Method for outlier detection: a tool to assess the consistency between laboratory data and ultraviolet–visible absorbance spectra in wastewater samples

D. Zamora and A. Torres

ABSTRACT

Reliable estimations of the evolution of water quality parameters by using *in situ* technologies make it possible to follow the operation of a wastewater treatment plant (WWTP), as well as improving the understanding and control of the operation, especially in the detection of disturbances. However, ultraviolet (UV)–Vis sensors have to be calibrated by means of a local fingerprint laboratory reference concentration-value data-set. The detection of outliers in these data-sets is therefore important. This paper presents a method for detecting outliers in UV–Vis absorbances coupled to water quality reference laboratory concentrations for samples used for calibration purposes. Application to samples from the influent of the San Fernando WWTP (Medellín, Colombia) is shown. After the removal of outliers, improvements in the predictability of the influent concentrations using absorbance spectra were found.

Key words | outliers, partial least squares (PLS) regression, pollutants, spectrometric probe (UV–Vis)

D. Zamora (corresponding author)
A. Torres
Research Group Ciencia e Ingeniería del Agua y el Ambiente,
Faculty of Engineering,
Pontificia Universidad Javeriana,
Carrera 7 No. 40-62,
Bogotá,
Colombia
E-mail: *david.zamora@javeriana.edu.co*

INTRODUCTION

The use of installable sensors *in situ*, which use continuous measurement technologies such as ultraviolet (UV)-Visible spectrometry, is one of the concepts of Instrumentation, Control and Automation (ICA). These concepts have been recognized as essential in water systems by different international organizations such as the International Water Association for more than three decades, as documented at the ICA conference in 2001 (Olsson *et al.* 2003; Olsson 2004). One of the main reasons to implement a control device in a wastewater treatment plant (WWTP) is the presence of perturbations, which must be compensated for to maintain the proper functioning of the treatment system (Olsson 2007). Indeed, the influent of a plant goes through considerable temporary variations, both in pollutant concentrations and in flow rates, during periods of time ranging from minutes to months (Bourgeois *et al.* 2001; Olsson 2007).

However, it is not only typical wastewater perturbations that bring challenges to ICA, but also the technologies implemented as UV–Vis spectrometers. These instruments can present problems associated with proper operation which limit potentially their application and affect directly the processes that are controlled by the information they generate (Vanrolleghem & Lee 2003; Olsson 2007). Therefore, continuous measurement sensors are not simply enhanced laboratory instruments: they need adapted servicing and in most cases a well-coordinated monitoring concept with clear decision-support rules, able to detect a bias with a high probability while minimizing measurement effort, allowing estimation of time and costs required for servicing actions (Rieger *et al.* 2004).

Experiences with these probes (Hofstaedter *et al.* 2003; Langergraber *et al.* 2003; Torres & Bertrand-Krajewski 2008) have shown that local calibration results are better and that their success, most of the time, is to ensure the quality of laboratory measurements (Hofstaedter *et al.* 2003; Winkler *et al.* 2008). However, the data used to perform such calibrations can contain outliers. Such data are characterized by the presence of unusually large or small magnitudes compared with others in the set of data, in the case of univariate data-sets (Seo 2006), or unusual relationships between variables in the case of multivariate data-sets. Outliers can have a negative effect on data analysis such as variance and regression analysis, or can provide useful information about the data as an unusual response in a given study, its detection being a fundamental part of data analysis (Fayyad *et al.* 1996).

doi: 10.2166/wst.2014.139

This paper presents the development of an alternative method for outlier detection based on a 1000-model generation of partial least squares (PLS) regressions which estimate the concentrations from the absorbance values of the UV–Visible spectra, using these data to define whether a sample characterized by the fingerprint and the concentration value from laboratory analysis is an outlier candidate. Therefore, the application of the method presented in this paper could be expanded, not only to the special case of the UV–Vis spectrometer data of the case study tackled in this work (San Fernando WWTP influent in Medellín, Colombia), but potentially also to other multivariate databases.

MATERIALS AND METHODS

UV–Visible spectroscopy

One of the latest techniques in continuous measurement, which reduces the disadvantages associated with laboratory testing, is UV–Vis spectroscopy *in situ* (Langergraber *et al.* 2003). UV–Vis spectrometers measure the absorbance of light by suspended or dissolved particles in wavelengths ranging from ultraviolet (200–400 nm) to visible (400–750 nm). The spectrometer sold by the company s::can, called spectro::lyser, is a submersible sensor that measures the attenuation of light between 200 nm and 750 nm in wavelength steps of 2.5 nm and is capable of providing results in real-time with a high temporal resolution (e.g. a measurement every minute) (for more details of the operation and maintenance of this sensor, see Langergraber *et al.* (2004), Hochedlinger (2005), Gruber *et al.* (2006) and Gamerith (2011)). These sensors have been tested non-exhaustively in various operating conditions including rivers (Staubmann *et al.* 2001), treatment plants (Winkler *et al.* 2002), spillways (Gruber *et al.* 2004) and sewage systems (Torres & Bertrand-Krajewski 2008).

Wastewater monitoring has to deal with an array of many dissolved and suspended compounds. The numerous overlapping absorbances of one substance can cause cross-sensitivity and lead to poor performance of the sensor. Chemometric models are used for this purpose. These models formalize the procedure correlating factors for the spectra and their relation to the concentration of the analytes (Langergraber *et al.* 2003). However, direct models of chemometrics can only be used if the spectra of all components are known and the Lambert–Beer law is valid. These conditions are not fulfilled in the case of wastewater, where a large number of unknown compounds are present (Langergraber *et al.* 2003). Choosing how many and which wavelengths are more related to the contaminant is thus an important factor in estimating the concentration versus absorbance.

Therefore, the manufacturer offers an equation based on the statistical technique of PLS in order to assess the concentration of different pollutants in accordance with the absorbance of UV–Vis spectra (Hochedlinger 2005). This equation provides a comprehensive calibration, used for the typical composition of the studied water system. Due to the fact that the composition of the wastewater is highly variable, the manufacturer suggests adapting the overall calibration to the specific quality of the water of the studied water system through a local calibration (Fleischmann *et al.* 2001; Hochedlinger 2005).

PLS regression

Based on the program OPP (OTHU PLS Program) developed by Torres & Bertrand-Krajewski (2008) in the MatLab platform, based on the NIPALS (Nonlinear estimation by Iterative Partial Least Squares) algorithm, a new code was written on the platform R (R Development Core Team 2012) with the following changes: (i) the PLS package was employed (Mevik & Wehrens 2007), using the Wide Kernel algorithm (suitable for many observations and few variables) (Rännar *et al.* 1994) (according to Mevik & Wehrens (2007), the Kernel algorithm and the orthogonal scores algorithm implemented in NIPALS produce the same results, however Kernel is faster to solve most problems); (ii) the optimum number of latent variables is determined by means of cross-validation of Jackknife or Leave One Out type, but selection and classification with respect to the relevance of the independent variables are not determined by the correlation coefficient obtained between each independent variable and dependent within the calibration data-set, but from a new method developed by Zamora & Torres (2013) called ZATO (Zig-zagged graphical Analysis and Treatment of UV–Vis Outliers). This method makes it possible to detect, in a multivariate way, the wavelengths most closely related to the target pollutants, and which can therefore represent in a better way the interactions of the water system with the beam in the spectroscopic process, carried out by the measurement instrument.

The ZATO method considers five bivariate functions in order to calibrate regression models: linear, polynomial of second and third degrees, logarithmic and power. 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. Then, the sum of squared errors (SSE) between reference laboratory concentrations and equivalent concentrations obtained from each model is computed (Zamora & Torres 2013).

After having computed SSE values for each regression model, each wavelength is ordered based on the following criterion: for a particular regression 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 describes the affinity between the wavelengths and the target pollutant (for more details see Zamora & Torres (2013)).

Outliers

The procedure for the detection of outliers is (i) establish the possible criteria for defining an outlier in a data-set, and then (ii) implement a method for identifying those values. Some of the frequently used methods for detecting outliers are statistics-based, such as those developed by Tukey (1977), who introduced several methods for analysis of univariate data, including the Boxplot, which does not assume a normal distribution of the data-set (without making any assumptions of the underlying statistical distribution), and is less sensitive to outliers. Therefore, with this method, x is declared an extreme outlier if it is out of the range $(Q_1 - 3 \times IQR,\ Q_3 + 3 \times IQR)$, where Q_1 is the first quartile, Q_3 is the third quartile and IQR is the interquartile range calculated as $Q_3 - Q_1$ (Acuña & Rodriguez 2004). On the other hand, x is declared a mild outlier if it is outside the range $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)$ (Acuña & Rodriguez 2004).

Several authors have implemented different methods for detecting outliers of the concentration or absorbance values of UV–Visible spectra but not both at the same time. Lorenz *et al.* (2002) presented detection of outliers as a step prior to calibration of PLS models to improve their performance. This detection was performed on the set of laboratory concentrations of combined sewer systems. The authors suggested some methods to detect outliers such as F- and T-Tests, Cook's distance, and other aspects such as level of significance and maximum number of outliers but did not show the specific results of this step.

With squared residuals between concentrations measured in the laboratory (samples of combined sewer overflow) and predicted concentrations from the wavelengths used and their absorbance, positive and negative deviations can be considered equally. With this method Hochedlinger *et al.* (2006) suggested that outliers can only be identified visually by the user, but it is not possible to determine if the outlier is a measurement failure or only an unexpected but correct value. Outliers greatly affect the simple linear regression method owing to the squared distance, which has a major influence for points far off the regression line (Hochedlinger *et al.* 2006).

On the other hand, Thomas *et al.* (2005) proposed a non-parametric measurement, based on comparison of the UV absorption spectra of samples with the characterization of wastewater quality variability. The presence of isosbestic points, occurring in the set of spectra either directly or indirectly after normalization, allows quantification of the variability of given water or effluent. They conducted the search for outliers responsible for a coefficient of variation greater than the fixed value based on a statistical test (the Dixon test). The UV spectra eliminated following this test are considered as not representative of the studied flux (Thomas *et al.* 2005).

Outlier detection based on UV–Visible spectra

As mentioned above, the local calibration of the spectro::lyser probe requires sample collection and subsequent laboratory analysis in order to assess the concentration values of the pollutants of interest, coupled with the measurement of the absorbance spectra of these samples. Then, it is important to detect which of the values of the calibration data-set (spectra and concentrations) are outliers, in order to find better models giving more precise results and not affected by outliers associated with unusual behaviour of the water system or errors associated with laboratory tests or sample collection.

The method for detecting outliers is based in the results of PLS models generated by 1000 Monte Carlo simulations, some elements of the methodology proposed by Tukey (1977) and the root mean square error (RMSE). The execution of 1000 PLS models was carried out taking into account the results obtained by Winkler *et al.* (2008), who evaluated the uncertainties in model development by Torres & Bertrand-Krajewski (2008) and in the equivalent concentrations due to uncertainties in both reference spectra and chemical oxygen demand (COD) values.

Therefore, the following method was developed in order to detect the outliers.

(i) Calibration of 1000 PLS models with 67% randomly selected spectrometry data and the corresponding sample concentrations, leaving the rest for validation.

(ii) Calculation of the concentration values for the calibration and validation data-sets used for each one of the PLS models.

(iii) Assessing of the first condition to determine if a datum is an outlier: a spectrum-concentration set will be an outlier if the distance between the 50% quantile (Q_2) and $EO = Q_{1-3} \pm 3 \times IQR$ does not intersect the bisector in a scatter graphic of concentration measured in the laboratory versus predicted concentrations by PLS models (see Figure 1). In the equation for EO, Q_1 is the first percentile, Q_3 is the third percentile and IQR is the inter-quantile range computed as $(Q_3 - Q_1)$ (Tukey 1977).

(iv) Even if an intersection exists as defined above, a datum is also considered as an outlier if the root mean square of the error (RMSE_L) (Equation (1)) from the concentrations estimated by the models for a measured concentration is higher than the RMSE_G (Equation (2)) of all the data from the 1000 executions obtained (Figure 1 (left)). This second condition is proposed because some data obtained from regression models can show very high variabilities in comparison with the computed average for the whole data-set, such as shown in Figure 2.

$$\text{RMSE}_L = \sqrt{\frac{\sum_{j=1}^{1000} (y_j - \hat{y}_j)^2}{m}} \tag{1}$$

$$\text{RMSE}_G = \sqrt{\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{1000} (y_{ji} - \hat{y}_{ji})^2 \right)}{N}} \tag{2}$$

In Equations (1) and (2), n is the number of samples, N is the total number of data and m is the number of times that data in random generations become part of the calibration or validation set. Consequently, a datum is defined as an outlier if

$$\frac{\text{RMSE}_L}{\text{RMSE}_G} \geq 1 \tag{3}$$

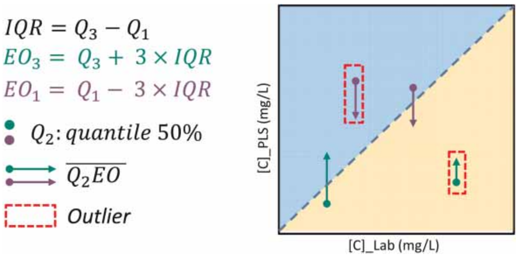

Figure 1 | Graphical representation of the first detection condition.

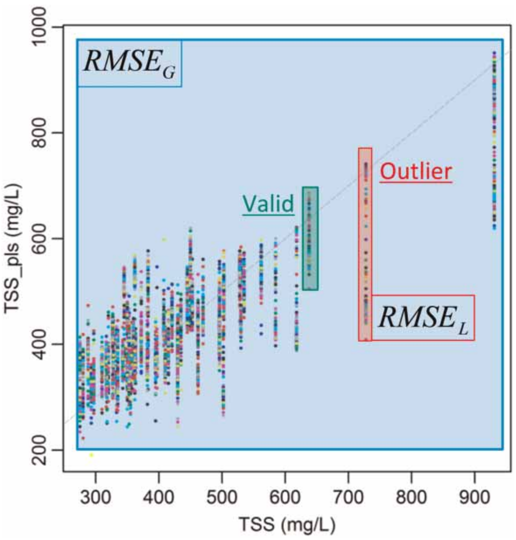

Figure 2 | Graphical representation of the second detection condition.

This last condition suggests that the concentration value obtained in the laboratory is not in accordance with the water quality of the sample and thus the fingerprint associated with it is not the one that physically or chemically represents the sample or vice versa. Therefore, this is not about finding the source of an outlier, because it might only be an error generated in the laboratory or related to the operation and/or maintenance of the probe. The outlier detection method seeks to establish a set of spectra associated with the concentrations to represent the dynamic behaviour of water quality in a reliable and accurate manner.

Case study

The information provided by the Public Enterprises of Medellín, Colombia for the influent of the San Fernando WWTP corresponds to the absorbance spectra, recorded by a spectro::lyser probe with light path of 2 mm (Figure 3 (left)), and concentrations of total suspended solids (TSS) and COD for 124 samples taken in the influent (Figure 3 (right)). This information was originally recorded for the purpose of achieving a local calibration of the spectro:: lyser probe. In order to validate the results of the outlier detection method, calibration and validation data subsets

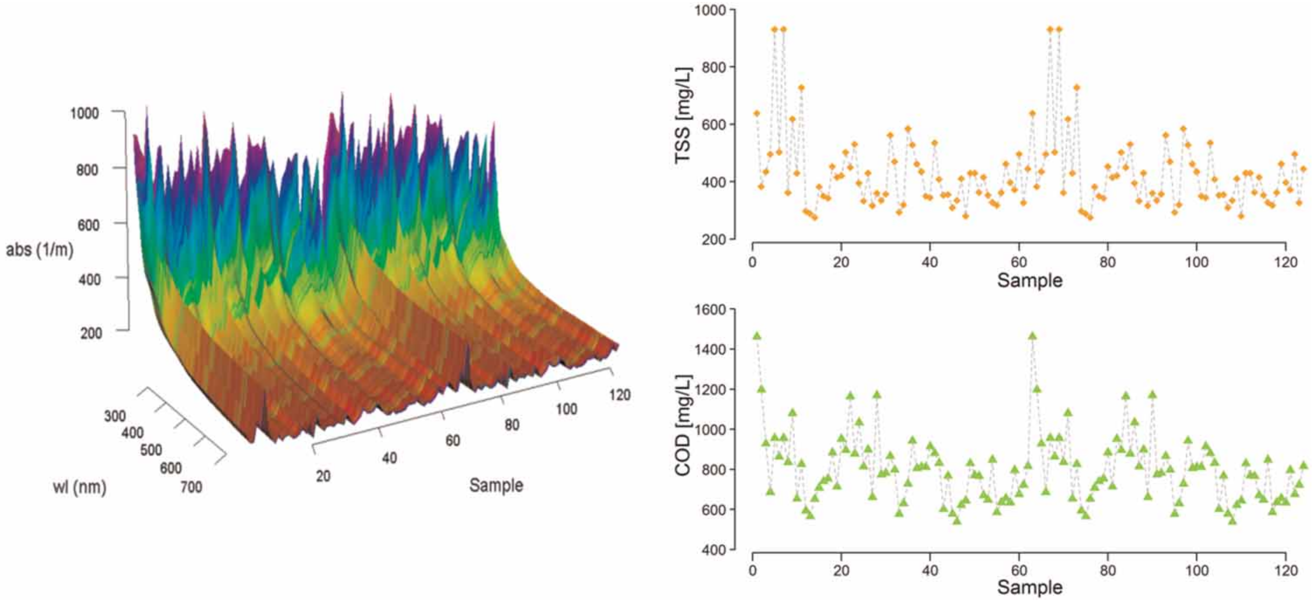

Figure 3 | Absorbance spectra (left) and TSS and COD concentrations (right) for the influent of San Fernando WWTP.

with no outliers were created, where the adjustments of concentrations estimated through PLS regression models and laboratory data were tested.

RESULTS AND DISCUSSION

Graphical representation of the conditions for the detection of outliers

The implementation of the 1000 Monte Carlo simulations is represented in the scatter plots in Figures 4 and 5, where the ordinate corresponds to the COD and TSS equivalent concentrations from PLS models and the abscissa corresponds to the reference concentrations from laboratory analysis. In these graphs can be observed the variability that the value of a pollutant concentration can have, and even how it varies when it is part of the calibration or validation of a model. Therefore, the variability in the prediction errors will determine if a pairing of spectrum and concentration is an outlier, as defined by the second detection condition.

Outliers detected

After viewing how it is determined whether a pair of data is an outlier by employing both detection conditions, it is necessary to know which pairs are classified as outliers. Therefore, in Figures 6 and 7, three kinds of graphs are shown: the first and the second from left to right (TSS and COD) represent the 50% quantile value on the ordinate, and on the abscissa is the concentration of the TSS and COD values measured in the laboratory. In these graphs, it is determined which data were classified for each condition as outliers in the calibration or validation stages only (triangles and squares respectively), which data were detected as outliers in both stages (circles), and which data were classified as not being outliers or as valid data. The third graph summarizes and compares which samples were classified as outliers for each condition and stage.

Figure 6 shows that the largest number of outliers detected was found in the TSS data-set (55), both for the calibration and validation stages, according to the first condition (bisector). With the exception of four samples, all the data detected as outliers for calibration were also detected as outliers for the validation data-set: samples 63 and 91 were detected as outliers only for the calibration data-set, while samples 47 and 109 were detected as outliers only for the validation data-set. On the other hand, the number of outliers detected according to the second condition (residuals) was lower even between stages: 30 samples were detected as outliers in the calibration data-set, while 36 were detected as outliers in the validation data-set.

For the COD data-set, in general terms fewer outliers were detected for both conditions, but a higher number was detected for the first condition and the same number for both stages (38 outliers, see Figure 7). Furthermore, as obtained for the TSS data-set, and with the exception of

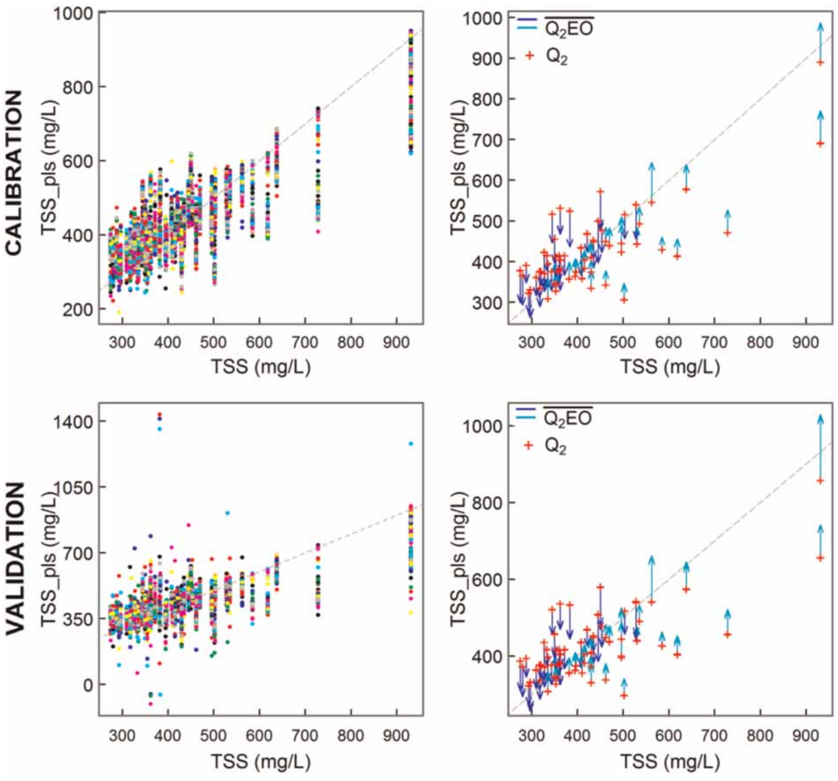

Figure 4 | 1000 Partial Least Squares model executions and outlier detections with first (right) and second (left) criteria proposed in the TSS data-set.

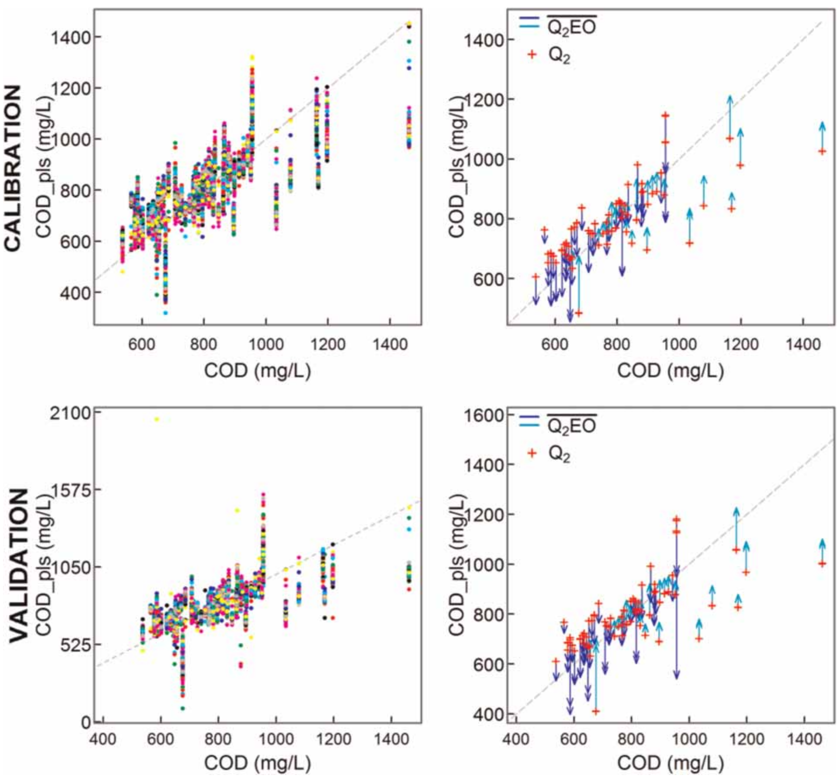

Figure 5 | 1000 Partial Least Squares model executions and outlier detections with first (right) and second (left) criteria proposed in the COD data-set.

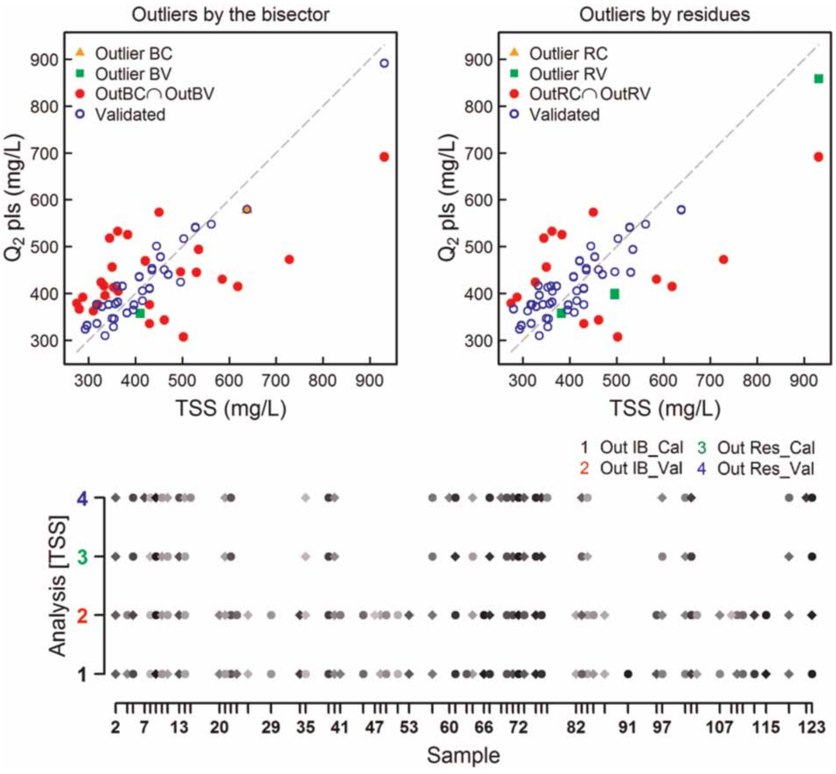

Figure 6 | Outliers detected in the TSS data-set by means of both criteria proposed.

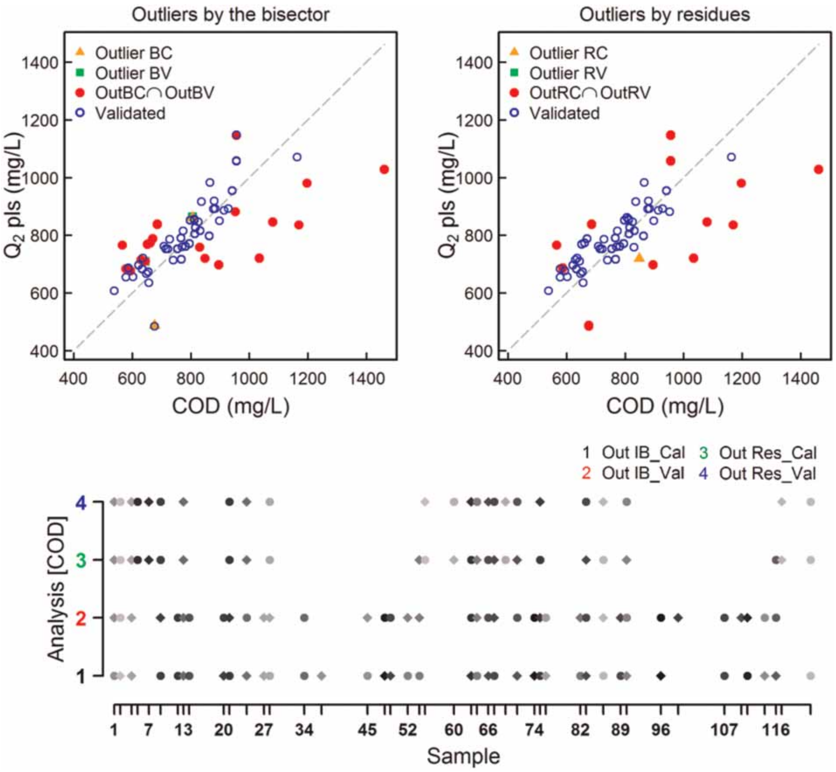

Figure 7 | Outliers detected in the COD data-set by means of both criteria proposed.

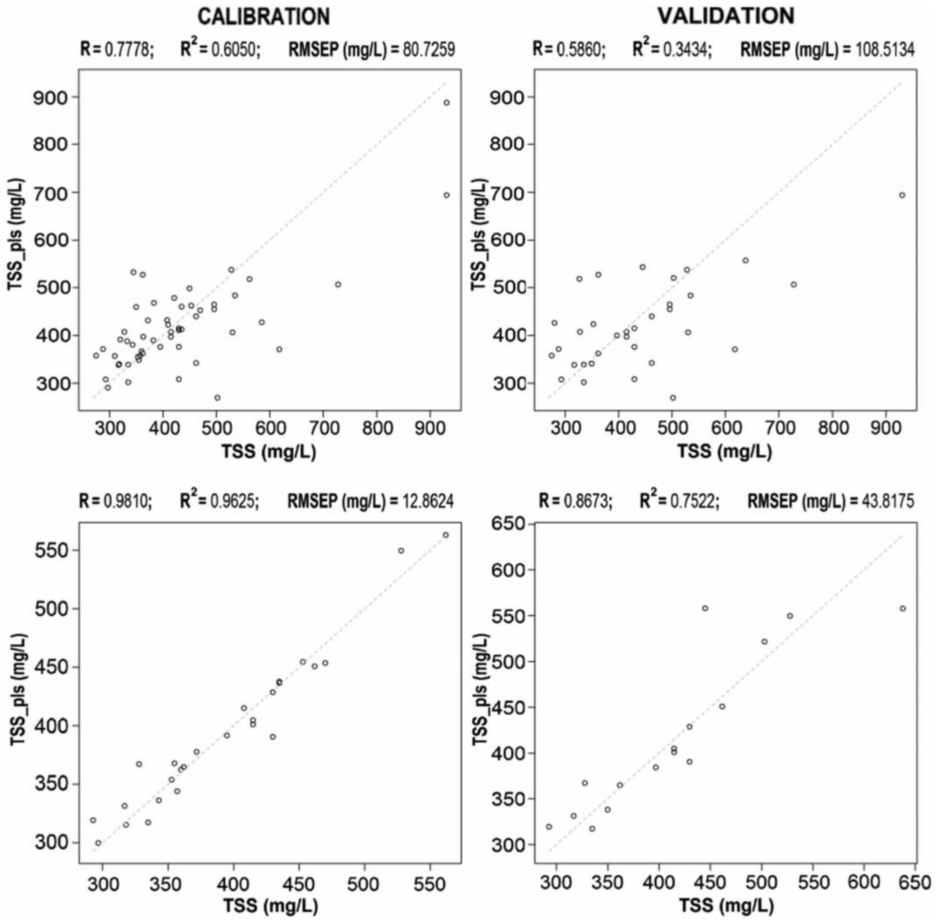

Figure 8 | Results of PLS models calibrated for TSS with outliers (top) and without outliers (bottom) (right: calibration; left: validation).

four samples, all the data detected as outliers for calibration were also detected as outliers for the validation data-set: samples 37 and 122 were detected as outliers only for the calibration data-set, while samples 99 and 110 were detected as outliers only for the validation data-set. Finally, the number of outliers detected by the second condition was lower and different between stages: for the calibration data-set 26 samples were detected as outliers, for the validation data-set 24 samples were detected as outliers. For this case study the data detected by both conditions were assumed to be outliers in both TSS and COD data-sets, and the remaining data, hereinafter referred to as validated data (VD), were used for PLS regression models and, in this way, the predictability of the models with and without outliers was compared.

Post-outlier PLS models

In Figures 8 and 9, the results of PLS models for calibration (left) and validation (right) data-sets are shown, constructed with (top) and without outliers (bottom). In these figures equivalent PLS concentrations (y-axis) are compared to reference concentrations obtained from laboratory analysis (x-axis). On top of each figure, the results of the evaluated metric R, R^2 and RMSEP, are shown, for the models constructed with and without outliers. The adjustment level reached in the calibration process (Figures 8 and 9 (left)) for TSS and COD VD groups is better for those results generated by the PLS calibration model constructed after the exclusion of outliers (WoOM). By applying the proposed method, the predictability of the influent concentrations by means of the absorbance spectra was improved.

CONCLUSIONS

The proposed detection method of outliers becomes an alternative to identifying outliers without checking the normality of data and evaluating prediction variability in multiple PLS models across two conditions: the probability that a combination of varying spectra-concentrations in a

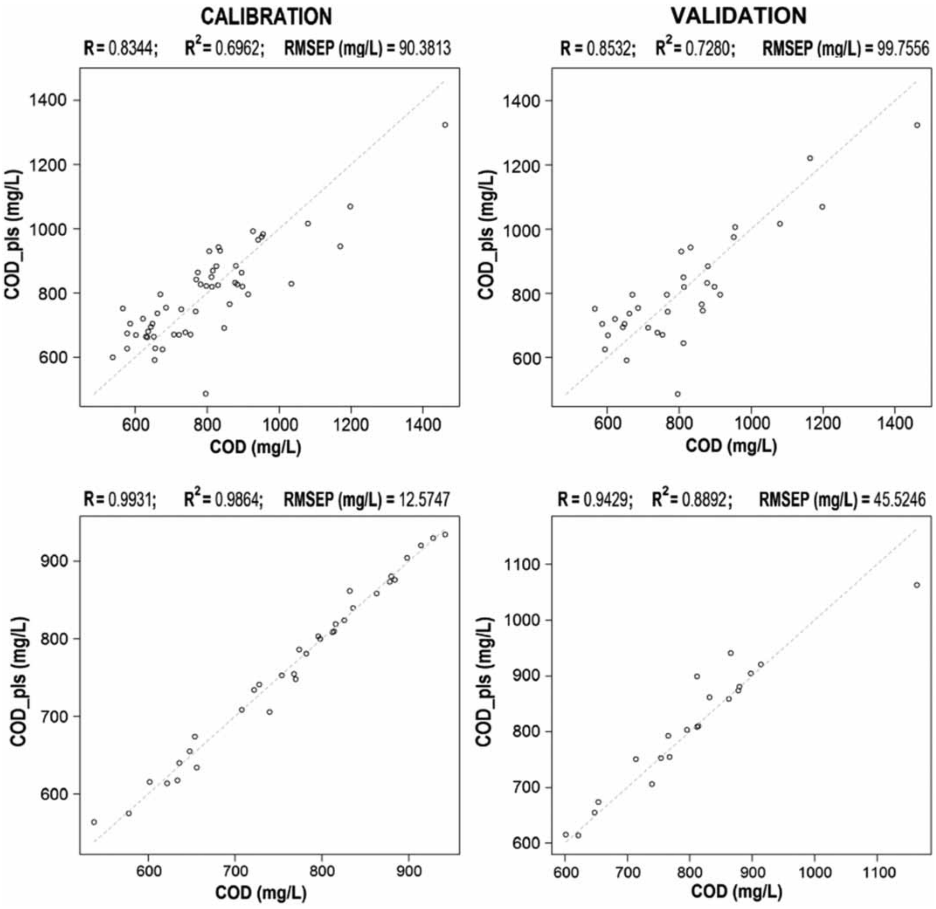

Figure 9 | Results of PLS models calibrated for COD with outliers (top) and without outliers (bottom) (right: calibration; left: validation).

PLS model generates the same values measured in the laboratory and/or that the relation of the RMSE_L of multiple prediction values of a sample in regards to the RMSE_G of all samples and predictions can be less or equal to one. Therefore, the method does not only depend on the magnitude of the dependent variables but also on the relevance of their relationship with the independent variables, since a weak relation between the absorbances and the values of concentration of a pollutant can cause the PLS models always to have a number of different wavelengths and not always the same wavelengths used for calibration. The experimental results demonstrate the effect of removing outliers in the performance of the PLS regression models from data in the influent of the San Fernando WWTP (Medellín, Colombia). Finally, the variability of results obtained from the PLS models constructed after the exclusion of outliers should be evaluated independently for each condition, as well as for calibration and validation groups, in order to define the advantages and disadvantages of each of these.

REFERENCES

Acuña, E. & Rodriguez, C. 2004 On detection of outliers and their effect in supervised classification. http://academic.uprm.edu/~eacuna/vene31.pdf (accessed 26 September 2012).

Bourgeois, W., Burgess, J. E. & Stuetz, R. M. 2001 On-line monitoring of wastewater quality: a review. *J. Chem. Tech. Biotech.* **76**, 337–348.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996 Knowledge discovery and data mining: towards a unifying framework. In: *Discovery and Data Mining, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, CA, pp. 82–88.

Fleischmann, N., Langergraber, G., Weingartner, A., Hofstaedter, F., Nusch, S. & Maurer, P. 2001 On-line and in-situ measurement of turbidity and COD in wastewater using UV/VIS spectrometry. In: *Proceedings of the 2nd IWA World Water Congress, 15–19 October 2001, Berlin, Germany*, paper no. B1375.

Gamerith, V. 2011 *High Resolution Online Data in Sewer Water Quality Modeling*. PhD thesis, Faculty of Civil Engineering, University of Technology Graz, Austria.

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

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

Hochedlinger, M. 2005 *Assessment of Combined Sewer Overflow Emissions*. PhD thesis, Faculty of Civil Engineering, University of Technology Graz, Austria.

Hochedlinger, M., Kainz, H. & Rauch, W. 2006 Assessment of CSO loads – based on UV/VIS-spectroscopy by means of different regression methods. *Water Science and Technology* **54** (6–7), 239–246.

Hofstaedter, F., Ertl, T., Langergraber, G., Lettl, W. & Weingartner, A. 2003 On-line nitrate monitoring in sewers using UV/VIS spectroscopy. In: *Proceedings of the Fifth International Conference of ACE CR 'Odpadni vody–Wastewater 2003'*, J. Wanner & V. Sykora (eds), pp. 341–344.

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

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

Lorenz, U., Dettmar, J. & Fleischmann, N. 2002 Adaptation of a new online probe for qualitative measurements to combined sewer systems. In: *Proceedings of the Ninth International Conference on Urban Drainage*, E. W. Stricker & W. C. Huber (eds), ASCE, Reston, VA, pp. 427–428.

Mevik, B. H. & Wehrens, R. 2007 The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software* **18**, 1–24.

Olsson, G. 2004 Current status of instrumentation, control and automation in wastewater treatment operations. *EICA* **9** (3), 2–14.

Olsson, G. 2007 Automation development in water and wastewater systems. *Environmental Engineering Research* **12** (5), 197–200.

Olsson, G., Newell, B., Rosen, C. & Ingildsen, P. 2003 Application of information technology to decision support in treatment plant operation. *Water Science and Technology* **47** (12), 35–42.

R Development Core Team 2012 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rännar, S., Lindgren, F., Geladi, P. & Wold, S. 1994 A PLS Kernel algorithm for data sets with many variables and fewer objects. Part 1: theory and algorithm. *Journal of Chemometrics* **8**, 111–125.

Rieger, L., Thomann, M., Joss, A., Gujer, W. & Siegrist, H. 2004 Computer-aided monitoring and operation of continuous measuring devices. *Water Science and Technology* **50** (11), 31–39.

Seo, S. 2006 *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. Masters thesis, University of Pittsburgh, Department of Biostatistics, USA.

Staubmann, K., Fleischmann, N. & Langergraber, G. 2001 UV/VIS spectroscopy for the monitoring of testfilters In: *Proceedings of the 2nd IWA World Water Congress, 15–19 October 2001, Berlin, Germany*, paper no. B1378.

Thomas, O., Baurès, E. & Pouet, M. 2005 UV spectrophotometry as a non-parametric measurement of water and wastewater quality variability. *Water Quality Research Journal of Canada* **40** (1), 51–58.

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.

Vanrolleghem, P. & Lee, D. S. 2003 On-line monitoring equipment for wastewater treatment processes: state of the art. *Water Science and Technology* **47** (2), 1–34.

Tukey, J. W. 1977 *Exploratory Data Analysis*. Addison-Wesley, Massachusetts, USA.

Winkler, S., Rieger, L., Thomann, M., Siegrist, H., Bornemann, C. & Fleischmann, N. 2002 In-line monitoring of COD and COD-fractionation: improving dynamic simulation data quality. International IWA World Water Congress, 7–12 April, Melbourne, Australia.

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

Zamora, D. & Torres, A. 2013 Proposal for recurrence, level of importance and quality detection of UV-Vis spectra and target pollutant dataset. In: *Proceedings of the Eleventh Latin American and Caribbean Conference for Engineering and Technology 14–16 August 2013, Cancún City, México (LACCEI 2013)*, paper no. 194.

First received 18 March 2013; accepted in revised form 4 March 2014. Available online 25 March 2014