as little loss of useful information as

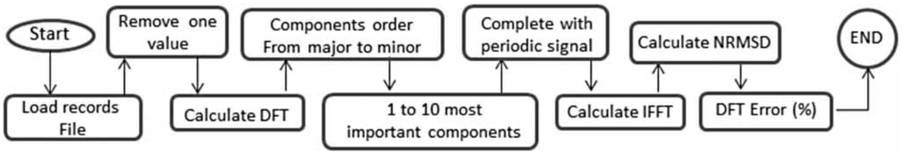
Start
Load records File
Remove one value
Calculate DFT
Components order From major to minor
1 to 10 most important components
Complete with periodic signal
Calculate IFFT
Calculate NRMSD
DFT Error (%)
END

Figure 1 | DFT analysis process flow diagram.

possible using the covariance matrix (Equation (2). Data for analysis must be average zero, which is obtained by subtracting the average of each column from each data, as indicated by Equation (3). However, in order to standardize the data, it is recommended to have the data as auto-scaled values computed as indicated by Equation (4). It is then necessary to obtain eigenvalues and eigenvectors, in order to obtain the principal components as indicated by Equation (5). After the above procedure, it is necessary to reconstruct the data (with original dimensions) by using principal components with Equation (6):

$$\mathrm{COV}(S_{xy}) = \frac{1}{N-1}\sum_{n=1}^{N}\left[\left(x(n)-\bar{X}\right)\left(y(n)-\bar{Y}\right)\right] \quad (2)$$

$$\mathrm{Adjust\_Data} = X_i - \bar{X} \quad i = 1, \ldots, n \quad (3)$$

$$\mathrm{Auto-scale\_Data} = \frac{X_i - \bar{X}}{\sigma} \quad i = 1, \ldots, n \quad (4)$$

$$\mathrm{PC} = \mathrm{eigvector}(\mathrm{Cov}(X))^T \cdot \mathrm{Auto-scale\_Data} \quad (5)$$

$$\mathrm{Data}_{\mathrm{PC}} = \left(\mathrm{Eigen\_Vectors}^T\right)^{-1} \cdot PC + \mathrm{mean} \quad (6)$$

PCA/DFT procedure

After obtaining principal components that correspond to eigenvalues of the covariance matrix over one (1), PCA/DFT consists of applying the DFT procedure explained above to the values obtained with Equation (1) (see flowchart in Figure 2).

Time series forecasting is undertaken for the time series of the main components (PCA) instead of the time series of wavelengths (in order to pass from time series of 219 wavelengths to time series of a few main components – for example 2 and 3). Then, the removal of from one to half of the values of the time series of PCA is done: for each time series obtained, DFT is applied, obtaining a periodic equation used to calculate the deleted values from the original PCA series. Finally, a return to the time domain is performed using IFFT (Proakis & Manolakis 2007).

For each result, the number of PCA and DFT components, the number of records used (N), the number of forecasting values (n) needed to obtain the minimum forecasting error and the processing time are then reported.

All the procedures explained above are undertaken for wavelength specific ranges representing water pollutant parameters commonly used: 13 wavelength groups (Table 1) were identified from van den Broeke (2007) and Thomas & Burgess (2007).

RESULTS AND DISCUSSION

Figure 3 shows a comparison between forecasting results and the original time series in terms of percentage errors for Salitre WWTP: for each wavelength (200–745 nm with steps of 2.5 nm), the lowest (Figure 3 top-left) and the highest (Figure 3 top-right) errors obtained, considering all the models constructed (1–10 DFT components) and all the forecasting times (1–525 min), are shown. It can be observed that both maximum and minimum errors obtained with DFT and PCA/DFT procedures are almost the same for the UV range. However, lower errors are obtained with the PCA/DFT procedure for the visible range of the spectra, in comparison to those obtained with the DFT procedure. In addition, the variability (calculated as the difference between the maximum and the minimum errors obtained for each wavelength) of PCA/DFT errors obtained is also lower than that obtained with the DFT procedure for the visible range of the spectra, as shown in Figure 3 (down). Similar results were obtained for the GPS and San Fernando WWTP study sites.

To generalize the results obtained and to know in which cases it is recommended to use the DFT or PCA/DFT forecasting procedures proposed, a systematic comparison between DFT and PCA/DFT forecasting errors for each wavelength and each study site was undertaken by means of boxplots and t-tests. In order to summarize the results obtained for all the UV-Vis range studied, and for all the study sites, Table 2 shows, for specific wavelength ranges

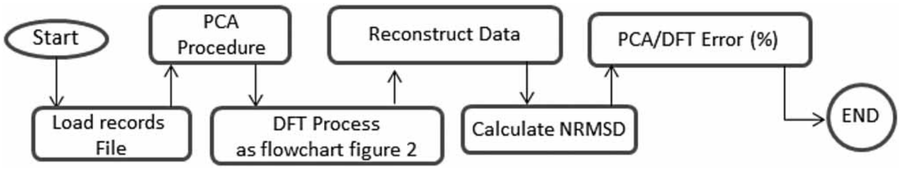

Figure 2 | PCA/DFT procedure flowchart.

Table 1 | UV-Visible ranges subgroups that represent pollutants' range and wavelength

Spectrum	Parameters	Wavelength range (nm)
UV	NO_2 Nitrites and NO_3 Nitrates, Detergents (benzenic forms) at 225 nm	200–250
	COD-1 Acetone 266 nm	252.5–267.5
	Phenols Acetaldehyde 277 nm	270–286
	COD-2 (Phenols), presence of hypochlorite ion 290 nm	287.5–357.5
	Formaldehydes	360–380
VISIBLE	DOC	382.5–427.5
	Violet	430–477.5
	Blue	480–537.5
	Green	540–577.5
	Yellow	580–617.5
	Orange	620–647.5
	Red	650–687.5
	TSS	690–745

DOC: dissolved organic carbon.

between 200 and 745 nm chosen in accordance with Table 1, the percentage of wavelengths for which the DFT forecasting errors are significantly lower, using the p-value of Student's t-test, than the PCA/DFT ones ('DFT' columns in Table 2) and vice versa ('PCA/DFT' columns in Table 2), as well as the percentage of wavelengths for which no significant differences between DFT and PCA/DFT forecasting errors (*p*-value >0.05 from Student's t-test) were obtained ('Indifferent' columns in Table 2). No general results can be deduced from this table for UV wavelength ranges, which means that the forecasting procedure has to be chosen for each water matrix dynamics. However, for visible wavelength ranges, the errors obtained by means of the PCA/DFT forecasting procedure gives systematically lower errors than those obtained with the DFT procedure.

A similar approach as that presented above was used to assess the most convenient method for different forecasting times. Table 3 presents the forecasting time ranges in which each forecasting procedure (DFT or PCA/DFT) gave lower errors than the other one. In this table, it can be observed that results obtained are hardly generalizable, as they seem to be highly dependent of each study site. However, some trends can be outlined: (i) generally, at all three study sites, the PCA/DFT procedure gives lower errors than the DFT one for the visible part of the wavelengths; (ii) for forecasting times under 2 hours, the DFT procedure seems to give lower errors for the UV range and the PCA/DFT procedure seems to give lower errors for the visible range; (iii) according to results obtained for GPS, absorbances in the visible range seem to be easier to forecast for forecasting times over 6 hours, but no recommendation about the procedure can be outlined.

The processing time for each forecasting procedure proposed was taken into account, in order to compare the performance of each one, using the Sys.time() command in R (R Development Core Team 2013) with an Intel® Celeron® Processor B820 ×64 Dual core (2M Cache,

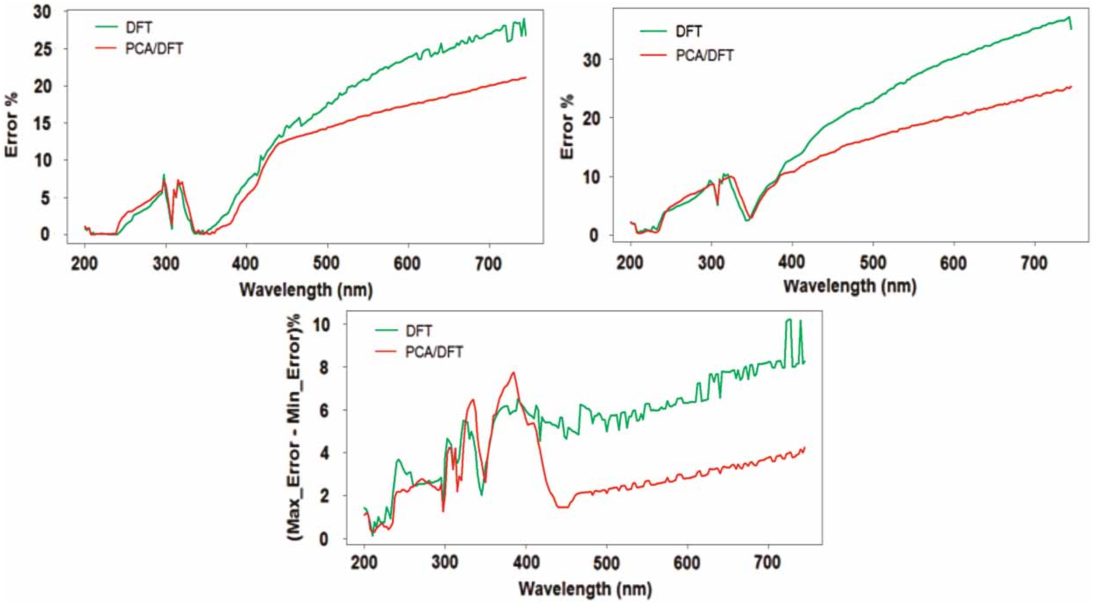

Figure 3 | Forecasting errors (minimum: top-left; maximum: top-right) and error variability (bottom) obtained with DFT (green) and PCA/DFT (red) procedures for each wavelength (lambda). The full colour version of this figure is available online at http://www.iwaponline.com/wst/toc.htm.

Table 2 | Comparison between DFT and PCA/DFT forecasting procedures in terms of the percentage of wavelengths for which corresponding forecasting errors are significantly lower for specific wavelength ranges at the three study sites. 'Indifferent' means that no significant difference was obtained

	Salitre WWTP			GPS			San Fernando WWTP		
Wavelength range (nm)	**DFT**	**PCA/DFT**	**Indifferent**	**DFT**	**PCA/DFT**	**Indifferent**	**DFT**	**PCA/DFT**	**Indifferent**
200–250	9.5	4.8	85.7	85.7	4.8	9.5	100.0	0.0	0.0
252.5–267.5	85.7	0.0	14.3	0.0	0.0	100.0	100.0	0.0	0.0
270–286	0.0	0.0	100	0.0	0.0	100.0	100.0	0.0	0.0
287.5–357.5	31.0	13.8	55.2	0.0	79.3	20.7	100.0	0.0	0.0
360–380	0.0	100.0	0.0	0.0	100.0	0.0	44.4	0.0	55.6
382.5–427.5	0.0	100.0	0.0	0.0	100.0	0.0	0.0	0.0	100.0
430–477.5	0.0	100.0	0.0	0.0	100.0	0.0	0.0	80.0	20.0
480–537.5	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0
540–577.5	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0
580–617.5	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0
620–647.5	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0
650–687.5	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0
690–745	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0

Table 3 | Forecasting time and wavelength ranges for which each forecasting procedure (DFT or PCA/DFT) gives more accurate results. In grey or no UV or Vis label: no significant difference between procedures

<table>
<tr><th>Forecasting time (min)</th><th>Salitre WWTP</th><th>GPS</th><th>San Fernando WWTP</th></tr>
<tr><td>1–30</td><td></td><td rowspan="4">DFT (UV)
PCA/DFT (Vis)</td><td rowspan="3">DFT (UV)
PCA/DFT (Vis)</td></tr>
<tr><td>30–120</td><td>DFT (UV)
PCA/DFT (Vis)</td></tr>
<tr><td>120–150</td><td rowspan="3">PCA/DFT (UV-Vis)</td></tr>
<tr><td>150–180</td><td rowspan="3">DFT (UV-Vis)</td></tr>
<tr><td>180–350</td><td rowspan="2">PCA/DFT (Vis)</td></tr>
<tr><td>350–390</td><td rowspan="3"></td></tr>
<tr><td>390–420</td><td>DFT (Vis)</td><td rowspan="4">PCA/DFT (UV-Vis)</td></tr>
<tr><td>420–450</td><td rowspan="2">PCA/DFT (Vis)</td></tr>
<tr><td>450–455</td><td rowspan="2">DFT (UV)
PCA/DFT (Vis)</td></tr>
<tr><td>455–525</td><td>DFT (Vis)</td></tr>
</table>

1.70 GHz), 4GB-RAM personal computer: (i) for both the DFT and PCA/DFT procedures, the processing time for one component, one wavelength and forecasting times between 1 and 15 mins is one second (forecasting times under 15 min give processing times less than one second, which is the minimum resolution value given); (ii) for both the DFT and PCA/DFT procedures, the processing time for one component, one wavelength and forecasting times between 1 and 525 min, is 36 s; (iii) the total DFT processing time (one component, 219 wavelengths and forecasting times between 1 and 525 min) was about 3.35 hours (for 10 components the processing time was about 33.5 hours); (iv) the DFT processing time (for 10 components and 219 wavelengths) for a forecasting time of one minute was about 3.8 min, which is much higher than the sampling time (1 min); (v) the total PCA/DFT processing time (one component, three PCs, and forecasting times between 1 and 525 min) was about 108 s (for 10 components the processing time was about 18 min); (vi) the PCA/DFT processing time (for 10 components and three PCs) for a forecasting time from 1 to 292 min was about 1 min, which represents a forecasting time much higher (approximately 4.9 hours) than the sampling time (1 min).

CONCLUSIONS

In general terms, and if no specific forecasting times are required, it can be concluded that the DFT and PCA/DFT procedures proposed give similar results for the UV range, which implies that the forecasting procedure has to be chosen for each water matrix dynamic if the target pollutants belong to the UV range (e.g. nitrites, nitrates, COD, etc.). However, for the visible wavelength range (e.g. DOC, TSS, etc.), the PCA/DFT forecasting procedure proposed gives systematically lower forecasting errors and variability than those obtained with the DFT procedure.

It seems important to highlight that although water quality and pollutant concentrations are different for the three study sites and the absorbance time series were not taken simultaneously, the results obtained indicate that for the UV range, the DFT procedure proposed shows generally lower errors than those obtained using the PCA/DFT procedure. In contrast, for the three study sites the results obtained from the PCA/DFT procedure show lower errors for the visible range than those obtained using the DFT procedure.

On the other hand, for forecasting times under 2 hours, the DFT procedure seems to give lower errors for the UV range but the PCA/DFT procedure seems to give lower errors for the visible range. In addition, absorbances in the visible range seem to be easier to forecast for forecasting times over 6 hours, but no recommendation about the procedure can be outlined. Therefore, the results obtained are hardly generalizable, as they seem to be highly dependent on each study site, which implies that the choice of the forecasting procedure (DFT or PCA/DFT) has to be part of the whole forecasting system, applied for specific water system dynamics.

The performance of the PCA/DFT procedure in terms of processing times was found to be much better than that of the DFT procedure. In fact, the PCA/DFT procedure provides values of forecasting times much higher than the sampling time (number of minutes of forecasting between two sampling values), whereas DFT uses values of processing time much higher than the sampling time to obtain one forecasting value. This result implies that, even if for some UV-Vis ranges the DFT procedure gives more accurate forecasting results than the PCA/DFT one, for short forecasting times the PCA/DFT procedure proposed is more suitable than the DFT procedure.

This research will continue for longer time series (weeks, months, years) and using different forecasting techniques (spectral estimation, linear and nonlinear autoregressive methods, etc.) that will allow a more thorough understanding of the studied phenomena and which can be used to construct decision support tools leading to the optimization of urban water systems operation.

ACKNOWLEDGEMENTS

The authors acknowledge the Bogota Water and Sewage Company (Empresa de Acueducto y Alcantarillado de Bogotá – EAAB) (under Administrative Contract No. 9-07-25100-0763-2010) and Medellín Water and Sewage Company (Empresas Públicas de Medellín – EPM) for providing the information used in this research.

REFERENCES

Box, G., Jenkins, G. & Reinsel, G. 1994 *Time Series Analysis: Forecasting and Control*. 3rd edn. Prentice Hall, Englewood Cliffs.

Faruk, D. Ö. 2010 A hybrid neural network and ARIMA model for water quality time series prediction. *Engineering Applications of Artificial Intelligence* **23** (4), 586–594.

Gruber, G., Bertrand-Krajewski, J.-L., De Beneditis, J., Hochedlinger, M. & Lettl, W. 2005 Practical aspects, experiences and strategies by using UV/VIS sensors for long-term sewer monitoring. In: *Proceedings of 10th International Conference on Urban Drainage*, Copenhagen, Denmark, 21–26 August 2005. pp. 1–8.

Halliday, S. J., Wade, A. J., Skeffington, R. A., Neal, C., Reynolds, B., Rowland, P., Neal, M. & Norris, D. 2012 An analysis of long-term trends, seasonality and short-term dynamics in water quality data from Plynlimon, Wales. *Science of The Total Environment* **434**, 186–200.

Langergraber, G., Fleischmann, N., Hofstaedter, F. & Weingartner, A. 2004 Monitoring of a paper mill wastewater treatment plant using UV/VIS spectroscopy. *IWA Water Science and Technology* **49** (1), 9–14.

Plazas-Nossa, L. & Torres, A. 2013 Fourier analysis as a forecasting tool for absorbance time series received by UV-Vis probes installed on urban sewer systems, Novatech 2013.

Proakis, J. & Manolakis, D. 2007 *Digital Signal Processing Principles, Algorithms, and Applications*. 4th edn. Pearson Prentice Hall, New Jersey.

R Development Core Team R 2013 *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna (Austria). http://www.R-project.org/.

Rieger, L., Langergraber, G., Thomann, M., Fleischmann, N. & Siegrist, H. 2004 Spectral in-situ analysis of NO2, NO3, COD, DOC and TSS in the effluent of a WWTP. *AutMoNet 2004 – Proceedings of the International Conference on Automation in Water Quality Monitoring*, Vienna, Austria, 19–20 April 2004. pp. 29–36.

Shlens, J. 2009 *A Tutorial on Principal Component Analysis*. Center for Neural Science and Systems Neurobiology Laboratory, Salk Insitute for Biological Studies, pp. 1–12.

Thomas, O. & Burgess, C. 2007 *UV-Visible Spectrophotometry of Water and Wastewater*. Elsevier B.V, Amsterdam.

Torres, A. & Bertrand-Krajewski, J.-L. 2008 Partial Least Squares local calibration of a UV–visible spectrometer used for in situ measurements of COD and TSS concentrations in urban drainage systems. *Water Science and Technology* **57** (4), 581–588.

van den Broeke, J. 2007 On-line and In-situ UV/Vis Spectroscopy: Real Time Multi Parameter Measurements with a Single Instrument. *AWE International* **10** (March 2007), 54–57.

Yan, H., Zou, Z. & Wang, H. 2010 Adaptive neuro fuzzy inference system for classification of water quality status. *Journal of Environmental Sciences* **22** (12), 1891–1896.

First received 1 July 2013; accepted in revised form 3 January 2014. Available online 21 January 2014