Proposal For Recurrence, Level Of Importance And Quality Detection Of Uv-Vis Spectra And Target Pollutant Dataset – An Omar Rayo's esthetics inspiration for UV-Vis graphical representation –

David Zamora
Grupo de Investigación Ciencia e Ingeniería del Agua y el Ambiente, Facultad de Ingeniería, Pontificia Universidad Javeriana, Carrera 7 No. 40 – 62, Bogotá, Colombia, david.zamora@javeriana.edu.co

Andrés Torres
Grupo de Investigación Ciencia e Ingeniería del Agua y el Ambiente, Facultad de Ingeniería, Pontificia Universidad Javeriana, Carrera 7 No. 40 – 62, Bogotá, Colombia, andres.torres@javeriana.edu.co

ABSTRACT

In situ UV-Visible spectrometric probes offers the opportunity to follow the behavior of water pollutants through the use of mathematics models. Hence, the results of these models depend on the quality of the input data and its significance will be framed by how representative they are of the water quality of the system: absorbance spectrum and pollutants concentrations obtained from standard laboratory analysis. Therefore, a new method is being developed and presented in this paper that enables to detect, in a multivariate way, the wavelengths most closely related to the target pollutants, and hence that can represent in a better way the interactions of the water system with the beam in the spectroscopic process, carried out by the measurement instrument. In the case study of this research, absorbance spectra and TSS, COD, and filtered COD concentrations of grab samples obtained in the affluent and effluent of San Fernando WWTP (Medellín, Colombia) were used. These results allow identifying the quality of the data beforehand and establishing the low predictability of some regressive models.

Keywords: Wavelengths in UV-Visible Spectra, Pollutants, Affinity, Data Quality.

1. INTRODUCTION

Information is the raw material to meet the simplicity or complexity of a process in the different fields of knowledge. Therefore, quantify and qualify the water quality is a task that has allowed visualize roughly the dynamics of the different water systems and try to understand the interaction of water with both natural and artificial systems. Then, in the specific case of wastewater treatment plants (WWTP) where the matrix of both influent and effluent are highly complex and with a high time-space variability, high frequency representative information of the dynamics related to pollutant flows will be required, even more when the plant influent correspond to a combined sewer system and on which waste waters of industrial processes generate large fluctuations characterized by volumetric flow rates, temperatures and pollutant loads (Hoppe *et al*., 2009).

Against this requirement, measuring instruments *in situ* and continuously become a metrology tool that allows a near-real-time observation of the pollutants dynamics: UV-Visible spectrometry probes help track the fingerprint and quantify, as well as qualify, the presence of contaminants of interest through chemometric models. These models formalize the process of correlating the physical, chemical or biological properties present in the spectra. In order to do this, spectra are measured by establishing a training set of samples of those values that are wanted to have several properties (from standard or reference values obtained in the laboratory) which are known to have been measured by conventional means (DiFoggio, 2000). Taking into account this information and using

regression functions, a calibrated equation is created which will determine and quantify the properties of the sample to be represented across the spectrum. Finally, the value of property can be predicted in an unknown sample by assessing the same spectrum using the mathematical calibrated function.

Therefore, the results of these models depend on the intrinsic quality of the input data and its significance will be framed by the representativeness of the water quality of the system: absorbance spectra and concentrations of contaminants derived from laboratory analysis. But it is a narrow scope to use the spectral richness only in the estimation of a set of simple or compound pollutants (NO_2 or Chemical Oxygen Demand respectively), due to the fact that, with this real-time information, alert and alarm states can be generated (*e.g.* Gruber *et al*., 2004; Langergraber *et al*., 2004; Rieger *et al*., 2008)), as well as control rules and why not the prediction of the digital footprint and thus the pollutants matrix in the near future.

In order to develop these activities using the spectral information it is necessary to know which wavelengths in the UV-Visible have a significant relationship with the property or pollutant of interest over time and which affinity will allow model results to be accurate.

Due to the fact that the wastewater monitoring has to deal with a matrix of many dissolved and suspended compounds, the superposition of many substances in individual absorbances ─even at times with superimposed peaks─ may cause cross sensitivity and lead to poor performance of the sensor (Langergraber *et al*., 2003). Normally, in order to establish such relationship, a correlation coefficient R^2 is used, assuming linearity between pollutant concentration and the absorbance values. However, some direct chemometric models may be used only if the spectra of all the components are known and it can be said that the measured spectrum will be the sum of the hypothetical individual spectra of each of the sample constituents, making the Lambert-Beer's law valid, which is not satisfied in the case of waste water, where a large number of unknown compounds are present (Langergraber *et al*., 2003; Escalas *et al*., 2003; Vargas and Buitrón, 2006). In addition, by using linear correlations, many local solutions with similar performances can be found, making difficult to identify only one wavelength with the highest relevance (Lorenz *et al*., 2002). In the present study a method to detect in a multivariate way the wavelengths most closely related with the pollutant concentrations and to assess the quality of the data, taking into account issues such as cross-sensitivity (sensitivity to a substance that predisposes the sample to be sensitive to other substances listed by their chemical structure (Fleischmann *et al*., 2001)) and non-linearity between absorbance values and pollutant concentrations was developed.

2. MATERIALS AND METHODS

2.1 METHOD OF SELECTION OF THE WAVELENGTHS MOST CLOSELY CORRELATED WITH THE PARAMETER ANALYZED – ZATO

The selection of the wavelengths most closely related with a target pollutant is crucial for assessment of equivalent concentrations based on fingerprints. Statistical regression models allow to transform the absorbances of multiple wavelengths in equivalent concentration values when carrying out a local calibration, which is able to determine the specific wastewater composition and the possible effects that they generate (Gruber *et al*., 2006). In the present study, a method to detect wavelengths related to pollutant concentrations was developed. The method, called ZATO, considers five bivariate functions in order to calibrate regressive models: linear, polynomial of second and third degree, logarithmic and power (see Eq. 1 to 5). The variable to be estimated is the concentration of the target pollutant based on absorbance values for a particular wavelength of the spectra. Therefore, it would generate the same number of models for each regression function as the number of wavelengths with absorbance values.

$$\hat{y}_i = \sum_{k=1}^{1000} \left[A_i + B_i \cdot x(\lambda_{[200-750nm]}) \right]$$

Equation 1: Linear model

$$\hat{y}_i = \sum_{k=1}^{1000} \left[A_i + B_i \cdot x(\lambda_{[200-750nm]}) + B_i \cdot x^2(\lambda_{[200-750nm]}) \right]$$

Equation 2: Polynomial of second degree

$$\hat{y}_i = \sum_{k=1}^{1000} \left[A_i + B_i \cdot x(\lambda_{[200-750nm]}) + C_i \cdot x^2(\lambda_{[200-750nm]}) + D_i \cdot x^3(\lambda_{[200-750nm]}) \right]$$

Equation 3: Polynomial of third degree

$$\hat{y}_i = \sum_{k=1}^{1000} \left[A_i + B_i \cdot \log(x(\lambda_{[200-750nm]})) \right]$$

Equation 4: Logarithmic model

$$\hat{y}_i = \sum_{k=1}^{1000} \left[E_i \cdot x^{F_i}(\lambda_{[200-750nm]}) \right]$$

Equation 5: Power model

In Equations 1 to 5, A_i, B_i, C_i, D_i, E_i and F_i are the coefficients and exponent respectively, calibrated in each of the k executions, x_i are the absorbances of each wavelength $\lambda_{[200-750\ nm]}$ of the i analized samples and $\hat{y}_i$ are the set of equivalent concentrations of a contaminant estimated by each model in each execution.

Then, sum of squared errors (SSE) between reference laboratory concentrations and equivalent concentrations obtained from each model are computed according to Eq. [6].

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i\lambda} \right)^2$$

Equation 6: Sum of Squared Errors

After having SSE values for each regressive model, each wavelength is ordered based on the following criterion: for a particular regressive model, wavelengths with low SSE values are considered more relevant than SSE with higher ones. Based on the above procedure used to calculate the prediction errors, an Importance Factor (IF) is proposed, which explains the affinity between the wavelengths and the target pollutant. In order to avoid bias because of the presence of possible outliers in the data, a random selection of a fraction (e.g. 67 % of the data) of the dataset n number of times is proposed. With this method, the recurrence of a wavelength to occupy a certain position out of 220 possible positions, based on IF indicator, can be computed. This recurrence, analyzed graphically for now, is used for two purposes: (i) to establish the level of importance of a wavelength on the target pollutant; (ii) to assess the quality of the data, since a greater dispersion of recurrences at different levels of importance of many wavelengths can confirm the low affinity between laboratory concentrations obtained and the UV-Vis spectra related. In fact, these possible benefits were confirmed numerically (not shown in this paper) for both selection of wavelengths to construct chemometric models and outliers detection.

2.2 ARTISTIC SPECTRAL REPRESENTATION

Finally, the design of the graphs for the results representation is inspired on the work of Colombian geometric-optic artist Omar Rayo (1928-2010). The method of spectral analysis and graphs that represent it were developed and programmed into the platform *R* (R Development Core Team, 2012).

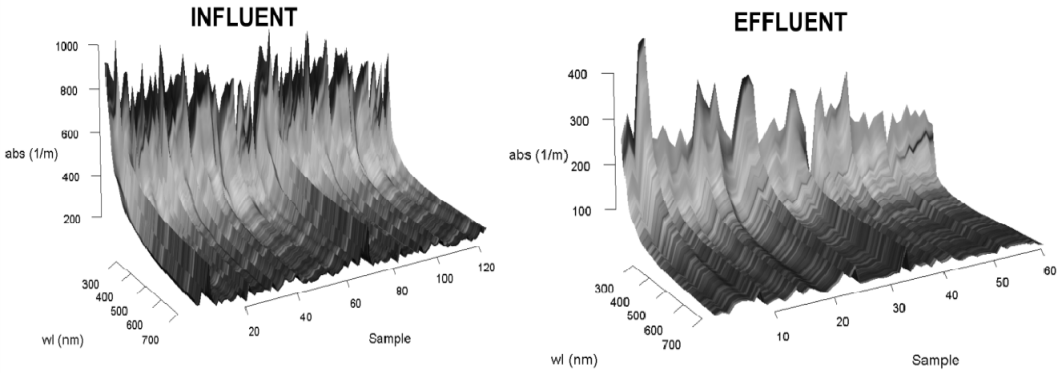

Figure 1: Absorbance spectra for influent (left) and effluent (right.) of San Fernando WWTP

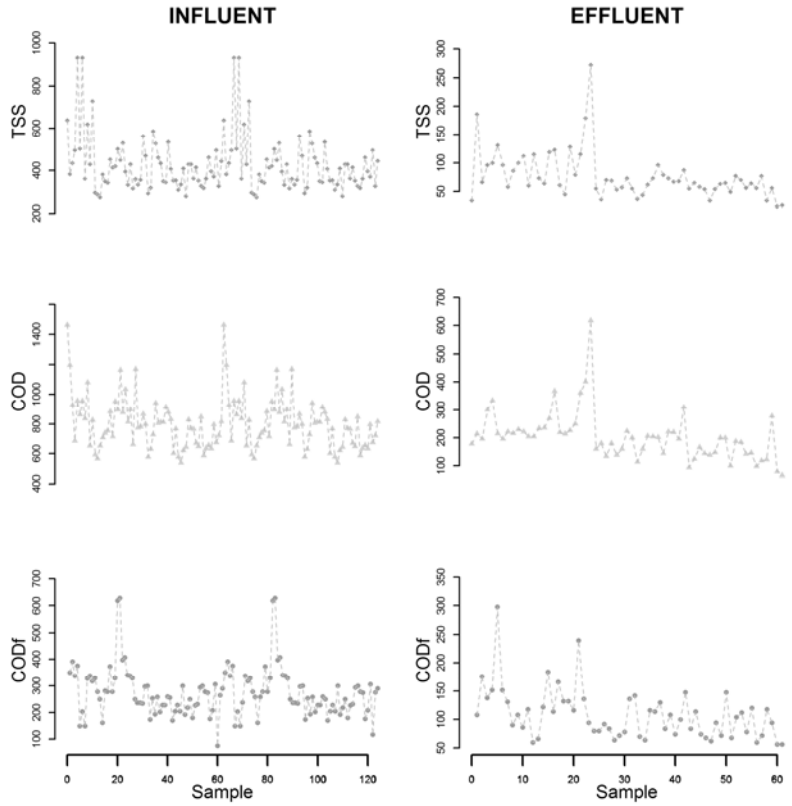

Figure 2: TSS, COD and CODf concentrations for influent (left) and effluent (right.) of San Fernando WWTP

3. CASE STUDY

In the case study of this research, absorbance spectra and Chemical Oxygen Demand (COD), filtered COD and Total Suspended Solids (TSS) concentrations of grab samples obtained in the affluent and effluent of San Fernando wastewater treatment plant (WWTP) (Medellín, Colombia) were analyzed. These samples were originally taken in order to achieve a local calibration of the probes *spectro::lyser* used in influent and effluent of the WWTP.

In Figures 1 and 2 the concentration data and spectra is shown, these last two probes measured with *spectro::lyser* of different pathlengths: 2 mm for influent and 5 mm for effluent.

4. RESULTS AND DISCUSSION

Figures 3 to 8 were obtained using ZATO method. From these figures, it can be observed that the wavelengths belonging the UV-visible spectrum range are located in x-axis, whereas in y-axis it can be observed the level of importance of the relationship of the contaminant with each of the wavelengths of the spectrum. Finally, the color bar on the right of the figures allows to visualize the number of times where in the n Monte Carlo simulations, fewer errors are generated in the prediction of a pollutant concentration. Therefore, these last two elements determine the affinity, defining this as the ability of the specific properties of the absorbance at multiple wavelengths to be related to the presence of a contaminant in a water matrix.

In the case of the TSS from the affluent (Fig. 3), there is a very close range of wavelengths (725 nm – 750 nm) where the recurrence of 1000 random executions (selection of 67 % of the input data) is over 800. In this figure, the "dust" data represent the degree of affinity of some of the couples spectra-concentration which must be detected and defined as outliers. On the other side, in the effluent (Fig. 4) it can be observed lower recurrence values in the diagonal for the visible part of the spectrum (400 nm-750 nm) and the "dust" data are present in all the figure, from which can be inferred that a very important part of the dataset has low quality. Notwithstanding, the recurrence and the levels of importance are still present in the visible part of the spectrum, which suggests that a physical representation of the spectroscopic process for some couples spectra-analyte still exists.

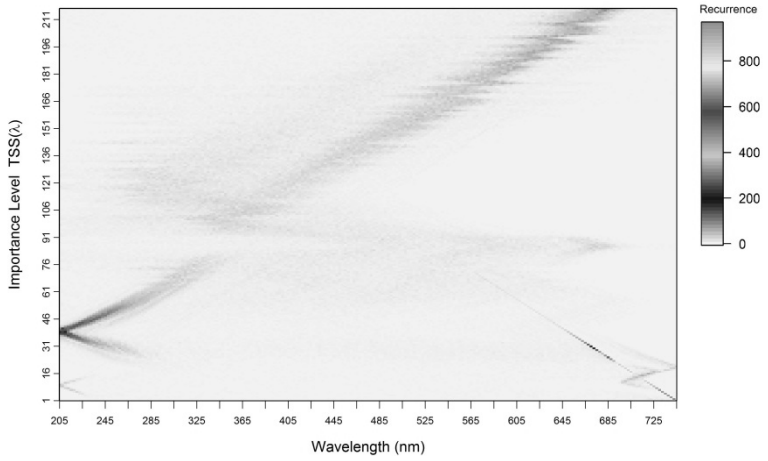

Figure 3: Rayo Affinity: recurrence, level of importance and quality of the data on the relationship spectrum-concentration for influent-TSS

A similar behavior observed for the TSS occurs also for the COD graphs for both influent and effluent. In Figure 5 it can be observed how a greater number of wavelengths allow to validate the presence of the pollutant in the sample, relating absorbance values in the UV range, as well as in the Visible. This allows to confirm the physical influence of suspended solids on the UV range of the spectrum and with this, major fractions of organic carbon that demands dissolved or suspended oxygen, which absorb light in the Visible range, and are related to the presence of COD (Vaillant *et al*., 2002). However, it is important to remember that other major fractions of organic carbon present do not show absorption in the UV-Vis spectrum, such as short chain fatty acids, sugars and starch and therefore it is not possible to quantify the total COD by a spectrophotometric method, as a standard laboratory analysis can do it (Ojeda and Rojas, 2009).

In Figure 6 it can be observed that a range of wavelengths could be associated *a priori* with COD, in a similar way that the one observed in Figure 4 (effluent-TSS), but with a lower recurrence. This shows that the data used have a low quality and suggests a lack of affinity between the pollutant concentration and the spectrum. It also could even suggest that the matrix of the treated wastewater compound in the effluent is less complex and therefore the organic or inorganic substances that demand oxygen are less, then there should be a smaller number of wavelengths associated with the pollutant and with a higher recurrence.

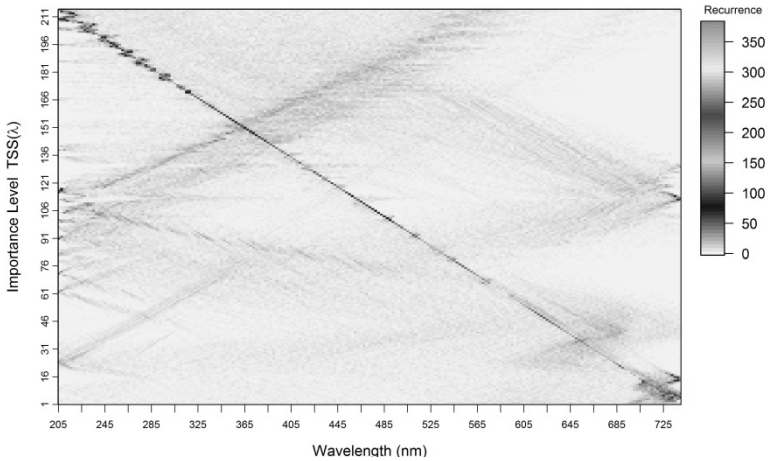

Figure 4: Rayo Affinity: recurrence, level of importance and quality of the data on the relationship spectrum-concentration for effluent-TSS

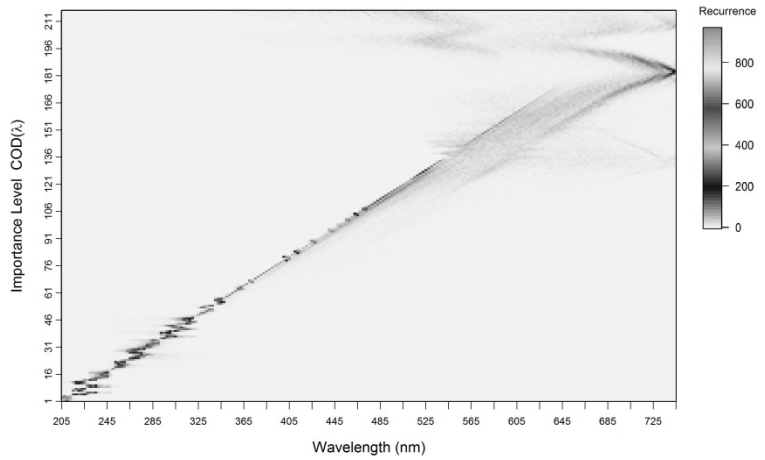

Figure 5: Rayo Affinity: recurrence, level of importance and quality of the data on the relationship spectrum-concentration for influent-COD

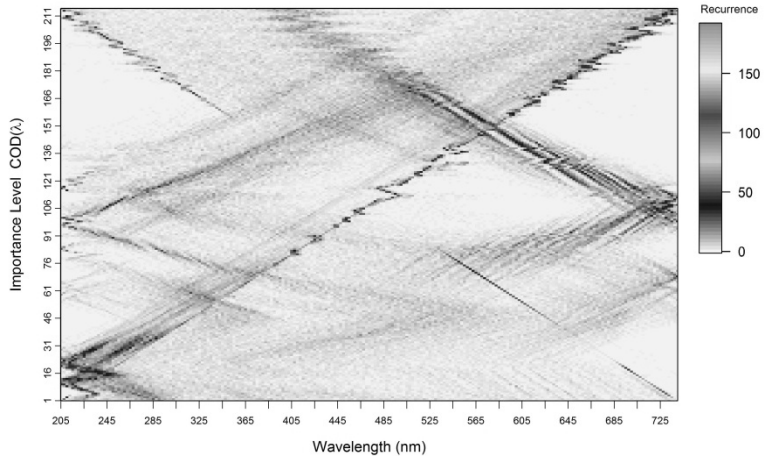

Figure 6: Rayo Affinity: recurrence, level of importance and quality of the data on the relationship spectrum-concentration for effluent-COD

Finally, the filtered COD is the only pollutant analyzed for which the influent data present quality problems (Fig. 7), due to the fact that the presence of "dust" dispersed around the diagonal represents a lower degree of relationship between the absorbance and the concentration values of the samples. For this case it can be assumed that the data from the laboratory tests are responsible of the low affinity between the wavelengths and the CODf concentrations in the influent. However, the assumption that the spectra do not represent the fingerprint sample quality is not valid from the visual analysis, because when comparing the results obtained for the same spectra related to COD and TSS concentrations, high recurrence values could be determined in ranges of more defined wavelengths (Fig. 3 and 4). It might even be defined that the wrong quantification of the CODf in the influent is linked to a poor laboratory procedure or problems related to the elements used in the filtration of the suspended material, because after the 400 nm in the Visible range, there are important recurrences and others of less magnitude around the diagonal. This suggests an influence of suspended material on the soluble part of the COD, which should have been removed during the filtration procedure.

What was stated before can be compared with the CODf results obtained for the effluent, where there are more recurrent wavelengths from the UV part of the spectrum, in a similar range to that reported by Rieger *et al.* (2004).

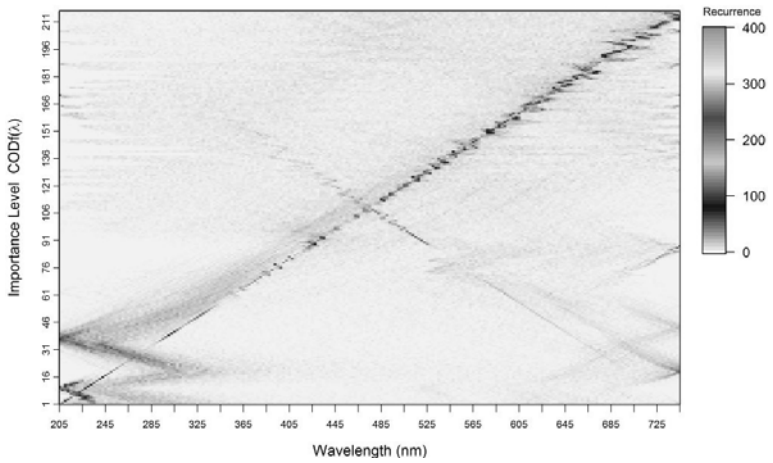

Figure 7: Rayo Affinity: recurrence, level of importance and quality of the data on the relationship spectrum-concentration for influent-CODf

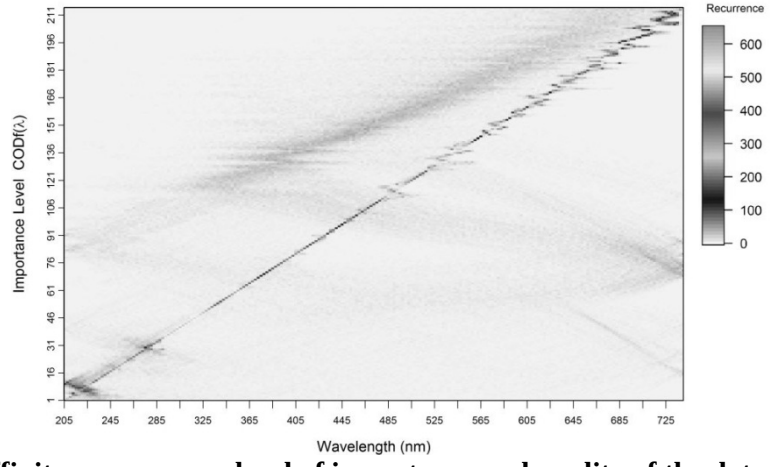

Figure 8: Rayo Affinity: recurrence, level of importance and quality of the data on the relationship spectrum-concentration for effluent-CODf

CONCLUSIONS

The method presented in this paper can be useful to determine the degree of affinity between the absorbance values at single or multiple wavelengths in the spectrum and the pollutant concentrations. This method allows to identify beforehand the quality of the data and to establish the degree of predictability of the regressive models. On the other hand, the detection of which and how many wavelengths have a higher affinity with the pollutant will reduce computation times used in the prediction of concentrations when generating Monte Carlo simulations and cross validation processes.

Finally, it is important to implement image analyses methods to exploit the information presented in this paper, as well as to compare the results with other methods such as the correlation coefficient (R) or the Maximal Information Coefficient (MIC). Finally, it is expected to consolidate the method and apply it to different databases in other fields of knowledge.

REFERENCES

DiFoggio, R. (2000). Guidelines for applying chemometrics to spectra: feasibility and error propagation. *Applied Spectroscopy*. Vol. 54(3). pp. 94A–114A.

Escalas, A., Droguet, M., Guadayol, J. M., and Caixach, J. (2003). Estimating DOC regime in a wastewater treatment plant by UV deconvolution. *Water Research*. Vol. 37(11). pp. 2627–35.

Fleischmann, N., Langergraber, G., Weingartner, A., Hofstaedter, F., Nusch, S. and Maurer, P. On-line and in-situ measurement of turbidity and COD in wastewater using UV/VIS spectrometry. *Proceedings of the 2nd IWA World Water Congress*. Berlin, Germany. 2001. Paper No. B1375.

Gruber, G., Bertrand-Krajewski, J.-L., De Beneditis, J., Hochedlinger, M., and Lettl, W. (2006). Practical aspects, experiences and strategies by using UV/VIS sensors for long-term sewer monitoring". *Water Practice and Technology*. Vol. 1(1). pp. 1-8.

Gruber, G., Winkler, S., and Pressl, A. (2004). Quantification of pollution loads from CSOs into surface water bodies by means of online techniques. *Water Science and Technology*. Vol. 50(11). pp. 73–80.

Hochedlinger, M. (2005). Assessment of combined sewer overflow emissions. PhD thesis: Faculty of Civil Engineering, University of Technology Graz (Austria), June 2005, 174 p. þ annexes.

Hoppe, H., Messmann, S., Giga, A., and Grüning, H. (2009). Options and limits of quantitative and qualitative online-monitoring of industrial discharges into municipal sewage systems. *Water Science and Technology*. Vol. 60(4). pp. 859–67.

Langergraber, G., Fleischmann, N, and Hofstädter, F. (2003). A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. Water Science and Technology. Vol. 47(2). pp. 63–71.

Langergraber, G., Weingartner, A., and Fleischmann, N (2004). Time-resolved delta spectrometry: a method to define alarm parameters from spectral data. *Water Science and Technology.* Vol. 50(11). pp. 13–20.

Langergraber, G., Fleischmann, N., Hofstaedter, F., y Weingartner, A. (2004b) Monitoring of a paper mill wastewater treatment plant using UV/VIS spectroscopy. *Trends in Sustainable Production.* Vol. 49(1). pp. 9–14.

Lorenz, U., Fleischmann, Nikolaus, and Dettmar, J. (2002). Adaptation of a new online probe for qualitative measurement to combined sewer systems (W. C. (eds) In: Stricker, E.W., Huber, ed.). , 427–428.

Ojeda, C. and Rojas, F. (2009). Process analytical chemistry: applications of ultraviolet/visible spectrometry in environmental analysis: an overview. *Applied Spectroscopy Reviews*. Vol. 44(3). pp. 245–265.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rieger, L., Vanrolleghem, P. a, Langergraber, G., Kaelin, D., and Siegrist, H. (2008). Long-term evaluation of a spectral sensor for nitrite and nitrate. *Water Science and Technology.* Vol. 57(10). pp. 1563–1569.

Vaillant, S., Pouet, M. and Thomas, O. (2002). Basic handling of UV spectra for urban water quality monitoring. *Urban Water.* Vol. 4(3). pp. 273–281.

Vargas, A. and Buitrón, G. (2006). On-line concentration measurements in wastewater using nonlinear deconvolution and partial least squares of spectrophotometric data. *Water Science and Technology*. Vol. 53(4-5). pp. 457–463.

Winkler, S., Saracevic, E., Bertrand-Krajewski, J.-L., and Torres, A. (2008). Benefits, limitations and uncertainty of in situ spectrometry. *Water Science and Technology.* Vol. 57(10). pp. 1651–1658.

Authorization and Disclaimer

Authors authorize LACCEI to publish the paper in the conference proceedings. Neither LACCEI nor the editors are responsible either for the content or for the implications of what is expressed in the paper.