Detection method in outliers of spectrometry UV-Visible databases: preliminary phase for calibration models applied to regressive real-time monitoring of water quality

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

Estimate reliably and through technologies *in situ* the time evolution of quality parameters allow monitor the status of the proceedings of a Waste Water Treatment Plant (WWTP), promoting the understanding and control over them, especially in the detection of disturbances. However, these technologies have problems related to proper operation and maintenance, which reduce the potential for its application. Then the presence of outliers in longitudinal series is an important phase in the data analysis, as a preliminary to calibrate regression models. Therefore, this article presents a method for detecting outliers, whose application is expanding not only the particular case of the spectrometry UV-Visible data of influent and effluent of the WWTP-San Fernando in Colombia if not to different databases number. To validate the results for detection of outliers, were formed subsets of calibration and validation without outliers data, which evaluated the fit between the estimated concentrations using regression models of partial least squares (PLS) and laboratory data, finding improvements in the predictability of the influent and effluent concentrations using absorbance spectra.

Keywords: Outliers, Spectrometric Probe (UV-Vis), Pollutants, PLS regression

RESUMEN

Estimar de forma fiable y a través de tecnologías *in situ* la evolución temporal de diferentes parámetros de calidad permite monitorear el estado de los procesos de una Planta de Tratamiento de Aguas Residuales (PTAR), favoreciendo la comprensión y el control sobre éstos, especialmente en la detección de perturbaciones. No obstante, dichas tecnologías tienen problemas ligados a su correcta operación y mantenimiento, los cuales reducen el potencial de su aplicación. Luego, detectar la presencia de *outliers* en las series longitudinales es una fase importante en el análisis de datos, como fase preliminar para la calibración de modelos regresivos. Por lo tanto, este artículo presenta un método de detección de *outliers*, cuya aplicación se expande no solamente al caso particular de los datos de espectrometría UV-Visible del afluente y efluente de la PTAR-San Fernando en Colombia, sino a diferentes bases de datos numéricas. Para validar los resultados del método de detección de *outliers*, se conformaron subconjuntos de datos de calibración y validación sin *outliers*, donde se evaluó el ajuste entre las concentraciones estimadas por medio de modelos regresivos de mínimos cuadrados parciales (*PLS*) y datos de laboratorio, encontrando mejoras en la predictibilidad de las concentraciones del afluente y efluente por medio de los espectros de absorbancia.

Palabras claves: *Outliers*, Sonda de Espectrometría (UV-Vis), Contaminantes, Regresión *PLS*

1. INTRODUCTION

Estimar a través de tecnologías *in situ* la evolución temporal de diferentes parámetros de calidad permite monitorear el estado de diferentes hidrosistemas, así como detectar el impacto de las aguas lluvias o vertimientos clandestinos que pueden afectar el medio ambiente y/o condicionar la operación de los componentes del sistema de saneamiento urbano (Ruban *et al*., 2001; Langergraber *et al*., 2004a; Hur *et al*., 2010). No obstante, hasta hace relativamente poco tiempo, las concentraciones de los Sólidos Suspendidos Totales (SST) y de la Demanda Química de Oxígeno (DQO), por ejemplo, eran estimadas a partir de análisis de laboratorio efectuados sobre muestras puntuales recolectadas *in situ*. Esta práctica presenta varios inconvenientes demostrados entre los que se encuentran la baja representatividad espacio-temporal de los resultados, ya que debido al costo elevado asociado a la recolección y análisis de las muestras en laboratorio sólo es posible recolectar un número relativamente pequeño de muestras durante periodos prolongados de tiempo, así como el transporte de las muestras, almacenamiento y conservación de las mismas y los plazos prolongados para la obtención de resultados (Winkler *et al*., 2008).

Una de las alternativas posibles para limitar dichas dificultades consiste en utilizar captores instalables *in situ*, los cuales utilizan tecnologías de medición en continuo, como la espectrometría UV-Visible. Estos captores son capaces de proporcionar informaciones del orden de una medición por minuto, que pueden traducirse en términos de concentraciones equivalentes de contaminantes como la sonda *spectro::lyser* de la sociedad s::can. La utilización de dichos captores se enmarca dentro de los conceptos de Instrumentación, Control y Automatización (ICA). Dichos conceptos han sido reconocidos como esenciales en los sistemas de agua por diferentes organizaciones internacionales como la Asociación Internacional del Agua (IWA) y por más de tres décadas, como se documenta en la conferencia de la ICA en 2001 (Olsson, 2004) y en diferentes libros y publicaciones.

Una de las principales razones para implementar el control en una PTAR es la presencia de perturbaciones, que deben ser compensadas para mantener el correcto funcionamiento del sistema de tratamiento (Olsson, 2007). En efecto, el afluente de una planta varía temporalmente de forma considerable, tanto en su concentración como en su composición y caudal; durante periodos de tiempo que van desde la fracción de horas a meses (Bourgeois *et al*., 2001; Olsson, 2007). Por otra parte, eventos discretos, tales como tormentas, derrames de sustancias tóxicas y picos de caudal también pueden ocurrir de vez en cuando a la entrada de las PTAR (Bourgeois *et al*., 2001; Olsson, 2007). No obstante, no solamente las perturbaciones típicas de las aguas residuales representan desafíos en la ICA, sino también las tecnologías implementadas como los espectrómetros UV-Vis. Estos instrumentos pueden presentar problemas ligados a su correcta operación y mantenimiento, los cuales limitan potencialmente su aplicación y afectan de forma directa los procesos que son controlados por la información que éstos generan (Vanrolleghem y Lee, 2003).

Experiencias con este captor (Hofstaedter *et al*., 2003; Langergraber *et al*., 2003; Torres y Bertrand-Krajewski, 2008) han demostrado que los resultados de la calibración local son mejores que la global. Además, establecen que el éxito de esta calibración la mayoría de casos radica en garantizar la calidad de las mediciones de laboratorio (en relación con el método de análisis, rango de medición, errores de muestreo, identidad de las muestras y la asignación al azar del muestreo) (Hofstaedter *et al*., 2003; Winkler *et al*., 2008).

Sin embargo, los datos utilizados para realizar dichas calibraciones pueden contener valores atípicos (*outliers*). Tales datos están caracterizados por presentar magnitudes inusualmente grandes o pequeñas en comparación con los demás en el conjunto de datos en el caso de conjuntos de datos univariados o relaciones inusuales entre variables en el caso de conjuntos de datos multivariados. Los *outliers* pueden causar un efecto negativo en análisis de datos tales como análisis de varianza y regresión, o pueden proporcionar información útil acerca de los datos cuando se fija una respuesta inusual de un estudio determinado, constituyéndose su detección en una parte fundamental del análisis de datos (Seo, 2006). En este artículo se presenta el desarrollo de un método alternativo para detección de outliers basado en los cuantiles y regresiones polinomiales de segundo grado para datos de espectrometría UV-Visible. La aplicación de este método se expande no solamente al caso particular de los registros de una PTAR estudiado en este artículo, sino a diferentes bases de datos numéricas.

2. MATERIALES Y MÉTODOS

2.1 ESPECTROMETRÍA UV-VISIBLE

Una de las técnicas más recientes de medición en continuo, que permite reducir los inconvenientes asociados a los ensayos de laboratorio es la espectrometría UV-Visible *in situ*. Los espectrómetros UV-visibles realizan una medición de la absorbancia de la luz generada por las partículas disueltas o en suspensión en longitudes de onda que van desde el rango ultravioleta hasta el visible. Estos captores son capaces de proporcionar informaciones del orden de una medición por minuto, que pueden traducirse en términos de concentraciones equivalentes de contaminantes (SST, DQO y nitratos). El espectrómetro comercializado por la sociedad s::can, llamado spectro::lyser, es un captor sumergible, que mide la atenuación de la luz entre 200 nm y 750 nm en pasos de longitud de onda de 2.5 nm, y es capaz de otorgar resultados en tiempo real (Langergraber *et al*., 2004b; Hochedlinger, 2005). La medición se realiza directamente *in situ* sin necesidad de muestreo o de tratamiento de las muestras y por lo tanto algunos errores experimentales con el captor se consideran mucho menores que aquellos asociados a los ensayos estándares de laboratorio (Langergraber *et al*., 2003).

Utilizar los captores implica correlacionar las absorbancias con la concentración de los diferentes contaminantes que puedan ser detectados en las longitudes de onda del espectro UV-Visible presentes en el hidrosistema analizado. La compañía fabricante ofrece una ecuación que se basa en la técnica estadística de mínimos cuadrados parciales (*PLS: Partial Least Squares*). Dicha ecuación ofrece una calibración global para una serie de parámetros válidos para la composición típica del hidrosistema estudiado. Por lo general, al utilizar esta ecuación se obtienen coeficientes de correlación altos (R^2= 0.90 - 0.95 (Hofstaedter *et al*., 2003)) para un conjunto de parámetros estándar (SST, DQO, DBO, etc.), ofreciendo resultados de calidad suficiente para muchos propósitos, tales como el control de una planta de tratamiento de aguas residuales (Fleischmann *et al*., 2001). Debido a la composición de las aguas residuales, las cuales presentan propiedades específicas que varían de acuerdo a la clase de vertimiento en las redes de alcantarillado (por ejemplo vertimientos industriales), las concentraciones de los diferentes compuestos son variables temporalmente, especialmente los orgánicos (Hochedlinger, 2005). Por consiguiente, el fabricante sugiere adaptar la calibración global a la calidad del hidrosistema estudiado por medio de una calibración local.

2.2 *OUTLIERS*

El procedimiento para la detección de *outliers* consiste, primero en definir cuáles serían los posibles criterios para que un dato dentro de un conjunto de datos dado reciba el calificativo de *outlier*, y luego en aplicar un método para identificar dichos valores.

Por lo tanto, cuando se tiene un conjunto de datos con n observaciones de una variable x, donde $\bar{x}$ es la media y S es la desviación estándar de la distribución de los datos, una observación se declara como *outlier* si se encuentra fuera del intervalo $(\bar{x} - kS, \bar{x} + kS)$, donde el valor del coeficiente k es usualmente 2 ó 3 (Acuña y Rodríguez, 2004). Estos valores se justifican en el hecho que al suponer una distribución normal se espera contar con un porcentaje del 95 % ó 99 % respectivamente, de los datos en el intervalo centrado en la media, con una longitud aproximadamente igual a dos o tres veces la desviación estándar respectivamente. El problema de este método es que asume la distribución normal de la información, que con frecuencia es algo que no ocurre, y además tanto la media como la desviación estándar son muy sensibles a los valores atípicos de magnitudes significativas (Chen *et al*., 1996). En respuesta a esto, *Tukey* (1970) introdujo varios métodos para el análisis de datos univariados, entre los que se encuentra el *Boxplot*, el cual, al no suponer una distribución normal del conjunto de datos, son menos sensibles a valores extremos (Acuña y Rodríguez, 2004; Seo, 2006).

Un dato x se declara *extreme outlier*, si se encuentra fuera del intervalo $(Q_1 - 3 \times IQR, Q_3 + 3 \times IQR)$, donde Q_1 es el primer cuartil, Q_3 es el tercer cuartil e IQR recibe el nombre de rango intercuartil calculado como $Q_3 - Q_1$. Un dato x se declara *mild outlier* si se encuentra fuera del intervalo $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)$ (Acuña y Rodríguez, 2004).

2.3 DETECCIÓN DE *OUTLIERS* EN FUNCIÓN DE ESPECTROS UV-VISIBLES

Como se mencionó anteriormente, la calibración local de la sonda *spectro::lyser* requiere la recolección de muestras y su posterior análisis en laboratorio a través de ensayos estándar de los contaminantes de interés y la medición de los espectros de absorbancia de estas muestras. Luego, es importante detectar cuáles valores del conjunto de datos de calibración (espectros y concentraciones) son *outliers*, con el fin de encontrar mejores modelos cuyos resultados sean más precisos y no se vean afectados por valores atípicos asociados a un comportamiento inusual del hidrosistema o errores ligados a los ensayos de laboratorio.

Por lo tanto, utilizando la metodología concebida por Tukey (1970), se desarrolló el método descrito a continuación para la detección de *outliers*: (i) calcular el coeficiente de correlación (r) entre los valores de absorbancia de cada longitud de onda del espectro y la concentración del contaminante por cada muestra. Aquí, los autores suponen que la atenuación de la radiación en una longitud de onda específica puede ser medida en el espectro, y que el valor de su absorbancia tiene una relación lineal con la concentración. Por lo tanto, la absorbancia aumenta con la concentración del analito (DQO, SST *etc*.) asumiendo así que la ley de Beer-Lambert es válida, con lo cual se define el rango de longitudes de onda en función de la absorbancia para las cuales son válidas las concentraciones de las muestras. (ii) Seleccionar la longitud de onda con el mayor coeficiente de correlación entre los valores de absorbancia y los valores de concentración del parámetro estudiado, denominada *miw* (*most important wavelength*); (iii) conformar dos grupos de datos: absorbancias asociadas a *miw* y concentraciones correspondientes obtenidas en laboratorio; (iv) de dichos grupos se selecciona el 67 % de los datos de forma aleatoria, los cuales se usan para calibrar los coeficientes de un modelo de regresión lineal, repitiendo el proceso 50000 veces, utilizando la ecuación 1; (v) a partir de los coeficientes calculados utilizando la Ecuación 1, se estiman las concentraciones para cada una de las ejecuciones aleatorias en función de las absorbancias de la *miw*, obtenidas en el paso anterior; (vi) Se calculan los cuartiles Q_1, Q_2 y Q_3 de las concentraciones estimadas, conformando así una matriz de dimensiones $n \times 1$ por cada cuantil; (vii) en función de las *miw* y los cuantiles calculados en el paso anterior, se calibra un modelo regresivo de carácter polinomial de segundo grado por cuantil, los cuales tienen por fin modelar el comportamiento de los tres cuantiles del conjunto de datos $n \times k$ (ver Ecuación 2); (viii) calibradas las ecuaciones polinomiales, se calculan los límites y rangos para la detección de los *mild outliers* $(MQ_1 - 1{,}5 \times IQR, MQ_3 + 1{,}5 \times IQR)$, *extreme outliers* $(MQ_1 - 3 \times IQR, MQ_3 + 3 \times IQR)$ y la tendencia central de los datos (Q_2 cuantil 50 %), con el rango intercuantil calculado como la diferencia entre MQ_3 y MQ_1.

$$\hat{y}_i = \sum_{k=1}^{50000} m_k \cdot x_i(\lambda_{miw}) + b_k$$

Ecuación 1: Función regresiva lineal

$$MQ_{1,2,3} = \sum_{i=1}^{n} C_{i(1,2,3)} \cdot x_{i(1,2,3)}(\lambda_{miw})^2 + D_{i(1,2,3)} \cdot x_{i(1,2,3)}(\lambda_{miw}) + E_{i(1,2,3)}$$

Ecuación 2: Función polinomial de segundo grado

donde m_k y b_k son los coeficientes calibrados en cada una de las k ejecuciones, x_i son las absorbancias de la longitud de onda más importante, $\hat{y}_i$ son las concentraciones calculadas de la ecuación lineal (con $i = 1,2,\ldots n$), los subíndices de $MQ_{1,2,3}$ hacen referencia al modelo independiente para primer cuartil (MQ_1), segundo cuartil (MQ_2) y tercer cuartil (MQ_3), $x_i(\lambda_{miw})$ son los valores de absorbancia correspondientes a las *miw* y $C_{i(1,2,3)}$, $D_{i(1,2,3)}$ y $E_{i(1,2,3)}$ son los coeficientes que se calibran para cada modelo.

2.4 REGRESIÓN *PLS*

Con base en el programa *OPP* (*OTHU PLS Program*) desarrollado por Torres y Bertrand-Krajewski (2008) en la plataforma MatLab y basado en el algoritmo *NIPALS* (*Non linear estimation by Iterative Partial Least Squares*), se rescribió el código en la plataforma *R* (R Development Core Team, 2012) con los siguientes cambios: (i) Se

utilizó el paquete *pls* (Mevik y Wehrens, 2007) de *R* (R Development Core Team, 2012). (ii) El algoritmo *PLS* utilizado es *Wide Kernel* (apropiado para muchas observaciones y pocas variables) (Rännar *et al*., 1994) –según Mevik y Wehrens (2007), el algoritmo *Kernel* y el algoritmo de puntuaciones ortogonales implementado en *NIPALS* generan los mismos resultados; no obstante *Kernel* es más rápido para resolver la mayoría de problemas. (iii) El número óptimo de variables latentes se determina por medio de validación cruzada tipo *Jackknife* o *Leave One Out*.

2.5 VALIDACIÓN DEL MÉTODO DE DETECCIÓN DE *OUTLIERS*

La forma de validar los resultados del método de detección de *outliers* se encuentra en el ajuste que se alcance entre los valores estimados por la regresión *PLS* y las concentraciones obtenidas en laboratorio, para lo cual se emplearon dos métricas: la primera métrica es la raíz cuadrada del error cuadrático medio *RMSE* (*Root Mean Square Error*), el cual es más sensible a los datos de alta magnitud, que generalmente presentan mayores errores y poco sensible a los valores de baja magnitud. En esta métrica un valor igual a 0 indica un modelo perfecto (Ecuación 1). La otra métrica es el coeficiente de eficiencia de Nash y Sutcliffe (1970), el cual puede variar entre 0 y 1, no obstante los valores negativos también se pueden presentar (Ecuación 2): 1 representa un modelo perfecto, 0 indica que las predicciones del modelo son tan precisas como la media de los datos observados y un valor negativo indica que la media observada es un mejor indicador que el modelo, o visto de otra forma, se puede decir que la varianza de los residuos de la predicción $(y_i - \hat{y}_i)$ es mayor que la varianza de los datos $(y_i - \bar{y})$ (Dawson *et al.*, 2007).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

Ecuación 3: Root Mean Square Error

$$NSC = 1 - \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Ecuación 4: Coeficiente de Nash-Sutcliffe (1970)

El método propone generar modelos *PLS* independientes, calibrados con el 70 % de los datos con y sin *outliers* para cada contaminante y el 30 % restante para la validación, conformados de la misma forma. Para comparar el nivel de ajuste entre los modelos, tanto en la etapa de calibración como de validación, los datos que conforman los grupos de dichas etapas deben ser los mismos. Esto quiere decir que si el grupo $F = [Z_1, Z_2, Z_3, Z_4, Z_5]$ contiene *outliers*, el grupo sin *outliers* será $D - [Z_2, Z_4, Z_5,]$ y de la misma forma para los grupos de validación.

Por último, los espectros *outliers* son utilizados para estimar, por medio de los modelos *PLS* calibrados sin *outliers*, la concentración de los contaminantes homólogos a dichos espectros, y comparar los resultados obtenidos del $RMSE$ y NSC calculados para los datos *outliers* con los modelos *PLS* calibrados con y sin *outliers*.

3. CASO DE ESTUDIO

La Planta de Tratamiento de Aguas Residuales San Fernando (PTAR) localizada en el municipio de ltagüí, Colombia. Esta PTAR recibe, para su tratamiento, las aguas residuales de tipo industrial y residencial de los municipios de Envigado, Itagüí, Sabaneta, La Estrella y parte del sur de Medellín. La información suministrada por las Empresas Públicas de Medellín (EPM) del afluente de la PTAR San Fernando corresponde a los espectros de absorbancia y las concentraciones de SST, DQO y Demanda Química de Oxígeno filtrada (DQOf) para un total de 124 muestras del afluente tomadas en diferentes tiempos. Estas muestras fueron tomadas originalmente con el propósito de lograr una calibración local de la sonda *spectro::lyser* utilizada en el afluente de la PTAR.

En la figura 1 se muestran los datos de concentraciones y espectros de las muestras del afluente de izquierda a derecha respectivamente, estos últimos fueron medidos una *spectro::lyser* de un paso de luz de 2.

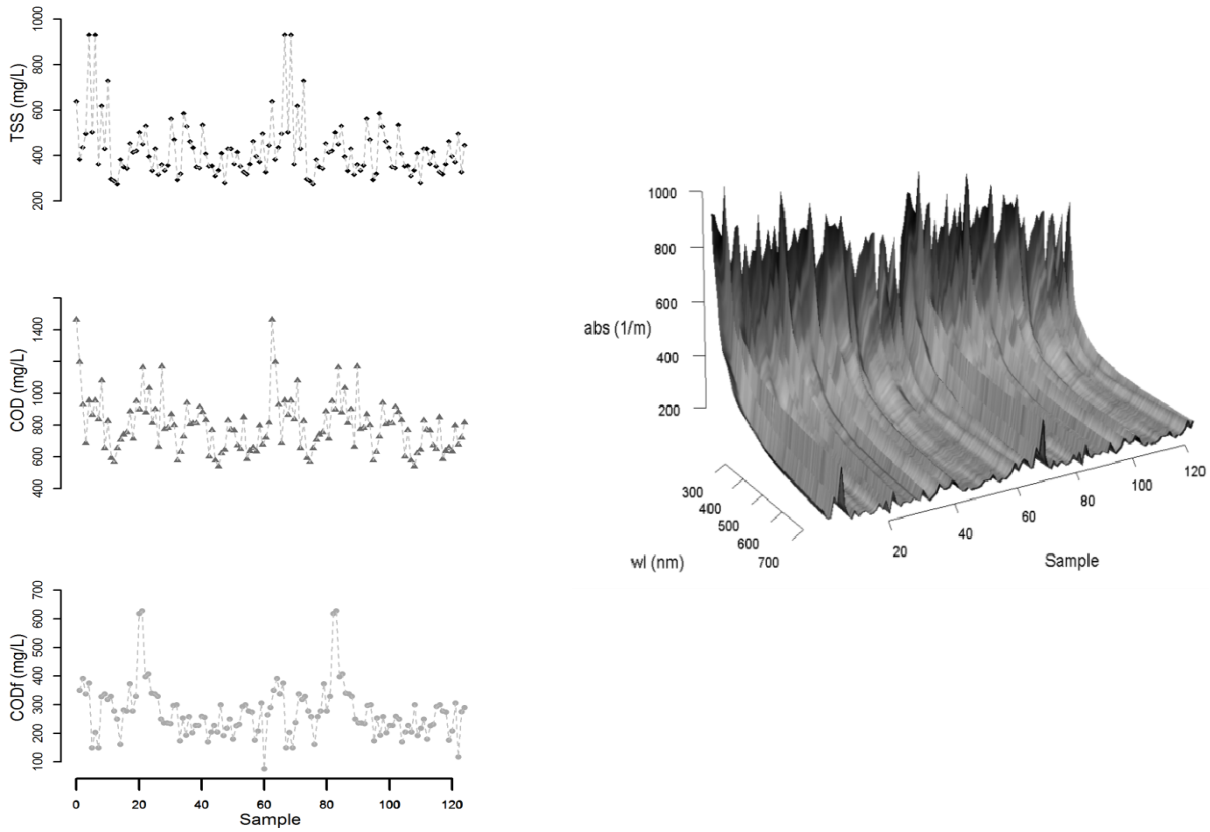

Figura 1: Concentraciones de SST, DQO y DQOf (izq.) y espectros de absorbancia (der.) del afluente de la PTAR San Fernando

4. RESULTADOS

4.1 *OUTLIERS* DETECTADOS

En la figura 2 presenta los resultados de la detección de *outliers* del afluente. A la izquierda de esta figura en las ordenadas se muestran las gráficas de detección de *outliers* para las concentraciones de contaminantes SST, DQO y DQOf, presentado en el eje *y* sus concentraciones en mg/L en función de las absorbancias (Abs/m) de las *miw* para cada muestra. Además se establecen los límites y rangos de detección de los datos validados, *mild outliers* y *extreme outliers*. En las gráficas de barras ubicadas a la derecha de esta figura, se consolida la cantidad y porcentaje sobre el total de datos validados, *mild outliers* y *extreme outliers*.

Para los datos del afluente (Figura 2) se detectó en general un mayor porcentaje de *Extreme outliers (Eo)* en relación con los detectados en el efluente: el mayor de estos porcentajes está en la DQOf con 75,58 %, mientras que los *Mild outliers* (*Mo*) y los datos validados corresponden al 12,1 % y 15,32 % de la totalidad de los datos respectivamente para este contaminante. Por otra parte, la DQO tiene el mayor porcentaje de datos validados (22,58 %) y la menor parte de *Eo* correspondiente al 55 % de la totalidad de los datos. En cuanto a los SST, se logró establecer que el porcentaje de *Eo* de este contaminante tiene un valor cuya magnitud (66,13 %) se encuentra entre los valores de los contaminantes DQO y DQOf.

Por otra parte, la estrecha diferencia entre el límite de los *Mo* y el límite de los *Eo* observada en la Figura 4 para los SST, en particular para el rango de absorbancias (100 Abs/m a 150 Abs/m), puede probablemente aumentar al incrementar el número de datos considerados en el análisis, ya que esto permitiría incrementar la distancia de los límites *Mo* y *Eo* respecto a la curva correspondiente a la MQ_2. Con lo cual un mayor número de datos quedarían definidos como datos validados y/o *Mild outliers*. Un comportamiento similar puede esperarse para la DQOf (Figura 2, inferior), ya que ésta tiene un rango más o menos definido entre 700 y 900 Abs/m; a diferencia del

comportamiento de la DQO cuyas absorbancias más importantes se encuentran tanto en la parte UV como en la Visible, ya que este parámetro está relacionado con oxidación química de la materia disuelta y en suspensión.

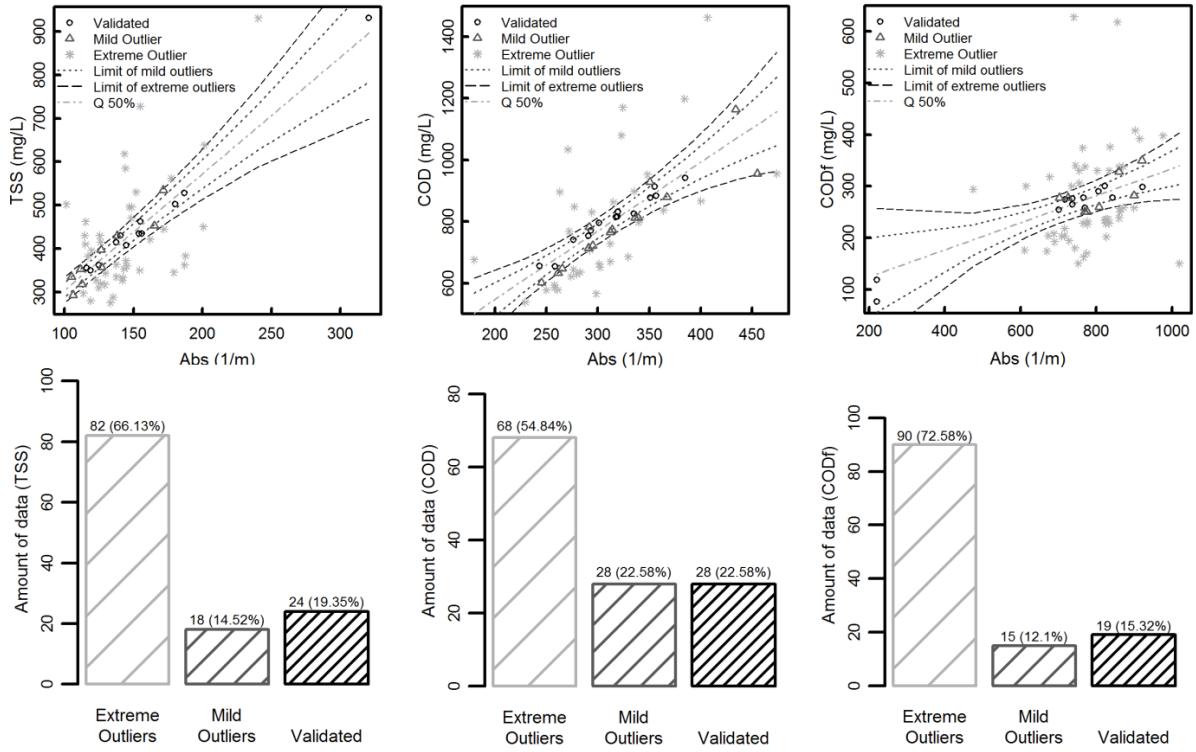

Figura 2: (Arriba) Detección de los datos *mild outliers*, *extreme outliers* y validados; (Abajo) cantidad y porcentaje de los datos detectados en los conjuntos de datos de SST, DQO y DQOf del afluente

Es importante aclarar que para este caso particular de aplicación del método de detección de outliers, se unieron los subconjuntos de datos *Mo* y validados, creando un único conjunto de datos validados denominado en adelante Datos de Regresión (DR), que a su vez es divido en un subconjunto de calibración y otro de validación para generar los modelos de regresión *PLS* en los datos del afluente.

4.2 MODELOS DE REGRESIÓN *PLS* Y VALIDACIÓN DEL MÉTODO DE DETECCIÓN DE *OUTLIERS*

Inicialmente se muestran los resultados del proceso de calibración de los modelos de *PLS* con y sin *outliers*, considerando que el conjunto de datos *outliers* de los espectros de absorbancia son utilizados para cuantificar concentraciones utilizando el modelo *PLS* sin *outliers* y así evaluar su ajuste con respecto a las concentraciones de laboratorio.

En las Figuras 3 y 4 se presentan los resultados de calibración y validación de los modelos con y sin *outliers* de izquierda a derecha, para el afluente. En las ordenadas de estos gráficos se presentan lo valores de las concentraciones calculadas por los modelos regresivos *PLS* y en las abscisas los valores de las concentraciones obtenidas en laboratorio. En la parte superior de cada gráfico se presentan los resultados de las métricas evaluadas *RMSE* y *NSC* para los DR y *outliers*, de los modelos entrenados con y sin *outliers*.

El nivel de ajuste alcanzado en el proceso de calibración (Figura 3) para los grupos de DR de los SST y la DQOf es mejor para los resultados generados con el modelo *PLS* calibrado sin considerar *outliers* (*WoOM*), pero esto no ocurre de la misma forma para la DQO, ya que por 6 mg/L el error generado por el modelo *PLS* calibrado considerando o*utliers* (*WOM*) es menor que el error generado al aplicar *WoOM*.

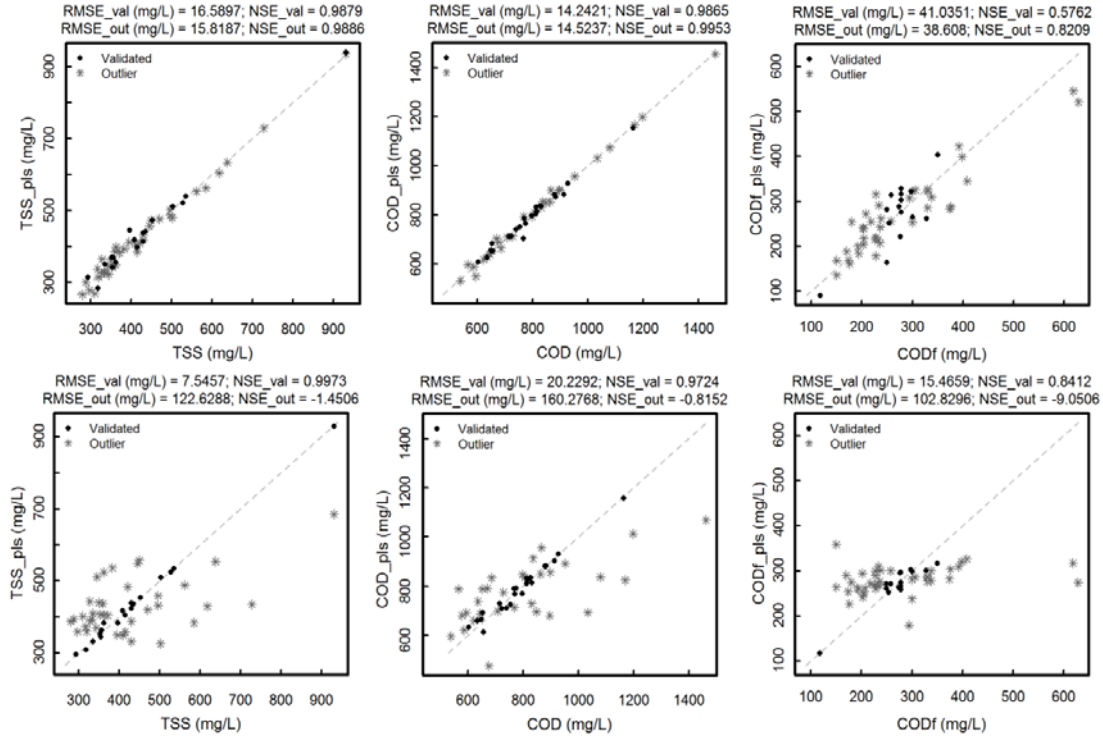

Figura 3: Comparación entre las métricas de ajuste *RMSE* y NSC calculadas para la calibración de los modelos *PLS* (arriba) con y sin *outliers* (abajo) del afluente

Con respecto al grupo de datos *outliers*, se encontraron mejores resultados con el modelo *WOM*, ya que las métricas evaluadas presentan menores errores y valores cercanos a 1 en el coeficiente *NSC*, lo cual es la tendencia general en todos los contaminantes. El alcance predictivo de los resultados del grupo de *outliers* en el modelo que prescindía de éstos para su calibración, se debe a que el pronóstico de las concentraciones de los contaminantes en función de las absorbancias más importantes no supera el pronóstico por inercia del modelo calibrado y por lo tanto sus ajustes con respecto a los datos de laboratorio serán pobres.

Por lo tanto, este resultado permite respaldar en primera instancia la detección de ese conjunto de datos como *outliers* en el afluente, ya que si se hubiera generado un error menor o igual al calculado en los *outliers* con el modelo *WoOM*, se podría inferir que en ambos casos los *outliers* no son magnitudes atípicas tanto en la concentración de laboratorio como en el espectro de absorbancia y éstos no incrementarán el error de predicción ni afectarán la capacidad predictiva del modelo. Por otra parte los valores negativos del coeficiente NSC para los *outliers* en el modelo *WoOM* denotan que los residuos de los errores generan una mayor varianza que la varianza de los datos originales.

En la validación del conjunto de DR del afluente (Figura 4), se encontró que el error en mg/L de los SST es 10 veces menor en el modelo *WoOM*, en comparación al *WOM*. No obstante, para la DQO y la DQOf el decrecimiento en el error no fue tan sustancial como en los SST, ya que para estos contaminantes se redujo el error 26,21 % y 18,72 % respectivamente con respecto al modelo *WOM*. Con relación al conjunto de datos *outliers*, los errores en mg/L obtenidos de estos datos en el *WoOM* es mayor con respecto al *WOM*, con una diferencia importante en la DQO (105 mg/L). Sin embargo, la menor diferencia del error entre ambos modelos para los datos *outliers* de SST y DQOf respaldan la premisa que el número de datos validados podría ser mayor, sobre todo los datos más cercanos al límite de los *Extreme outliers* como aparecen en la Figura 2. Con esto se espera que la suma de los residuos cuadráticos incremente su valor, ya que la falta de predictibilidad de los *outliers* en el modelo calibrado sin estos datos probablemente genere mayores residuos, y al dividirlos en un menor número de datos el valor del *RMSE* será mayor. Por consiguiente, estos resultados respaldan la detección del grupo *outliers* en el afluente y se corrobora con la mejora en la capacidad predictiva del modelo *PLS* para cada contaminante.

5. CONCLUSIONES

El método de detección de *outliers* propuesto resulta una alternativa sencilla y práctica para identificar los *outliers* sin necesidad de comprobar la normalidad de los datos y generando límites de detección en función del conjunto datos de las variables independientes. Por lo tanto, el método no depende únicamente de la magnitud de las variables dependientes sino de la relevancia de su relación con las variables independientes.

Los resultados experimentales evidencian el efecto de la supresión de los *outliers* en el rendimiento de los modelos regresivos *PLS*, sobre todo en los datos del afluente de la PTAR San Fernando. Aunque el número de *outliers* detectados por la metodología parece significativamente alto, los resultados de las métricas de ajuste justifican su eliminación, ya que permiten que el modelo de regresión implementado sin considerar los *outliers* mejore sustancialmente su capacidad predictiva.

De los resultados alcanzados por el método de detección de *outliers* en la base datos del efluente, la falta de predictibilidad y los errores importantes en el ajuste conducen a revisar los datos iniciales, y por ende la forma en la cual éstos son obtenidos. Por lo tanto, se deben revisar la forma de operación y mantenimiento de este captor, así como las técnicas y procedimientos en los ensayos laboratorio, para detectar y mitigar la fuente de error que repercute en la predictibilidad de los modelos.

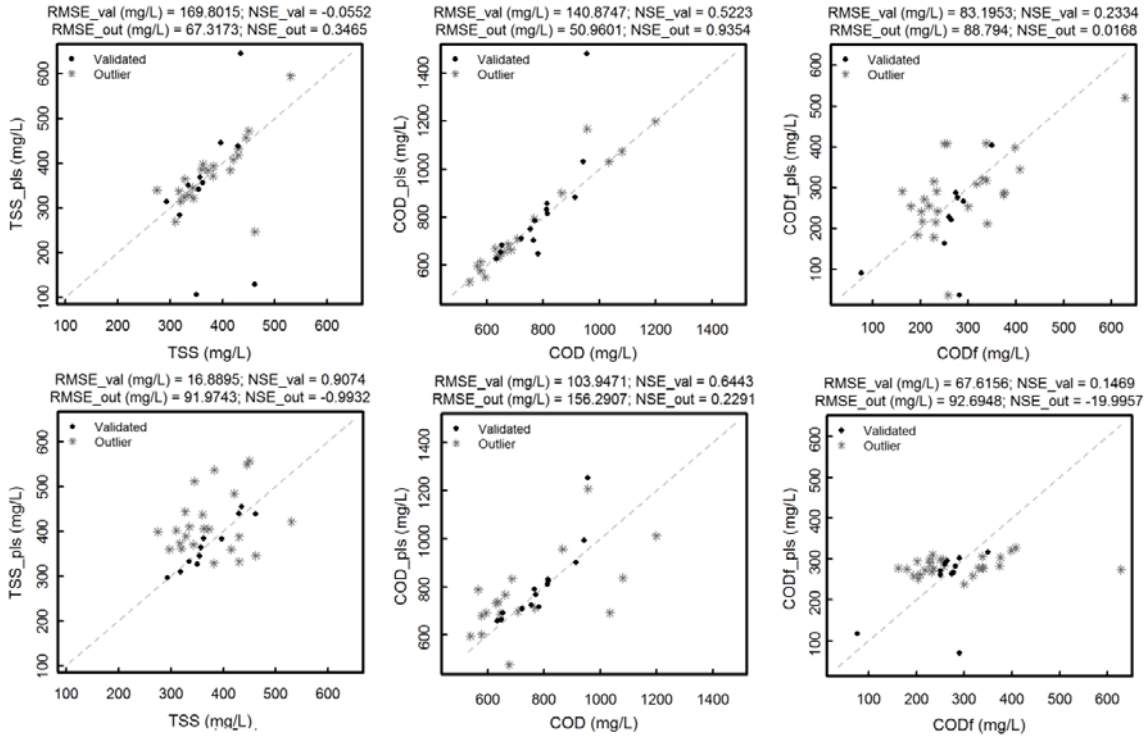

Figura 4: Comparación entre las métricas de ajuste *RMSE* y NSC calculadas para la validación de los modelos *PLS* con (arriba) y sin *outliers* (abajo) del afluente

REFERENCIAS

Acuña E., and Rodríguez C. (2004). On Detection Of Outliers And Their Effect In Supervised Classification.

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

Chen M. S., Han J., and Yu P.S. (1996). "Data mining: an overview from a database perspective", IEEE Transactions on Knowledge and Data Engineering.

Dawson, C. W., Abrahart, R. J., and See, L. M. (2007). HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts, Environ. Modell. Software, 22(4), 1034–1052.

Fleischmann, N., Langergraber, G., Weingartner, A., Hofstaedter, F., Nusch, S., y Maurer, P. (2001) On-line and in-situ measurement of turbidity and COD in wastewater using UV/VIS spectrometry.

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.

Hofstaedter, F., Ertl, T., Langergraber, G., Lettl, W. y Weingartner, A. (2003). On-line nitrate monitoring in sewers using UV/VIS spectroscopy. In: Wanner, J., Sykora, V. (eds): Proceedings of the 5th International Conference of ACE CR "Odpadni vody – Wastewater 2003", 13–15 May 2003, Olomouc, Czech Republic, pp. 341–344.

Hur Jin, Bo-Mi Lee, Tae-Hwan Lee y Dae-Hee Park. (2010). Estimation of Biological Oxygen Demand and Chemical Oxygen Demand for Combined Sewer Systems Using Synchronous Fluorescence Spectra. Journal Sensors 10, 2460-2471.

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, 49(1), 9–14.

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

Langergraber, G., Weingartner, A., y Fleischmann, N. (2004a). Time-resolved delta spectrometry: a method to define alarm parameters from spectral data, Water science & technology, 50(11), 13–20.

Mevik, B.H., and R.Wehrens, 2007 The pls package: principal component and partial least squares regression in R. J. Stat. Softw. 18: 1–24.

Nash, J. E., and J. V. Sutcliffe. (1970). River flow forecasting through conceptual models, 1, A discussion of principles, J. Hydrol., 10, 282- 290.

Olsson G., Nielsen M., Yuan Z., Lynggaard-Jensen A., and Steyer J. P. (2004) Instrumentation, control and automation in wastewater systems. State-of-the-art book, IWA Publ., London.

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. Vol. 12, No. 5, pp. 197-200.

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/.

Rännar, S., Lindgren, F., Geladi, P. and 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.

Ruban, G., Ruperd, Y., Laveau, B. y Lucas, E. (2001). Self-monitoring of water quality in sewer systems using absorbance of ultraviolet and visible light. Water science & technology, 44(2-3), 269-276.

Seo S. (2006). A review and comparison of methods for detecting outlier in univariate data sets. PhD thesis. University of Pittsburgh, Department of Biostatistics.

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, 581–588.

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesely.

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

Winkler, S., Saracevic, E., Bertrand-Krajewski, J. L., y Torres, A. (2008). Benefits, limitations and uncertainty of in situ spectrometry. Water science and technology: a journal of the International Association on Water Pollution Research, 57(10), 1651.

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.