MAESTRÍA EN HIDROSISTEMAS

PONTIFICIA UNIVERSIDAD JAVERIANA

David Zamora

Métodos *Machine Learning* aplicados para estimar la concentración de los contaminantes de la DQO y de los SST en hidrosistemas de saneamiento urbano a partir de espectrometría UV-Visible

Director:

Andrés Torres

Bogotá D.C., 2013

MAESTRÍA EN HIDROSISTEMAS

PONTIFICIA UNIVERSIDAD JAVERIANA

David Zamora

Machine Learning Methods applied to estimate the concentration of the pollutants of COD and TSS in urban sanitation hydrosystems from UV-Visible spectrometry

Director:
Andrés Torres

Evaluadores:
Luis Alejandro Camacho Botero, Profesor Asociado, Universidad de los Andes
Nelson Obregón Neira, Profesor Titular, Pontificia Universidad Javeriana (Bogotá)
William A. Ocampo D., Profesor Asistente, Pontificia Universidad Javeriana (Cali)

Bogotá D.C., 2013

AGRADECIMIENTOS

A mi tutor quien siempre me brindó tres cosas: tiempo, conocimiento y disertación.

La Empresa de Acueducto de Bogotá (EAAB-ESP) y la Pontificia Universidad Javeriana, quienes en el marco de los Convenios de investigación y desarrollo No. 9-07-26100-1060-2008 y No. 9-07-25100-0763-2010, quienes brindaron información de la calidad del afluente de la Estación Elevadora Gibraltar.

A los ingenieros Jean-luc Bertrand-Krajewski y Mathieu Lepot del *INSA* de Lyon, Francia, quienes brindaron información de la calidad del afluente de la planta de tratamiento de aguas residuales de *Fontaines-sur-Saône*, *Grand Lyon*.

A las Empresas Públicas de Medellín en especial al ingeniero Antonio Quintero, jefe de operación de la planta de tratamiento de aguas residuales San Fernando en Medellín; quien brindo información de la calidad del afluente de dicha planta.

A CeiBA (Centro de Estudios Interdisciplinarios Básicos y Aplicados en Complejidad) quienes respaldaron mis estudios de maestría y a Colciencias quienes por medio del programa Jóvenes Investigadores e Innovadores (Virginia Gutiérrez de Pineda-510 del 2010) patrocinaron mi manutención en el periodo marzo 2011 a marzo 2012 en el desarrollo de mi investigación.

Al ingeniero Jaime Velandia gerente de la empresa ByC Biosciences quien compartió su conocimiento de espectrometría UV-Visible y de mediciones de calidad del agua *in situ* y continuo.

Al Laboratorio de Pruebas y Ensayos de la Universidad, quienes realizaron los diferentes análisis de laboratorio a las muestras de las aguas residuales del afluente de la Estación Elevadora de Gibraltar.

Al Centro de Alto Rendimiento Computacional-ZINE de la Universidad Javeriana en el especial al ingeniero Alexander Herrera, quien puso a disposición los recursos computacionales para ejecutar de forma paralela y múltiple los modelos de redes neuronales artificiales y *support vector machine*.

A María Victoria Rocha y Gloria Chacón personal administrativo del Departamento de Ingeniería Civil y Facultad de Ingeniería de la Universidad respectivamente, quienes siempre estuvieron dispuestas a colaborar en los trámites administrativos de la investigación.

A Alejandra Ruiz y María Isabel Rivero quienes realizaron conmigo las diferentes campañas de monitoreo y muestreo.

RESUMEN

El conocimiento de la concentración de determinantes (*i.e.* contaminantes) de calidad del agua representa un insumo fundamental para el desarrollo de la gestión de los sistemas de saneamiento urbano (SSU). Sin embargo, hasta hace relativamente poco tiempo, las concentraciones de Sólidos Suspendidos Totales (SST) y Demanda Química de Oxígeno (DQO) eran estimadas a partir de análisis de laboratorio efectuados sobre muestras recolectadas *in situ*. Esta práctica presenta varios inconvenientes demostrados, entre los que se encuentra la baja representatividad espacio-temporal de los resultados. Por lo tanto, una de las alternativas posibles para reducir dichas dificultades consiste en utilizar captores instalables *in situ*, que utilizan tecnologías de medición en continuo, como la espectrometría UV-Visible. Esta tecnología reciente es capaz de proporcionar informaciones a alta frecuencia, que pueden traducirse en términos de concentraciones equivalentes de SST y DQO (total o filtrada), lo cual permite monitorear el estado de los flujos contaminantes favoreciendo la comprensión y el control sobre éstos, especialmente en la detección de perturbaciones. Dado que estos captores no proporcionan directamente valores de concentraciones de determinantes, se deben desarrollar y poner en práctica métodos específicos para evaluar las concentraciones equivalentes y sus incertidumbres.

Por lo tanto, el presente trabajo tuvo como objetivo desarrollar nuevas metodologías basadas en métodos *machine learning*, para lo cual se implementaron tres técnicas de inteligencia artificial denominadas: *Support Vector Machine* (*SVM*), Redes Neuronales Artificiales (RNA) y algoritmos evolutivos. Éste último fue empleado para realizar una optimización multiobjetivo de los parámetros *SVM* y *RNA* con el fin de estimar concentraciones equivalentes de determinates en continuo asociadas a las aguas de drenaje urbano mediante datos de espectrometría UV-visible *in situ.* Adicionalmente, para comprender mejor la relación entre el espectro de absorbancias y presencia-magnitud de los determinantes objeto de estudio (SST y DQO (total o filtrada)), se desarrollaron varias metodologías que abarcan los siguientes puntos importantes para consolidar y evaluar un modelo quimiométrico, orientas a: evaluar la incertidumbre de los datos medidos *in situ* y de ensayos de laboratorio (Ley de la propagación de la incertidumbre y métodos Monte Carlo), establecer la recurrencia y la relevancia de las longitudes de onda del espectro UV-Visible en su relación con la presencia de un determinante, y por último evaluar la calidad y representatividad de un par de datos espectro-concentración (*outliers*).

Las metodologías propuestas fueron implementadas en las bases datos de espectrometría UV-Visible y concentraciones de laboratorio de tres sistemas de saneamiento: Planta de Tratamiento de Aguas Residuales (PTAR) de San Fernando en Medellín, Colombia (escenario piloto), PTAR de *Fontaines-sur-Saône* (en tiempo seco y lluvia), Francia, y la estación elevadora de aguas residuales Gibraltar (EEG) en Bogotá, Colombia.

En general los resultados alcanzados de la implementación de las metodologías propuestas usando métodos *machine learning* fueron relativamente buenas, ya que la capacidad y desempeño de estos métodos depende directamente de una gran cantidad de datos y de lo representativos que éstos sean de un amplio universo del fenómeno estudiado. En particular, esto se evidenció en los conjuntos de datos del afluente de la PTAR *Fontaines-sur-Saône* en tiempo seco (94 muestras), donde los resultados presentaron bajos errores entre los valores observados con incertidumbre y las estimaciones obtenidas por los modelos *SVM*, tanto en la etapa de calibración como de validación. Por otra parte, las estimaciones de las concentraciones equivalentes obtenidas por medio de *SVM* para los conjuntos de datos del afluente de la EEG (41 muestras), fueron pobres en múltiples modelos ejecutados con diferentes arquitecturas (número de predictores y valor de los parámetros del modelo). Dichos resultados están probablemente asociados a la baja calidad de los datos de laboratorio (SST, DQO y DQOf) y a la selección del paso de luz de la sonda empleada para medición del espectro de UV-Vis, esto último afecta notablemente la relación entre la matriz de compuestos y el haz de luz de instrumento de medición, lo cual se representa a través de la absorbancia que define la presencia de un determinante.

Por otra parte, los resultados alcanzados para las muestras del afluente de la PTAR *Fontaines-sur-Saône* en tiempo de lluvia (44 muestras), evidenciaron resultados satisfactorios para ciertos rangos de concentración en el caso de los SST y de la DQOf, y mejores en el caso de la DQO para todo el rango de concentración en el cual fue evaluado este determinante. Esto puede demostrar que el modelo *SVM* no depende del número de datos empleados para generar buenos resultados, pero sí de su calidad y representatividad de la variabilidad del determinante analizado. Además, Los resultados fueron comparados con un modelo quimiométrico ampliamente conocido como mínimos cuadrados parciales (*Partial Least Squares-PLS*).

Finalmente, el modelo *feed-forward* de RNA empleado en la investigación mantuvo una tendencia a presentar sobreajuste en las calibraciones y altos errores en las etapas de validación de los modelos evaluados con diferentes arquitecturas (entradas, salidas, neuronas en la capa oculta, *etc*.).

Los conocimientos adquiridos proporcionan información para otros investigadores e ingenieros en la estimación de contaminantes por medio de tecnologías de medición *in situ* y en continuo de la calidad de agua y también puede ser la base para un posible control futuro en tiempo real de los sistemas de saneamiento urbano.

ABSTRACT

The knowledge of the pollutant concentration values represents a significant input to the improvement in the management of the urban sanitation systems (USS). However, a short while ago, the concentration of Total Suspended Solids (TSS) and Chemical Oxygen Demand (COD) were estimated by means of laboratory analysis carried out with *in situ* collected samples. This method has some proven drawbacks, such as the low time-space representativeness of the outcome. Therefore, one of the possible alternatives to reduce the drawbacks mentioned before is the use of installable sensors *in situ*. These sensors use on line measurement technology such as the UV-Visible spectrometry. This new technology is able to provide information in high frequency in terms of equivalent concentrations of TSS and COD (completed of filtered). This, in turn, allows the monitoring of quality of a water system providing benefits related with the understanding and the control of it, especially perturbation detection. Since these sensors do not provide direct pollutant concentration values, there has to be developed and implemented new methods to assess the equivalent concentrations and their uncertainties.

Henceforth, the main objective of this research was to develop new methods based on machine learning methods. Three artificial intelligence techniques were developed: support vector machine (SVM), artificial neural networks (ANN) and evolutionary algorithms (EA). The last one was used to carry out a multi-object optimization of the SVM and ANN parameters to estimate equivalent concentrations of pollutants associated with wastewaters through *in situ* UV-Vis spectrometers. Additionally, to have a better understanding of the relation between absorbance spectrum and magnitudes-presence of the target pollutant (TSS and COD (total and filtered)) several methods were developed which covered the following important topics to consolidate and assess a chemometric model: assess the uncertainty of the *in situ* collected data and the laboratory analysis (uncertainty propagation law and Monte Carlo method), determine the recurrence and relevance of the UV-Vis fingerprint related to the presence of the pollutants, and finally to evaluate the quality and representativeness of couples of concentration-fingerprint data (outliers).

The presented methods were implemented on the data set of UV-Vis spectroscopy and laboratory concentrations in three sanitation systems: Waste Water Treatment Plant (WWTP) San Fernando in Medellín, Colombia (pilot scenario) WWTP of *Fontaines-sur-Saône* (dry and wet), France, and the Pumping Station Gibraltar (PSG) in Bogotá, Colombia.

In general, the outcomes reached after the implementation of the presented methods using machine learning methods were relatively good, because the capacity and performance of such methods depends directly on the amount of data and the representativeness it might have in a broad universe of the study phenomenon. Therefore, this was observed in the data set of the tributary of the WWTP *Fontaines-sur-*

Saône in dry weather (94 samples), where the results showed low error rate between observed values with uncertainty and the obtained estimations through the SVM technique; both in the calibration and validation stage. On the other hand, estimations on the obtained equivalent concentrations through SVM technique for the data set of the tributary of PSG (41 samples) and the tributary of the WWTP *Fontaines-sur- Saône* in wet weather (41 samples) showed a significant variance in different models on the calibration adjustment which had a negative impact in the generation of error rates significantly high for the data set implemented in the validation stage. Besides, the results were compared to a chemometric model known as the partial least squares (PLS).

Finally, the feed-forward model in the ANN technique implemented in the research maintained a tendency in presenting an overfitting in the calibrations and high error rates in the validation stages of the tested models with different shapes (input, output, neuron in the hidden layer, etc.)

The acquired knowledge provides valuable information to other researchers and engineers in the estimation of pollutants by means of *in situ* and on line water quality measurement technologies and also this research could be the base of a possible future real-time control of the urban sanitation system.

TABLA DE CONTENIDO

ÍNDICE DE ECUACIONES

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

La rápida urbanización trae consigo varios retos relacionados con los problemas de calidad de agua y saneamiento. Los principales avances en el uso de instalaciones mejoradas de saneamiento en las últimas décadas se ven socavados por el rápido crecimiento de la población urbana. Hoy, 789 millones de habitantes urbanos viven sin acceso a instalaciones mejoradas de saneamiento (ONU, 2010).

En la mayoría de los países de ingresos bajos y medios las aguas residuales se vierten directamente a los ríos o al mar sin tratamiento alguno. Muchas de las grandes ciudades, como Bogotá, no tienen plantas de tratamiento o las plantas existentes se revelan rápidamente como insuficientes debido a que la población urbana supera el crecimiento de las inversiones. La descarga de aguas residuales no tratadas ocasiona problemas a las zonas situadas río abajo. La buena gestión de las aguas residuales puede, en vez de ser una fuente de problemas, ser una cuestión positiva para el medio ambiente y conducir a mejorar el desarrollo económico (ONU, 2010).

La contaminación del agua está, a pesar de las mejoras en algunas regiones, en aumento a nivel mundial, razón por la cual hay una mayor persistencia e incremento de las sustancias que afectan la calidad del agua. Estas sustancias afectan la composición química, física y biológica del agua, los cuales de acuerdo a su magnitud pueden ser nocivos para la flora y fauna de un medio acuático, así como para los asentamientos humanos que se abastezcan de esa fuente y para el ambiente circundante (ONU, 2010).

Aunque se hagan progresos sustanciales en la regulación y la implementación, se espera que aumente la contaminación como consecuencia del desarrollo económico impulsado por la urbanización, las industrias y los sistemas de agricultura intensiva. La contaminación del agua generada por el hombre es una amenaza grave para la salud humana y del ecosistema, pero su impacto es difícil de cuantificar. Los asentamientos urbanos son el principal causante de la contaminación de las fuentes de agua (ONU, 2010).

Entonces, la implementación exitosa de estrategias para mejorar la calidad del medio ambiente acuático dependerá de la disponibilidad y confiabilidad de la información de la calidad del agua que las entidades encargadas de la gestión y manejo del recurso hídrico tengan. Por esta razón estas entidades han venido caracterizando la composición del agua en diferentes hidrosistemas (*e.g.* redes de alcantarillado residual y pluvial, ríos, acuíferos, *etc.*). Por lo general, los determinantes (*i.e.* contaminantes) se han clasificado sobre la base de las propiedades físicas y químicas, sus fuentes y sus efectos típicos (Fletcher *et al*., 2013).

Por lo tanto, en la década de 1970, se propusieron analizadores en línea para la medición remota aplicados a procesos industriales. Los primeros instrumentos de medición de COT,

DQO y nutrientes fueron adaptados de los instrumentos empleados en los procedimientos para el laboratorio, y por lo tanto no tenían la posibilidad de un registro automático, y sin verificación en continuo. Esta es la razón por la cual el éxito de estos dispositivos es limitado, sobre todo la obstrucción en las muestras en línea y problemas electrónicos. Sin embargo, los sensores diseñados para el control de procesos como caudalímetros, oxímetros o detectores de manto de lodo fueron bien aceptados y su uso fue popularizado. En la década de 1980 se desarrollaron otros sensores, tales como los sistemas multiparamétricos (para la medición de la temperatura, pH, conductividad y oxígeno), y turbidímetros para la medición de la turbidez en el agua y para la estimación de sólidos suspendidos de las aguas residuales, con un éxito limitado para este último uso. Luego, desde la década de 1990, una gran cantidad de nuevos métodos y dispositivos se han diseñado para el monitoreo de la calidad de aguas residuales *on-site/on-line* (Bourgeois *et al*., 2001; Vanrolleghem y Lee, 2003).

Por consiguiente, la medición de la concentración de los determinantes en continuo resulta una alternativa interesante para el conocimiento de los fenómenos asociados a dinámicas de flujos de determinantes a diferentes escalas de tiempo. Por lo tanto, estimar de forma fiable y a través de tecnologías *in situ* la evolución temporal de diferentes determinantes de calidad permite monitorear el estado de los diferentes Hidrosistemas de Saneamiento Urbano (HSU), así como detectar el impacto de las aguas lluvias o vertimientos clandestinos que pueden afectar el medio ambiente y/o modificar la eficiencia del tratamiento de las aguas residuales (Ruiz *et al*., 2011). No obstante, hasta hace relativamente poco tiempo, las concentraciones de Sólidos Suspendidos Totales y de Demanda Química de Oxígeno 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 al análisis de las muestras en laboratorio, sólo es posible recolectar un número relativamente pequeño de muestras durante periodos prolongados de tiempo, el transporte de las muestras del lugar de recolección al laboratorio, almacenamiento y conservación de las mismas y los plazos prolongados para la obtención de resultados (Winkler *et al.*, 2008). Los resultados obtenidos bajo esas condiciones no pueden proporcionar una información precisa, suficientemente representativa de la dinámica de los fenómenos ni completa para estimar de manera fiable los flujos de determinantes a diferentes escalas de tiempo.

Una de las técnicas más recientes de medición en continuo, que permite reducir dichos 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 ultravioleta hasta el visible. Dichos captores son capaces de proporcionar informaciones del orden de una medición por minuto, que pueden traducirse en términos de concentraciones equivalentes de SST y DQO. 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 deltas de longitud de onda de 2.5 nm, y otorga resultados en tiempo real (Langergraber *et al.*, 2004; 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).

La mayoría de experiencias utilizando estos captores reportados en la literatura utilizan la calibración inicial (Staubmann *et al*., 2001; Winkler *et al*., 2002; Fleischmann *et al*., 2002), o en el mejor de los casos curvas de calibración entre los resultados proporcionados por el captor y aquellos obtenidos en el laboratorio para muestras puntuales y siguiendo protocolos estándar de laboratorio. Una de las razones que podrían explicar lo que se menciona anteriormente es la dificultad numérica para el tratamiento de espectros, al tener un número elevado de valores de absorbancia (entre 200 y 750 nm, con pasos de 2.5 nm: 220 valores de absorbancia) para cada muestra (Langergraber et al., 2003). Adicionalmente, aún no se han desarrollado, excepto de manera preliminar (Langergraber et al., 2003; Torres y Bertrand-Krajewski, 2008), métodos específicos de explotación y análisis de los datos proporcionados por estos captores con el fin de: (i) calibrar los espectros proporcionados con el fin de interpretarlos de manera precisa en términos de concentraciones de determinantes; (ii) analizar las series de tiempo de las concentraciones obtenidas con el sensor (una medición por minuto, es decir 1440 mediciones de concentraciones por día); (iii) evaluar la precisión y las incertidumbres asociadas a las mediciones por espectrometría UV-visible *in situ* en HSU en operación; (iv) determinar los requerimientos operacionales y de mantenimiento para su uso apropiado en HSU. Adicionalmente, la mayoría de métodos de calibración se basan en métodos estadísticos comúnmente utilizados en quimiometría como *PLS (*en inglés *Partial Least Squares)* (Lorenz *et al*., 2002; Rieger *et al*., 2004; Torres y J. L. Bertrand-Krajewski, 2008; Sutherland-Stacey *et al*., 2008), y son casi nulas las experiencias reportadas con métodos *machine learning*. Sin embargo, análisis exploratorio con redes neuronales (Zamora *et al.,* 2010) han mostrado que dichos métodos podrían otorgar resultados interesantes, en particular para el caso de las aguas residuales, caso en el que la ley de Beer-Lambert no es válida (Thomas y Burgess, 2007) y se presentan fenómenos como sensibilidad cruzada (Fleischmann *et al*., 2001; Langergraber *et al*., 2004a), la no linealidad entre valores de absorbancia y los valores de las concentraciones de determinantes (DiFoggio, 2000).

Por lo tanto, el presente proyecto pretende contribuir en el desarrollo de dos metodologías basadas en métodos m*achine learning* para estimar concentraciones de contaminantes en continuo asociadas a las aguas de HSU mediante datos de espectrometría UV-visible *in situ*.

El presente documento está conformado por las siguientes secciones: (i) Introducción, que justifica y explica brevemente el enfoque y el alcance de la presente investigación, (ii) definición y explicación de conceptos y teorías del fundamento de la investigación, así como la revisión del estado del arte de la presente tesis, (iii) materiales y métodos, una

descripción de los casos estudio y recopilación de datos, espectrofotómetros utilizados y su funcionamiento, ensayos de laboratorio y métodos de análisis, (iv) metodologías desarrolladas para abordar el problema objeto de investigación (incertidumbre, detección de *outliers,* métodos regresivos, algoritmos de evaluación), (v) los resultados obtenidos de la aplicación las metodologías propuestas a los datos experimentales, y (vi) las conclusiones y perspectivas sobre los resultados obtenidos.

1. MARCO TEÓRICO

1.1. DETERMINANTES EN SISTEMAS DE DRENAJE URBANO (SDU)

En el caso puntual de los sistemas de saneamiento urbano, se presenta en la Figura 1 el funcionamiento de una red de alcantarillado combinado y un sistema de alcantarillado separado para las aguas de origen sanitario y pluvial. En un sistema de alcantarillado combinado se transporta por una misma tubería las aguas residuales de origen sanitario y lluvia, incluso aguas de origen industrial, lo cual genera una mezcla de compuestos y aumento en el caudal trasportado. Como la capacidad de las plantas de tratamiento de aguas residuales (PTAR) es limitada para garantizar su adecuado funcionamiento, muchas veces esta capacidad es superada por las condiciones del alcantarillado combinado, de ahí que este tipo de alcantarillado tiene que ser manejado de otra manera para mitigar impactos negativos en los procesos de tratamiento (Gamerith, 2011).

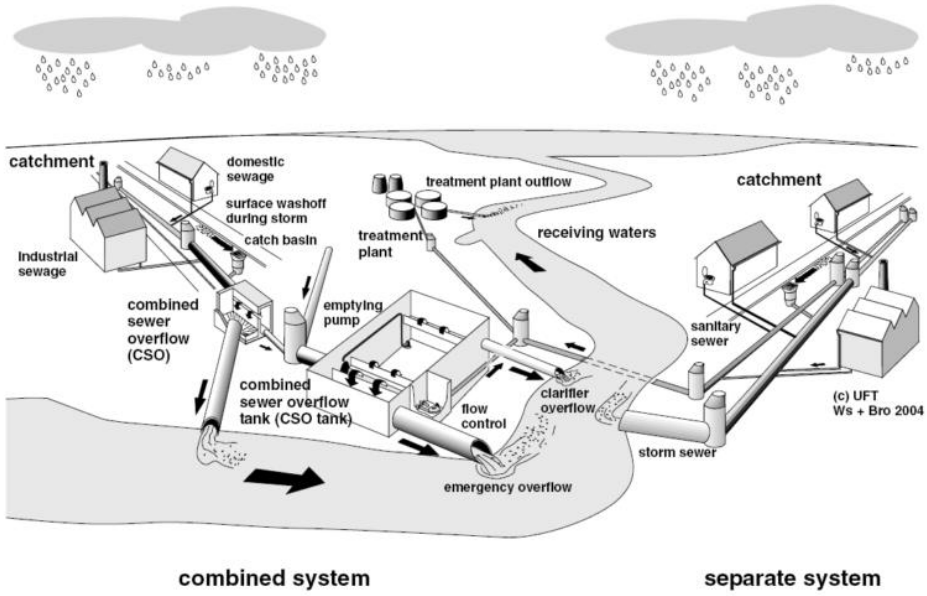


Figura 1- Comparación entre un sistema de alcantarillado combinado y separado (Brombach *et al*., 2005)

Sin embargo, en las últimas décadas mejores prácticas han permitido reducir el ingreso de aguas lluvias a las redes sanitarias, tales como sistemas independientes para el transporte de aguas lluvias. No obstante, en zonas densamente urbanizadas o de espacio limitado para la infiltración o retención del agua lluvia, ésta sólo se puede ser almacenada o desbordarse de las redes sanitarias por medio de estructuras de control. Por lo tanto, los contaminantes vertidos tienen un impacto importante en la calidad de las aguas receptoras, sobre todo el ecosistema y el medio ambiente acuático (Gamerith, 2011).

1.1.1. Principales determinantes de calidad del agua en sistemas combinados de alcantarillado

El agua residual está integrada por componentes físicos, químicos y biológicos. Es una mezcla de materiales orgánicos e inorgánicos, suspendidos o disueltos en el agua. La mayor parte de la materia orgánica consiste en residuos alimenticios, heces, material vegetal, sales minerales, materiales orgánicos y materiales diversos como jabones y detergentes sintéticos. Las proteínas son el principal componente del organismo animal, pero también están presentes en los vegetales. El gas sulfuro de hidrógeno presente en las aguas residuales proviene del Azufre de las proteínas.

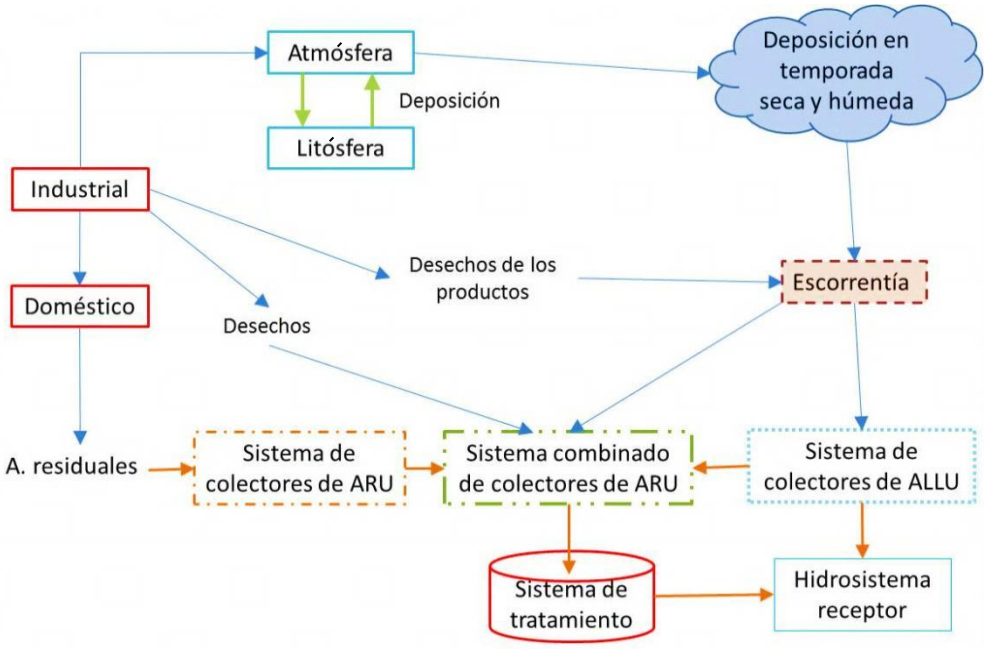


Figura 2- Fuentes de los compuestos y sustancias presentes en aguas residuales

En las aguas lluvias, los determinantes (*i.e.* contaminantes, sustancias, compuestos, elementos) se acumulan durante todo el proceso de la precipitación y la escorrentía en las superficies y en las redes de alcantarillado. En la atmósfera las gotas de agua acumulan los compuestos o elementos suspendidos (aerosoles). En la superficie estos compuestos son depositados y movilizados por i) el impacto de la gota de agua, y ii) la escorrentía superficial (ver Figura 2). Además, se pueden presentar posibles depósitos en el sistema de alcantarillado generados por pequeñas descargas en periodos secos y acumulables en el tiempo. Estos depósitos pueden ser resuspendidos o movilizados si una nueva descarga en el alcantarillo es lo suficientemente alta (Burton y Pitt, 2002; Hochedlinger, 2005; Gamerith, 2011). Entonces, durante los eventos de precipitación, los contaminantes en sistemas combinados se originan a partir de: i) las aguas residuales sanitarias (doméstica y residual) y ii) agua de lluvia.

Varios estudios exhaustivos se han llevado a cabo para caracterizar los y determinantes en los sistemas de alcantarillado. En los Estados Unidos (EE.UU.) durante las décadas de 1970

y 1980 se realizó un trabajo intensivo en esta materia, destacando a Pitt y Amy (1973), Sartor *et al.* (1974) y la EPA de EE.UU. (1983). Los principales hallazgos fueron recopilados recientemente en una base de datos de Maestre y Pitt (2005), quienes hicieron seguimiento durante 10 años a las caracterizaciones de la calidad de las aguas lluvias en 200 municipios de EE.UU. Una descripción completa de los determinantes y las concentraciones basada en datos de Alemania se puede encontrar en Brombach *et al.* (2005).

Las principales características de los determinantes presentes en las aguas residuales son:

- *Sustancias químicas:* Se encuentran presentes en sólidos en suspensión y disueltos, que pueden ser orgánicos e inorgánicos. Los sólidos inorgánicos están formados principalmente por nitrógeno, fósforo, cloruros, sulfatos, entre otros y algunas sustancias tóxicas como metales pesados, tales como arsénico, cianuro, cadmio, cromo y mercurio. Los sólidos orgánicos se pueden clasificar en nitrogenados y no nitrogenados. Los nitrogenados, es decir, los que contienen nitrógeno en su molécula, son proteínas, ureas, aminas y aminoácidos. Los no nitrogenados son principalmente grasas y jabones (Butler and Davies, 2011).
- *Características bacteriológicas:* Las aguas residuales contienen numerosos microorganismos entre los cuales se encuentran los agentes patógenos (e.g. virus, bacterias, parásitos). Por ejemplo en 1 g de heces de un enfermo de hepatitis existen entre 10 y 106 dosis infecciosas del virus de esta enfermedad. Por otra parte, el tracto intestinal del hombre contiene numerosas bacterias conocidas como organismos coliformes. Cada individuo evacúa de 10^5 y $4x10^5$ millones de coliformes por día, que aunque no son dañinos, se utilizan como indicadores de contaminación debido a que su presencia indica la posibilidad de que existan gérmenes patógenos de más difícil detección (Muttamara, 1996).
- *Materia en suspensión y materia disuelta:* La materia en suspensión es aquella que por su propio peso tiende a sedimentarse en el tiempo, cuya separación de la fracción líquida puede ser llevada por gravedad. Por otra parte, la materia disuelta puede ser tanto orgánica como inorgánica y su interacción con las moléculas de agua pueden generar compuestos peligrosos y los procesos de separación son más complejos y costosos (Mujeriego *et al*., 1982).

Los principales determinantes analizados que generalmente se incluyen en la caracterización de aguas residuales son:

- Materia orgánica: Demanda Biológica de Oxígeno (DBO_n), Demanda Química de Oxígeno (DQO) y Carbón Orgánico Total (TOC).
- Sólidos Suspendidos Totales (SST).
- Nitrógeno (NH_4-N: Amonio, NO_3: Nitrato, Nitrógeno Total)
- Fósforo
- Metales pesados como: Cadmio, Cromo, Plomo, Níquel, Cobre, Mercurio, Zinc, *etc*.

Descripciones exhaustivas de estos determinantes se pueden encontrar en la literatura básica de drenaje urbano como por ejemplo Butler y Davies (2011) y Gujer (2008). Este trabajo se centra en el análisis de tres determinantes: Sólidos Suspendidos Totales (SST), Demanda Química de Oxígeno bruta (DQO) y filtrada (DQOf), los cuales serán explicados a continuación.

1.1.2. Sólidos Suspendidos Totales

Los tipos de sólidos de interés en las aguas residuales y pluviales en términos generales se pueden clasificar en cuatro categorías: gruesos, arena, suspendidos y disueltos (ver Tabla 1). Los sólidos gruesos y suspendidos pueden subdividirse de acuerdo a su origen: aguas residuales y pluviales (Butler y Davies, 2011).

Tipo de sólido	Tamaño (μm)	SG (–)
Grueso	>6000	0.9 – 1.2
Arena	>150	2.6
Suspendido	≥0.45	1.4 – 2.0
Disuelto	<0.45	–

Tabla 1- Clasificación básica de sólidos (Butler y Davies, 2011)

Los sólidos suspendidos totales relacionan la presencia de partículas de materia orgánica o de minerales en el agua. Luego, en el caso de las aguas residuales la mayor parte de los sólidos suspendidos provienen de los desechos humanos, desperdicios de alimentos, desechos industriales y células biológicas que forman una masa de partículas suspendidas, por efecto de absorción de las partículas de materiales inertes (*e.g.* minerales) presentes en la agua. Existen, además otras fuentes de sólidos en suspensión, tales como la erosión del suelo a partir de un sitio de construcción y de actividades agrícolas, que por efecto de la escorrentía superficial son arrastrados a las redes de drenaje (Muttamara, 1996).

Por lo tanto, en un cuerpo de agua donde la presencia de los SST aumenta, la capacidad para soportar una diversidad de vida acuática disminuye. Los sólidos en suspensión absorben el calor de la luz solar, lo que aumenta la temperatura del agua y, posteriormente, disminuye los niveles de oxígeno disuelto (el agua a mayor temperatura contiene menos oxígeno que el agua a menor temperatura). Ciertas especies de aguas frías, como la trucha y plecóptera, son especialmente sensibles a los cambios en el oxígeno disuelto. La fotosíntesis disminuye también, ya que menos luz penetra en el agua, y en consecuencia menos oxígeno es producido por las plantas y algas. Los SST también pueden destruir el hábitat de los peces debido a que los sólidos en suspensión se depositan en el fondo y puede eventualmente cubren el lecho del río. Los sólidos en suspensión pueden sofocar los huevos de peces e insectos acuáticos y pueden sofocar las larvas recién nacidas de insectos. Los sólidos en suspensión también pueden dañar a los peces directamente al obstruir las branquias, reducir las tasas de crecimiento y disminuir de la resistencia a las enfermedades, así mismo los movimientos naturales y las migraciones de poblaciones acuáticas pueden ser interrumpidas (Pescod, 1992).

1.1.3. Demanda Química de Oxígeno

La Demanda Química de Oxígeno (DQO) es la cantidad de oxígeno que se requiere para oxidar químicamente el material orgánico. Difiere de la Demanda Biológica de Oxígeno en que en esta última prueba solo se detecta el material orgánico degradado biológicamente o que es biodegradable. En la determinación de DQO todo el material orgánico biodegradable y no biodegradable es químicamente oxidado por el dicromato de potasio ($Cr_2O_7K_2$) en medio ácido en la presencia de un catalizador. Para esto se emplea una mezcla de ácido sulfúrico y dicromato de potasio con iones plata como catalizador. En estas condiciones, en un tiempo de dos horas de digestión, a una temperatura de 150 °C, el Cromo (VI) pasa a estado de oxidación Cromo (III) oxidando la materia orgánica (Ecuación 1). Cuando se termina el proceso de oxidación, la concentración de compuestos orgánicos en la muestra se calcula mediante la medición de la cantidad de oxidante restante en la solución. La DQO se expresa en mg O_2/L, que indica la masa de oxígeno consumido por litro de solución (Muttamara, 1996; Gonzalez *et al*., 2009; Butler y Davies, 2011).

$$C_6H_{12}O_6 + 4Cr_2O_7{}^{-2} + 32H^+ \xrightarrow{\Delta} 6CO_2 + 8Cr^{3+} + 22H_2O$$

Ecuación 1-

La ventaja de las mediciones de DQO es que los resultados se obtienen rápidamente (2 a 3 horas), pero tienen la desventaja de que no ofrecen ninguna información de la proporción del agua residual que puede ser oxidada por las bacterias ni de la velocidad del proceso de biooxidación, sumado a las interferencias por la presencia de sustancias inorgánicas susceptibles de ser oxidadas (sulfuros, sulfitos, yoduros, entre otros elementos), que también se reflejan en el resultado de la medición (Clair *et al*., 2003; Gonzalez *et al*., 2009).

Este método es aplicable a las muestras de agua provenientes diferentes tipo de fuente, tales como ríos, lagos o acuíferos, aguas residuales, aguas pluviales o agua de cualquier otra procedencia que pueda contener una cantidad apreciable de materia orgánica. La DQO varía en función de las características de las materias presentes, de sus proporciones respectivas, de sus posibilidades de oxidación y de otras variables. Es por esto que la reproductividad de los resultados y su interpretación no pueden ser satisfechos más que en condiciones de metodología de ensayo bien definidas y estrictamente respetadas (Clair *et al*., 2003).

1.1.4. Demanda Química de Oxígeno filtrada

Ciertos componentes de la DQO del agua residual no son solubles. Adicionalmente, las células producidas por la degradación anaeróbica de la DQO total son insolubles (ver numeral 1.1.3). La Figura 3 ilustra las categorías de DQO basadas en la solubilidad y el tamaño de las partículas. Normalmente, la muestra es filtrada a través de un papel filtro;

la parte de la muestra que es retenida por el papel filtro contiene los sólidos suspendidos y la parte que pasa el filtro es la DQO filtrada (Field, 1987).

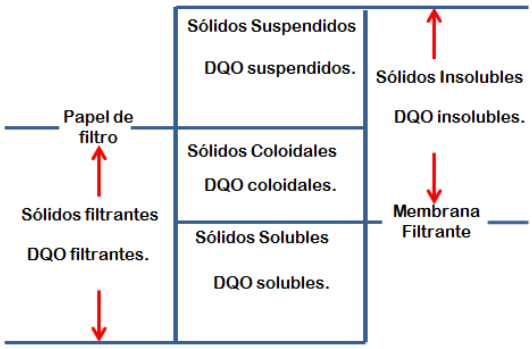


Figura 3- Clasificación de la DQO basada en la solubilidad y filtración (Field, 1987)

La DQO filtrada contiene compuestos insolubles, principalmente de forma coloidal y compuestos solubles. Estos últimos representan el sustrato que realmente los microrganismos no están consumiendo en el agua residual, como por ejemplo Manganeso, Hierro, Plomo, contaminantes emergentes, entre otros y cuyos tamaños de partículas son inferiores a 0.1 µm (Clair *et al*., 2003; Quevauviller *et al*., 2006)

En el ANEXO E-1 se presenta una descripción detallada de la actividad de muestreo y análisis de laboratorio para la caracterización de la calidad del agua.

1.2. MEDICIONES EN CONTINUO

El principal procedimiento de control de la calidad de aguas residuales se basa en los siguientes pasos: i) muestreo (puntual o compuesto), ii) conservación y almacenamiento de la muestras (por lo general a baja temperatura), iii) transporte, y iv) análisis de laboratorio. Para esto existen varias normas que definen los diferentes pasos que podrían garantizar la fiabilidad de los resultados obtenidos de una caracterización, pero existen límites o inconvenientes, con respecto a algunos modalidades de monitoreo. Las modalidades de monitoreo en aguas residuales pueden ser numerosas, pero principalmente existen tres que abarcan las demás: control y vigilancia para evaluar los cambios a largo plazo, vigilancia operativa sobre las masas de agua para proporcionar datos adicionales que permitan evaluar riesgo o el no cumplimiento de los objetivos ambientales y monitoreo para determinar las causas desconocidas que generan incumplimiento (Quevauviller *et al*., 2006). En aguas residuales, las dos últimas modalidades son las más usadas, pero éstas conducen a las siguientes limitaciones:

- El primer límite está relacionado con el retraso en el procedimiento, desde el muestreo a los resultados. En general, y dependiendo del tipo de análisis, se requiere un retraso de al menos 1 ó 2 semanas para los resultados. Este retraso se puede acortar si es necesario, pero con esto se genera un aumento en el costo del

análisis de laboratorio. Aun así, un retraso de varios días puede ser problemático en algunos casos (*e.g.* control operativo e investigación).

- La información proporcionada por una muestrea en un instante de tiempo podría o no estar asociada a fluctuaciones generadas por eventos episódicos o bien llevar a conclusiones elaboradas sobre la base de errores de medición. Luego, el costo de la información incorrecta puede ser muy alto y no hay, por lo tanto, una necesidad de mejorar las metodologías de detección en laboratorio (Gonzalez *et al*., 2009).
- El tercer límite es la relevancia de los resultados en relación con los objetivos de seguimiento. Los determinantes de calidad, ya sean agregados (*e.g.* DBO, DQO, toxicidad, *etc.*) o específicos (carbono orgánico total (TOC), las formas de nitrógeno, compuestos orgánicos, *etc*.) no pueden ser analizados para cada muestra, ya que el costo sería muy alto y de igual forma las jornadas de monitoreo. Además, el número resultados estaría limitado a la capacidad operativa del laboratorio (Quevauviller *et al*., 2006).

Por lo tanto, en la actualidad existe la necesidad del desarrollo y la validación de tecnologías rentables y metodologías que puedan ser ampliamente adoptadas para el control y gestión constante de las aguas residuales y en general de todo tipo de aguas, los cuales puedan detectar contaminantes que se encuentran en rastros o ultra-traza (µg/L o pg/L – conocidos como contaminantes emergentes) (Bester y Schäfer, 2009; Becouze *et al*., 2011).

Sin embargo, las herramientas de monitoreo en continuo serán útiles sólo si son asequibles, fiables y capaces de producir datos de calidad comparables entre tiempos y lugares. Soluciones disponibles existen comercialmente a bajo costo y gran flexibilidad en su uso, algunos de estos son: kits de prueba de campo para contaminantes específicos, equipos portátiles de análisis toxicológicos, una amplia gama de sensores, ensayos toxicológicos directos, y muestreadores más sofisticados.

1.2.1. Medición en continuo de la calidad del agua residual

Existen dos formas de llevar a cabo el monitoreo en continuo *in situ*. La primera de ellas es embeber los instrumentos de medición en la muestra en línea (cuerpo superficial, alcantarillado, acuíferos, *etc.*) y registrar la información la cual puede ser almacenada o transmitida en 'tiempo real'; a esta forma de monitoreo en continuo se le conoce como *in situ/on line.* La segunda forma es capturar la muestra en línea, por ejemplo por medio de un sistema de bombeo constante se envía la muestra (muestreo) a un tanque donde los instrumentos realizan la medición; a esta forma de monitoreo en se le conoce como *in situ*/*off line* (ver Figura 4). Esta última forma de monitoreo es muy frecuente en los sistemas que transportan agua residual o lluvia, ya que las basuras, sedimentos, hojas, entre otras generan frecuentemente obstrucciones en los instrumentos de medición, lo cual se reduce en cierta medida con esta forma de monitoreo. Por último, existe otra forma de medir en 'continuo' denominada *in situ*/*on field* en la cual se toma de forma manual una muestra puntual y se miden los determinantes de interés sobre ésta, pero la

frecuencia de medición será menor si se compara con las anteriores, ya que los costos del monitoreo incrementarán con el número de veces que el personal técnico tome una muestra (adaptado de Gonzalez *et al*., 2009). En la Tabla 2 se presenta un cuadro comparativo entre las diferentes formas de realizar monitoreo en continuo, donde significativamente los monitoreos *on line* y *off line* permiten la caracterización de la calidad del agua con mayor frecuencia y obtener resultados en un menor tiempo respuesta, incluso reduce los errores en la medida asociados a la actividad de muestreo especialmente en la medición *on line*.

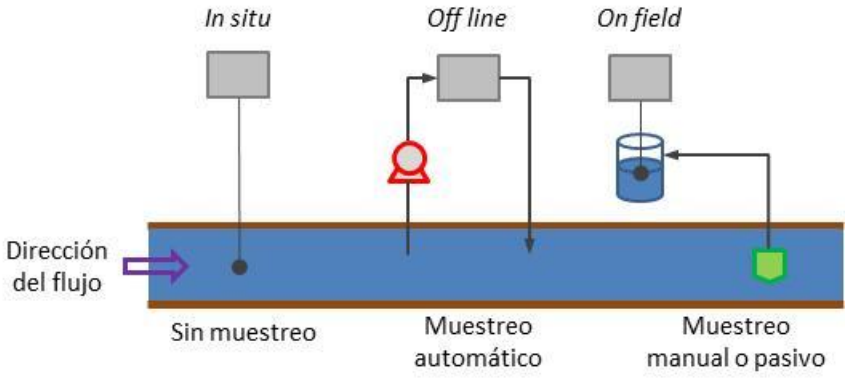


Figura 4- Tipos de muestreo y monitoreo (adaptado de González *et al*., 2009)

	Técnicas clásicas	Técnicas emergentes		
	Métodos de laboratorio	*In situ*	*Off-line*	*On-field*
Cuantificación	+ + +	+ +	+ +	+ +
Identificación	+ + +	+ +	+	+
Muestreo	M	Ninguno	MA	M
Frecuencia de medición	– –	+ + +	+ + +	+ + +
Velocidad de respuesta	– – –	+ + +	+ + +	+ + +
Sensibilidad	+ + +	_	+/-	+/-
Especificidad	+ + +	+ +	+/-	+/-

M= muestreo in situ, MA= muestreo automático

Tabla 2- Comparación de propiedades entre las técnicas de monitoreo clásicas y emergentes (Gonzalez *et al*., 2009)

En el ANEXO E-2 se presenta una descripción detallada de las clases de monitoreo, así como los principios de análisis instrumental y equipos de monitoreo *in situ* (*on line*) para la detección y cuantificación de determinantes en aguas residuales.

1.3. ESPECTROMETRÍA ULTRAVIOLETA-VISIBLE (UV-VIS)

Debido a la importancia adquirida por la química analítica en los años recientes, las técnicas ópticas son muy útiles ya que permiten evaluar los procesos para su control en

tiempo real, requisito fundamental para mantener las condiciones óptimas de operación en términos de calidad, cantidad, costos y eficiencia de un sistema. El análisis en línea de la composición de una muestra basada en la absorción selectiva de UV-Visible se ha usado para supervisar y controlar los procesos por más de 60 años. Los criterios para el diseño de analizadores de procesos UV-Vis son algo diferentes a los de los instrumentos analizadores de laboratorio porque a menudo deben trabajar en condiciones ambientales adversas, tales como vibraciones, cambios de temperatura, ambientes corrosivos, *etc.* (Ojeda y Rojas, 2009).

En la actualidad, nuevas y mejores fuentes de luz UV y Vis, detectores, filtros de interferencia, y materiales ópticos permiten la expansión de las aplicaciones de los analizadores de UV-Vis, y en este sentido, los procesos de analizador que utilizan un detector de fotodiodos y rejillas holográficas se pueden utilizar para el análisis multicomponente, que anteriormente no era posible. En general, las medidas de absorción de radiación UV-Vis han encontrado una gran aplicación en la determinación cuantitativa de una amplia variedad de especies, tanto inorgánicas como orgánicas (Lynggaard-jensen, 1999; Ojeda y Rojas, 2009).

En el ANEXO E-3 se presenta las descricipcion del proceso espectroscópico (luz-medio), y se explica el origen de un espectro: absorción de la radiación por átomos, iones y moléculas.

1.3.1.1. Leyes para cuantificar la atenuación de la luz

En la interacción luz-medio, se pueden definir las siguientes variables que representan el proceso: I_0 es la intensidad de un haz paralelo de la radiación en una longitud de onda λ incidente y que pasa a través de una cubeta de ancho *b*, la cual contiene una solución. Haciendo caso omiso a las pérdidas por dispersión o reflexión, el haz emergente ha sido atenuado por el proceso de absorción a una intensidad *I* (Figura 5).

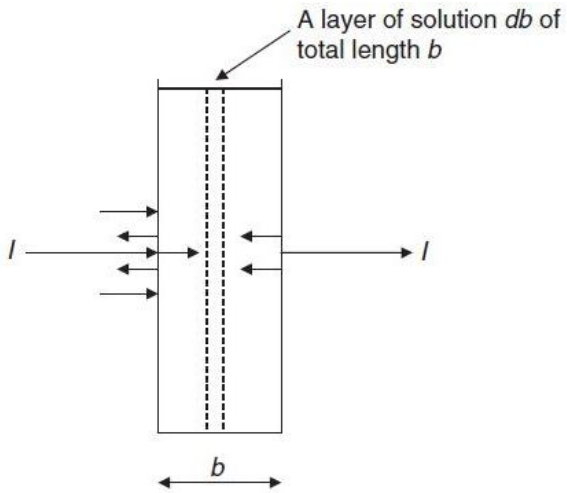


Figura 5- Atenuación de la radiación por una cubeta que contiene una solución (Thomas y Burgess, 2007)

El cambio en la intensidad del haz incidente *dI* causada por el espesor db de la solución absorbente está dada por la ley de Lambert:

$$dI = -k_{\lambda} db$$

Ecuación 2-

donde k_{λ} es una constante que depende de la longitud de onda.

De una manera similar, el cambio en la intensidad del haz incidente, *dI*, causada por el incremento de la concentración de un material absorbente, *dM*, en el espesor de la solución, *db*, está dado por la ley de Beer, donde *M* es la concentración del material absorbente en moles por dm^3:

$$dI = -k_{\lambda} dM$$

Ecuación 3-

Estas dos leyes combinadas forman la ley de Beer-Lambert, la cual establece que la absorbancia de una solución es directamente proporcional a la concentración de la solución. Por tanto, la espectrometría UV-Vis puede usarse para determinar la concentración de una solución (Thomas y Burgess, 2007).

$$A = -\log_{10}(I / I_0) = \frac{k_{\lambda}' bM}{2.303}$$

Ecuación 4-

Para cada componente y longitud de onda, $k_{\lambda}'/2.303$ es una constante conocida como absortividad molar o coeficiente de extinción. Esta constante es una propiedad fundamental molecular en un solvente dado, a una temperatura y presión particular, y tiene como unidades $1/M \cdot cm$ (Thomas y Burgess, 2007).
Luego, por definición la ley de Beer-Lambert establece que para una longitud de onda dada y un único componente, la absorbancia es una función lineal de la concentración del componente. Sin embargo, esto no sirve como relación universal para la concentración y absorción de todas las sustancias. En moléculas complejas de gran tamaño, como los tintes orgánicos (Xylenol Naranja o Rojo Neutro, por ejemplo), a veces se encuentra una relación polinómica de segundo orden entre la absorción y la concentración, incluso se requieren modelos de mayor complejidad cuando la matriz de compuestos es mayor (*e.g.* aguas residuales). Por lo tanto, la se basa en una serie de suposiciones, incluyendo (O Thomas y Burgess, 2007):

- La radiación es perfectamente monocromática.
- No hay compensación de pérdidas debido a dispersión o reflexión.
- El haz de radiación que golpea una cubeta genera incidencia normal.
- La temperatura se mantiene constante.

- Se asume que no existen interacciones moleculares entre la molécula que absorbe y las otras presentes en la solución.

Estos supuestos no siempre se cumplen y causan variaciones del comportamiento ideal de la ley Beer-Lambert, como en el caso del agua y en particular de aguas residuales monitoreadas mediante espectros UV-Visible.

La naturaleza química y la concentración de los componentes disueltos absorbentes junto con las características físicas y la concentración de material heterogéneo son los dos fenómenos responsables de la forma del espectro de UV-Visible en una muestra de agua. Por consiguiente, la espectroscopia directa implica dos fenómenos principales: el mecanismo de absorción química, que se explica por la ley de Beer-Lambert, y el efecto de dispersión y su difusión asociada a la presencia de sólidos suspendidos y coloidales (Thomas *et al.*, 1996.)

Debido a que los compuestos orgánicos (oxidables) de origen antropogénico y naturales contienen grupos cromóforos (conjunto de átomos de una molécula responsable de su calor), pueden ser detectados por espectrofotometría UV-Visible (Thomas *et al*., 1996). La región UV concentra una parte de la información espectral relevante que se puede utilizar para la caracterización de aguas residuales, como se muestra en la Figura 6.

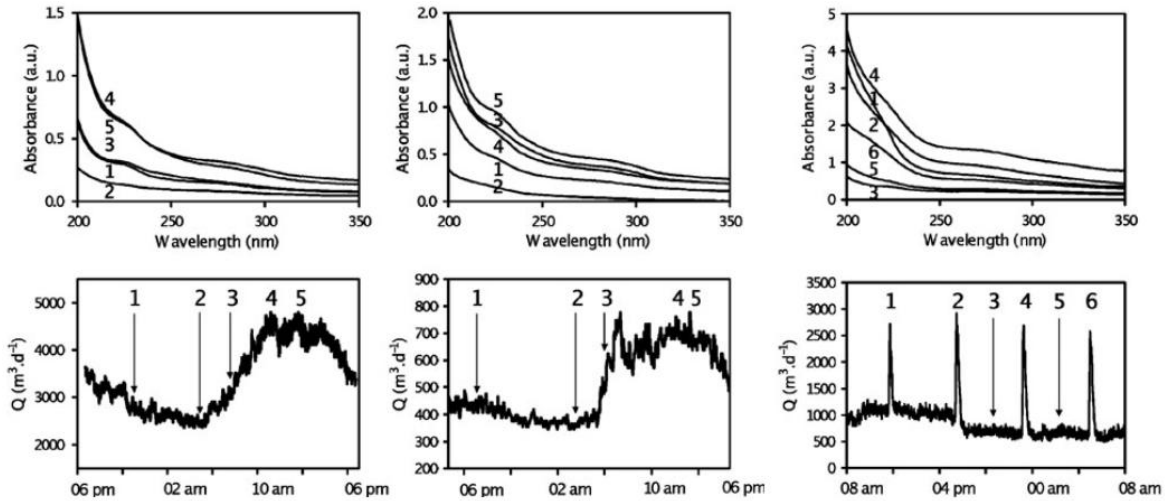


Figura 6- Relación entre caudal y espectros UV horarios obtenidos de aguas residuales de un hospital, comercio e industria de izquierda a derecha (Baurès *et al*., 2007)

En el ANEXO E-4 ANEXO Ese presenta se describen los elementos y funcionamiento de un espctrofotometro UV-Visible.

1.3.2. Aplicación de la espectrometría UV-Vis en el agua residual

Durante la década de 1980, la introducción de detectores de diodos miniatura, combinada con potentes microprocesadores y herramientas matemáticas, dio lugar a un

renacimiento de la espectrometría UV-Vis y permitió la rápida propagación de esta tecnología, que redujo el costo y aumento el uso de máquinas UV-Vis potentes en los laboratorios. Durante la década de 1990, la tecnología se trasladó del laboratorio al campo, pero todavía en forma de analizadores *off-line* relativamente grandes, complicados y costosos (Ojeda y Rojas, 2009). Sin embargo, el mercado de sensores/sondas ópticas *in situ* era todavía, y aun es, dominado por fotómetros relativamente simples, que son capaces de medir una o dos longitudes de onda a la vez. Estos instrumentos limitan la medición de una única sustancia o compuesto y, en general emplean métodos muy básicos e inestables para compensar las sensibilidades cruzadas a variaciones, por ejemplo, en una matriz de agua donde la turbiedad incrementa la sensibilidad de la técnica, la cual puede ser compensada con una segunda longitud de onda (Wu *et al*., 2006). No obstante, un mayor grado de compensación podrá ser alcanzado si se tiene la totalidad del espectro, y así reducir la sensibilidad cruzada. El desarrollo de la electrónica y la óptica en los últimos años ha permitido afianzar la relación entre el espectro UV-Vis completo y los instrumentos a pequeña escala. Por otra parte, estos desarrollos han permitido el diseño de analizadores espectroscópicos robustos que pueden funcionar en ambientes hostiles, tales como el interior del alcantarillado y vertimientos industriales, con poco o ningún mantenimiento, ya que ni productos químicos ni piezas móviles para la limpieza son necesarios para su funcionamiento.

Por lo tanto, la espectroscopia UV-Visible es una técnica analítica madura, base de varias aplicaciones establecidas. Aunque es evidente su utilidad, esta técnica es aún poco explotada en varios campos (Thomas y Burgess, 2007). Sin embargo, no es una novedad el estudio de espectroscopia UV-Visible como un método alternativo y rápido para obtener información sobre la calidad del agua y de aguas residuales (Da Silva, 2008).

La principal aplicación de la técnica es correlacionar la respuesta de UV-visible (por ejemplo absorbancia) con la sustancia o determinante a ser estimado. Teniendo en cuenta sólo espectroscopia UV, el rango de 200-300 nm se ha considerado particularmente interesante para este propósito. Por ejemplo, la absorbancia a 254 nm fue correlacionada con la DQO por Mrkva (1975) y el COT por Dobbs *et al*. (1972), para las aguas residuales municipales e industriales. Estos mismos determinantes fueron detectados por Matsche y Stumwöhrer (1996), mediante la absorbancia a 254 nm junto con la absorbancia a 350 nm, para la corrección por los sólidos en suspensión. Luego, la combinación de estas mediciones conduce a resultados fiables, que tienen una buena correlación con DQO y TOC para aguas residuales domésticas (Ojeda y Rojas, 2009).

A pesar de que es muy interesante el uso rápido y sencillo de una medición UV con una o dos longitudes de onda en lugar de un análisis de laboratorio para determinar la DQO o la DBO, los coeficientes de la ecuación deben ser calibrados para garantizar buenos resultados (Thomas *et al*., 1993). Por otra parte, un enfoque univariado se basa en el hecho de que la contaminación orgánica presente en una muestra de aguas residuales tiene un pico de absorbancia máxima. Sin embargo, este valor puede variar, dependiendo de la matriz de composición (Fogelman *et al*., 2006). Además, el fenómeno de

envejecimiento de las muestras reduce la fiabilidad de las mediciones a través de esta técnica, y por tanto puede plantear algunos problemas con respecto al cumplimiento de la regulación o el control de procesos.

1.3.2.1. Espectro de absorbancia o huella digital del agua residual

Las aguas residuales de las ciudades es una mezcla de las aguas residuales de las construcciones residenciales y oficinas, restaurantes, escuelas, hospitales, industrias, *etc*. Sus variaciones en el tiempo reflejan la actividad humana, modificadas por las condiciones climáticas. Durante muchos años la espectroscopia UV-Vis se ha propuesto como una manera alternativa y rápida para estimar el nivel de contaminación de las aguas residuales y el agua superficial. En este sentido, las muestras de aguas residuales fueron recogidas por Wu *et al*. (2006) en tres comunidades diferentes y sus huellas digitales o espectro de absorbancia, medidas por medio de la combinación de los siguientes métodos ópticos: espectrometría UV-Vis, espectrometría de fluorescencia sincrónica y turbiedad.

Los espectros, referidos como espectros de absorbancia, se obtienen de espectrómetros en línea (*on-line*) y se utilizan para la caracterización de una muestra de agua. Dentro de estos espectros de absorbancia a veces casi sin rasgos distintivos, se puede encontrar una gran cantidad de información acerca de la composición del agua. Los espectros de absorbancia se utilizan para controlar los cambios en la composición del agua a través del análisis de la forma general del espectro o de absorbancia a una longitud de onda específica (ver Figura 7). Por otra parte, se utilizan para derivar determinantes más específicos, tales como la turbidez, la concentración de nitrato, y determinantes conjugados tales como coeficiente de absorción espectral a 254 nm (SAC254), de carbono orgánico total (COT), y el carbono orgánico disuelto (COD), que son, comúnmente utilizados en el análisis de calidad de aguas (Mrkva, 1975; Rieger *et al*., 2004).

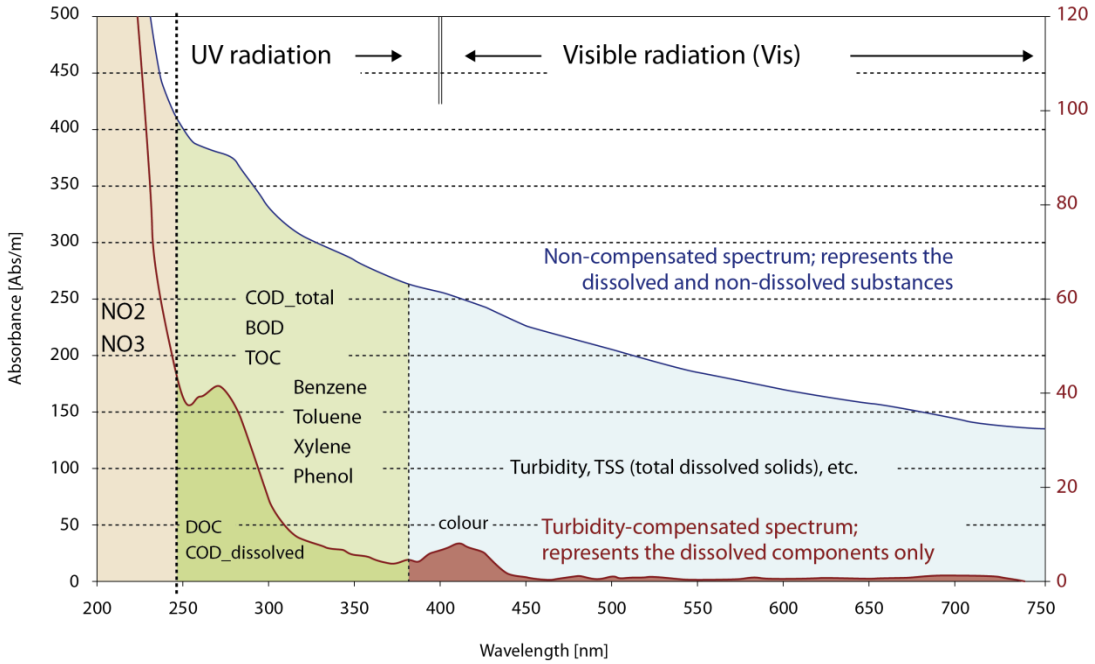


Figura 7- Detección de diferentes parámetros de monitoreo en aguas a través del rango espectral UV-Visible (s::can Messtechnik GmbH, Viena, Austria).

La turbidez, generada por las sustancias en suspensión, provoca dispersión de la luz y sombra, influyendo así en la absorbancia en todo el espectro. Este es un factor importante que influye en las mediciones *in situ* y requiere una compensación con el fin de obtener lecturas fiables y reproducibles. Un algoritmo para la compensación de la turbidez se ha desarrollado, sobre la base de una relación entre la intensidad y la longitud de onda de dispersión como una función del diámetro de las partículas y la forma espectral causada por los sólidos en suspensión (Huber y Frost, 1998). La dispersión es dependiente de la longitud de onda (Langergraber *et al*., 2004).

El desarrollo de nuevas tecnologías y el aumento en la potencia de cálculo observada en los últimos años, han permitido un cambio hacia un enfoque de múltiples longitudes de onda (Fogelman *et al*., 2006). A pesar de que el equipo necesario puede ser más complejo, los resultados son más robustos (Thomas *et al*., 1993). De hecho, un enfoque de múltiples longitudes de onda puede lograr mejores resultados en comparación con el uso de los procedimientos de una sola longitud de onda, sobre todo para los efluentes cuyas características de composición varían constantemente (Rieger *et al.*, 2004; Langergraber *et al.*, 2004a).

Por lo tanto, actual en el mercado existen tres espectrómetros UV-Visible sumergibles de aplicación *in situ* y *on-line* para las mediciones de la calidad del agua, que se basan en determinar la correlación entre la absorbancia variable indirecta y la equivalencia de los determinantes resultantes (concentración). La primera denominada STIP-scan es una sonda fabricada por la empresa Endress+Hauser la cual opera en el rango UV-Visible. Por otra parte, la fábrica alemana TriOS ofrece dos sondas las cuales operan de forma

independe el rango UV y Visible, comercializadas con los nombres de ProPS-UV y VIPER respectivamente. La última denominada spectro::lyser de la fábrica austriaca s::can la cual funciona en el rango UV-Visible del espectro.

Estas sondas han demostrado ser un instrumento prometedor para cuantificar las concentraciones de determinantes de aguas residuales para la gestión integrada y el control de las redes de alcantarillado, así como la operación y control de plantas de tratamiento, y en general para todo tipo de hidrosistemas.

STIP-scan-Endress+Houser	ProPS-UV-TriOS	VIPER-TriOS	spectro::lyser-s::can
			

Figura 8- Sonda disponibles en mercado para la medición *in situ* del espectro UV-Visible

Marca	Referencia	Rango del espectro	Ancho de banda espectral	Fuente de luz	Paso de luz
TriOS	ProPS-UV	190 - 360 nm	0.75 nm	Lámpara de deuterio	1 – 100 mm
TriOS	VIPER - hyperspectral VIS	360 – 720 nm	0.5 nm	*LED*	50 – 250 mm
s::can	spectro::lyser	200 – 750 nm	2.0 nm	Lámpara de destellos de Xenón	0.5 – 100 mm
Endress+Houser	STIP-scan	200 – 680 nm		Lámpara de destellos de Xenón	N/A

Tabla 3- Características tecnológicas de las sondas para medición *in situ* y en continuo del espectro UV-Visible en el agua

1.3.2.2. Mediciones multiparamétricas *in situ* y *on-line* por medio de espectroscopia UV-Vis

La Espectrofotometría es ampliamente utilizada en la quimiometría[1] para determinar la concentración de ciertas sustancias en las muestras. Por lo tanto, para este fin se utilizan métodos quimiométricos de los cuales el más simple es la ley de Beer-Lambert. Sin embargo, este método no puede ser generalizado, ya que presenta una serie de limitaciones (ver numeral 1.3.1.1) y no permite la explotación de la información presente en un espectro de absorbancia de una forma simultánea reduciendo la capacidad

[1] En 1975, La *International Chemometrics Society* (*ICS*) la definió como: *"..la disciplina química que utiliza métodos matemáticos y estadísticos para diseñar o seleccionar procedimientos de medida y experimentos óptimos, y para proporcionar la máxima información química mediante el análisis de datos químicos."*

predictiva del método, debido a que no se presenta una única relación lineal entre la medición de la absorbancia a una determinada longitud de onda y la concentración de la sustancia analizada (adaptado de Dahlén *et al*., 2000).

Entonces, los datos de múltiples longitudes de onda se pueden utilizar para mejorar las mediciones de la concentración. En este sentido, diferentes procedimientos matemáticos han sido utilizados para el procesamiento de la información espectral y poder cuantificar múltiples determinantes con una sola medición. A continuación se presentan diferentes métodos quimiométricos, y determinantes detectados y cuantificados por medio del espectro de absorbancia UV-Vis.

Usando el rango espectral de 205 a 330 nm y un método de deconvolución[2] para la determinación del carbono orgánico disuelto (COD), DQO, COT, DBO, SST, y nitrato, Thomas *et al*. (1996) demostraron que es posible obtener muy buenas correlaciones para todos los determinantes mencionados, con el fin de mejorar el control de una PTAR. Khorassani *et al*. (1999) también llegaron a la conclusión que mediante el uso del método de deconvolución determinista y el rango espectral UV es posible lograr buenos resultados de la calibración para determinar DQO, COT, SST, nitrato y de cromo IV, presentes en diferentes aguas residuales industriales. (Escalas *et al*., 2003) utilizaron un método de deconvolución UV modificado para estimar DQO de muestras crudas y diluida de una PTAR.

Una técnica ampliamente utilizada es la regresión por mínimos cuadrados parciales (sigla en inglés *PLS*), en la que la concentración en una muestra se calcula a partir de un vector de datos de absorbancia a diferentes longitudes de onda mediante la extracción de los componentes principales y combinaciones lineales de las absorbancias medidas (Dahlén *et al*., 2000). Karlsson *et al*. (1995) utilizaron espectroscopia UV-Visible junto con la calibración multivariada de *PLS* para determinar la concentración de nitratos, obteniendo valores entre 0,5 y 13,7 mg/L en muestras provenientes de tres PTAR diferentes y durante un período de más de un año. Los coeficientes de determinación (R^2) para la calibración *PLS*, fueron siempre muy altos y cercanos a la unidad. En el caso del nitrógeno total presente en las aguas residuales Ferrée y Shannon (2001) estudiaron el uso de un segundo método derivado para la determinación de nitratos y nitrógeno total, mediante la oxidación de compuestos nitrogenados en nitrato por *auto-claving*, alcanzando un coeficiente de correlación de 0.99, a pesar de que los resultados obtenidos fueron para concentraciones de NO_3-N entre 0.1 y 3 mg/L. Estos ejemplos muestran que la espectroscopia UV se puede considerar un método alternativo para la supervisión nitrato sin el uso de reactivos peligrosos (por ejemplo, técnica de reducción de cadmio) o equipos costosos (por ejemplo, cromatografía iónica) (Ferrée y Shannon, 2001).

[2] Se refiere a las operaciones matemáticas empleadas en restauración de señales para recuperar datos que han sido degradados por un proceso físico que puede describirse mediante la operación inversa a una convolución. En óptica se basa en expresar el espectro de absorbancia medido como una combinación lineal de unos pocos espectros base.

La influencia de la turbidez en la cuantificación de la DQO en aguas residuales grises y efluentes (retretes) utilizando el rango UV y redes neuronales artificiales (RNA) fue investigado por. En dicha investigación se obtuvieron los mejores resultados entre 190 y 350 nm, cuando se compara con el rango de absorbancia entre 200 y 350 nm. Por otra parte, los autores concluyeron que para las aguas grises, mejores correlaciones se lograron sin filtración de la muestra sólo cuando la turbidez no fue superior a 150 NTU. Además, el valor de turbiedad como un vector de entrada en modelo RNA no mejora su predictibilidad, incluso decrece su precisión (Fogelman *et al*., 2006).

Todas las aplicaciones anteriores requieren toma de muestras para el análisis espectral *off-line*. Mientras tanto, nuevos desarrollos logrados mediante la construcción de equipos sumergibles, como los presentados en el numeral 1.3.2.1 pueden realizar un análisis de espectros directamente en medios líquidos. El uso de este tipo de espectrómetros *in situ* para la determinación de varios determinantes ha tenido varias aplicaciones, por ejemplo en el efluente de una PTAR, determinando con éxito la presencia y concentración de DQO, SST, nitratos y nitritos, o utilizando espectros UV en el rango 200-400 nm (Rieger *et al*., 2004). Otra aplicación de este tipo de sondas operando rango UV-visible fue realizada para controlar la PTAR de una fábrica de papel (Langergraber *et al*., 2004) determinando los valores de DQO filtrado, DQO y nitrato. Más recientemente, (Maribas *et al*., 2008) utilizaron un espectrofotómetro UV-Visible sumergible para monitorear los cambios rápidos en la DQO total y los SST, probando tres lugares diferentes en una unidad de pretratamiento de una PTAR. Estos investigadores concluyeron que nuevas calibraciones de los modelos quimiométricos se necesitan cada vez que se produce una variación repentina en la matriz de composición del agua residual. Por lo tanto, los resultados muestran que no es fácil de tomar en cuenta las grandes variaciones en la matriz de las aguas residuales como también afirmaron (Rieger *et al*., 2006) en un estudio donde se analizaron los resultados de múltiples calibraciones de modelos *PLS* para seis PTAR utilizando un espectrómetro UV-Visible.

Otros métodos para determinar la coherencia entre la absorbancia y determinantes equivalentes obtenidos a través de esta información (DQO, DQO filtrado, SST) fue estudiado por (Hochedlinger, 2005) con diferentes métodos de regresión. La correlación de la absorbancia y la concentración se analizan sobre la base de los métodos de regresión lineal, regresión árbol modelo, métodos de regresión multivariados, y *support vector machine* utilizando un algoritmo de optimización mínima secuencial.

En los numerales 2.2 y 2.3 se explicará ampliamente el funcionamiento y aplicaciones de la sonda de espectrometría UV-Vis spectro::lyser, la cual fue empleada en esta investigación.

1.4. ANÁLISIS DE DATOS

1.4.1. Incertidumbre en las mediciones

"*Las mediciones junto con la observación son la base para documentar, cuantificar, describir y entender los experimentos, los sistemas y sus variables. Sin embargo, el método de medición está sujeto a incertidumbres y errores que influyen en los resultados y deben ser considerados en la interpretación de los valores medidos*" (Gujer, 2008).

En primer lugar, y antes de presentar el procedimiento para la evaluación de la incertidumbre, es conveniente recordar algunas definiciones.

La trazabilidad se define como 'la propiedad del resultado de una medición o el valor de una norma que puede ser relacionado con las referencias indicadas, generalmente con normas nacionales o internacionales, a través de una cadena ininterrumpida de comparaciones teniendo en cuenta todas las incertidumbres indicadas' (ISO, 1993). La cadena ininterrumpida de comparación implica que no debe producirse ninguna pérdida de información durante el procedimiento analítico. Por último, la trazabilidad implica, en teoría, que la incertidumbre de todas las referencias establecidas que contribuyen a la medición son consideradas debidamente (lo que significa que cuanto menor es la cadena de comparación más pequeña es la incertidumbre del resultado final). Además, cabe señalar que la trazabilidad no debe confundirse con la exactitud que abarca los términos de veracidad (el grado de coincidencia de los valores medidos con el 'valor verdadero') y la precisión (el grado de concordancia entre los resultados obtenidos mediante la aplicación de los mismos procedimientos experimentales en varias ocasiones en condiciones establecidas). En otras palabras, un método que es trazable sobre una referencia indicada no es necesariamente exacto (es decir, la referencia indicada no necesariamente corresponder con el 'valor verdadero'), mientras que un método exacto es siempre trazable a lo que se considera sea la mejor aproximación al valor real (que se define como 'un valor que se obtendría mediante la medición si la cantidad puede definirse completamente si todas las imperfecciones de la medición son eliminadas') (Gonzalez *et al*., 2009).

1.4.1.1. Evaluación de la incertidumbre

Existe principalmente dos tipos de incertidumbre catalogadas como Tipo A y Tipo B (Miranda, 2003).

a) *Tipo A*

La primera está asociada con el resultado de una medición, que caracteriza la dispersión de los valores que se podrían atribuir razonablemente al medido, calculado por medio del análisis estadístico directo de una serie de observaciones o mediciones repetidas. En este caso se asume que las mediciones presentan una distribución normal. Esto significa que los valores individuales de x_i de la variable aleatoria X están sujetos a cierta función de

probabilidad de densidad $f(x)$ con un valor medio del valor μ_x (el valor "verdadero de la medida", Ecuación 5) y la varianza σ_x^2 (Ecuación 6)

$$\mu_x = \int_{\infty}^{-\infty} x \cdot f(x) \cdot dx$$

Ecuación 5-

$$\sigma_x^2 = \int_{\infty}^{-\infty} (x - \mu_x) \cdot f(x) \cdot dx$$

Ecuación 6-

La distribución empírica de una variable aleatoria se puede determinar aproximadamente por n múltiples mediciones de exactamente el mismo objeto (repeticiones). En ese caso, la media aritmética (Ecuación 7) se utiliza como una aproximación de μ_x y la desviación estándar experimental s_x^2 (Ecuación 8) la cual sustituye σ_x^2.

$$m_x = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Ecuación 7-

$$s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - m_x)^2$$

Ecuación 8-

Por otro lado, la mejor estimación de la varianza de la media se define como $\sigma^2(\mathrm{m_x}) = \sigma^2/\mathrm{n}$. Lo anterior, cuantifica qué tan bien m_x estima el valor esperado de x, y se puede utilizar como una medida de la incertidumbre ($u(x_i)$) de m_x.

$$u(x_i) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - m_x)}{n(n-1)}}$$

Ecuación 9-

b) Tipo B

Es la incertidumbre de una medida asociada a cantidades cuyos valores se introducen en la medición a través de fuentes externas, tales como cantidades asociadas a patrones de medición calibrados, materiales de referencia certificados o datos de referencia obtenidos de manuales. Además, cuando existe o no correlación entre las cantidades que aparecen en una medición, se debe utilizar un procedimiento para obtener la incertidumbre estándar compuesta basado en las incertidumbres estándares de las cantidades originales y alguna relación funcional entre ellas $y = f(x_1, x_2, \dots, x_n)$, de la cual se obtiene la nueva cantidad (Miranda, 2003). Por lo tanto, este tipo de incertidumbre se evalúa por medio de

la ecuación conocida como la ley de la propagación de la incertidumbre, la cual se describe a continuación:

$$u^2(y) = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 u^2(x_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} u(x_i, x_j)$$

Ecuación 10-

donde $u(x_i)$ es la incertidumbre estándar asociada a x_i y $u(x_i, x_j)$ es la covarianza estimada entre x_i y x_j. Las derivadas parciales que aparecen en la Ecuación 10 están evaluadas en $X_i = x_i$, y frecuentemente se les llama coeficientes de sensibilidad, los cuales describen cómo cambia la estimación de salida y con cambios en las estimaciones de entrada $x_1, x_2, \ldots, x_n$.

1.4.2. Simulaciones de Monte Carlo

Los métodos de simulación estadísticos tal vez contrastan con los métodos numéricos de discretización convencionales, que normalmente aplican ecuaciones diferenciales: ordinarias o parciales, las cuales describen algún sistema físico o matemático subyacente. En muchas aplicaciones de Monte Carlo el proceso físico se simula directamente, y no hay necesidad utilizar las ecuaciones diferenciales que describen el comportamiento del sistema (Figura 9). El único requisito es que el sistema físico (o matemático) se describa por medio de funciones de densidad de probabilidad (en inglés *pdf*).

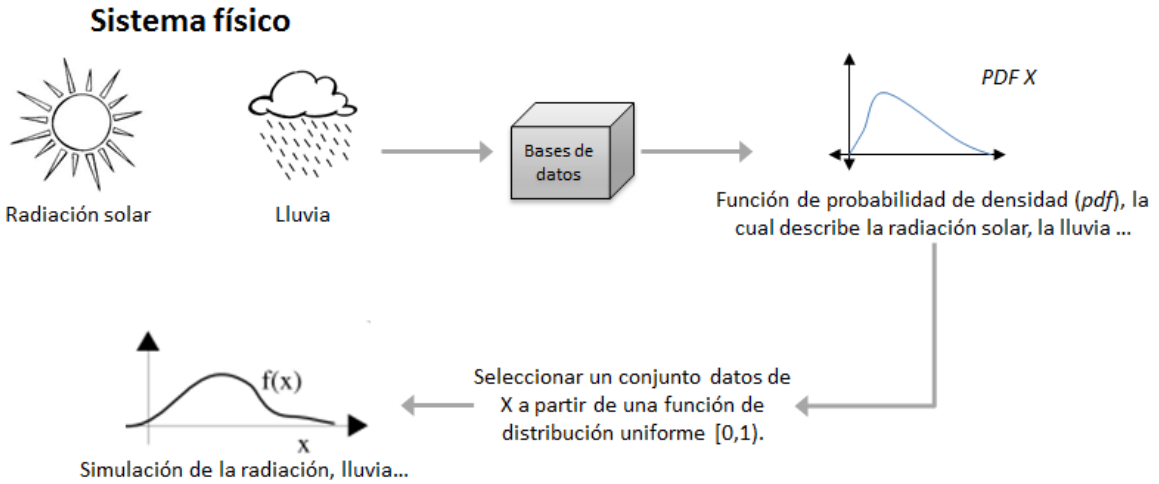


Figura 9- Esquema general de las simulaciones Monte Carlo

La simulación estadística de los métodos Monte Carlo tienen por fin generar digitalmente un gran conjunto de datos para el estudio de la propagación de las funciones de densidad de probabilidad entre las variables de entrada X_i (variables y parámetros) y la salida y. Las simulaciones de Monte Carlo se utilizan generalmente cuando no se cumplen las condiciones de la ley de propagación de la incertidumbre (no hay un modelo analítico, distribución asimétrica,...) por lo tanto son el método de referencia. La Figura 10 ilustra el

principio general de la estimación de la incertidumbre por simulaciones de Monte Carlo (Lepot, 2012).

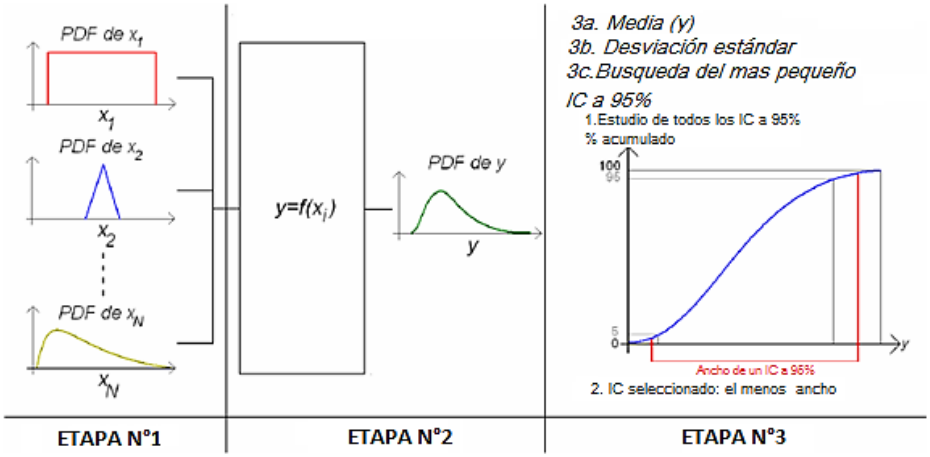


Figura 10- Pasos para generar una simulación Monte Carlo (Lepot, 2012)

Los elementos principales de un método de simulación de Monte Carlo son los siguientes:

- Identificar las variables explicativas del fenómeno X_i, teniendo en cuenta la presencia o ausencia de correlaciones (covarianza cero) necesarias para conectar las variables explicativas para obtener el resultado deseado $y = f(X_i)$.
- *Funciones de probabilidad de densidad-pdf*: el sistema físico o matemático debe ser descrito por un conjunto de *pdf*.
- *Generador de números aleatorios*: consiste en una fuente de números pseudo-aleatorios distribuidos uniformemente en un intervalo de [0,1), lo cual permite que cualquier valor de una variable explicativa en un universo (*pdf*) del fenómeno y tenga la misma probabilidad de ser seleccionada.
- *Muestreo*: con base en paso anterior se generan múltiples escenarios posibles y consistentes basados en los supuestos establecidos en el modelo.
- *Estimación de error*: una estimación del error estadístico (varianza) como una función del número de ensayos y otras cantidades deben ser determinadas, tales como la media, la desviación estándar y un intervalo de confianza (normalmente 95 %) - Figura 10.
- *Paralelización y vectorización*: algoritmos que permitan que los métodos de Monte Carlo sean implementados de manera eficiente en arquitecturas informáticas avanzadas.

1.4.3. Detección de *outliers*

La mayor parte de los conjuntos de datos del mundo real contienen valores atípicos (*outliers*) y las aguas residuales no son la excepción. Tales datos están caracterizados por presentar magnitudes inusualmente grandes o pequeñas, en comparación con los demás en el conjunto de datos (Seo, 2006), tal como se muestra en la Figura 11. Los *outliers*

pueden generar valores errados 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. Además, por el incremento significativo en la cantidad de datos medidos y almacenados es importante identificar las observaciones inesperadas o inusuales. La detección de valores denominados *outliers* es una tarea de minería de datos que permite detectar objetos desviados, eventos extraños y/o excepcionales. Las causas de los *outliers* se pueden clasificar en dos: los derivados de errores en los datos y los derivados de la variabilidad inherente de los datos (Preetha y Radha, 2011). Luego, la detección de *outliers* es una parte importante del análisis de datos en los dos casos anteriores, aumentando la necesidad de métodos de análisis, para hacer uso de la información contenida de manera implícita en una base de datos (Fayyad *et al*., 1996).

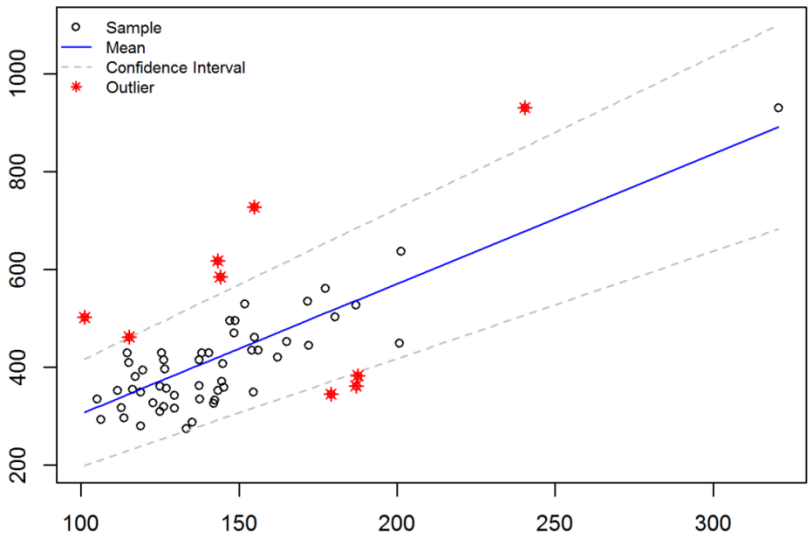


Figura 11- Detección de *outliers* en un conjunto de datos por medio de análisis bivariado

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 aplicar un método para identificar dichos valores. Los métodos para la detección de *outliers* se basan en estadísticas tales como la distancia entre valores, la desviación estándar y/o análisis basado en las distribuciones de densidad de los datos.

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 (Acuña y Rodríguez, 2004), $(\bar{x} - kS, \bar{x} + kS)$ donde el valor del coeficiente k es usualmente 2 ó 3. 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. Por consiguiente, la variable x es considerada *outlier* si: $(x - \bar{x})/S > k$ (Acuña y Rodriguez, 2004).

El problema del método anterior es que asume la distribución normal de la información, esperando formas de campana y simetría razonable en los datos, que con frecuencia es algo que no ocurre. Además, la media y desviación estándar son muy sensibles a los valores atípicos de magnitudes significativas (Iglewicz y Hoaglin, 1993; Chen *et al*., 1996). En respuesta a esto, John Tukey en 1977 introdujo varios métodos para el análisis de datos, uno de ellos fue el *Boxplot*. Ésta es una conocida herramienta gráfica sencilla, que se utiliza con el propósito de mostrar información continua acerca de los datos univariados como la media, los *mild outliers* y los *extreme outliers* de un conjunto de datos (ver Figura 12). Este método es menos sensible a valores extremos de los datos que aquellos métodos que se basan en la media y la desviación estándar, ya que utiliza los cuartiles, los cuales son consistentes ante los valores extremos (Acuña and Rodriguez, 2004; Seo, 2006).

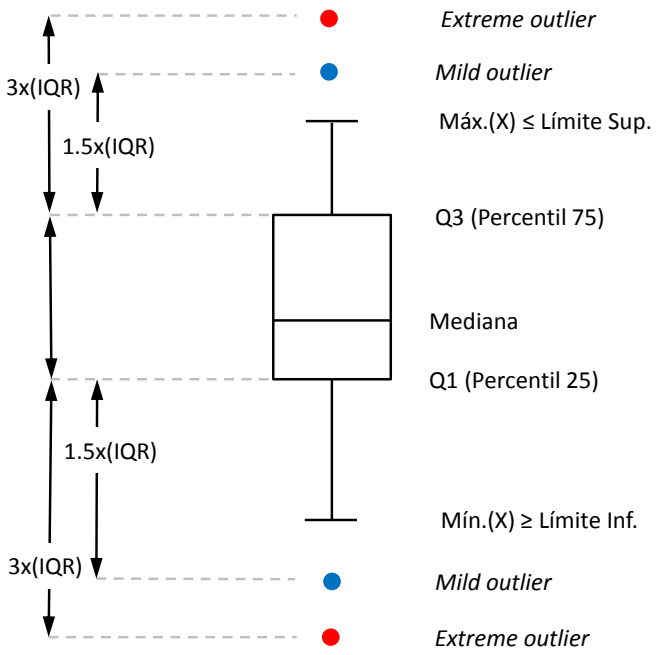


Figura 12- Boxplot o diagrama de caja

Un dato x se declara un *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 (en inglés *interquatil range*) calculado como $Q_3 - Q_1$ (Acuña y Rodríguez, 2004). 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).

1.4.4. *Nonlinear Least Squares (NLS)*

Para calibrar los coeficientes o pesos de las ecuaciones 5, 6 y 7 se usó la función *nls (nonlinear least squares)* o mínimos cuadrados no lineales. Esta técnica estima los valores de los coeficientes de una función m representados como $\theta = (A, B, C \ldots)$ que minimicen la suma de los residuos cuadrados (Bates y Watts, 1988; Bates y Chambers, 1992):

$$S(\theta) = \sum w[y - m(\theta, x)]^2$$

Ecuación 11-

donde y es la variable a predecir en función del variable independiente x y el conjunto de parámetros θ de la función m, y w son los pesos conocidos no negativos. En la mayoría de aplicaciones $w = 1$ para todas las observaciones (Bates y Watts, 1988; Bates y Chambers, 1992).

Este algoritmo simple oculta por lo menos tres consideraciones importantes. En primer lugar, se pretende garantizar que en cada paso el método obtenga un valor menor de S o por lo menos que S no aumente. Hay muchos algoritmos de mínimos cuadrados no lineales; véase, por ejemplo, Bates y Watts (1988). El algoritmo por defecto en *nls* utiliza una forma de iteración de Gauss-Newton. En segundo lugar, la suma de cuadrados de la función S puede ser una función con múltiples mínimos. Como consecuencia, los supuestos de mínimos cuadrados estimados podrían ser un local en lugar de un mínimo global de S. En tercer lugar, el algoritmo continua buscado mejores soluciones siempre que el valor de S sea menor en cada paso. Por lo tanto, se puede limitar el número máximo de iteraciones a un valor fijo, y/o a una tolerancia del mejor mínimo de la función S con respecto a su valor en el paso anterior (Bates y Watts, 1988; Bates y Chambers, 1992).

1.4.5. *Partial Least Squares* (*PLS*)

Un método ampliamente utilizado para estimar las concentraciones u otros variables a partir de datos espectrofotométricos es la regresión de mínimos cuadrados parciales (*Partial Least Squares*-PLS). El siguiente párrafo presenta aspectos generales del método PLS basado en las descripciones realizadas por Abdi (2003) y (Varmuza y Filzmoser, 2009).

El método PLS original se desarrolló alrededor de 1975 por el estadístico Herman Wold para un tratamiento de cadenas de matrices y aplicaciones en econometría. Su hijo, Svante Wold y otros introdujeron la idea PLS en Quimiometría. Sin embargo, la PLS fue durante mucho tiempo bastante desconocido para los estadísticos (Varmuza y Filzmoser, 2009). Fue presentado por Wold et al. (1983), como un método que fusiona y generaliza el análisis de componentes principales (en inglés PCA) y los métodos de regresión múltiple (Abdi, 2003). Es especialmente útil en casos muy comunes donde el número de variables es igual o mayor que el número de observaciones y/o donde hay otros factores que conducen a correlaciones entre las variables (VCCL - Virtual Laboratorio de Química

Computacional 2005). Por lo tanto, es ampliamente utilizada en la química, especialmente en aplicaciones de cromatografía y espectrometría (Tenenhaus 1998), donde el número de variables (es decir, longitudes de onda) que caracteriza a los espectros son por lo general muy grandes en comparación con el número de observaciones. No obstante, el énfasis se hace a menudo en la predicción de las observaciones en lugar de la comprensión de las relaciones entre las variables (Tobias, 1995).

El objetivo de *PLS* es predecir una variable *Y* a partir de una matriz de variables *X* y describir su estructura común. Si el número de variables independientes que componen la matriz *X* es muy grande en comparación con las observaciones, una regresión múltiple estándar no es suficiente para lograr esta tarea, debido al riesgo de singularidad y multicolinealidad de *X*. Con el fin de evitar este problema, algunos métodos alternativos se han desarrollado (Torres y Bertrand-Krajewski, 2008).

Uno de esos métodos tiene la intención de eliminar algunas de las variables independientes mediante el uso de métodos por pasos. Otro, llamado regresión de componentes principales, lleva a cabo un análisis de componentes principales (sigla en inglés *PCA*) de la matriz ***X*** con el fin de utilizar sólo los componentes principales de ***X*** con el fin de predecir ***Y***: la ortogonalidad de *PCA* elimina el problema de multicolinealidad, pero la elección del conjunto de las variables independientes es difícil, ya que no hay certeza de que los componentes principales que explican ***X*** también son pertinentes para explicar ***Y***. La finalidad de *PLS* es detectar los componentes de la matriz ***X*** que también sean pertinentes para explicar ***Y*** (ver Figura 13). Más exactamente, *PLS* busca un conjunto de componentes, llamado vectores latentes, que proporciona una descomposición simultánea de ***X*** y ***Y***, con la condición de que esos componentes explican lo mejor posible la covarianza entre ***X*** y ***Y***. Este paso lo generaliza la *PCA*. Luego, existe una etapa de regresión, donde se utiliza la descomposición de ***X*** para predecir ***Y***. La *PLS* está relacionada con otras técnicas como la correlación canónica y el análisis de múltiples factores. La principal novedad de *PLS* es preservar la asimetría de la relación entre las variables independientes ***X*** y las variables dependientes ***Y*** de pronóstico, mientras que estas otras técnicas la tratan simétricamente (Torres y Bertrand-Krajewski, 2008).

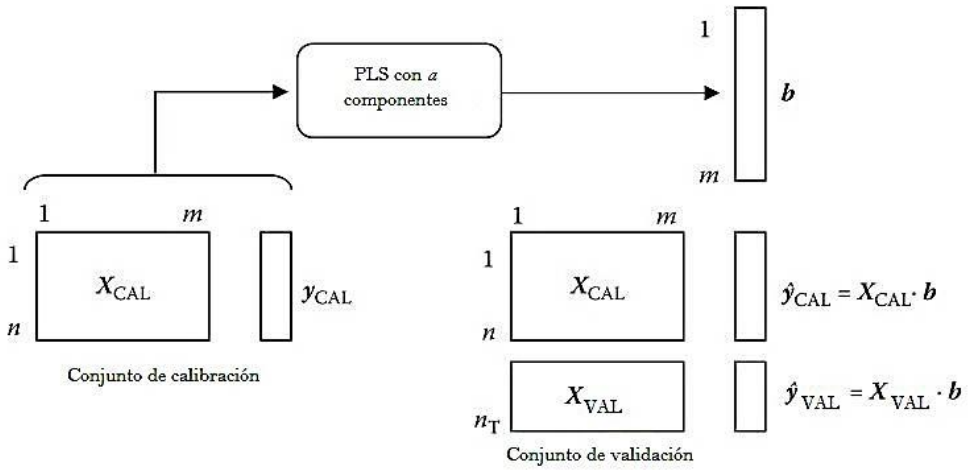


Figura 13- PLS como un método de regresión lineal múltiple para la predicción de la propiedad *y* desde las variables $X_1, ..., X_m$, aplicando los coeficientes de regresión $b_1, ..., b_m$. A partir de un conjunto de calibración, el modelo PLS se crea y aplica a los datos de calibración y de validación (Varmuza y Filzmoser, 2009).

1.4.6. Aspectos matemáticos de *PLS* (Lepot, 2012)

PLS fue utilizado en esta tesis para determinar y seleccionar la mejor relación (Ecuación 12) entre valores de absorbancia $X(N,n_X)$ proporcionadas por las longitudes de onda de los espectrómetros UV-Vis y las concentraciones de determinantes $y(n, 1)$ de n muestras analizadas.

$$y = k + \sum_{i=1}^{N_{VI}} b_i \cdot X_i + \varepsilon$$

Ecuación 12-

donde X_i los valores de absorbancia del espectro UV-Vis a N_{VI} longitudes de onda ($N_{VI} < n_x$), y ε es el residuo de la estimación y.

La principal ventaja de la regresión PLS es proporcionar una relación entre y y X (Ecuación 13), describiendo su estructura común por la búsqueda sus componentes (llamados vectores latentes) comunes a las estructuras de y y X al tiempo que se maximiza la covarianza entre y y X.

$$y = X \cdot b + f$$

Ecuación 13-

donde b es el vector que contiene los coeficientes de la regresión *PLS*, $f(N, 1)$ es vector de los residuos entre las predicciones y los valores observados.

La etapa clave de este método es reducir la matriz $X(N, n_X)$ a una matriz $T(N, n_T)$, donde las n_T columnas de T $(n_T < n_x)$ son combinaciones lineales de las columnas de X, como se presenta en la Ecuación 14.

$$T = X \cdot W$$

Ecuación 14-

donde $W(n_X, n_T)$ es la matriz de pesos. Contiene los coeficientes de las combinaciones lineales.

Las columnas de la matriz T se denominan vectores latentes $(n_T = N_{VL})$. Si n_T es igual al rango de la matriz X, T proporciona una descomposición exacta de X.

Por lo tanto, es posible escribir y y X siguiendo la Ecuación 15 a y b

$$X = T \cdot P + R \qquad \text{a}$$

$$y = T \cdot q + f \qquad \text{b}$$

Ecuación 15-

donde $R(N, n_X)$ es la matriz de residuos asociada a la predicción X, $f(N, 1)$ es el vector de residuos asociado con la predicción y.

El cálculo de los componentes de la matriz $P(n_T, n_X)$ y el vector $q(1, n_T)$ se detalla a continuación.

La descomposición simultánea de X y y es iterativa, como puede verse en la Figura 14. En el paso k, la descomposición está descrita por la Ecuación 16 a y b:

$$X = t_1 \cdot p_1 + t_2 \cdot p_2 + \ldots + t_k \cdot p_k + R_k \qquad \text{a}$$

$$y = t_1 \cdot q_1 + t_2 \cdot q_2 + \ldots + t_k \cdot q_k + f_k \qquad \text{b}$$

Ecuación 16-

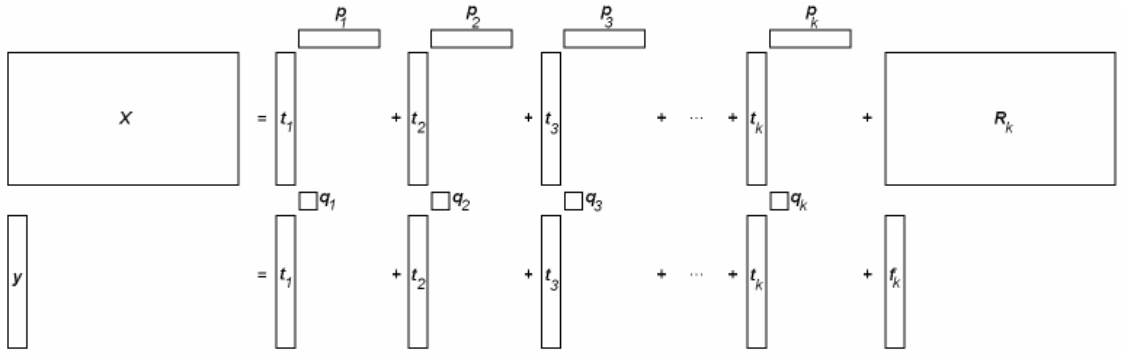


Figura 14- Ilustración de la descomposición simultánea e iterativa del método PLS (Lepot, 2012)

En el paso $k + 1$, el vector latente t_{k+1}, el vector p_{k+1} y el número q_{k+1}, se calculan a partir de residuos de R_k y f_k del paso anterior, de acuerdo con la:

$$t_{k+1} = R_k \cdot W_{k+1}$$

Ecuación 17-

El cálculo de W_{k+1} denominado puntuaciones no correlacionadas, actualmente se puede lograr mediante tres algoritmos: Kernel, NIPALS y SIMPLS.

q_{k+1} se calcula mediante una regresión lineal simple entre f_k y t_{k+1}. Del mismo modo, las regresiones lineales n_X se realizan entre cada columna de R_k y t_{k+1}, para calcular R_{k+1} y p_{k+1}.

La estimación de y en el modelo puede escribirse como:

$$y = X \cdot [W \cdot (P \cdot W)^{-1} \cdot q] + f$$

Ecuación 18-

Este proceso iterativo se detiene cuando la varianza entre los valores estimados y reales es satisfactoria. El criterio de parada, se basa en la precisión de una predicción aceptable, lo cual puede conducir a la creación de numerosos modelos (en términos de número de variables predictoras y vectores latentes seleccionadas) más o menos eficaces en la predicción. La optimización entre la robustez y la precisión del modelo se tratará en el numeral 3.2 sobre la estimación de las concentraciones de determinantes de calidad del agua.

1.4.7. Métodos *machine learning*

Los métodos *Machine Learning* nombre en inglés que se refiere a los métodos que se basan en el proceso de aprendizaje de una máquina, consisten en tener una base de datos que se coloca como entrada a un algoritmo. Dicho algoritmo se ejecuta dentro de una máquina entrenada que procesa datos para obtener una salida. Este es un campo altamente interdisciplinario que toma prestado y se basa en las ideas de la estadística, la informática (inteligencia artificial), la ingeniería, la ciencia cognitiva, la teoría de optimización y muchas otras disciplinas de la ciencia y las matemáticas (Bousquet *et al*., 2004). Las aplicaciones de este método son: la clasificación de problemas, la predicción de resultados con base en datos de entrada, regresión y *clustering*. Existen varios tipos de aprendizaje que caracterizan estos métodos de los cuales dos son los más comunes: aprendizaje supervisado y no supervisado (Michalski *et al.*, 1985).

Los modelos de aprendizaje supervisado implementados en esta investigación se derivan de experimentos previamente ejecutados, los cuales son alimentados por pares de datos de entrada (*input*) y salida (*output*). A partir de estos datos y de alguna técnica de inteligencia artificial, la máquina puede generar regresiones o clasificar los datos de entrada o salida en ausencia de alguno de éstos, creando así un nuevo algoritmo con base en la experiencia generada en los procesos. Existen dos tipos de modelos para desarrollar el aprendizaje supervisado: (i) globales, los cuales son utilizados de forma general y (ii)

locales, los cuales son implementados en casos particulares (por ejemplo, los casos basados en el razonamiento con lo que se pretende dar solución a un problema a partir de soluciones similares observadas en el pasado) (Nilsson, 1996) (Michalski *et al.*, 1985). Algunos modelos fundamentados en este tipo de aprendizaje son: redes neuronales artificiales, árboles de decisión, máquina de vectores de soporte (en inglés *Support Vector Machine*), lógica difusa y métodos Kernel, entre otros.

Por otra parte, los modelos de aprendizaje no supervisado son un método de aprendizaje, donde un modelo es ajustado a las observaciones. Se distingue del Aprendizaje supervisado por el hecho de no haber un conocimiento *a priori*. El aprendizaje no supervisado se utiliza principalmente para clasificar y diferenciar rasgos significativos de un conjunto de datos no clasificado *a priori*, dado que la red internamente intenta encontrar redundancias y rasgos significativos para agrupar los datos (Nilsson, 1996). De los métodos desarrollados hasta el momento en esta técnica de aprendizaje están: K-*means*, análisis de componentes principales (en inglés PCA), *clustering,* cadenas de markov y análisis de componentes independientes entre otros métodos.

1.4.7.1. Redes Neuronales Artificiales (RNA)

Las RNA se inspiran en hallazgos biológicos relacionados con el comportamiento del sistema nervioso humano, el cual está conformado por una red de unidades llamadas neuronas. Las RNA al margen de "parecerse" al cerebro presentan una serie de características propias del cerebro. Por ejemplo, aprenden de la experiencia, generalizan de ejemplos previos a ejemplos nuevos y abstraen las características principales de una serie de datos (Olabe, 1998).

Por lo tanto, en 1943, el neurobiólogo Warren McCulloch, y el estadístico Walter Pitss, publicaron el artículo "*A logical calculus of Ideas Imminent in Nervous Activity*". Este artículo constituyó la base y el inicio del desarrollo en diferentes campos como son los Ordenadores Digitales (John Von Neuman), la Inteligencia Artificial (Marvin Minsky con los Sistemas Expertos) y el funcionamiento del ojo (Frank Rosenblatt con la famosa red llamada Perceptron).

Nathaural Rochester del equipo de investigación de IBM presentó en 1956 el modelo de una red neuronal que él mismo realizó y puede considerarse como el primer software de simulación de redes neuronales artificiales.

En 1957, Frank Rosenblatt publicó el mayor trabajo de investigación en computación neuronal realizado hasta esas fechas. Su trabajo consistía en el desarrollo de un elemento llamado "*Perceptron*". El primer *perceptron* era capaz de aprender algo y era robusto, de forma que su comportamiento variaba sólo si resultaban dañados los componentes del sistema, este modelo de RNA fue originalmente diseñado para el reconocimiento óptico de patrones. Este modelo presenta algunas limitaciones debido a que se trataba de un dispositivo en desarrollo. La mayor limitación la reflejaron Minsky y Papert años más

tarde, y ponían de manifiesto la incapacidad del *perceptron* en resolver algunas tareas o problemas sencillos como por ejemplo la función lógica *OR* exclusivo.

Uno de los mayores cambios realizados en el *perceptron* de Rossenblatt a lo largo de la década de los 60 ha sido el desarrollo de sistemas multicapa que pueden aprender y categorizar datos complejos.

Teuvo Kohonen, de la Universidad de Helsinki, es uno de los mayores impulsores de la computación neuronal de la década de los 70. De su trabajo de investigación destacan dos aportaciones: la primera es la descripción y análisis de una clase grande de reglas adaptativas, reglas en las que las conexiones ponderadas se modifican de una forma dependiente de los valores anteriores y posteriores de las sinapsis. Y la segunda aportación es el principio de aprendizaje competitivo en el que los elementos compiten por responder a un estímulo de entrada, y el ganador se adapta él mismo para responder con mayor efecto al estímulo.

En 1982 John Hopfield con la publicación del artículo Hopfield *Model* o *Crossbar Associative Network*, junto con la invención del algoritmo de retropropagación (*backpropagation)*. Hopfield presenta un sistema de computación neuronal consistente en elementos procesadores interconectados que buscan y tienden a un mínimo de energía.

En la actualidad, se utiliza ampliamente un modelo de RNA de capas múltiples, el cual utiliza el método de retropropagación (*backpropagation*). Éste es un método de aprendizaje supervisado que intenta minimizar los errores promedios mediante las derivadas parciales de dichos errores, actualizando los pesos de la red neuronal en la dirección en la que la función de desempeño disminuye más rápidamente, el gradiente descendiente (Demuth *et al*., 2005). Este tipo de RNA trabaja desde la capa de salida hacia la capa de entrada. Después de cada entrenamiento se comparan los datos de salida con los valores deseados, para calcular el error de cada una de las salidas. A partir de esto se estima el peso a ser ajustado que se adapta dentro de las neuronas para reducir los errores. Ahora se repite el proceso anteriormente descrito para la capa anterior, hasta llegar a la capa de entrada (Witten *et al*., 2005).

1.4.7.2. Estructura y algoritmia del modelo *Feed Forward* con múltiples neuronas en una capa oculta de RNA

Las RNA se componen de neuronas artificiales, cada una de las cuales está conformada por tres partes (*MIT*, 2002): (i) Sinapsis artificiales, o enlaces de conexión a otras neuronas: agregan pesos a los valores de entrada, es decir multiplican cada valor de entrada por un número que indica la importancia del valor para la función que calcula. (ii) Un sumador: adiciona los valores de entrada, sumando además un número específico de la neurona a los valores. (iii) Una función de activación: proyecta los valores de entrada a los valores de salida, aplicando pesos a los datos entrantes (ver Figura 15).

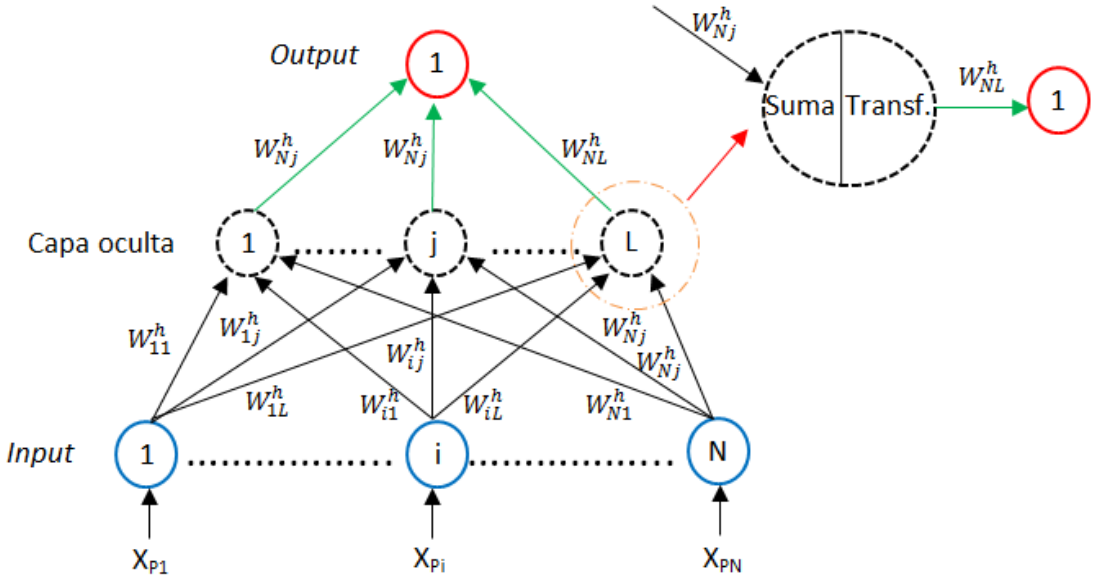


Figura 15- Esquema de una red neuronal con una sola capa oculta

Las redes neuronales artificiales tipo *feed-forward* proporcionan una forma flexible para generalizar funciones de regresión lineal y no lineal. El modelo más sencillo es una red con una sola capa (Figura 15). En este tipo de redes de capa sencilla, el aprendizaje funciona de tal manera que entran a la red neuronal datos de entrenamiento uno a la vez y los pesos de la neurona son revisados después de cada caso, en un intento por minimizar el error cuadrado medio. Este proceso de ajuste gradual de los pesos se basa en el error cometido en los casos de entrenamiento. Se usan principalmente algoritmos de regresión lineal para el aprendizaje (*MIT*, 2002).

El algoritmo y procedimiento matemático se describe a continuación (Venables y Ripley, 1994):

Las unidades de entrada ofrecen un "abanico" y que se distribuye en las unidades "ocultas" en la segunda capa. Estas unidades suman sus entradas, se adiciona una constante denominada ("sesgo" α_k) y toman una función fija para el resultado denominado ϕ_h. Las unidades de salida (*output*) operan de la misma forma, pero con la función de salida ϕ_o:

$$y_k = \phi_o(\alpha_k + \sum_i W_{ij}\phi_h(\alpha_i + \sum_j W_{j1}x_j))$$

Ecuación 19-

La función de activación (de un nodo define la salida de un nodo dada una entrada o un conjunto de entradas) de las unidades de la capa oculta ϕ_h, casi siempre es una función logística:

$$l(z) = \frac{e^z}{1+e^z}$$

Ecuación 20-

y las unidades de salida son lineales, logísticas o unidades umbral. (Este último tiene $\phi_o(x) = I(x > 0)$). La definición general permite tener más de neurona en la capa oculta, y también permite conexiones tipo "*skip-layer*" que conectan las variables de entada y salida de entrada de forma directa a salida cuando se tiene:

$$y_k = \phi_o(\alpha_k + \sum_{i \to k} W_{ik} x_i + \sum_{i \to j} W_{ij} \phi_h(\alpha_i + \sum_{j \to k} W_{j1} x_j))$$

Ecuación 21-

La Ecuación 21, permite que las unidades no lineales perturben la forma de la funcional lineal. El sesgo α_i puede ser eliminado mediante la introducción de una unidad de entrada igual a 0, la cual está permanentemente en la posición +1 y se alimenta de todas las demás unidades. La función de regresión f es parametrizada por el conjunto de pesos W_{ij}, uno para cada enlace en la red (o cero para los enlaces que están ausentes).

El modelo *Feed-forward* igualmente puede ser visto como una manera de parametrizar una función no lineal bastante general. Se ha demostrado que las RNA con unidades de salida lineales pueden realizar aproximaciones de cualquier función continua uniforme f en conjuntos compactos, mediante el aumento en la cantidad de capas ocultas.

Los pesos en el modelo tienen que ser elegidos para minimizar la diferencia entre las estimaciones y los valores observados, para lo cual normalmente se utiliza el criterio de mínimos cuadrados:

$$E = \sum_p \left\| t^p - y^p \right\|^2$$

Ecuación 22-

donde t^p son los valores observados y y^p la salida para el patrón del ejemplo p. Se han propuesto otras medidas, incluyendo máxima verosimilitud para $y \in [0,1]$ (menos el logaritmo de una probabilidad condicional) o su equivalente la distancia Kullback-Leibler, que equivale a minimizar:

$$E = \sum_p \sum_k \left[t_k^p \log \frac{t_k^p}{y_k^p} + (1 - t_k^p) \log \frac{1 - t_k^p}{1 - y_k^p} \right]$$

Ecuación 23-

Una forma de asegurar que a la función f se ajuste a los datos es restringir la clase de las estimaciones, lo cual se puede hacer mediante la regularización en el que el criterio de aceptación se altera de la siguiente forma:

$$E + \mu(f)$$

Ecuación 24-

El decaimiento o tasa de decaimiento del peso (μ), específica en las redes neuronales, la penalización en la suma de los cuadrados de los pesos W_{ij}. Esto sólo tiene sentido si las entradas del modelo se reescalan entre [0, 1] para sean comparables con las salidas de las unidades internas. El uso de la tasa de decaimiento del peso parece para ayudar tanto a la optimización de los procesos y como para evitar *overfitting*. El valor de μ normalmente oscila entre 10^{-4} y 0.1 en esos casos se puede utilizar la Ecuación 22, pero $\mu = 0$ entonces se debe usar la para optimizar (Venables y Ripley, 1994; Varmuza y Filzmoser, 2009).

1.4.7.3. *Support Vector Machine* (*SVM*)

Máquinas de Vectores Soporte (en inglés *Support Vector Machine-SVM*) son máquinas de aprendizaje que aplican el principio inductivo de minimización del riesgo estructural para obtener una buena generalización en un número limitado de patrones de aprendizaje. La minimización del riesgo estructural involucra evaluar de forma simultanea la minimización del riesgo teórico y la dimensión VC (Vapnik–Chervonenkis) que es un valor escalar que mide la capacidad de un conjunto de funciones. La teoría fue originalmente desarrollada por Vapnik y sus compañeros del *AT&T Bell Laboratories*. SVM implementa un algoritmo de aprendizaje, útil para el reconocimiento de patrones sutiles en los conjuntos de datos complejos. El algoritmo realiza la clasificación discriminatoria de aprendizaje por ejemplo para predecir la clasificación de los datos inéditos (Basak *et al*., 2007; Ghanty *et al*., 2009).

SVM es normalmente usado para problemas de clasificación y de regresión. En este trabajo se explicara el algoritmo del método *SVM* para regresiones (*SVM-R*) en dos partes: formulación matemática y selección de las constantes que controlan el desempeño del modelo.

1.4.7.4. Formulación matemática de *SVM-R*

Sea un conjunto de variables de salida y_t, con predictores x_t, para la cual se poseen D muestras representativas del fenómeno. Una *SVM* permite estimar y_t a través de la función (Basak *et al*., 2007; Velásquez *et al*., 2009):

$$\hat{y}_t = b + \sum_{d=1}^{D} w_d \cdot k(x_t, x_d)$$

Ecuación 25-

donde b es una constante y w_d son los factores de ponderación de la función de núcleo $k(.,.)$. Así, una *SVM* es la combinación lineal del mapeo de x_t en un espacio de características altamente no lineales definido por los puntos x_d y la función de transformación no lineal $k(.,.)$.

La estimación de la Ecuación 25 se basa en minimización de la función de riesgo regularizado, $R(C, \varepsilon)$, definida como:

$$R(C, \varepsilon) = C \frac{1}{D} \sum_{d=1}^{D} L_\varepsilon(y_d, \hat{y}_d) + \frac{1}{2} \sum_{d=1}^{D} w_d^2$$

Ecuación 26-

donde la primera sumatoria mide el error empírico entre el modelo y los datos, mientras que el segundo corresponde a la componente de regularización y depende únicamente de los pesos w_d. La constante de regularización C permite variar la importancia de cada una de las componentes; así, valores muy altos de C enfatizan el ajuste del modelo a los datos, sin que importe que tan grandes deban ser los pesos w_d para conseguirlo. Sin embargo, se sabe que el modelo pierde su capacidad de generalización de los datos a medida que los pesos w_d aumentan en magnitud, ya que ellos suelen causar una varianza excesiva en el modelo. Cuando C tiende a cero, la magnitud de la función $R(C, \varepsilon)$ depende únicamente de w_d, sin importar el ajuste a los datos, haciendo que los pesos w_d disminuyan tanto como sea posible. L_ε es la función de error ε-insensible de Vapnik (Vapnik, 1998) definida como:

$$L_\varepsilon(y_d, \hat{y}_d) = \begin{cases} |y_d - \hat{y}_d| - \varepsilon \_\_\_ para \_\_\_ |y_d - \hat{y}_d| > \varepsilon \\ 0 \_ en \_ otro \_ caso \end{cases}$$

Ecuación 27-

donde la constante ε es la precisión deseada, y_d representa el radio del tubo dentro del cual el error se considera cero. Este comportamiento es esquematizado en la Figura 16, donde los puntos negros representan los datos disponibles, que están alejados a una distancia mayor de ε unidades de la predicción del modelo, y para los cuales el error se considera superior a cero; los puntos blancos representan los datos dentro del túnel de tolerancia del modelo, para los cuales el error se considera cero. Mientras menor sea el valor dado a ε, mayor es la precisión exigida, ya que se tendrá en cuenta una mayor cantidad de puntos que contribuyen al error de ajuste.

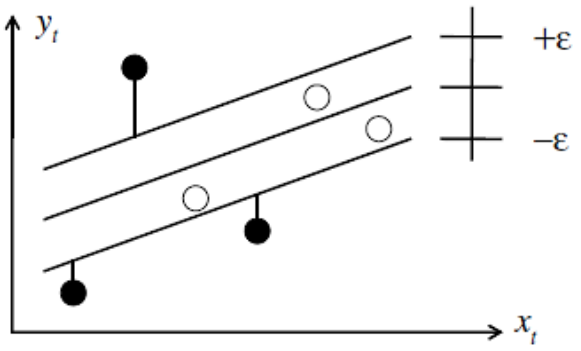


Figura 16- Función de error (Velásquez *et al*., 2009)

La solución de la Ecuación 26 puede ser obtenida mediante la teoría de multiplicadores de Lagrange, y es fácilmente demostrable que el problema original es equivalente a resolver el problema de programación cuadrática (Vapnik, 1998; Platt, 1998):

$$\min_z \frac{1}{2} z^T H z + f^T z$$

Ecuación 28-

Con restricciones lineales:

$$\sum_{d=1}^{D} (a_d - a_d^*) = 0$$

Ecuación 29-

$$0 \le a_d, a_d^* \le C$$

Ecuación 30-

donde a_d y a_d^* son los multiplicadores de Lagrange asociados a los datos, los cuales cumplen con las restricciones $a_d \times a_d^* = 0, a_d \ge 0$ y $a_d^* \ge 0; z = [a_d, a_d^*]'$: $f = [\varepsilon - y_t, \varepsilon + y_t]$; y finalmente,

$$H = \begin{bmatrix} -K & K \\ K & -K \end{bmatrix}$$

Ecuación 31-

En esta última ecuación, K es una matriz de orden $D \times D$, con $k_{ij} = k(x_i, x_j)$. Los pesos w_d se obtienen como:

$$w_d = a_d - a_d^*$$

Ecuación 32-

En la práctica, esto equivale a que $|w_d| \le C$.

La formulación como un problema de programación cuadrática implica: (i) que la solución obtenida es global, ya que la formulación del problema es convexa; (ii) que la solución es única, ya que la función de costo es estrictamente convexa, y (iii) que muchos de los pesos w_d serán cero; de tal forma, que solamente una fracción de los datos originales contribuyen en la función de error. Esta conclusión es basada en las condiciones de Karush-Kuhn-Tucker para la programación cuadrática. Los puntos para los cuales sus correspondientes multiplicadores de Lagrange son diferentes de cero (esto es, con $L_\varepsilon(.,.) > 0$) son llamados vectores de soporte, y son los que permiten realizar la aproximación de y_t en la Ecuación 25. Así, se espera que a medida que aumente el valor de ε, disminuya la cantidad de vectores de soporte; pero que también disminuya la capacidad de aproximación del modelo a los datos originales.

Funciones de núcleo (kernel)

En la Ecuación 25, $k(x, x_d)$ representa una función de núcleo que permite llevar el punto x a un espacio altamente dimensional parametrizado por los puntos x_d. Existen varias funciones que son típicamente utilizadas como núcleos (Velásquez *et al*., 2009):

Kernel	Ecuación	
Lineal	$k(x, x_d) = x^T \cdot x_d$	**Ecuación 33-**
Polinomial	$k(x, x_d) = (a_1 x^T \cdot x_d + a_2)^d, a > 0$	**Ecuación 34-**
Gaussiana o *RBF*	$k(x, x_d) = \exp(-1/(a_1^2 \times (x - x_d)^2)), a > 0$	**Ecuación 35-**
Exponencial	$k(x, x_d) = \exp(-1/(a_1^2 \times (x - x_d)))$	**Ecuación 36-**
Sigmoidal	$k(x, x_d) = Tanh(a_1 x^T \cdot x_d + a_2)$	**Ecuación 37-**

Tabla 4- Funciones Kernel

Las constantes a_1 y a_2 son dependientes del problema y sus valores deben ser ajustados por el modelador. La función gaussiana o *RBF* presenta un mejor desempeño para la solución de problemas de regresión y de predicción de series temporales. Consecuentemente, en este trabajo sólo se considerará el uso de esta función de núcleo.

1.4.7.5. Optimización de los parámetros de modelo *SVM-R*

A partir de un conjunto de datos (*output*), una *SVM* con función *kernel RBF* (cuya desviación estándar es a ó σ) y valores para las constantes C y ε es posible determinar los valores óptimos de los parámetros b y w_d de la *SVM* definida en la Ecuación 25, a partir de la solución del problema de programación cuadrática definido anteriormente.

Tal como ya se indicó, el modelo definido por las ecuaciones anteriores tiene una solución única y global, pero dependiente de los valores asumidos para σ, C y ε. La dificultad radica en que estos valores son dependientes del problema particular analizado, y en que no hay métodos heurísticos para la determinación de sus valores, por lo que el experto debe fijar sus valores de forma heurística (Cherkassky y Ma, 2004; Velásquez *et al*., 2009).

Considerando algunas fuentes existentes en la regresión *SVM* (Cherkassky y Mulier, 1998; Vapnik, 1998; Kwok, 2001) dan algunas recomendaciones sobre el ajuste adecuado de los parámetros de SVM, claramente no hay consenso y muchas opiniones son contradictorias. Por lo tanto, remuestreo (prueba y error) sigue siendo el método de elección para muchas aplicaciones.

Por consiguiente, para obtener los mejores parámetros del modelo *SVM* se asume que $\varepsilon = 0$, ya que se ha demostrado en problemas de regresión de baja y de alta dimensionalidad que la función de precisión o pérdida ε es insensible, ya que en realidad la estimación y generalización de este parámetro es considerada una función asintóticamente óptima para determinada densidad de ruido (Smola et al., 1998) (Cherkassky y Ma, 2004). Cherkassky y Ma (2004), sugieren en sus resultados empíricos que la elección de ε propuesto, el valor del parámetro de regularización C tiene sólo un efecto insignificante sobre la generalización de rendimiento de la regresión SVM (siempre y cuando C es mayor que un umbral determinado analíticamente a partir de los datos de entrenamiento).

Además, existe una regresión denominada SVM-ν en la cual el valor de ε es controlado por la variable $\nu \in (0,1)$. ν es el límite superior o el límite inferior de la fracción de los puntos de error dentro del tubo con ε-insensible (ver Figura 16). Por lo tanto, un buen ε se puede encontrar automáticamente escogiendo, un correcto ν que ajusta el nivel de exactitud de los datos observados. Por lo tanto, la función de minimización presentada en la Ecuación 26 es ajustada para esta nueva variable:

$$R(C,\varepsilon) = C\left(\nu \cdot y_d + \frac{1}{D}\sum_{d=1}^{D} L_\varepsilon(y_d, \hat{y}_d)\right) + \frac{1}{2}\sum_{d=1}^{D} w_d^2$$

Ecuación 38-

Por último, en este trabajo será implementado el modelo $\nu - SVM$, para el cual se asume el criterio de $\varepsilon = 0$ de (Cherkassky and Ma, 2004) y se optimizan lo valores de los parámetros ν y C, por medio de algoritmo evolutivo denominado Evolución Diferencial (Rainer Storn and K Price, 1997). Este último método será explicado en el numeral 1.4.8.2.

1.4.8. Desempeño de los modelos regresivos

Cualquier modelo o método para la predicción tiene sentido sólo si se definen y aplican criterios apropiados para medir su rendimiento (capacidad predictiva y ajuste). Para los modelos basados en la regresión, los residuos (errores de predicción) e_i:

$$e_i = y_i - \hat{y}_i$$

Ecuación 39-

es la base de las medidas de desempeño, donde y_i es valor observado (experimental, "verdadero") y $\hat{y}_i$ el valor estimado (modelado) de un objeto i. Los diferentes enfoques

para estimar el rendimiento de un modelo, utilizan diferentes estrategias para la selección de los objetos que se emplean para la generación y prueba de los modelos, y así definir diferentes medidas matemáticas derivadas de los errores de predicción. Uno de los objetivos de las medidas de rendimiento es seleccionar el mejor modelo aproximado. Otro objetivo es evaluar la capacidad predictiva del modelo frente a un nuevo conjunto datos. Por otra parte, se puede realizar la selección del mejor modelo en función de sus parámetros mediante un algoritmo genético (AG), por lo general varios cientos de modelos deben que ser comparados para este fin. Por último, el tiempo de cálculo no debe jugar un papel importante, ya que una buena estimación de las variables de salida en los nuevos casos justifica un alto esfuerzo computacional (Varmuza y Filzmoser, 2009).

En general, una medida de rendimiento que sólo caracteriza lo bien que se ajustan las predicciones del modelo a los datos observados no es aceptable, y en consecuencia una estimación realista del desempeño requiere probar el modelo con nuevos casos. A continuación, se presentan tres métodos que permiten reducir el tamaño de conjunto de datos (calibración, validación y prueba), evaluar múltiples escenarios y determinar capacidad predictiva y nivel de ajuste.

1.4.8.1. Validación cruzada

Un aspecto crucial de cualquier técnica estadística es replicar y generalizar los resultados a muestras de la misma población. En el caso de la regresión simple y múltiple el problema se agudiza por la selección aleatoria de las variables y su cantidad, lo cual en ocasiones produce, entre otros problemas, el aumento del número de predictores que no siempre todos aportan significativamente en mejorar las predicciones. Por lo tanto, es necesario, dividir las bases de datos de las variables independientes como dependientes en grupos de calibración (se utiliza para la formación del algoritmo), validación (para evaluar el rendimiento del algoritmo) y en algunos casos de prueba, seleccionados de forma aleatoria.

Sin embargo, en el proceso de formación del algoritmo (calibración) es necesario consolidar cuáles y cuántas muestras y variables predictoras representan el fenómeno estudiado, sin generar un sub-ajuste (*underfitting*) o sobre-ajuste (*overfitting*) en las predicciones y tener un modelo más o menos parsimonioso: optimizar la arquitectura del modelo y obtener una estimación realista del rendimiento de la predicción de los nuevos casos (Figura 17).

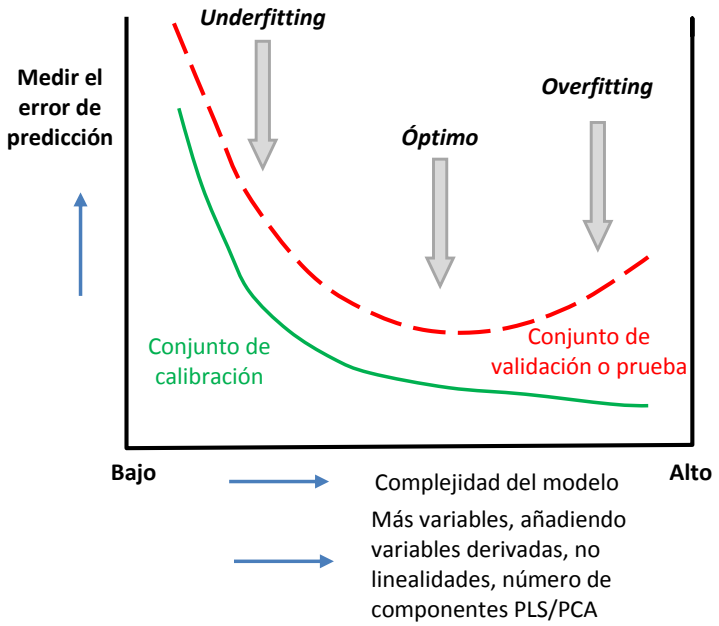


Figura 17- Complejidad del modelo en comparación con error de predicción para los conjuntos de calibración y validación (adaptado de Varmuza y Filzmoser, 2009)

Por lo tanto, técnicas como validación cruzada (VC) se pueden usar para tal fin, la cual consiste en dividir el conjunto de calibración en un conjunto de entrenamiento y otro de validación (por ejemplo para determinar número óptimo de componentes *PLS*). Entonces, el procedimiento de VC es seleccionar de forma aleatoria de un conjunto de n objetos s segmentos (partes), cada segmento con m muestras. El número de segmentos puede ser de 2 a n, como se muestra en la Figura 18, donde se divide en cuatro segmentos y selecciona tres para entrenar el modelo y una para validar su resultado. Este procedimiento se repite hasta que cada segmento ha sido un conjunto de validación. Pero más allá de validar el modelo, permite establecer qué segmento (conjunto de muestras) afecta notablemente el error en la predicción. A partir de esto, se constituye una matriz de residuos de predicción de la cual se selecciona el modelo que género el menor error. Una medida para evaluar el error, por ejemplo, es el error medio de los cuadrados (en inglés *mean squared errors-cross validation* MSE_{CV}). De acuerdo a la Figura 18, el menor resultado de MSE_{CV} o una medida similar, indicara la complejidad del modelo óptimo (Varmuza y Filzmoser, 2009).

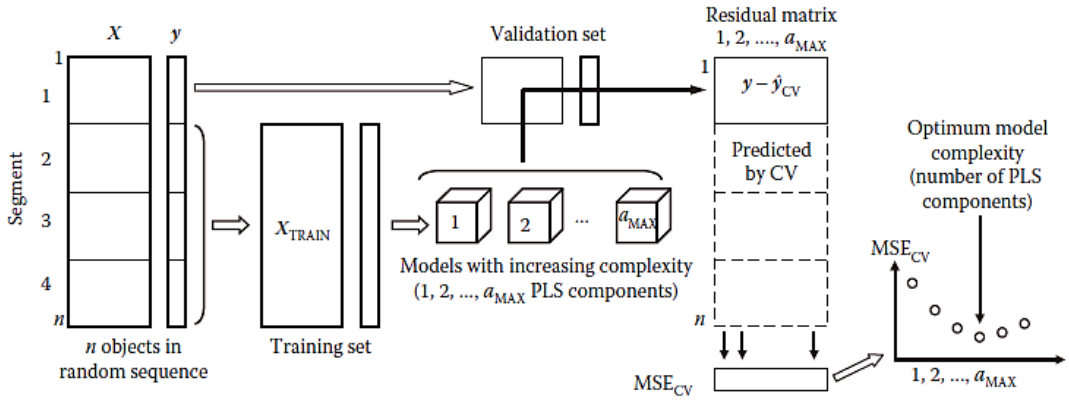


Figura 18- VC con cuatro segmentos (dejando un cuarto por fuera) se aplica a la estimación de la complejidad óptima del modelo (Varmuza y Filzmoser, 2009)

1.4.8.2. Algoritmos evolutivos

La selección de variables es un problema de optimización. Por lo tanto, algoritmos de optimización inspirados en la biología del proceso de la selección natural han estado en uso desde la década de 1950 (Mitchell, 1998), y se conocen a menudo como algoritmos evolutivos. El algoritmo genético (AG) es uno de esos métodos, y fue inventado por John Holland en 1960 (Holland, 1975), otro son las estrategias de evolución (EE) que fueron introducidas por Storn y Price, 1990 (Storn y Price, 1997). A continuación se realiza una breve descripción de estos métodos de optimización.

a) Algoritmos genéticos

Los algoritmos genéticos se aplican a operaciones lógicas, por lo general la selección particular de las variables puede ser denotado por un vector que consiste en componentes binarios: un "1" indica que se selecciona la variable, un "0"' no se selecciona. Tal vector de longitud m (el número total de variables) define uno de los posibles subconjuntos de variables y es simplemente una cadena de bits. Dicho vector recibe el nombre de cromosoma que contiene m genes. Un conjunto de cromosomas diferentes (cada uno en representación de los posibles subconjuntos de variables), se llama población (Figura 19). En el transcurso de sucesivas generaciones, los miembros de la población tienen más probabilidades de representar un mínimo de una función objetivo (Varmuza y Filzmoser, 2009; Ardia *et al*., 2011). Los algoritmos genéticos han demostrado ser útiles para los métodos heurísticos de optimización global, en particular para los problemas de optimización combinatoria (*e.g.* calibración de múltiples parámetros de un modelo de forma simultánea) (Mullen *et al*., 2011).

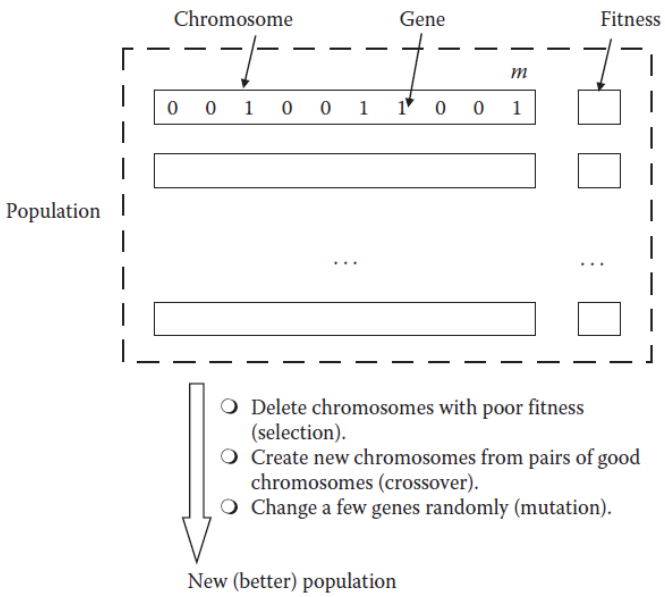


Figura 19- Esquema de un AG aplicado a la selección de variables. El primer cromosoma define un subconjunto de cuatro variables, seleccionadas de $m = 10$ variables. *Fitness* es una medida del desempeño de un modelo construido a partir del correspondiente subconjunto de variables

En general, la población contendrá cromosomas con diferente estado físico, y la estrategia de AG es producir mejores poblaciones. La siguiente población en cadena de la evolución se obtiene por acciones inspiradas en la biología de la siguiente manera:

- Algunos de los peores cromosomas se eliminan y sustituyen por nuevos cromosomas (Competencia).
- Nuevos cromosomas se derivan a partir de pares de cromosomas buenas, sobre todo por una llamada de cruce (Figura 4.21). La idea es que una combinación de dos buenos cromosomas puede producir una aún mejor.
- Un pequeño porcentaje de los genes se cambian aleatoriamente por mutación, es decir algunos "0" se cambian por "1" y viceversa. Esta acción aleatoria se debe evitar para ser atrapado en óptimos locales. La tasa de mutación puede disminuir durante la formación para lograr una mejor convergencia.

Determinar la aptitud o *fitness* para los nuevos cromosomas completa una generación de un entrenamiento del AG. El procedimiento se repite hasta que se alcanza un criterio de terminación (por ejemplo, ningún aumento en la aptitud de los mejores cromosomas o el máximo número definido de las generaciones alcanzadas).

b) Estrategias de evolución

En la década de 1990 Rainer Storn y Kenneth Price desarrollaron una estrategia de evolución que denominaron Evolución Diferencial (ED) (Storn y Price, 1997). ED está particularmente bien adaptado para encontrar el óptimo global de una función de valor real de los parámetros con valores reales, y no requiere que la función sea continua o diferenciable. En los casi 16 años desde su invención, ED ha sido aplicado con éxito en una

amplia variedad de campos, desde la física computacional a diferentes operaciones de investigación, como catalogaron (Price *et al*., 2005).

Este algoritmo es una técnica evolutiva que en cada generación transforma un conjunto de vectores de parámetros, denominado población, en otro conjunto de vectores de parámetros, cuyos miembros tienen más probabilidades de minimizar la función objetivo. Con el fin de generar un nuevo vector de parámetros, ED perturba un viejo vector de parámetros con la diferencia escalada entre dos vectores de parámetros seleccionados al azar (Ardia *et al*., 2011).

La variable NP representa el número de vectores de parámetros en una población. En la generación 0, NP supone que valor óptimo del vector de parámetros se realiza, o bien utilizando valores aleatorios entre los límites superior e inferior para cada parámetro o utilizando valores dados por el usuario. Cada generación implica la creación de una nueva población a partir de los miembros actuales de la población $x_{i,g}$, donde i hace referencia a los índices de los vectores que componen la población y g indexa la generación. Esto se logra mediante una mutación diferencial de los miembros de la población. Un vector de prueba de parámetros mutante $v_{i,g}$ es creado mediante la selección de forma aleatoria de tres miembros: $x_{r0,g}, x_{r1,g}$ y $x_{r2,g}$ (Price *et al*., 2005; Ardia *et al*., 2011). Entonces, $v_{i,g}$ es generado como:

$$v_{i,g} = x_{r0,g} + F \cdot (x_{r1,g} + x_{r2,g})$$

Ecuación 40-

donde $F \in (0,1+)$ es un factor escalar positivo, que controla la tasa en la cual el algoritmo evoluciona. Mientras que no se determine un límite superior en F, los valores efectivos son rara vez superiores a 1 (Price *et al*., 2005).

Después de completar la primera mutación, ésta continua hasta que las mutaciones equivalentes a la longitud de (x) se han hecho o hasta que $rand > CR$, donde CR es la probabilidad cruzada[3] $CR \in [0,1]$, y $rand$ se utiliza para denotar un numero aleatorio $\mathcal{U}(0,1)$. La probabilidad de cruce CR es un valor definido por el usuario que controla la fracción de los valores de los parámetros que se copian de los mutantes. Si el número aleatorio es menor o igual a CR, el parámetro de ensayo se hereda del mutante, $v_{i,g}$, de lo contrario, el parámetro se copia a partir del vector, $x_{i,g}$. CR es solo una aproximación de la verdadera probabilidad, pCR , pero no representa exactamente la probabilidad que el valor de un parámetro será heredado del mutante, ya que siempre se produce al menos una mutación. La mutación se aplica de esta manera a cada miembro de la población (Ardia *et al*., 2011).

[3] Probabilidad de cruce para indicar una relación de cuántos miembros de una población serán seleccionados como parejas para el apareamiento.

Si se encuentra un elemento del vector de parámetros v_j violando los límites después de la mutación y cruce, éste es reiniciado, donde j es el índice dentro de un vector de parámetro. Esto garantiza que los candidatos miembros de la población considerados como infractores se establezcan una cierta cantidad aleatoria lejos de los demás miembros, de tal manera que se garantice el cumplimiento de los límites (Ardia *et al*., 2011).

Luego, se determinan los valores de la función objetivo asociados con los herederos de v. Si un vector de ensayo $v_{i,g}$, tiene un valor en la función objetivo igual o menor que el vector $x_{i,g}$, éste se sustituye por $v_{i,g}$ en la población, de lo contrario $x_{i,g}$ permanece. El algoritmo se detiene después de un número determinado de generaciones, o después de que el valor de la función objetivo asociada con el mejor miembro se ha reducido por debajo de un umbral establecido, o si no es capaz de reducir el mejor miembro encontrado en las iteraciones establecidas (Ardia *et al*., 2011). Por último, Price *et al.* (2005) encontraron que la variación de los valores de NP y CR resultaron ser más efectivos en solución de una variedad de problemas.

1.4.8.3. Análisis de predictibilidad y ajuste

El proceso de evaluar el desempeño de un modelo matemático requiere realizar estimaciones tanto subjetivas como objetivas de la relación que existe entre el comportamiento de lo simulado con respecto a lo observado. El método más básico para evaluar el desempeño de un modelo en términos de comportamiento funcional es a través de una inspección visual de las diferencias entre las predicciones y observado. Al hacerlo, se puede formular una evaluación subjetiva del comportamiento del modelo con respecto a la sistemática (baja estimación o sobre estimación) (Dawson *et al*., 2007). Para la evaluación objetiva, se requiriere de la determinación de una o más estimaciones matemáticas del error derivadas de los residuos de predicción (Ecuación 39), los cuales permiten caracterizar el desempeño de un modelo y son conocidas como métricas.

Entonces las predicciones generadas por los modelos presentados en los numerales 1.4.5, 1.4.7.1 y 1.4.7.3 pueden ser evaluadas mediante las siguientes métricas: coeficiente de determinación, coeficiente de correlación de Spearman, la raíz del error cuadrático medio (*Root Mean Square Error-RMSE*), y la suma de los errores cuadrados de las predicciones (*Predicted Residual Error Sum of Squares-PRESS*). A continuación se realiza una breve descripción de cada métrica.

En las siguientes ecuaciones y_i es la variable observada o medida (concentraciones), $\hat{y}_i$ es la variable estimada por un modelo (donde $i = 1 \dots n$, siendo n la cantidad de datos) y $\bar{y}_i$ es la media de los valores medidos.

a) Coeficiente de correlación de Spearman

El análisis más común es el análisis de correlación (*r*) de Pearson (Ecuación 41). Este tipo de análisis presupone que las variables son ordinales o continuas y que la distribución de

estas variables se acerca a la distribución normal. Esta última condición no se cumple y por lo tanto se busca otro tipo de correlación que puede determinar el grado de relación entre caudal y nivel, cuando se desconoce el tipo de distribución que los datos generan.

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

Ecuación 41-

Por consiguiente, la correlación no paramétrica (ρ) de Spearman acepta variables de libre distribución, relaciones no lineales e incluso ordinales. El coeficiente de Spearman es el coeficiente de correlación de Pearson, pero aplicado después de transformar las puntuaciones originales en rangos. Toma valores entre -1 y 1, y se interpreta de la siguiente manera:

- Los valores cercanos a uno indican una correlación muy buena y los valores cercanos a cero indican una correlación mínima o nula.
- En cuanto al signo, si éste es positivo indica una correlación directa, mientras que un signo negativo indica correlación inversa entre las variables.

b) Coeficiente de determinación

El coeficiente de determinación R^2, mide la proporción de variabilidad total de la variable dependiente respecto a su media que es explicada por el modelo de regresión. A medida que R^2se acerca a 1, la ecuación de regresión es más confiable, y entre más cercano esté el R^2 de cero, la ecuación es menos confiable ya que la suma de cuadrados de la regresión tiende a cero (numerador de la Ecuación 42). Se puede calcular de la siguiente manera el valor de este coeficiente:

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

Ecuación 42-

c) Raíz del error cuadrático medio (RMSE)

Esta métrica permite determinar en las unidades reales de la variable el nivel de ajuste entre el conjunto de datos observados y las predicciones. Es una métrica no negativa que no tiene límite superior y para un modelo perfecto su resultado sería igual a cero. Se compone de una medida ponderada del error en el que las mayores desviaciones entre los valores observados y estimados aportan más (Dawson *et al*., 2007).

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

Ecuación 43-

d) Suma de los errores cuadrados de las predicciones (PRESS)

Es una forma de validación cruzada, la cual proporciona una medida del ajuste de un modelo cuando éste es entrenado sin una de las muestras del conjunto de calibración $y_{(i)}$, pero la predicción de este conjunto de datos se realiza teniendo en cuenta la variable eliminada. Este procedimiento se repite para todas las muestras de calibración, por lo tanto será una validación cruzada tipo *Leave-one-out*.

$$PRESS = \sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2$$

Ecuación 44-

La "mejor" regresión tendrá relativamente una menor suma de los errores cuadrados de predicción. Por otra parte, esta métrica permite determinar si un modelo está sobre-parametrizado, ya que los residuos de las predicciones de las muestras incluidas en la calibración del modelo son pequeños, pero mucho mayor para la observación que se excluye. Sin embargo, esta métrica no permite observar el error en las unidades reales de las variables. Por lo tanto, para este fin se implementó la raíz del error cuadrático medio con validación cruzada ($RMSE_{VC}$) definida por la siguiente ecuación:

$$RMSE_{VC} = \sqrt{\frac{PRESS}{n-1}}$$

Ecuación 45-

2. MATERIALES Y MÉTODOS

2.1. PUNTOS DE MONITOREO

Para evaluar los métodos, modelos y algoritmos, implementados y desarrollados en el marco del presente trabajo de investigación, se utilizó la información proveniente de las caracterizaciones de la calidad del agua residual obtenidas del monitoreo de tres sistemas de saneamiento urbano. Dos de ellos son el afluente a plantas de tratamiento de aguas residuales (PTAR): San Fernando ubicada en la ciudad de Medellín, Colombia, y *Fontaines-sur-Saône* que pertenece a la comunidad urbana de *Grand Lyon,* Francia. El tercer punto es afluente a la estación elevadora (EE) de aguas residuales denominada Gibraltar localizada en la ciudad de Bogotá, Colombia.

La caracterización de cada uno de los puntos de monitoreo está conformada por la siguiente información:

Punto de monitoreo	No. de muestras	Determinantes cuantificados en laboratorio				Medición de los espectros UV-Visible	Triplicado (Si/No)
		Tipo de monitoreo /muestra	DQO	DQOf	SST		
PTAR San Fernando	124	manual/ puntual	X	X	X	*on line/in situ*	No
PTAR *Fontaines-sur-Saône*	135	manual/ puntual	X	X	X	*off line/in situ*	Sí
EE Gibraltar	42	manual/ puntual	X	X	X	*off line/in situ*	Sí

Tabla 5- Características de las campañas de monitoreo de los casos de estudio

De esta información, se utilizó inicialmente los datos de la PTAR San Fernando, con los cuales se realizaron pruebas piloto de los métodos y algoritmos desarrollados, que serán presentados en numeral 3. Por otra parte, la información de los otros puntos de monitoreo fue implementada en un algoritmo que contempla el análisis de incertidumbre a partir de los ensayos y mediciones realizados por triplicado, detección de *outliers*, modelos regresivos y cuantificación del desempeño de los modelos. Dicho algoritmo será descrito en el numeral 3.5.

Los datos de DQO, DQOf, SST y espectros de absorbancia de cada punto de monitoreo se pueden observar en el numeral 2.1.1. A continuación se realiza una breve descripción de los puntos de monitoreo.

2.1.1. Planta Tratamiento de Aguas Residuales, San Fernando – Medellín, Colombia

El río Medellín nace aguas arriba del Municipio de Caldas, recibiendo a lo largo de 100 kilómetros de longitud la descarga de un poco más de 64 quebradas afluentes que recorren la zona urbana, densamente poblada. Anteriormente estas quebradas eran utilizadas como fuentes receptoras de contaminación doméstica, industrial y comercial, antes de la implementación del Programa de Saneamiento del Río Medellín y sus Quebradas Afluentes (EPM, 2009).

Dentro de dicho programa, se diseñó y construyó la Planta de Tratamiento de Aguas Residuales San Fernando (PTAR) localizada en el municipio de ltagüí, como se puede observar en la Figura 20. 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 planta se diseñó para un caudal de saturación de 4.8 m^3/s y se definió que su construcción sería por fases: la primera y de actual funcionamiento para un caudal de 1.8 m^3/s. San Fernando incluye los procesos de tratamiento preliminar, primario y secundario mediante lodos activados (espesados y estabilizados en digestores anaeróbicos y luego deshidratados y enviados a un relleno sanitario) (EPM, 2007). En el tratamiento secundario, se remueve entre el 80 % y el 85 % de la contaminación del agua residual antes de ser devuelta al río Medellín. Durante el año 2006 la planta trató un volumen de 39.4 millones de m^3, al tiempo que generó alrededor de 36000 toneladas de biosólidos (EPM, 2009).

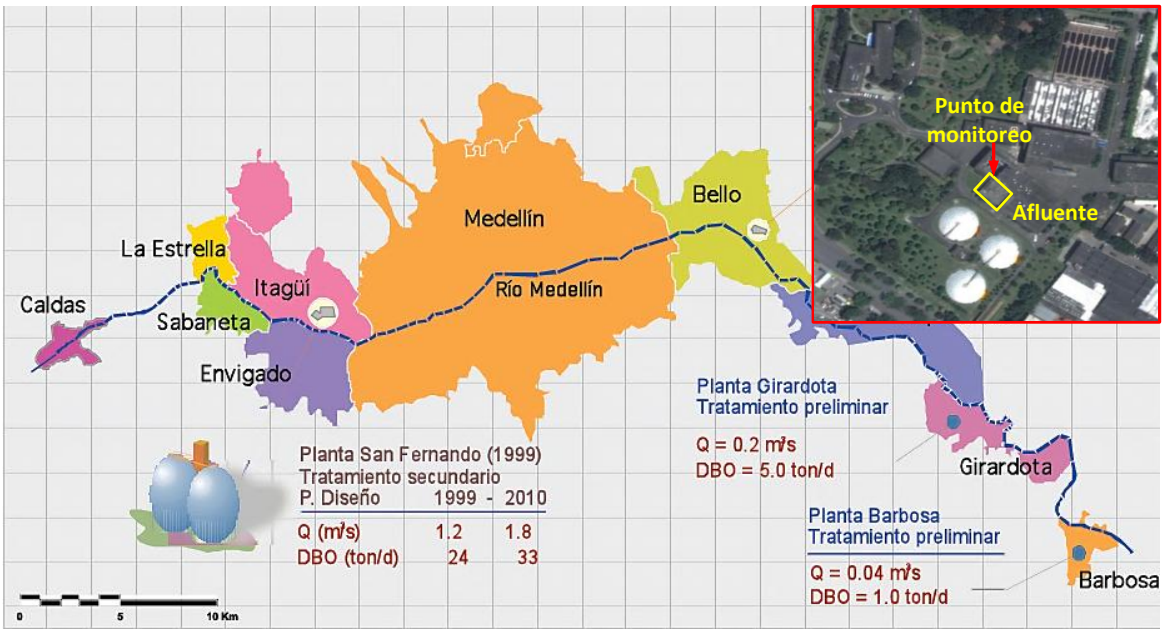


Figura 20- Ubicación de la PTAR San Fernando (EPM, 2009; Google Earth, 2012)

Figura 21- Punto de monitoreo (Izq.) y sistema de control de la sonda spectro::lyser en el afluente de la PTAR San Fernando

La información suministrada por la EPM del afluente de la PTAR San Fernando corresponde a los espectros de absorbancia y las concentraciones de SST, DQO y DQOf, que pertenecen a 124 muestras 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 . Los valores de concentración obtenidos en los ensayos estándar de laboratorio y las mediciones de los espectros de absorbancia UV-Vis serán presentados respectivamente a continuación:

Las ordenadas de la Figura 22, presentan los valores de concentración en mg/L y en el eje de las abscisas el índice de cada una de las muestras. Cada muestra fue tomada en diferentes instantes de tiempo para cada caso de estudio. Por otra parte, en las Figura 32 y Figura 33 se presenta con color rojo y azul el valor máximo y mínimo de la concentración para cada muestra respectivamente, y en color verde se presentan el valor intermedio de concentración que no corresponde a la media, pero se encuentra en el rango establecido por el valor máximo y mínimo de cada concentración.

Por otra parte, las Figura 23, Figura 34 y Figura 35 muestran los espectros UV-Vis medidos de las muestras tomadas en cada punto de monitoreo por medio de una sonda con un paso de luz de 2 mm. En dichos gráficos se presenta en el eje *x* las longitudes de onda en nanómetros [200-750 nm en pasos de 2.5 nm], en el eje *y* los índices de cada una las muestras a la cual pertenece un espectro, y por último en el eje *z* los valores de absorbancia de cada espectro en el rango UV-Vis.

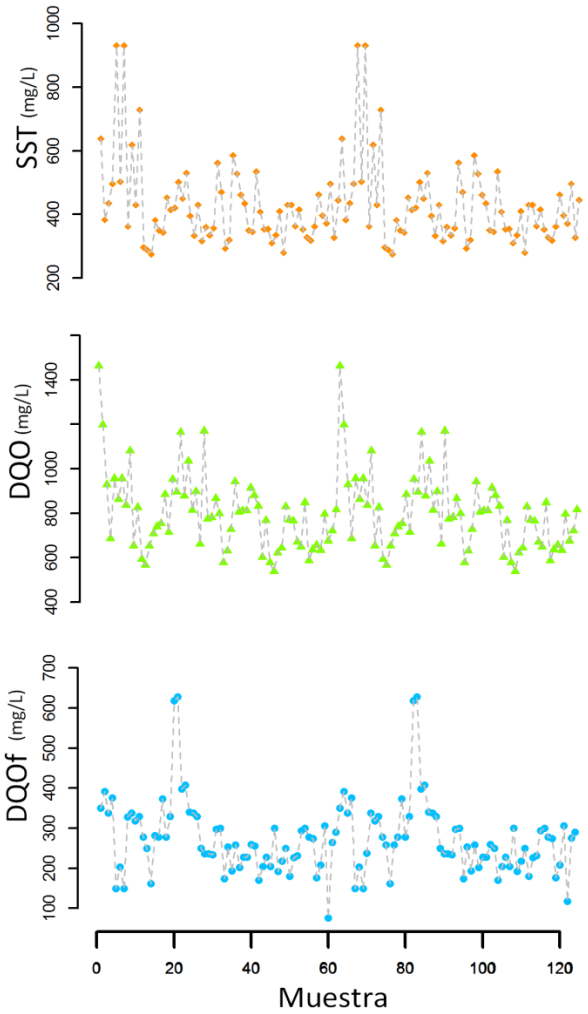


Figura 22- Valores de las concentraciones de los SST, la DQO y la DQO filtrada del afluente de la PTAR San Fernando (Enero a Septiembre de 2011)

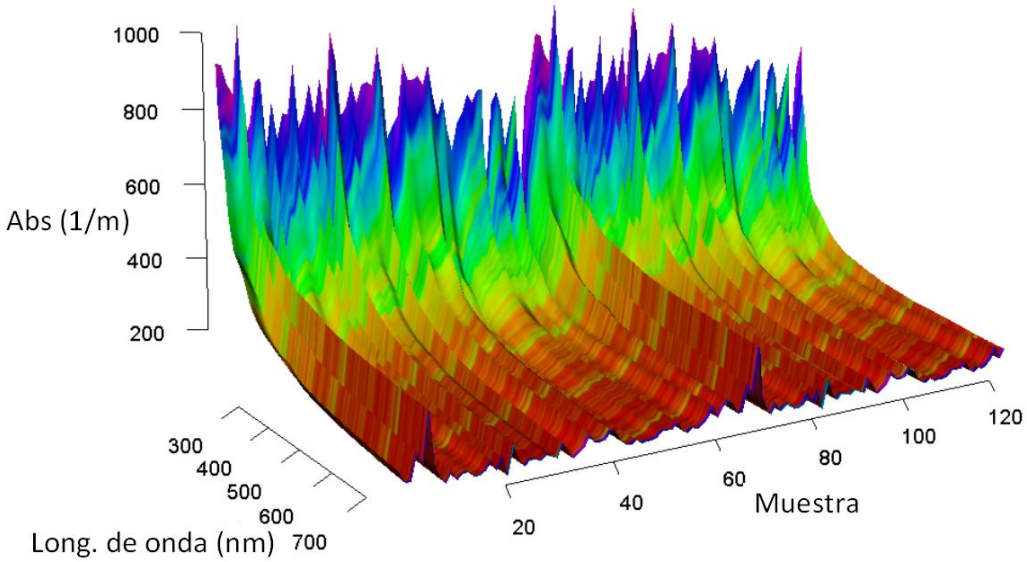


Figura 23- Espectros de absorbancia del afluente de la PTAR San Fernando (Enero a Septiembre de 2011)

2.1.2. Estación Elevadora de Gibraltar (EEG) – Bogotá, Colombia

La estación elevadora de Gibraltar se encuentra ubicada en el suroeste de la ciudad de Bogotá, Colombia. Esta estación tiene por fin bombear las aguas residuales y pluviales de la subcuenca El Tintal, pero recibe el 90 % de las aguas de esta cuenca (aproximadamente 2226 hectáreas), con una población de 423216 en el área aferente según el estudio realizado por HMV (2007).

Aunque la subcuenca aferente a Gibraltar cuenta con un sistema separado de redes (residuales y pluviales) se convierte en un sistema combinado debido a la presencia de: conexiones erradas y una separación inadecuada de las aguas residuales de las pluviales en las viviendas del sector. Por otra parte, la clasificación de tipo de aguas residuales se divide de la siguiente forma: (i) 83.13 % de las actividades domésticas; (ii) 0.31 % de las actividades comerciales y (iv) 3.18 % de vertederos industriales. Por otra parte, 1.03 % corresponde a las aguas de escorrentía superficial, mientras que del 12.35 % no se pudo establecer su origen (HMV, 2007).

El afluente es descargado al río Bogotá, para lo cual la estación cuenta con cuatro bombas tipo Tornillo de Arquímedes (o sin fin) con un capacidad individual de 1.27 m^3/s cada uno (ver Figura 24).

Figura 24- Ubicación geográfica de la Estación Elevadora de Gibraltar (Google Earth, 2012)

Con el fin de caracterizar las aguas residuales que recibe la estación, se realizó el monitoreo en continuo, *in situ* y *on line* de su afluente, y además se tomaron muestras puntuales y efectuaron ensayos laboratorio por triplicado a las mismas, para detectar y cuantificar los determinantes presentados en la

Tabla 5.

Por lo tanto, para realizar el monitoreo en continuo se diseñó, construyó e instaló en la cámara de recepción del afluente una estructura para soportar un bote (Figura 25), el cual a su vez soportaba un conjunto de sondas (Figura 26) para la medición de los siguientes determinantes de calidad de agua: oxígeno disuelto (FDO), turbiedad (Visoturb), sólidos suspendidos totales (Visolid), pH (Sensolyt), redox (Sensolyt-IQ) y espectro de absorbancia UV-Visible (*spectro::lyser*). En este orden, las cinco primeras sondas son de la empresa alemana WTW y la última de la empresa austriaca s::can.

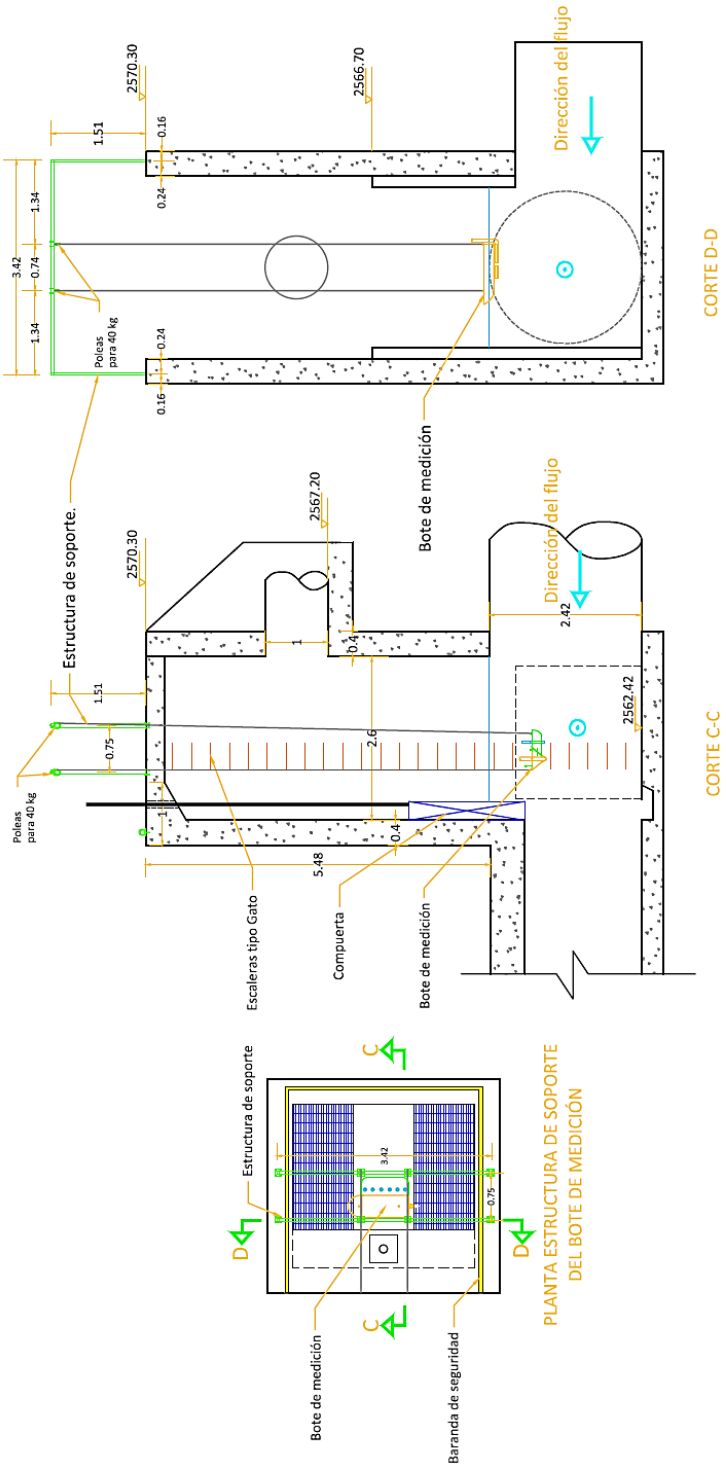


Figura 25- Planta y cortes de la cámara de recepción del afluente y estructura de soporte del sistema de medición (Autor, 2012)

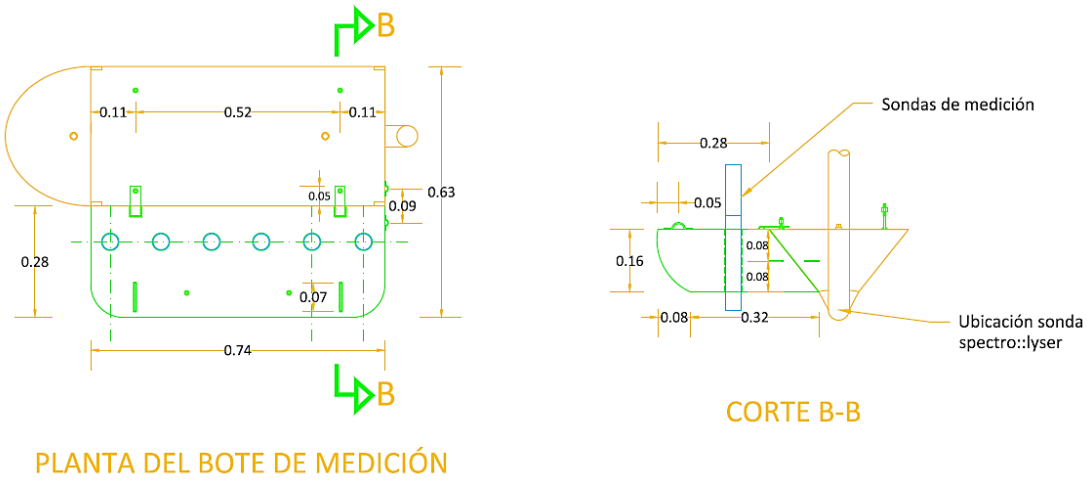


Figura 26- Bote de soporte de las sondas de medición (Autor, 2012)

De la información recopilada en la jornada de monitoreo y muestreo se utilizó para esta investigación los valores por triplicado obtenidos en los análisis de laboratorio de SST, DQO y DQOf, y los espectros UV-Vis medidos son una sonda *spectro::lyser* (paso de luz de 5 mm) de cada muestra puntual y por triplicado.

En la Figura 27 se presenta a la izquierda el bote y las sondas de medición en operación. A la derecha se puede observar la estructura de soporte, y el sistema de operación y almacenamiento de datos (SOAD) de las sondas.

Figura 27- Bote y sondas de medición (Izq.) y dispositivos para control y almacenamiento de datos de las sondas de monitoreo (Der.) (Autor, 2012)

Del afluente de la estación elevadora se tomaron un total de 41 muestras puntuales a las cuales se realizaron los ensayos de laboratorio por triplicado para cuantificar SST, DQO y DQOf (Figura 28). Por otra parte, la medición del espectro UV-Vis de cada muestra se realizó por triplicado de forma *in situ* y *off line*, por medio de una sonda spectro::lyser con un paso de luz de 5 mm (Figura 29).

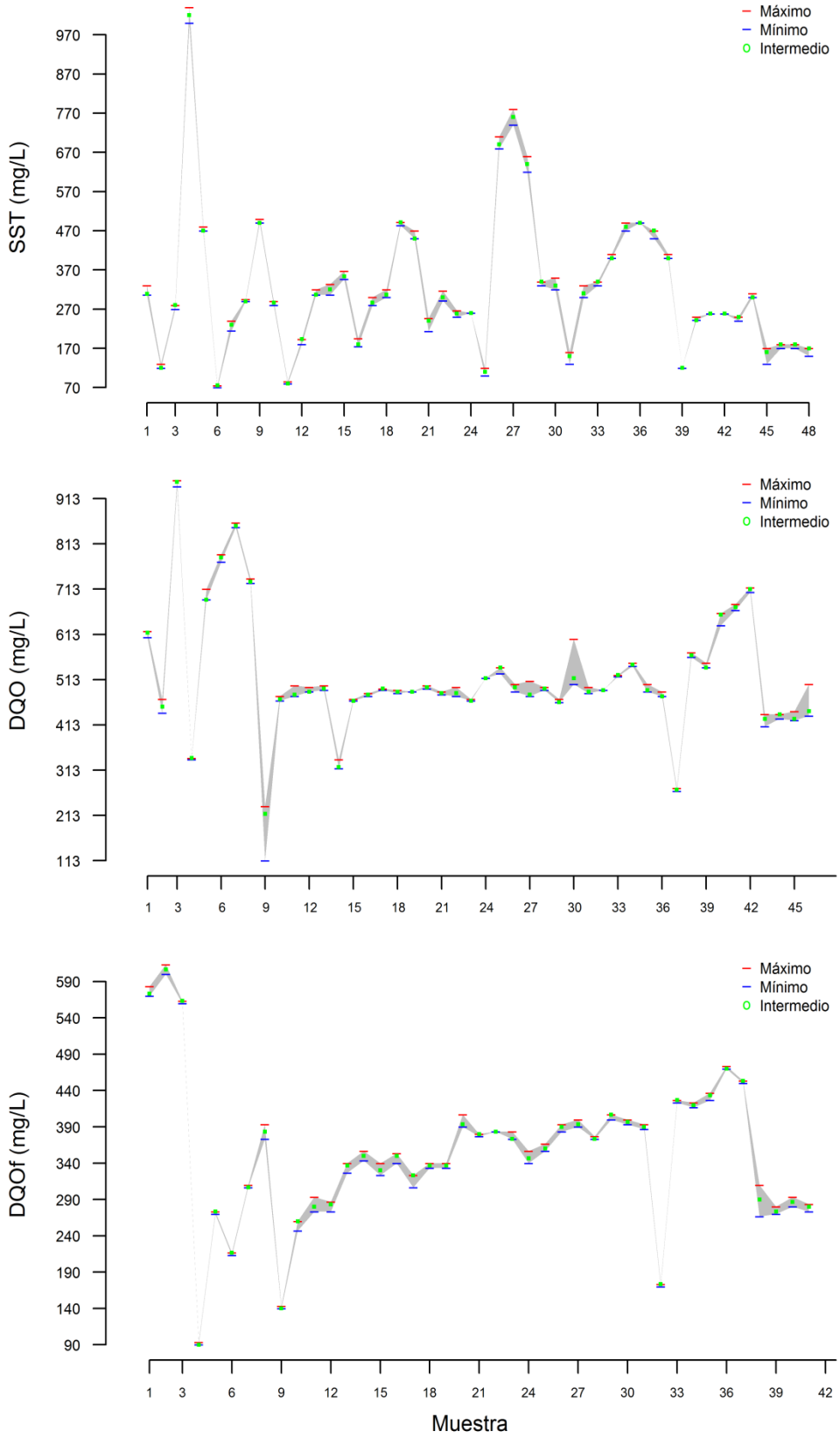


Figura 28- Valores de las concentraciones de los SST, la DQO y la DQO filtrada del afluente de la estación elevadora de Gibraltar

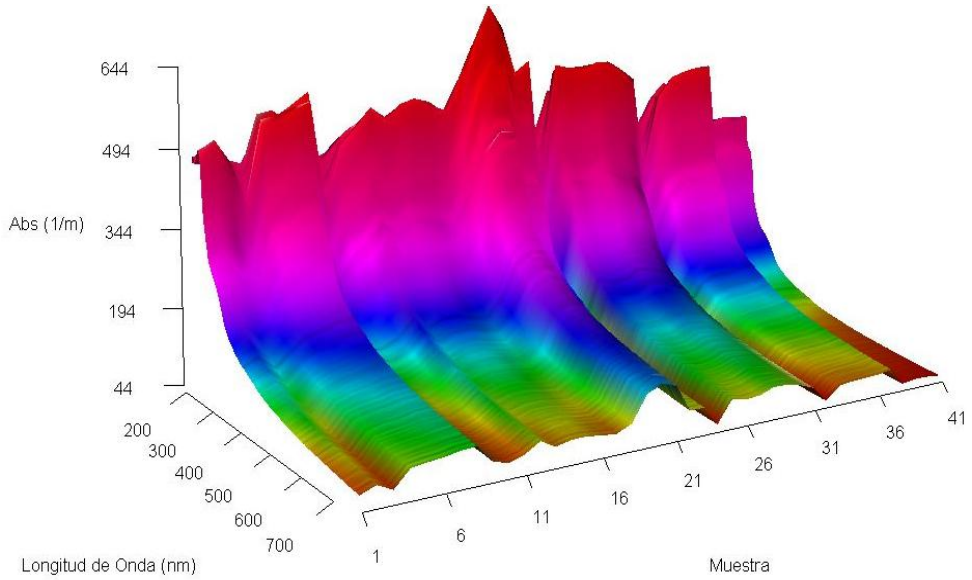


Figura 29- Espectros UV-Vis con los valores máximos valores de absorbancia del afluente de la estación elevadora de Gibraltar (18, 21 y 25 de octubre, y 2, 4, 8, 11, 26 y 29 de noviembre de 2011)

2.1.3. Planta de Tratamiento de Aguas Residuales de *Fontaines-sur-Saône*, Francia

Este sistema de saneamiento urbano ubicado en la ciudad de Lyon, Francia, trata el agua residual de la red de alcantarillado del sector *Fontaines-sur-Saône* de dicha ciudad. La planta tiene una capacidad de tratar el agua residual de 30000 personas (Lepot, 2012).

En este caso de estudio la información fue obtenida de la Tesis doctoral de Lepot (2012), quien realizó la toma de muestras del afluente a la PTAR de *Fontaines-sur-Saone,* después de haber pasado por el sistema de pretratamiento de la planta: cribado, desarenado y trampa de grasas. Para realizar la caracterización del agua residual, Lepot (2012) utilizó un banco de ensayo denominado Claude Chappe (ver Figura 30), en el cual realizó la toma de muestras puntuales por triplicado en tiempo seco y lluvia, así como la medición de los espectros UV-Vis. Además, se monitoreo continuo, *off line* e *in situ* los siguientes determinantes: turbiedad, caudal, conductividad, pH, temperatura y espectro de absorbancia UV-Visible (Lepot, 2012).

El banco de ensayo consiste en un circuito cerrado en el cual el agua se pone en movimiento por medio de una bomba centrífuga, controlada por un convertidor de frecuencia. El circuito comprende sucesivamente un recipiente de 650 L equipado con un agitador para asegurar concentraciones consistentes y homogéneas de los compuestos y sustancias en ese volumen. Después de esto el flujo es conducido por una tubería flexible de diámetro de 1” y 1/4”, y de allí es descargado a una bandeja tipo canal donde están las sondas de medición, como se muestra en la Figura 31 (adaptado de Lepot, 2012).

Figura 30- Banco de prueba (Lepot, 2012)

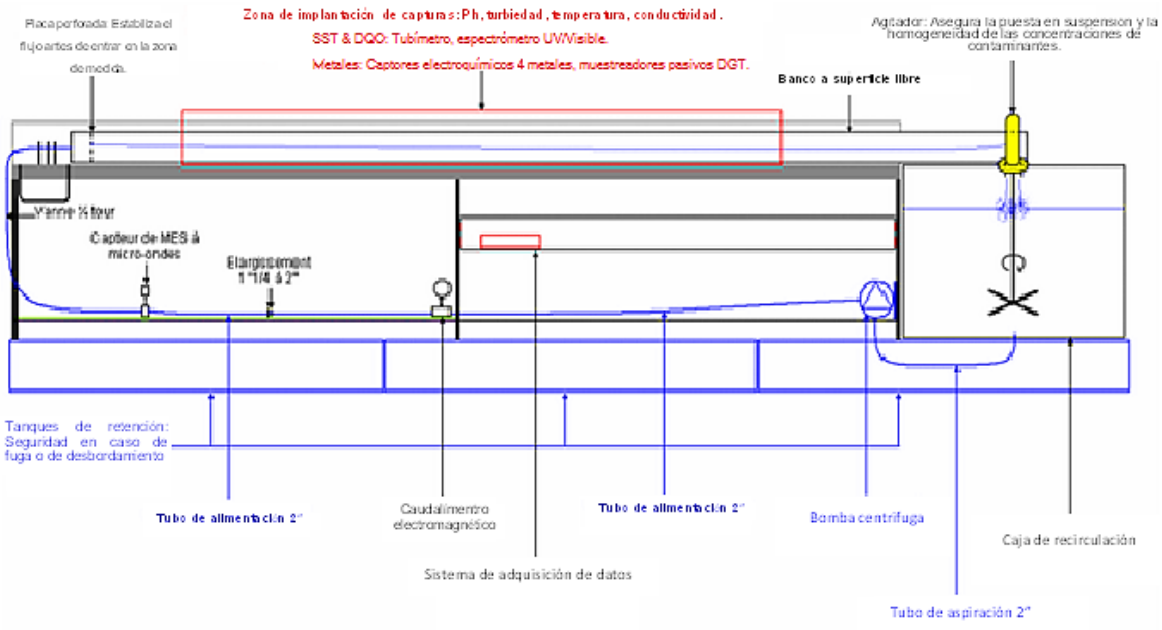


Figura 31- Detalles del banco de prueba empleado para monitorear el afluente de la PTAR de *Fontaines-sur-Saone* (adaptado de Lepot, 2012)

Para este caso se tomaron en total 135 muestras durante época seca (94) y época de lluvia (41) del afluente de la PTAR. Los ensayos de laboratorio para cuantificar la presencia de los determinantes objeto de estudio se realizaron por triplicado a cada una de las muestras. Los resultados del análisis de laboratorio son presentados para tiempo seco y lluvia en la Figura 32 y la Figura 33 respectivamente.

La medición del espectro UV-Vis de cada muestra se realizó por triplicado, para lo cual se empleó una sonda spectro::lyser con un paso de luz de 2 mm. Por lo tanto, en la Figura 34 y Figura 35 se presentan una medición del espectro por cada muestra para tiempo seco como de lluvia respectivamente. Los demás espectros pueden ser consultados en el ANEXO A.

Los procedimientos de los ensayos de laboratorio para cuantificar la concentración de SST, DQO, y DQOf están descritos en el ANEXO E-5.

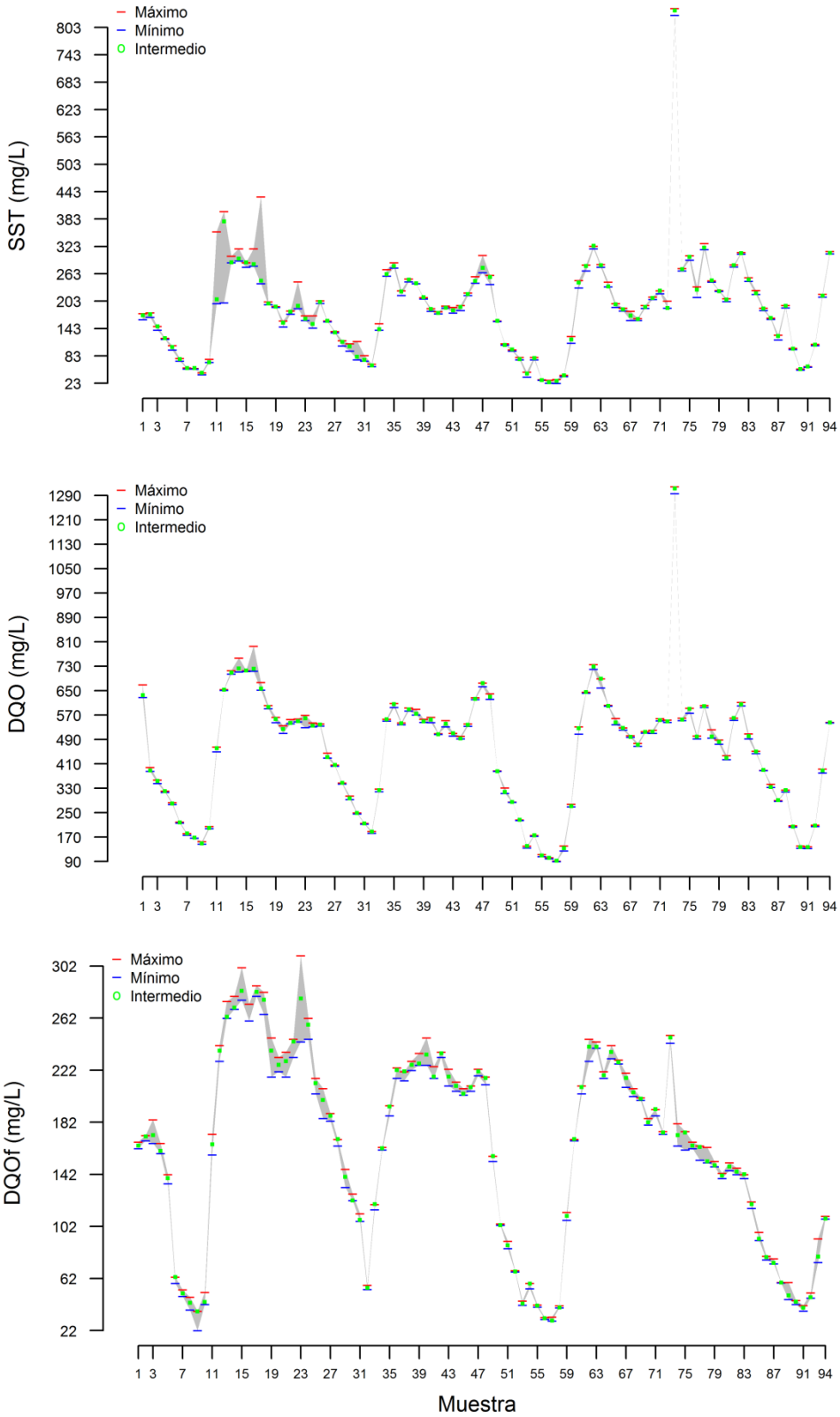


Figura 32- Valores de las concentraciones de los SST, la DQO y la DQO filtrada del afluente de la PTAR de *Fontaines-sur-Saône* (Época seca: 15,18 y 25 de enero, y 3 de mayo 2011 muestras bihorarias puntales)

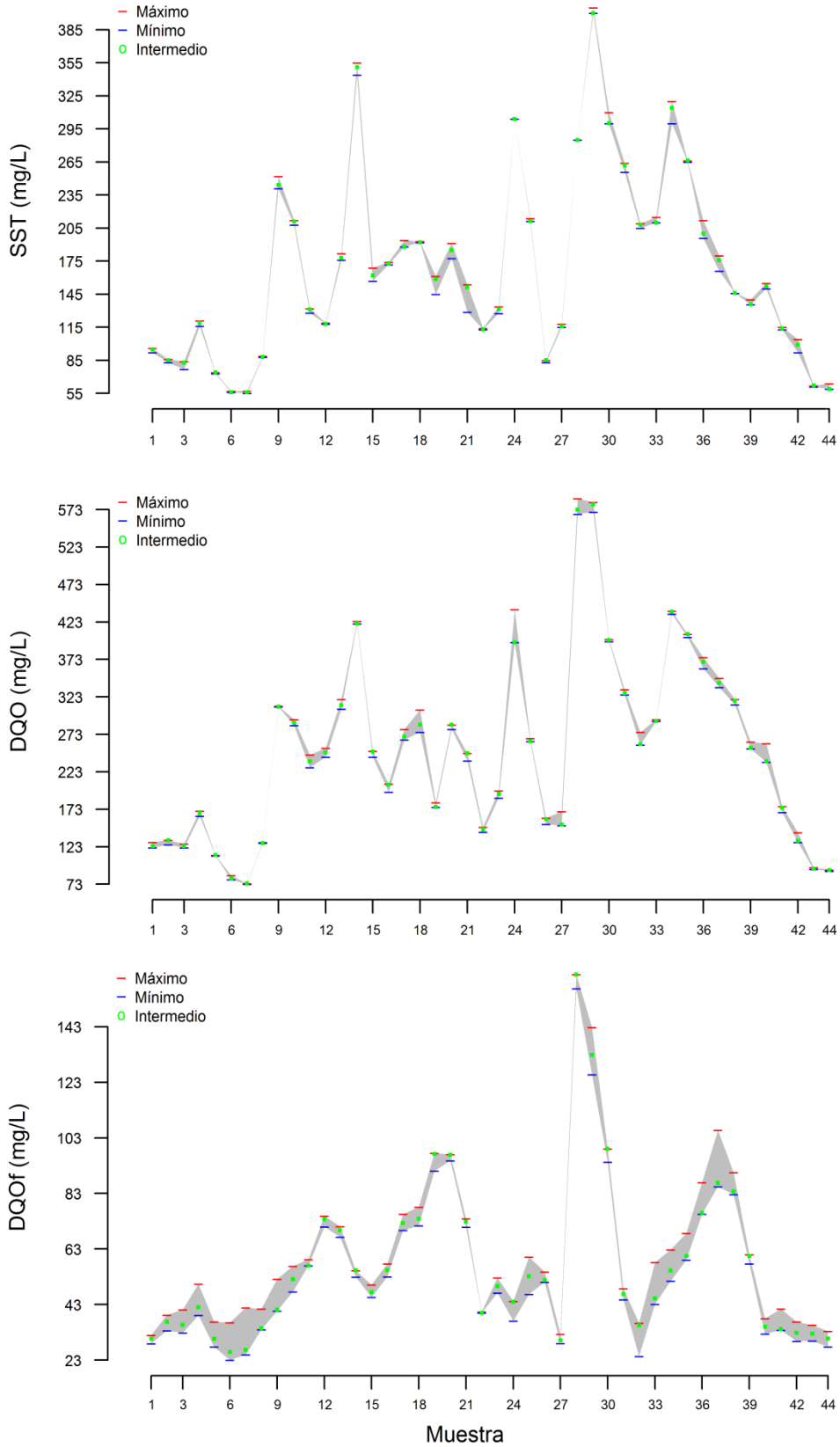


Figura 33- Valores de las concentraciones de los SST, la DQO y la DQO filtrada del afluente de la PTAR de *Fontaines-sur-Saône* (Epoca lluvia: 24 al 25 de octubre 2011 muestras bihorarias puntales)

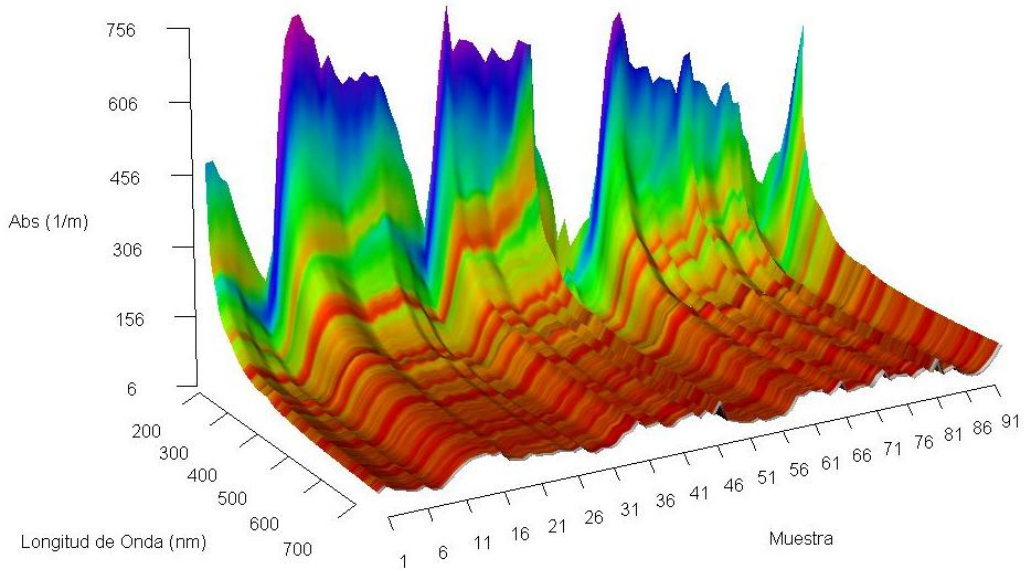


Figura 34- Espectros UV-Vis con los valores máximos valores de absorbancia del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seca: 15,18 y 25 de enero, y 3 de mayo 2011 muestras bihorarias puntales)

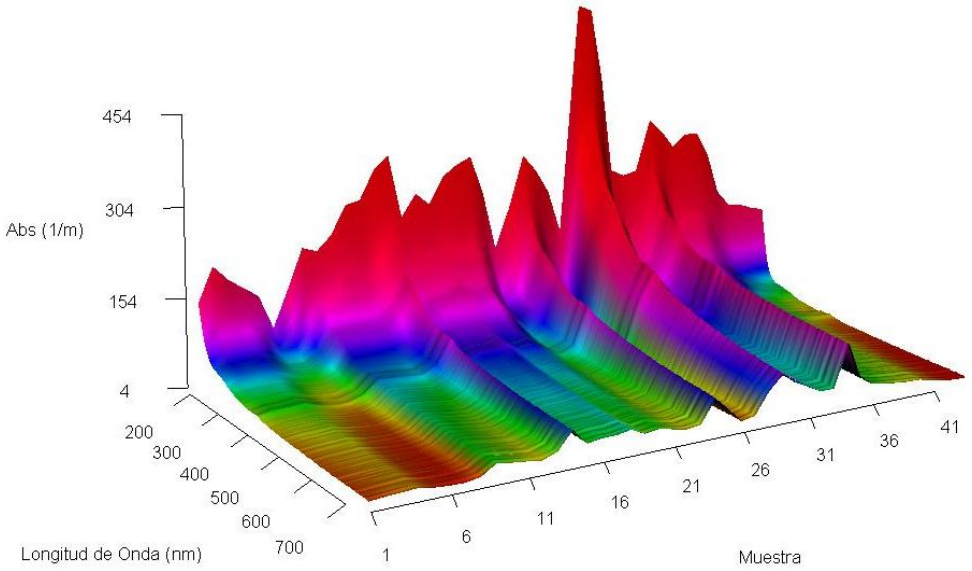


Figura 35- Espectros UV-Vis con los valores máximos de absorbancia de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia: 24 al 25 de octubre 2011 muestras bihorarias puntales)

2.2. ESPECTROMÉTRO UV-VISIBLE – spectro::lyser

Los espectrómetros UV-Visible 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 ultravioleta hasta el visible. Dichos captores son capaces de proporcionar información del orden de una medición por minuto, que pueden traducirse en términos de concentraciones equivalentes de SST, DQO, DQOf, Nitrato, Nitritos, entre otros. El espectrómetro comercializado por la sociedad s::can, llamado *spectro::lyser*, es un captor sumergible de 60 cm de longitud y 44 mm de diámetro que mide la atenuación de la luz

entre 200 nm y 750 nm en deltas de longitud de onda de 2.5 nm, y otorga resultados en tiempo real (Langergraber *et al*., 2004; Hochedlinger, 2005), además cuenta con un sistema de autolimpieza por medio de aire o agua que mantiene en mejores condiciones la ventana de medición de la sonda (Hochedlinger, 2005).

La medición se realiza directamente *in situ* sin necesidad de muestreo o tratamiento de las muestras y por lo tanto algunos errores experimentales con el captor se consideran mucho menores que aquellos asociados a los ensayos estándar de laboratorio (Langergraber *et al*., 2003). Además, el uso de esta y otras tecnologías de medición en continuo e *in situ* permiten reducir varios inconvenientes demostrados por Winkler *et al.* (2008), tales como la baja representatividad espacio-temporal de los resultados, puesto que debido al elevado costo asociado a la recolección y al análisis de las muestras en laboratorio sólo es posible recolectar un número relativamente pequeño de muestras durante periodos prolongados de tiempo, el transporte de las muestras del lugar de recolección al laboratorio, almacenamiento y conservación de las mismas y los plazos prolongados para la obtención de resultados.

Por lo tanto, una medición de alta resolución y fiable con la sonda *spectro::lyser* es posible, ya que una sola medición suele tardar unos 15 s. Además, el bajo consumo de energía facilita la aplicación en campo por medio de una fuente de alimentación o batería solar. La sonda tiene un registrador de datos a bordo, capaz de almacenar, por ejemplo, espectros de absorbancia completos durante un mes en un intervalo de medición de 30 min (Quevauviller *et al*., 2006).

Su diseño consiste en tres componentes principales: (i) dispositivo de emisión, (ii) ventana de medición y (iii) dispositivo de detección (Figura 36). El elemento central del emisor es una fuente de luz de flash de xenón, la cual se complementa con un sistema óptico que guía el haz de luz y un sistema de control electrónico para la operación de la lámpara. En la sección de medición la luz pasa a través de una ventana de medición que se llena con la sustancia en análisis que interactúa con él, un segundo haz de luz dentro de la sonda (haz de compensación) es guiado a la sección de comparación interna, lo cual permite identificar alteraciones durante el proceso de medición que son compensadas de forma automática. La unidad de detección consta de dos componentes principales: (i) el detector y (ii) la sección de operación electrónica. Un sistema óptico se centra en la medición y compensación en un puerto a la entrada del detector, donde la luz es recibida por éste y divide el haz de luz en longitudes de onda y guía los 256 fotodiodos, que transforman la señal, por lo que no es necesario el uso de componentes sensibles al movimiento. La parte electrónica de la sonda se encarga de controlar el proceso de medición y de las etapas de procesamiento para la edición y comprobación de la señal de medición y el cálculo de los deterinantes de calidad del agua (s::can, 2012).

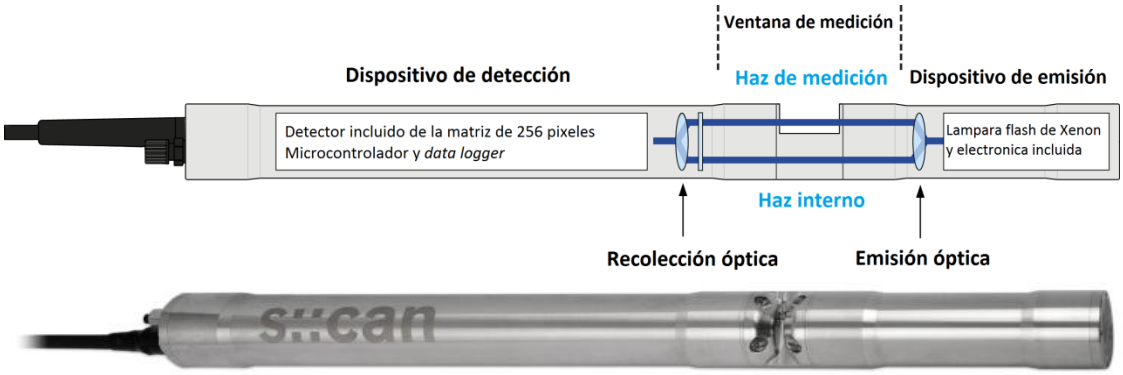


Figura 36- Partes de la sonda spectro::lyser (adaptado de s::can, 2012)

2.3. CALIBRACIÓN DEL SENSOR

Estos captores han sido probados de manera no exhaustiva en varias condiciones de funcionamiento incluyendo ríos (Staubmann *et al*., 2001), plantas de tratamiento (Winkler *et al*., 2002), aliviaderos (Gruber *et al*., 2004) y sistemas de alcantarillado (Torres y J.-L. Bertrand-Krajewski, 2008). Una de las principales ventajas de este tipo de captores es que un número importante de determinantes (SST, DQO y nitratos) pueden ser monitoreados en continuo utilizando un sólo instrumento de medición (Gruber *et al*., 2006) y la cantidad de mediciones de los mismos que pueden ser realizadas; por ejemplo si la resolución temporal es de 15 segundos se podrían realizar hasta 5760 mediciones en un día y 220 absorbancias por medición (Hochedlinger, 2005).

Por consiguiente es necesario correlacionar las absorbancias con la concentración de los diferentes determinantes que puedan ser detectados en las longitudes de onda del espectro UV-visible presentes en el hidrosistema analizado. Para lo cual la compañía fabricante ofrece una ecuación que se basa en la técnica estadística de *PLS* (Hochedlinger, 2005). Dicha ecuación ofrece una calibración global para una serie de determinantes válidos para la composición típica del hidrosistema estudiado. Por lo general, altos coeficientes de determinación (R^2= 0.90 - 0.98 (s::can, 2012)) se pueden lograr para un conjunto de determinantes estándar (SST, DQO, DBO *etc*.), ofreciendo resultados suficientes y de calidad para muchos propósitos, tales como el control de una planta de tratamiento de aguas residuales (Fleischmann *et al.*, 2001). No obstante, la composición del agua en la mayoría de hidrosistemas no es constante, por ejemplo, las aguas residuales tienen propiedades específicas que varían de acuerdo a la clase de vertimientos sean realizados en la red de alcantarillado (*e.g.* contribuciones industriales). Por tanto, las concentraciones de los diferentes compuestos cambian, especialmente los orgánicos (Hochedlinger, 2005). Entonces, el fabricante sugiere adaptar la calibración global a la calidad del hidrosistema estudiado por medio de una calibración local, la cual consiste en la toma de muestras y ensayos de laboratorio (por ejemplo DQO y/o SST) y obtener de las muestras el espectro de absorbancia, para que los resultados de las concentraciones estimadas por el programa sean mejores. El procedimiento de calibración local comprende las siguientes etapas: i) Compensación del espectro por la turbiedad, ii) eliminación de *outliers*, iii) validación de la verosimilitud del espectro para excluir

espectros equivocados, iv) un procedimiento de calibración multivariada mediante una regresión con *PLS* y v) validación de la función de calibración con análisis de sensibilidad. El procedimiento se repite hasta obtener un buen resultado e incluso se puede ejecutar durante la operación del captor y así controlar capacidad de predictiva del modelo regresivo (Fleischmann *et al.*, 2001; Langergraber *et al.*, 2003).

Diferentes autores han implementado *PLS* para construir modelos de calibración (calibración espectral) que conducen a la de determinación de coeficientes correlación buenos, entre los datos de los ensayos de laboratorio y las concentraciones estimadas a partir de la espectro de absorbancias, ya que hay una gran cantidad de variables independientes proporcionadas por el espectro UV-Visible cuando se quiere determinar la concentración de determinantes, como por ejemplo: DQOeq (Lorenz *et al*., 2002; Rieger *et al*., 2004; Torres y J. L. Bertrand-Krajewski, 2008), COT (Carbono Orgánico Total) (Fleischmann *et al.*, 2002), SSTeq (Torres y Bertrand-Krajewski, 2008), H_2S (Ácido Sulfhídrico) (Sutherland-Stacey *et al*., 2008) y hemoglobina (Sutherland-Stacey *et al.*, 2009). Incluso, se presentan métodos para la calibración local capaces de eliminar la sensibilidad cruzada de las sustancias con las mismas formas espectrales o de picos superpuestos, así como ciertas señales de fondo y se pueden visualizar los efectos de matriz de diferentes sustancias es decir, cambios de las reacciones a través de la forma espectral (Langergraber *et al.*, 2003). Sin embargo, la experiencia ha demostrado que la mayoría de las veces los datos de referencia son la parte crítica de todo el procedimiento de calibración. Luego, es esencial para garantizar la calidad de las medidas de referencia (en relación con el método de análisis de referencia, rango de medición, errores de muestreo, la identidad de las muestras y la asignación al azar del muestreo) para obtener buenos resultados de la calibración (Lorenz *et al.*, 2002; Langergraber *et al.*, 2003).

2.3.1. Aplicaciones con la sonda spectro::lyser en PTARs

Durante la última década, se han publicado muchos estudios sobre el uso de esta sonda en plantas de tratamiento de aguas residuales, pero pocos comparan el rendimiento de los métodos de calibración existentes: calibración global, calibración local con muestras tomadas en el sitio de medición (utilizando el software suministrado por el fabricante de la sonda) y la calibración espectral, el cual se lleva a cabo generalmente a través de métodos *PLS* a partir de los espectros medidos en muestras locales.

Cualquiera que sea el determinante seleccionado: SST (Tabla 6) o DQO (Tabla 7), la calibración global parece ser menos eficiente que la propia calibración local y mucho menos eficiente que la calibración espectral, como lo demostraron Rieger *et al*., 2006; Torres y J-L Bertrand-Krajewski, 2008; Maribas *et al*., 2008. Incluso, Rieger *et al*. (2004) probaron varios modelos de regresión espectral para los mismos conjuntos de datos obteniendo, mejores resultados en los modelos (de acuerdo con el criterio del coeficiente de determinación) cuando éstos son calibrados con un amplio grupo de longitudes de onda. Para SST se obtuvo un R^2= 0.326 cuando el modelo es calibrado con una longitud de onda, mientras que cuando es calibrado con siete esta métrica se incrementa

sustancialmente a 0.848. Para el caso de la DQOf los mismos autores, obtuvieron entre 0.316 y 0.905 en modelos calibrados con una y cinco longitudes de onda respectivamente. Por otra parte Torres y Bertrand-Krajewski (2008) utilizaron un gran número de longitudes de onda que cubre todo el espectro UV-visible, mientras que Rieger *et al*. (2004) utilizaron un menor número de longitudes de onda (en el UV lejano para SST-387 nm para el modelo con una sola longitud de onda y UV-230 nm en el UV cercano para DQOf para el modelo con una sola variable predictora) (Lepot, 2012).

No. de muestras	Origen de muestras	Calibración	No. Longitud de onda	R^2	Fuente
-	PTAR	Espectral	1	0.83	(Langergraber *et al*., 2003)
-	PTAR	Espectral	-	0.95	(Langergraber *et al*., 2003)
-	PTAR	Global	-	0.83	(Rieger *et al*., 2006)
26	PTAR	Local	-	0.83	(Rieger *et al*., 2006)
26	PTAR	Espectral	-	0.90	(Rieger *et al*., 2006)
44	Efluente PTAR	Espectro	7 (230 – 330 nm)	0.848	(Rieger *et al*., 2004)
44	Efluente PTAR	Espectro	1 (378 nm)	0.326	(Rieger *et al*., 2004
-	Piloto de PTAR	Espectro	-	0.995	(Langergraber *et al*., 2004b)
16	RS	Global	-	0.981	(Torres, 2008)
16	RS	Local	-	0.975 – 0.998	(Torres, 2008)
16	RS	Espectro	177 (190 – 730 nm)	0.996 – 0.996	(Torres, 2008)

Tabla 6- Algunos valores de coeficientes de determinación SST a partir de espectrometría UV-Vis empleando modelos *PLS* (Lepot, 2012)

	No. de muestras	Origen de muestras	Calibración	No. longitud de onda	R^2	Fuente
DQO	-	Afluente PTAR	Espectral	-	0.95	(Langergraber *et al*., 2004a)
	-	PTAR, Tiempo lluvia	Global	-	0.84	(Maribas *et al*., 2008)
	-	PTAR, Tiempo lluvia	Espectro	-	0.94 – 0.97	(Maribas *et al*., 2008)
	-	PTAR	Global	-	0.23 -0.83 -0.93	(Rieger *et al*., 2006)
	-	PTAR	Espectro	-	0.9 - 0.94 -0.97	(Rieger *et al*., 2006)
	-	PTAR	Espectro	1	0.88	(Langergraber *et al*., 2003)
	-	PTAR	Espectro	-	0.9	(Langergraber *et al*., 2003)
	-	Afluente PTAR	Espectro	-	0.9	(Langergraber et al., 2004a)
	16	Aguas lluvia	Global	-	0.911	(Torres, 2008)
	16	Aguas lluvia	Local	-	0.902 – 0.940	(Torres, 2008)
	16	Aguas lluvia	Espectral	210 (207.5 – 730 nm)	0.978 – 0.991	(Torres, 2008)
DQOf	-	Afluente PTAR	Espectro	-	0.95	(Langergraber *et al*., 2004a)
	-	STEP	Espectro	1	0.72	(Langergraber *et al*., 2003)
	-	STEP	Espectro	-	0.91	(Langergraber *et al*., 2003)
	-	Salida STEP	Espectro	-	0.91	(Langergraber *et al*., 2004a)
	12	Salida STEP	Espectro	5 (250 – 340 nm)	0.905	(Rieger *et al*., 2004)
	12	Salida STEP	Espectro	1 (230 nm)	0.316	(Rieger *et al*., 2004)
	-	Piloto STEP	Espectro	-	0.9	(Langergraber *et al*., 2004b)

Tabla 7- Algunos valores de coeficientes de determinación para DQO (total y filtrada) a partir de espectrometría UV-Vis empleando modelos *PLS* (Lepot, 2012)

DAC: Desbordamiento alcantarillado combinado
ARGD: Aguas residuales grises domésticas

Además de lo reportado por Lepot (2012), existen otros autores y métodos regresivos que han sido implementados para la estimación de determinantes a partir del espectro UV-Vis, diferentes a *PLS* y no necesariamente con mediciones en línea. Algunos reportan el número de longitudes de usadas por sus modelos y el nivel de ajuste. A continuación se presenta los autores que han trabajo de con la sonda objeto de la investigación del presente trabajo. No obstante, en la Tabla 8 se presentan otros autores y métodos que han trabajado con espectrometría UV-Vis de laboratorio para detección de SST, DQO y DQOf.

En importante aclarar dos aspectos. El primero tiene que ver con el tipo de calibración empleada por Folgelman *et al.* (2006), quienes no solamente usan como datos de entrada los valores de absorbancia en el rango UV del espectro sino que utilizan mediciones de turbiedad para calibrar un modelo de RNA y con esto estimar las concentraciones de DQO y DQOf presentes en las muestras de aguas residuales grises domésticas analizadas. En segundo lugar, Hochedlinger (2005) emplea el algoritmo optimización mínima secuencial (en inglés *Sequential Minimal Optimisation-SMO*) para resolver el problema de programación cuadrática en *SVM* (ver numeral 1.4.7.4) presentado en la Ecuación 28.

	No. de muestras	Origen de muestras	Calibración	Método	No. longitud de onda	R^2	Autores
SST	-	PTAR	Espectral	Deconvolución	4	0.938	(Thomas *et al.*, 1996)
DQO	-	PTAR	Espectral	Deconvolución	4	0.940	
DQO	40	ARGD	Espectral+ turbiedad	RNA	[190 a 350 nm]	0.726	(Fogelman *et al.*, 2006)
DQOf	31	ARGD	Espectral+ turbiedad	RNA	[190 a 350 nm]	0.88	
SST	25	DAC	Espectral	Regresión lineal simple	1	0.77	(Hochedlinger, 2005)
DQO	25	DAC	Espectral	Regresión lineal simple	1	0.848	
DQOf	25	DAC	Espectral	Regresión lineal simple	1	0.951	
SST	25	DAC	Espectral	Arboles de decisión M5 (Weka)	[660–670 nm]	0.765	
DQO	25	DAC	Espectral	Arboles de decisión M5 (Weka)	2	0.977	
DQOf	25	DAC	Espectral	Arboles de decisión M5 (Weka)	[230–500 nm]	0.837	
SST	25	DAC	Espectral	*SVM* usando *SMO*	[620–630 nm]	0.77	
DQO	25	DAC	Espectral	*SVM* usando *SMO*	[240–250 nm]	0.959	
DQOf	25	DAC	Espectral	*SVM* usando *SMO*	[260–270, 436 nm]	0.867	
DQO	14	PTAR	Espectral	*PLS*	145	0.82	(Sarraguça *et al.*, 2009)
SST	13	PTAR	Espectral	*PLS*	10	0.82	

DAC: Desbordamiento alcantarillado combinado
ARGD: Aguas residuales grises domésticas

Tabla 8- Coeficientes de determinación para SST, DQO y DQOf a partir de espectrometría UV-Vis empleando diferentes modelos

Por último, Gamerith (2011) presentó la optimización de la calibración local de la sonda empleando estrategias de evolución (ver numeral 1.4.8.2) por medio del *software* BlueM.OPT evaluando de forma individual diferentes eventos y usando diferentes rangos de longitudes de onda. De la aplicación de este método el autor concluye que calibrar la sonda spectro::lyser a nivel local con muestras de un evento o eventos (lluvia o seco) con intervalos de concentración similares. Sin embargo, en ese caso no todas las variaciones en la matriz de compuestos pueden ser evaluadas por los modelos. Para ese tipo de calibraciones se presentaron porcentajes de error menores al 10 %. Sin embargo, los errores relativos de las muestras de validación fueron de más de 100%. Esto significa que mayores errores totales que se obtienen cuando se utiliza la calibración global, la cual trata de estimar amplio rango de concentraciones. Para los datos disponibles el optimizador BlueM no genera una mejora significativa en los resultados.

2.4. MÉTODOS DE ANÁLISIS

2.4.1. Metodología para cuantificar las incertidumbres de SST, DQO, DQOf y espectro de absorbancia UV-Vis (Torres, 2011)

Torres (2011) define una metodología para estimar las incertidumbres asociadas a concentraciones de Sólidos Suspendidos Totales (SST) ligados a la fase de análisis de laboratorio, contemplando no solamente la precisión de los instrumentos de laboratorio

sino también el submuestreo y la manipulación de muestras y aparatos de laboratorio. La metodología propuesta tiene en cuenta el cálculo de la incertidumbre de réplicas por medio de métodos analíticos, la comparación entre los resultados de réplicas por medio de pruebas t y el cálculo de la incertidumbre compuesta por medio del método de Monte Carlo. Esta metodología fue adaptada para cuantificar las incertidumbres de la DQO, la DQOf y del espectro de absorbancias. A continuación se presentan los supuestos en los cuales se basa la metodología:

La metodología considera que todas las cantidades x_i y x_j se obtienen de forma independiente, las covarianzas $u(x_i, x_j)$ se pueden ignorar y por lo tanto la Ecuación 10 se simplifica para obtener la Ecuación 46:

$$u^2(y) = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 u^2(x_i)$$

Ecuación 46-

Con el objetivo de comparar dos resultados x_1 y x_2 de los que se conoce su incertidumbre $u(x_1)$ y $u(x_2)$, Torres (2011) utilizó la prueba t (Ruxton, 2006):

$$t = \frac{x_1 - x_2}{\sqrt{\frac{u^2(x_1)}{n_1} + \frac{u^2(x_2)}{n_2}}}$$

Ecuación 47-

donde n_1 y n_2 es el número de mediciones realizadas para obtener x_1 y x_2, respectivamente. Con ese parámetro t calculado se procede a consultar las tablas estadísticas para la distribución normal con los grados de libertad gl (estimador necesario para el cálculo de un estadístico particular, e indica el número de valores aleatorios que no pueden ser determinados o fijados mediante una ecuación matemática) (Ecuación 48):

$$gl = \frac{\left(\frac{u^2(x_1)}{n_1} + \frac{u^2(x_2)}{n_2} \right)^2}{\frac{\left(\frac{u^2(x_1)}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{u^2(x_2)}{n_2} \right)^2}{n_2 - 1}}$$

Ecuación 48-

Adicionalmente, se utilizó el método de Monte Carlo para el cálculo de incertidumbres teniendo en cuenta múltiples réplicas de cada ensayo por muestra.

2.4.1.1. Procedimiento para el cálculo de las incertidumbres

El cálculo de la incertidumbre asociada a cada réplica i es útil tanto para estimar la incertidumbre estándar compuesta como para depurar e interpretar los resultados en bruto obtenidos en laboratorio. Sobre este último aspecto, la metodología desarrollada por Torres (2011) contempla una comparación de los resultados de cada réplica teniendo en cuenta sus incertidumbres estándar asociadas. En la Tabla 9 se presentan las ecuaciones de la incertidumbre para los SST.

Si se asume que los resultados de las mediciones de cada réplica siguen una distribución normal, se propone una comparación por pares mediante pruebas t, utilizando la Ecuación 47 y la Ecuación 48. De acuerdo a los resultados de dichas pruebas, pueden existir tres posibilidades: (i) no se elimina ninguna réplica, al no detectar diferencias significativas (valores *p* entre parejas superiores a 0,05); (ii) no se elimina ninguna réplica porque se detectan diferencias significativas entre todas las réplicas (valores p entre parejas inferiores a 0,05); (iii) se elimina una sola réplica porque se detectan diferencias significativas entre una réplica y las dos restantes (Torres, 2011).

Una vez tomada la decisión sobre la aceptación o el rechazo de cada réplica, se propone el cálculo de la incertidumbre asociada a la muestra, considerando tanto la variabilidad entre los resultados de las réplicas como la incertidumbre de cada réplica considerada. El cálculo propuesto, explicado en seguida, se basa en el método de Monte Carlo: Teniendo en cuenta el valor medido y la incertidumbre asociada a la réplica 1 de la muestra j, se escoge aleatoriamente un valor de prueba $x_{1,j}$ de acuerdo con una distribución normal asumiendo el valor medido como el promedio en la distribución y la incertidumbre asociada como su desviación estándar (Ecuación 8). De manera similar se escogen valores de prueba $x_{2,j}$ y $x_{3,j}$ para las réplicas número 2 y 3, respectivamente. Para esos valores $x_{1,j}$, $x_{2,j}$ y $x_{3,j}$, se calcula la desviación estándar correspondiente σ_j. Posteriormente, se seleccionan aleatoriamente nuevos valores de prueba $x_{1,j+1}$, $x_{2,j+1}$ y $x_{3,j+1}$, de tal forma que se pueda calcular la desviación estándar entre esos resultados σ_{j+1} correspondientes a la iteración $j+1$. Al cabo de un número elevado n de iteraciones, se tienen n valores de desviaciones estándar entre las réplicas $(\sigma_j, \sigma_{j+1}, \dots \sigma_{j+n})$. El pro-medio de dichas desviaciones estándar se interpreta como la incertidumbre compuesta más probable u_{cj} asociada a la muestra j (ver Figura 37).

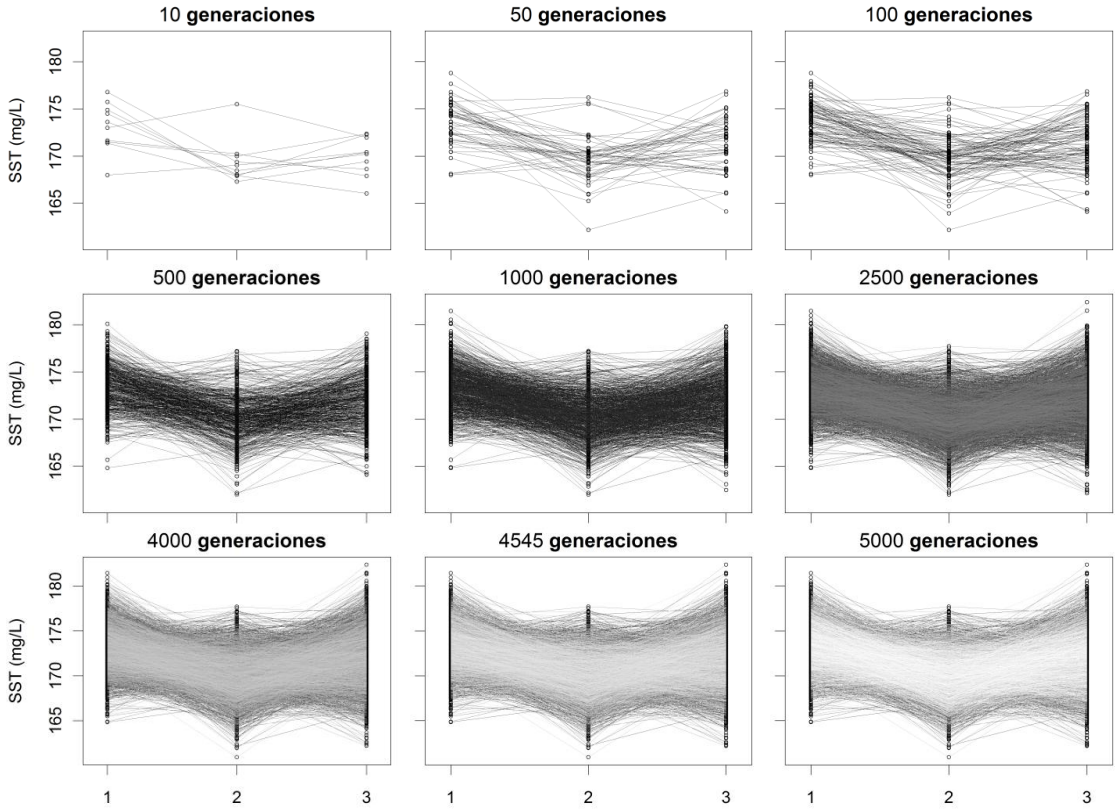


Figura 37- Simulación de Monte Carlo de 5000 ternas de réplicas para la muestra 1 (Torres, 2011)

A continuación, se presenta las ecuaciones para el cálculo de la incertidumbre de SST. En dicha ecuacion se considera que las mediciones de masa (balanza) y volumen (probeta), siguen una distribución normal, se puede asumir que las incertidumbres $u(m_s)$ y $u(m_p)$, asociadas a las mediciones de m_s, m_p, y V utilizando dichos aparatos, son iguales al mitad de la precisión (p) de p_m y p_V con un 95 % de confianza. Adicionalmente, se asume que las mediciones involucradas en cada ensayo estándar de laboratorio (ver numeral 7.2) son independientes, por ende la covariancia entre ellas tiende a ser igual a cero. Luego se puede utilizar la Ecuación 46 para el cálculo de la incertidumbre estándar compuesta.

Determinante	Ecuación	Variables
SST	$u(SST_i) = \frac{1}{V}\sqrt{u^2(m_{si}) + u^2(m_{pi}) + SST_i^2 u^2(V)}$	Datos Laboratorio Francia: $p_V = 0.03$ ml $p_m = 0.001g$ Datos Laboratorio Colombia: $p_V = 0.03$ ml $p_m = 0.001g$ $i = 1,2,$ y 3 replicado

Tabla 9- Ecuaciones para el cálculo de la incertidumbre de SST conforme a la metodología de Torres (2011)

2.5. HERRAMIENTAS INFORMÁTICAS

Un amplio número de software de programación orientada a objetos existe hoy en día. Sin embargo, la mayoría de éstos tienen un costo elevado y su soporte es limitado a un grupo de expertos.

No obstante, R-project proporciona un amplio abanico de herramientas matemáticas y gráficas, de forma gratuita y soportado por una amplia comunidad de científicos, quienes a su vez desarrollan nuevas librerías computacionales para ponerlas en libre distribución. R fue inicialmente diseñado por Robert Gentleman y Ross Ihaka, miembros del Departamento de Estadística de la Universidad de Auckland, en Nueva Zelanda. El código de R está disponible como software libre bajo las condiciones de la licencia GNU-GPL, y puede ser instalado tanto en sistemas operativos tipo Windows como en Linux o MacOS X (R *Development Core Team*, 2013).

La siguiente es una breve descripción de los paquetes y funciones del software R-project aplicadas en esta investigación:

a) Paquete *nls* (*nonlinear least squares*): determina los pesos (coeficientes) de un modelo no lineal por medio de mínimos cuadrados no lineales, por medio de un proceso iterativo, y por lo tanto valores semilla de los coeficientes deben ser suministrados para emprender la búsqueda de los valores que generen el mejor ajuste (ver numeral 1.4.4) (Bates y Watts, 1988).

b) Paquete *pls* (*partial least squares*): en este paquete están implementadas varias técnicas estadísticas como regresión de componentes principales (*PCR*) y regresión por mínimos cuadrados parciales (*PLS-R*) y mínimos cuadrados parciales de potencias canónicas (*CPPLS*). Además, tiene métodos para funciones genéricas para predecir (en función de un modelo calibrado), actualizar y modificar los coeficientes de la arquitectura de los modelos. También cuenta con funciones más especializadas como validación cruzada y evaluación del desempeño por medio de *RMSEP*, y funciones para graficar y visualizar el comportamiento de modelo (Mevik *et al*., 2012). La función *plsr*() puede ajustarse por medio de diferentes algoritmos según el argumento que se especifique en la función. Cuatro métodos están disponibles: algoritmo kernel (“kernelpls”), el algoritmo de kernel ancho ("widekernelpls"), SIMPLS ("simpls") y el algoritmo clásico de puntuaciones ortogonal ("oscorespls").

c) Paquete *kernlab*: contiene funciones primitivas de productos punto (kernel), implementación de máquinas de vectores soporte (*SVM*) y máquina de vector de relevancia, procesos Gaussianos, algoritmo de clasificación, PCA kernel, kernel CCA, análisis de función kernel, métodos kernel en línea y un algoritmo de *clustering* espectral. Por otra parte, los cálculos realizados por la mayoría de las

funciones son vectorizados, lo que permite garantizar un buen rendimiento y requerimientos de memoria aceptables.

Dentro de las funciones presentes en este paquete se encuentran *ksvm*, la cual incluye una versión muy eficiente de la optimización por minimización secuencial (sigla en inglés *SMO*). *SMO* descompone el problema cuadrático (*QP*) *SVM* sin necesidad de utilizar las medidas de optimización numéricas del *QP*. En su lugar, se resuelve el problema de optimización ajustando los pesos del modelo de forma iterativa y conjunta. Las implementaciones de *SVM* disponibles en *ksvm* incluyen el algoritmo de clasificación y regresión *C-SVM* junto con la formulación clasificación ε-SVM y ν-SVM (ν [0,1]), esta última proporcional a la fracción de vectores de soporte que se encuentran en el conjunto de datos y al error de entrenamiento (Karatzoglou, Smola, et al., 2004).

d) Función *nnet*: nnet proporciona funciones y conjuntos de datos para apoyar el entrenamiento de redes neuronales tipo *feed-forward* de una sola capa oculta con varias neuronas. Esta función permite realizar regresiones y clasificaciones en conjuntos de datos. Cuenta con diferentes criterios para minimizar el error, tales como: máxima verosimilitud (distancia de Kullback–Leibler), mínimos cuadrados y un modelo multinominal log-lineal. Permite, además el uso de una tasa decaimiento que penaliza el comportamiento de los pesos para ayudar en el proceso de optimización y evitar el *overfitting* (Venables y Ripley, 1994).

e) Paquete *DEoptim (Differential Evolution optimization)*: *DEoptim* implementa el algoritmo de evolución diferencial para la optimización global de una función real de un vector de parámetros de valor real. La implementación de la evolución diferencial con código C en las interfaces *DEoptim* lo hace eficiente. La función *DEoptim* en el paquete *DEoptim* busca los mínimos de una función objetivo entre los límites inferior y superior de cada parámetro para ser optimizado. Por lo tanto, al invocar la llamada *DEoptim* se especifican los vectores que comprenden los límites inferior y superior; estos vectores son de la misma longitud que el vector de parámetros. Además, la función permite cambiar diversos parámetros que modifican el desempeño del modelo, tales como: el número de poblaciones, el nivel de ajuste, la estrategia de optimización, y controlar los proceso de mutación y supervivencia entre otros (Mullen *et al*., 2011).

3. METODOLOGÍAS DESARROLLADAS

3.1. AMPLIACIÓN DEL ALGORITMO DESARROLLADO POR TORRES (2011) PARA CÁLCULO DE LA INCERTIDUMBRE DE LA DQO, LA DQOF Y ABSORBANCIAS DEL ESPECTRO UV-VIS

En la Tabla 1 se presentan las ecuaciones de las incertidumbres para: DQO, DQOf y absorbancias del espectro UV-Vis aplicando la ley de la propagación de la incertidumbre (ver numeral 1.4.1) y con base en los supuestos del método desarrollado Torres (2011). En dichas ecuaciones se considera que las mediciones de absorbancia a 620 nm (espectrofotómetro) y las absorbancias del espectro UV-Vis (sonda spectro::lyser) siguen una distribución normal, se puede asumir que las incertidumbres $u\big(Abs(620\,nm)\big)$ y $u(Abs(\lambda_{200-750\,nm}))$ asociadas a las mediciones de $Abs(620\,nm)$ y $Abs(\lambda_{200-750\,nm})$, utilizando dichos aparatos, son iguales al mitad de la precisión (p) de p_{abs} y $p_{abs(\lambda)}$, con un 95 % de confianza. Adicionalmente, se asume que las mediciones involucradas en cada ensayo estándar de laboratorio (ver numeral 7.2) son independientes, por ende la covariancia entre ellas tiende a ser igual a cero. Luego se puede utilizar la Ecuación 46 para el cálculo de la incertidumbre estándar compuesta.

Determinante	Ecuación	Variables
DQO y DQOf	$u(DQO_i) = u(Abs_{620nm}) = \frac{p_{abs}}{2 \cdot 0.003}$	$p_{abs} = 0.005\ Abs$ i = 1,2, y 3 replicado
Absorbancia(λ)	$u(Abs_i(\lambda_{200-750nm})) = \frac{p_{abs(\lambda)}}{2}$	Datos sonda spectro::lyser paso de luz de 2 mm (Francia) $p_{abs(\lambda)} = 0.35\ Abs/m$ Datos sonda spectro::lyser paso de luz de 5 mm (Colombia) $p_{abs(\lambda)} = 0.9\ Abs/m$ $i = 1,2, y\ 3$ replicado

Tabla 10- Ecuaciones para el cálculo de las incertidumbres de la DQO, DQOf y absorbancia del espectro UV-Vis conforme a la metodología de Torres (2011)

3.2. MODIFICACIONES AL ALGORITMO DEL *OPP* DE TORRES Y BERTRAND-KRAJEWSKI (2008)

Con base en el programa *OPP* (*OTHU PLS Program*-Figura 38) 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, 2013) con los siguientes cambios: (i) Se utilizó el paquete *pls* (Mevik y Wehrens, 2007) de *R* (R Development Core Team, 2013). (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*, pero la selección y clasificación de forma decreciente con respecto a la relevancia de las variables independientes no se determina a través del coeficiente de correlación, obtenido entre cada variable independiente y las dependientes dentro del conjunto de datos de calibración, sino a partir de un nuevo método desarrollado en la presente investigación y que será explicado en el numeral 3.1.

El algoritmo de calibración de *OPP* está conformado por los siguientes pasos (ver Figura 175):

i. Repartir de los datos X (espectros de absorbancia) y Y (concentraciones) en conjuntos de calibración ($X_c(N_c, n_x)$ y $Y_c(N_c, 1)$) y validación ($X_v(N_v, n_x)$ y $Y_c(N_v, 1)$) con el 67 % y 33 % de la información respectivamente. N_c y N_v es el número de muestras seleccionadas de forma aleatoria para conformar los conjuntos de calibración y validación, y n_x es número de longitudes de onda (variables independientes).
ii. Unificar de forma decreciente de acuerdo a la importancia de la relación (*IF*) entre cada longitud de onda de los espectros UV-Vis y las concentraciones de las muestras. Para esto se utiliza el método descrito en el numeral 3.1.
iii. Unificar los coeficientes de los modelos de *PLS* y la evaluación de los errores de predicción se realizan por medio de $RMSE_{VC}$, para cada par de longitud de onda (i) y vector latente (j) posible. No obstante, el máximo número de longitudes de onda (nl) que podrá evaluar un modelo *PLS* será igual al 67 % del total de las muestras analizadas (N). A las regresiones *PLS* se aplica el conjunto de calibración, donde la matriz X_c se reduce inicialmente a las dos longitudes de onda más importantes conforme a lo establecido en el paso dos y el número de muestras evaluadas con validación cruzada. El número de longitudes de onda incrementa conforme al *IF* hasta N.
iv. De todos los modelos evaluados se determina el mejor, es decir la arquitectura compuesta por (i,j): elegido por el que genera el menor $RMSE_{VC}$.
v. Por último, tomar nuevamente el conjunto de datos de calibración y se evalúa su desempeño por medio de las métricas presentadas en el numeral 1.4.8.3 y de la misma forma para el conjunto validación, y así se determinar la eficiencia y robustez del modelo.

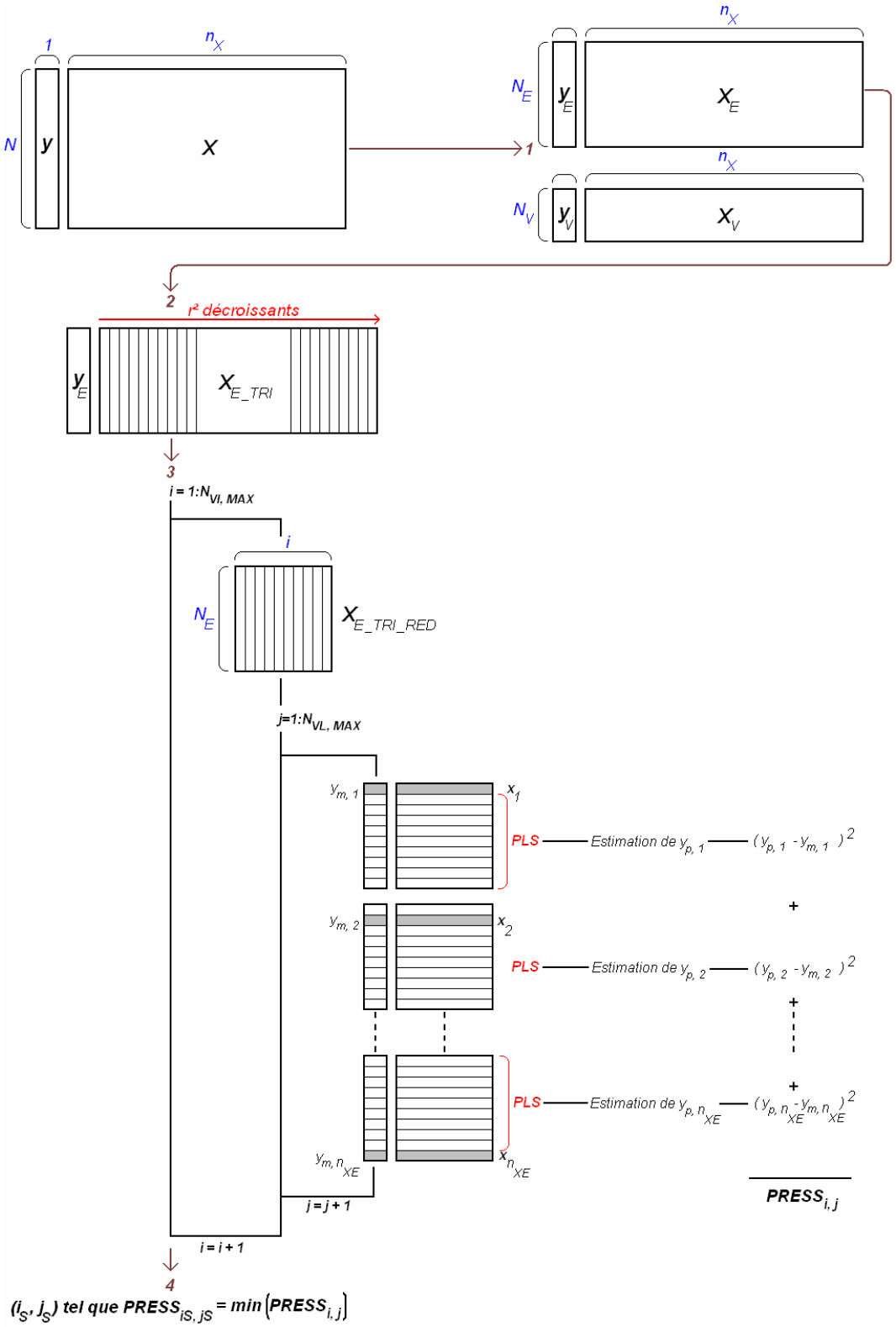


Figura 38- Algoritmo principal del programa *OPP* (Adaptado de Lepot, 2012)

3.3. SELECCIÓN DE LAS LONGITUDES DE ONDA MÁS CORRELACIONADAS CON EL DETERMINANTE ANALIZADO (Zamora y Torres, 2012a)

En el monitoreo de aguas residuales se tiene que lidiar con una matriz de numerosos compuestos disueltos y en suspensión. La superposición de numerosas absorbancias de una sola sustancia –incluso en algunos momentos con picos superpuestos– puede causar sensibilidad cruzada y conducir a malos resultados del sensor. En este caso el sensor puede calibrarse con la matriz de agua de interés (Langergraber *et al*., 2003). Los modelos quimiométricos se utilizan para este propósito. Estos modelos formalizan el procedimiento de la correlación de los factores determinantes para los espectros y su relación con la concentración de las sustancias analizadas (Langergraber *et al*., 2003). Sin embargo, los modelos directos de quimiometría sólo pueden utilizarse si los espectros de todos los componentes son conocidos y la ley de Lambert-Beer es válida, lo cual no se cumple en el caso de las aguas residuales, donde un gran número de compuestos desconocidos están presentes (Langergraber *et al*., 2003) (ver numeral 1.3.1.1). Por lo tanto, seleccionar cuántas y cuáles son las longitudes de onda que se relacionan más con el determinante es factor importante a la hora de estimar concentraciones en función de absorbancias. Sin embargo, por lo general las correlaciones lineales (R) encuentran varias soluciones locales con rendimientos similares, pero no sirven para identificar una longitud de onda de mayor relevancia, cuyo R tenga una diferencia significativa con aquel calculado para la siguiente longitud de onda, en orden de relevancia (Lorenz *et al*., 2002).

Por lo tanto, Zamora y Torres (2012a) desarrollaron un método para detectar de forma multivariada las longitudes de onda más relacionadas con las concentraciones del determinante y evaluar la calidad de los datos, teniendo en cuenta aspectos como la sensibilidad cruzada (sensibilidad a una sustancia que predispone la muestra a mostrarse sensible a otras sustancias relacionadas por su estructura química –Fleischmann *et al*., 2001) y la no linealidad entre valores de absorbancia y concentraciones de determinantes de caliad del agua.

El método, llamado *ZATO* (*Zig-zagged graphical Analysis and Treatment of UV-Vis Outliers*), considera cinco funciones bivariadas con el fin de calibrar los modelos regresivos: lineal, polinomio de segundo y tercer grado, logarítmica y potencial (véase Ecuación 49 a Ecuación 53). La variable a estimar es la concentración del determinante objetivo sobre la base de los valores de absorbancia por cada longitud de onda de los espectros. Por lo tanto, se generan el mismo número de modelos para cada función de regresión como número de longitudes de onda (nl) con valores de absorbancia se presenten.

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

Ecuación 49-

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

Ecuación 50-

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

Ecuación 51-

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

Ecuación 52-

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

Ecuación 53-

donde $A_i, B_i, C_i, D_i\ E_i$ son los coeficientes y exponentes respectivamente, calibrados en cada una de las k ejecuciones, x_j son las absorbancias de cada longitud de onda $\lambda_{[200-750\ nm]}$ de las i muestras analizadas y $\hat{y}_i$ son el conjunto de concentraciones equivalentes de un determinante estimadas por cada modelo en cada ejecución. Para calibrar los coeficientes o pesos de los modelos empleados por el método se usó la función de mínimos cuadrados no lineales *(nonlinear least squares-nls)* del programa R-*project*.

Luego, se calcula la suma de errores cuadrados (*sum of squared errors-SSE*) entre las concentraciones de laboratorios y_i y las concentraciones equivalentes obtenidas de cada modelo $\hat{y}_i$ de acuerdo a la Ecuación 54.

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

Ecuación 54-

Después de obtener los valores *SSE* para cada modelo regresivo, cada longitud de onda se ordena sobre la base del siguiente criterio: para un modelo regresivo en particular, las longitudes de onda con valores bajos de *SSE* se consideran más relevantes que los *SSE* con valores superiores. Por lo tanto, basado en el anterior procedimiento, se propone un Factor de Importancia (*IF*), que explica la afinidad entre las longitudes de onda y el determinante objetivo (Ecuación 55).

$$IF = 10^6 \times \left(1/SSE\right)$$

Ecuación 55-

donde 10^6 (mg^2/L^2) es un factor de mayoración, seleccionado arbitrariamente para reescalar los datos de la suma de los errores al cuadrado y poder establecer la importancia de cada longitud de onda dentro de los espectros de absorbancia.

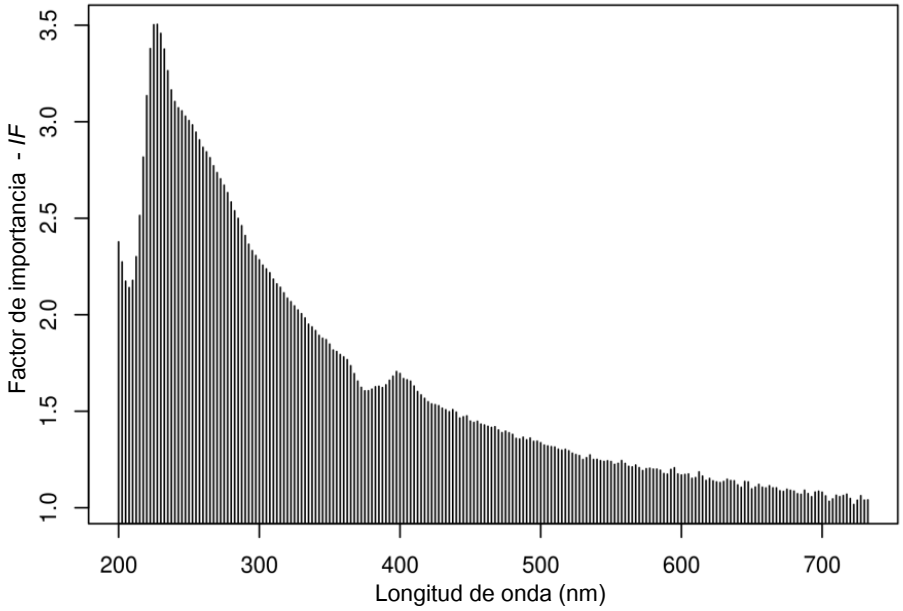


Figura 39- Factor de importancia: afinidad entre las longitudes de onda del espectro UV-Vis y la concentración del determinante objetivo (DQO)

Para evitar el sesgo debido a la presencia de posibles valores atípicos en los datos, se realiza una selección aleatoria de una fracción del conjunto de datos para calibrar las cinco funciones propuestas (*e.g.* 67 % de los datos), lo cual se repite un número k de veces. Con este método, la recurrencia de una longitud de onda a ocupar cierta posición en 220 posiciones posibles, sobre la base del indicador *IF* se puede calcular. Esta recurrencia, se utiliza para dos fines: (i) establecer el nivel de importancia de una longitud de onda en el determinante objetivo, (ii) para evaluar la calidad de los datos, ya que una mayor dispersión o polvo (en inglés *dust*) de las recurrencias en diferentes niveles de importancia de muchas longitudes de onda pueden confirmar la baja afinidad entre las concentraciones obtenidas en laboratorio y los espectros UV-Vis relacionados. De hecho, se confirmaron estos posibles beneficios numéricamente, tanto para la selección de longitudes de onda como para la construcción de modelos quimiométricos y detección de *outliers*, como se muestra en numeral 3.3.1.

Por último, el diseño de los gráficos para la representación de los resultados cuando se realizan k iteraciones se inspira en la obra del artista colombiano Omar Rayo. Es un pintor, dibujante, grabador y diseñador, adscrito al movimiento Op Art. Nació y murió en Roldanillo Valle del Cauca, Colombia (1928-2010). Su trabajo está dedicado a las figuras geométricas pintadas con imágenes claras, característico del arte geométrico-óptico centrado en cuadrados, rectángulos y zig-zags, ejecutado en blanco, negro y rojo. El

método de análisis espectral y gráficos que la representan se han desarrollado y programado en la plataforma R (R Development Core Team, 2013).

3.3.1. Aplicación y validación del método *ZATO*

A continuación se presenta una aplicación con los datos del afluente de la PTAR San Fernando analizando el comportamiento de los SST y sus respectivos espectros UV-Vis (ver numeral 2.1.1). Además, se comprobó visual y numéricamente la calidad de los datos (presencia de *outliers*), mejoras en la predicción y parsimonia del modelo (selección de variables predictoras).

En el caso de los SST del afluente (Figura 40), hay una gama muy estrecha de longitudes de onda (725 - 750 nm), donde la recurrencia de 1000 ejecuciones aleatorias (selección de 67% de los datos de entrada) es más de 800. En esta figura, el "*dust*" representa el grado de afinidad de algunas de las parejas de datos espectros-concentración que deben ser detectados y definidos como *outliers*. Por lo tanto, para establecer numéricamente si existen o no *outliers* en las bases de datos, y así como su nivel de impacto tiene en un modelo regresivo, se implementó el método de detección de *outliers* desarrollado por Zamora y Torres (2012b) (ver numeral 3.4.2), y el programa para estimar la concentración de determinantes *OPP* con las modificaciones propuestas en este trabajo.

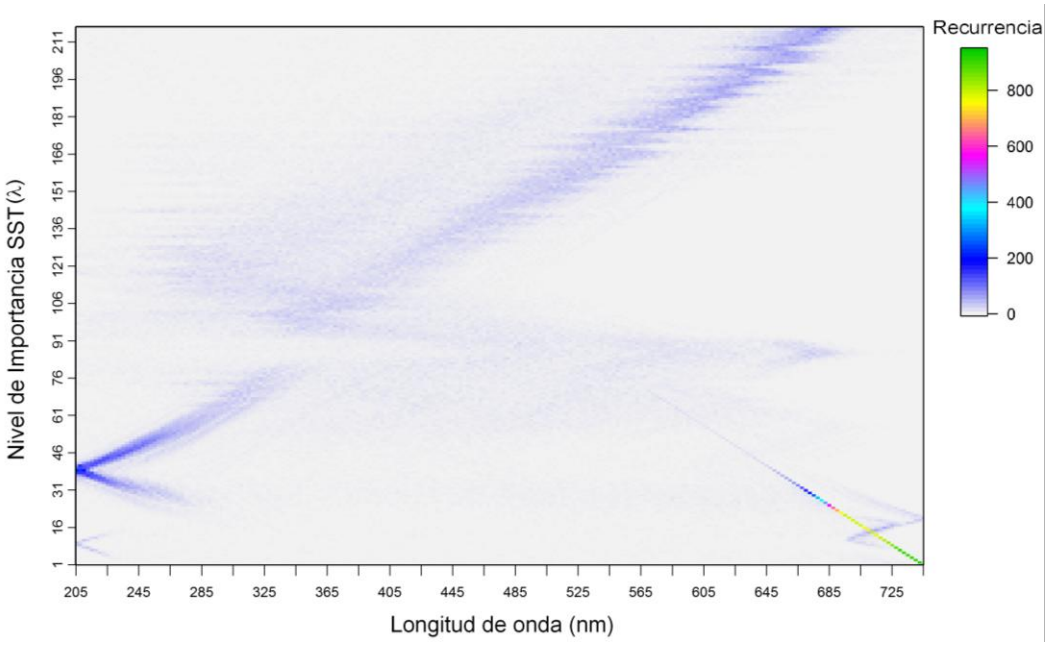


Figura 40- Rayo Afinidad: recurrencia, grado de importancia y calidad de los datos sobre la relación espectro-concentración del afluente-SST (Zamora y Torres, 2012b)

Entonces, para evaluar las premisas propuestas en el anterior párrafo se realizaron los siguientes pasos:

i. Se calibraron 1000 modelos *PLS* con el programa *OPP* utilizando las bases generadas de forma aleatoria por el método *ZATO.*
ii. Se determinó cuál de los 1000 *PLS* generó el menor *RMSEP,* de lo cual este se abstrae cuántas y cuáles longitudes de onda fueron usadas para tal fin (Figura 41).
iii. Luego se comparó las longitudes de onda establecidas con mayor porcentaje de importancia por el método *ZATO* versus las seleccionadas en por el mejor modelo *PLS* corridas en el programa *OPP* (Figura 42).
iv. Se detectaron y eliminaron los *outliers* de las bases de datos por medio del método mencionado. A partir de esta nueva base de datos sin *outliers*, se constituyeron 1000 conjuntos de datos generados de forma aleatoria con el 67 % de la información.
v. El método *ZATO* es corrido nuevamente con las bases de datos generadas en el paso anterior, cuyo resultado es la gráfica de la Figura 43.
vi. Se repitió el paso i y ii, pero con las bases de datos sin *outliers* (Figura 44).
vii. Por último, se repitió el paso iii con los resultados del paso v y vi, cuyo resultado es la Figura 45.

A continuación se presentan los resultados para el ejemplo específico de los SST y espectros de absorbancia de las muestras del afluente de la PTAR San Fernando.

En las Figuras 57 y 60 se presentan los resultados de la evaluación de la métrica *RMSEP* para los 1000 modelos *PLS* entrenados con (*WOM*) y sin *outliers* (*WoOM*) respectivamente. En estos gráficos el eje de la ordenas representa los valores de *RMSEP* en mg/L y en las abscisas el número de longitudes de onda (nl) utilizadas por cada modelo. Por otra parte, se presenta en rojo y azul los modelos con mayor y menor valor de RMSEP respectivamente, asociado al número de longitudes de onda utilizadas por cada modelo.

Resulta interesante como los modelos *WOM* y *WoOM* entrenados con alrededor de 80 o 40 longitudes de onda respectivamente, presenta errores de magnitud similar a modelos entrenados con dos longitudes de onda. No obstante, el menor *RMSEP* generado por el mejor modelo *WOM* no se presenta para modelos entrenados con dos longitudes de onda (Figura 41); pero en el caso de los modelos *WoOM* es probable encontrar modelos con un menor número de longitudes de onda y cuyos valores de *RMSEP* sean similares a mejor modelo entrenado sin *outliers* (Figura 44).

Por otra parte, se puede observar una amplia dispersión en los errores de predicción asociados al número de longitudes de onda (nl=2 a 82) usadas por modelos *WOM.* Por el contrario, los resultados obtenidos por los modelos *WoOM* (nl=2 a 41) presentan una menor variación, con una máxima diferencia de 13.02 mg/L entre el máximo y el mínimo *RMSEP* en comparación con 94.73 mg/L en el caso de los modelos *WoOM*. Aunque el menor *RMSEP*= 7.94 mg/L se presenta en los modelos *WOM* este es menos parsimonioso si se le compara con el mejor modelo *WoOM,* el cual es entrenado con 40 longitudes de onda que es la mitad de las usadas por el mejor modelo *WOM* (*RMSEP*= 21.72 mg/L). Lo anterior obedece a la restricción impuesta por el algoritmo *OPP*_modificado (ver numeral

3.2) de entrenar un modelo *PLS* con un máximo longitudes de onda igual a la cantidad de muestras analizadas.

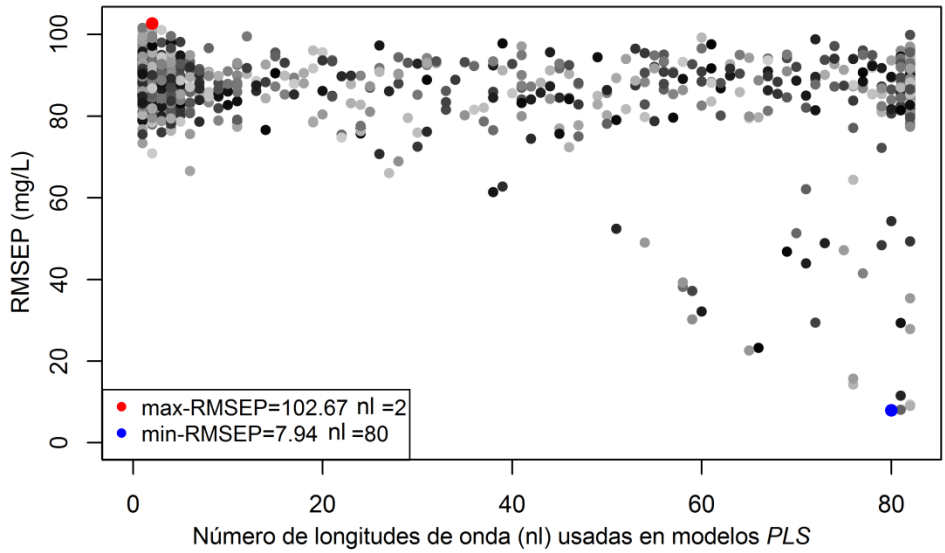


Figura 41- Número de longitudes de onda versus los valores de *RMSEP* de los modelos *PLS* calibrados con *outliers* (*WOM*)

Lo más importante del análisis anterior no es la reducción de la variación del error de predicción en múltiples modelos sin *outliers,* sino cómo antes de entrenar dichos modelos se puede establecer la presencia de *outliers* al relacionar la dispersión o *dust* presente en la Figura 40 y recurrencias similares con un bajo porcentaje en el nivel de relevancia (Figura 42) que tienden a generar buenos resultados, pero seleccionan una gran cantidad de longitudes de onda que no representan el proceso físico de la interacción de la luz con el determinante objetivo, como en el caso de los SST donde la presencia de este determinante en el espectro UV-Vis tiende a estar relacionada con la parte Visible. Este fenómeno lo puede representar el método *ZATO* aún en presencia de *outliers*, como se puede observar en las Figuras 56 y 58, o sin ellos Figura 43 y 61.

En la Figura 42 y Figura 45 el eje de las ordenadas representa el porcentaje de relevancia de la relación de los valores de absorbancia de una longitud de onda en los espectros UV-Vis con las concentraciones de un determinante (SST). En el eje de las abscisas se presentan las longitudes de onda del espectro en nanómetros. Dentro de estas gráficas las líneas multicolor representan la relevancia de cada longitud de onda determinada por el método *ZATO* y en color negro los modelos *WOM* y *WoOM* con los menores valores de *RMSEP*.

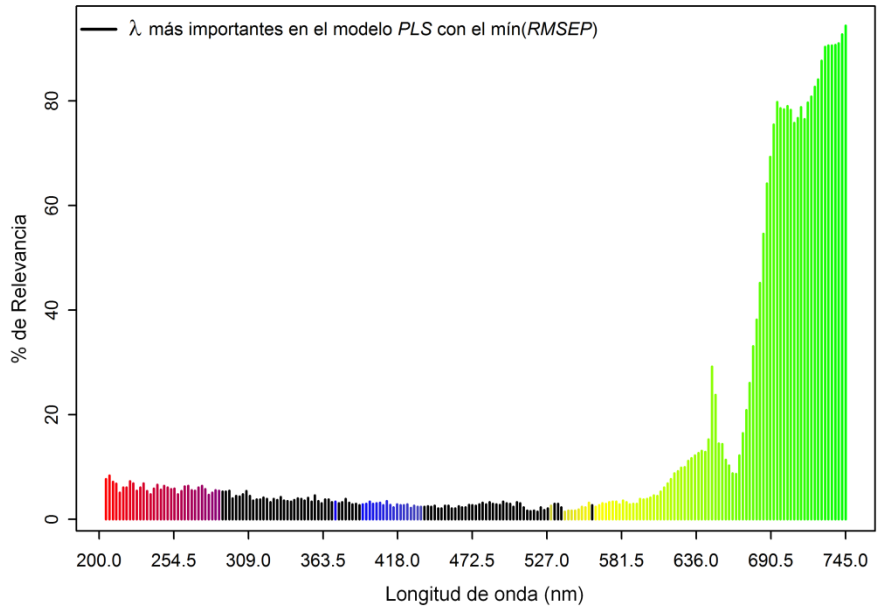


Figura 42- Porcentaje de relevancia de las longitudes de onda que revelan la presencia de las concentraciones de un determinante (SST) en el espectro UV-Vis para 1000 conjuntos de datos con *outliers* seleccionados de forma aleatoria

Además, el método revela desde el primer análisis a la base de datos con *outliers* las longitudes de onda que deberían ser utilizadas para que los modelos, en este caso *PLS*, estimen las concentraciones de un determinante de calidad del agua con errores menores luego de su eliminación. Esto se puede observar al comparar los gráficos de las Figuras 58 y 61, donde las longitudes de onda que más se relacionan con la presencia del determinante en los modelos *WOM* y *WoOM* están en el rango de 690.5 nm a 745 nm. Incluso después de la eliminación de *outliers* el nivel de relevancia (Figura 45) tiene mayores diferencias entre longitudes de onda y cuya selección para la calibración de un modelo *PLS* representa menores valores *RMSEP.*

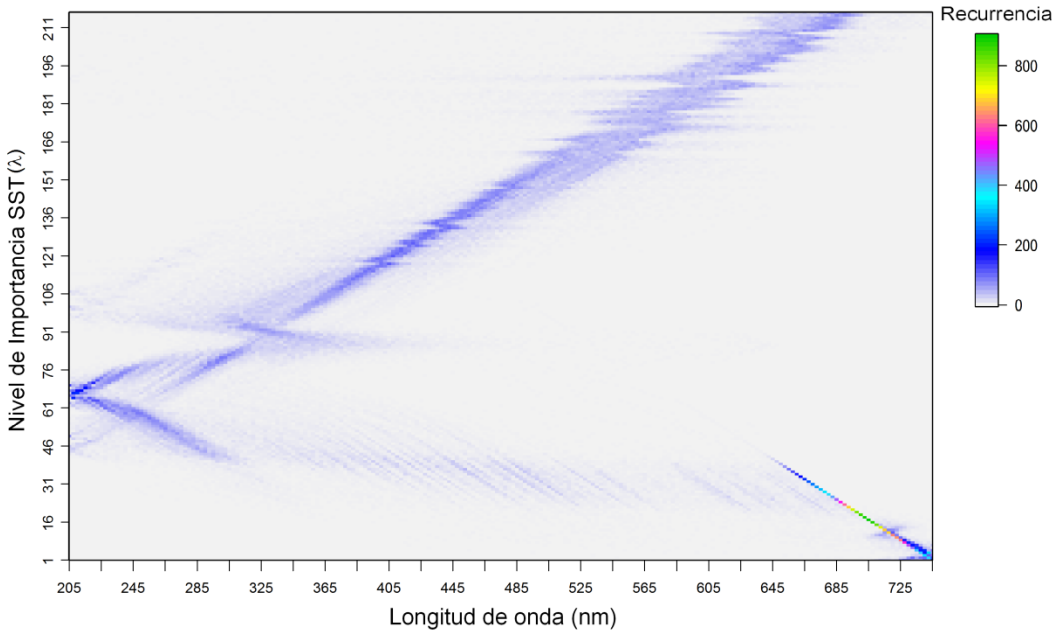


Figura 43- Rayo Afinidad: recurrencia, grado de importancia y calidad de los datos sobre la relación espectro-concentración del afluente-SST sin o*utliers*

Finalmente, como producto de la eliminación de *outliers* y de la selección de las longitudes de onda más importantes, los valores *RMSEP* para cualquier conjunto de datos empleado en la calibración de un modelo *PLS* tiende a ser menor y permanecer en un rango cuya variación es menor en magnitud, como se puede comparar entre la Figura 41 y Figura 44 con las bases de datos con y sin *outliers* respectivamente.

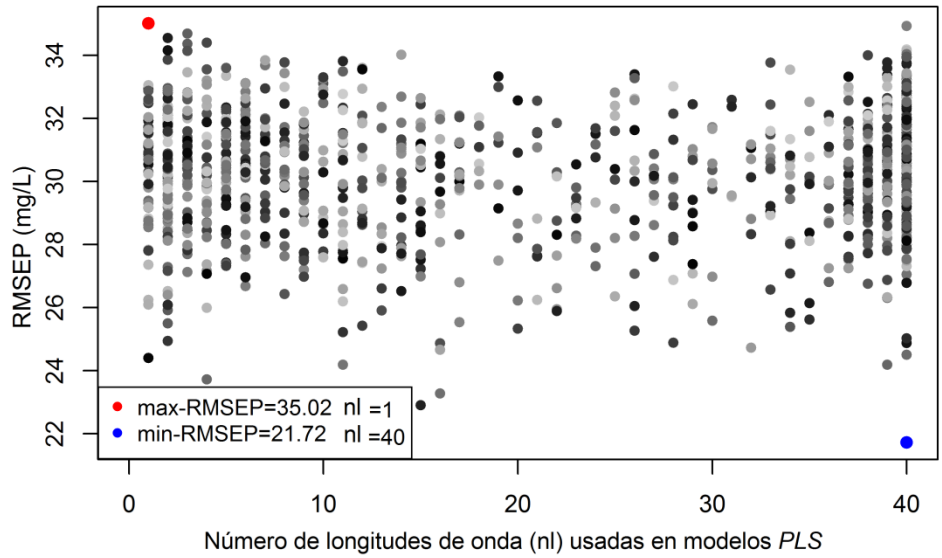


Figura 44- Número de longitudes de onda versus los valores de *RMSEP* de los modelos *PLS* calibrados sin *outliers* (*WoOM*)

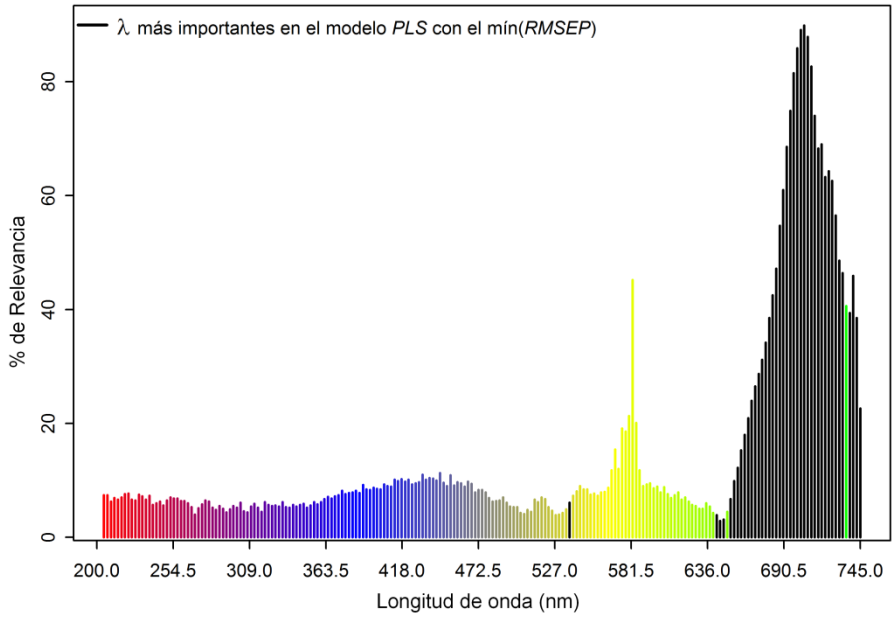


Figura 45- Porcentaje de relevancia de las longitudes de onda que revelan la presencia de las concentraciones de un determinante (SST) en el espectro UV-Vis para 1000 conjuntos de datos sin *outliers* seleccionados de forma aleatoria

3.4. METODOLOGÍAS DESARROLLADAS PARA LA DETECCIÓN DE *OUTLIERS*

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 determinantes de interés y la medición de los espectros de absorbancia de estas muestras. De allí que la información obtenida en el laboratorio sea un insumo fundamental para que las mediciones *in situ* y en continuo realizadas por la sonda generen resultados satisfactorios que representen la dinámica de los determinantes y cuyas magnitudes correspondan al estado físico-químico capturado en espectro de absorbancias representada a partir del espectro de absorbancias, conjunto de variables independientes a través de la cual se estiman las concentraciones. 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 las prácticas de laboratorio.

Dos métodos fueron desarrollados por Zamora y Torres (2012b; 2013), para la detección de *outliers*, con el fin de eliminar de las bases de datos los espectros de absorbancia y sus respectivas concentraciones de laboratorio declarados como *outliers*. A continuación se describen en detalle los pasos de cada método de detección de forma independiente.

3.4.1. Función cuantil con un polinomio de segundo grado (Zamora y Torres, 2013)

Este método se basa en la metodología de John Tukey (1977). Se proponen siete pasos sucesivos, los cuales se describen a continuación:

i. Calcular el coeficiente de correlación (r) entre los valores de absorbancia de cada longitud de onda del espectro y la concentración del determinante 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 determinante estudiado, denominado *miw* (*most important wavelength*), y a partir de esto conformar dos grupos de datos: uno con las absorbancias asociadas a dicha longitud de onda y otro con su correspondiente concentración.

iii. 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. Este proceso se repite 50000 veces, utilizando la Ecuación 56.

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

Ecuación 56-

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 y $\hat{y}_i$ son las concentraciones calculadas de la ecuación lineal (con $i = 1,2, \ldots . n$).

iv. A partir de los coeficientes calculados, se estiman las concentraciones para cada una de las ejecuciones aleatorias en función de las absorbancias de la *miw*, obtenidas en el paso anterior (Figura 46).

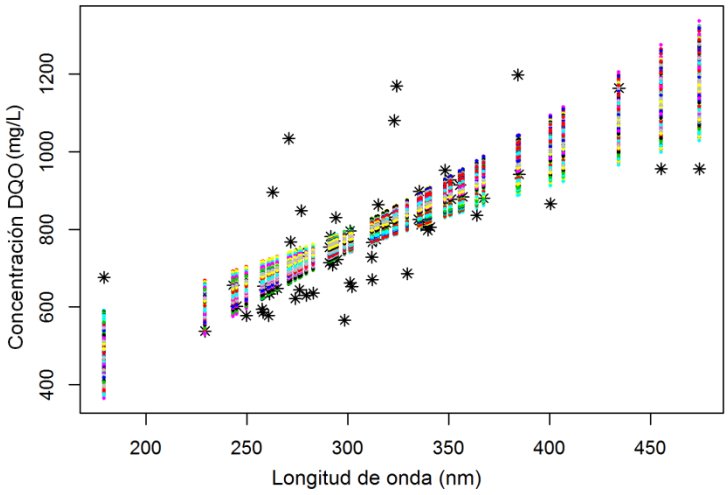


Figura 46- Concentraciones estimadas para cada una de las ejecuciones aleatorias en función de las absorbancias de la *miw* por muestra

v. 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.

vi. Luego, 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 $n \times k$ (Ecuación 57).

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

Ecuación 57-

donde 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. La calibración de estos coeficientes se realizó por medio de la función *nls* del programa R-project (ver numerales 1.4.4 y 2.1.1).

vii. 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 %). Aquí el rango intercuantil es la diferencia entre MQ_3 y MQ_1.

A manera de ejemplo se presentan los resultados de la implementación del método con el conjunto de datos de espectros UV-Vis y concentraciones de la SST, DQO y DQOf (ver numeral 2.1.1) de las muestras del afluente de la PTAR San Fernando (Zamora y Torres, 2013).

En la Figura 47 se presentan los resultados de la detección de outliers del afluente respectivamente. A la izquierda de estas figuras en las ordenadas se muestran las gráficas de detección de *outliers* para las concentraciones de SST, y 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 estas figuras, se consolida la cantidad y porcentaje sobre el total de datos validados, *mild outliers* y *extreme outliers* (Zamora y Torres, 2013).

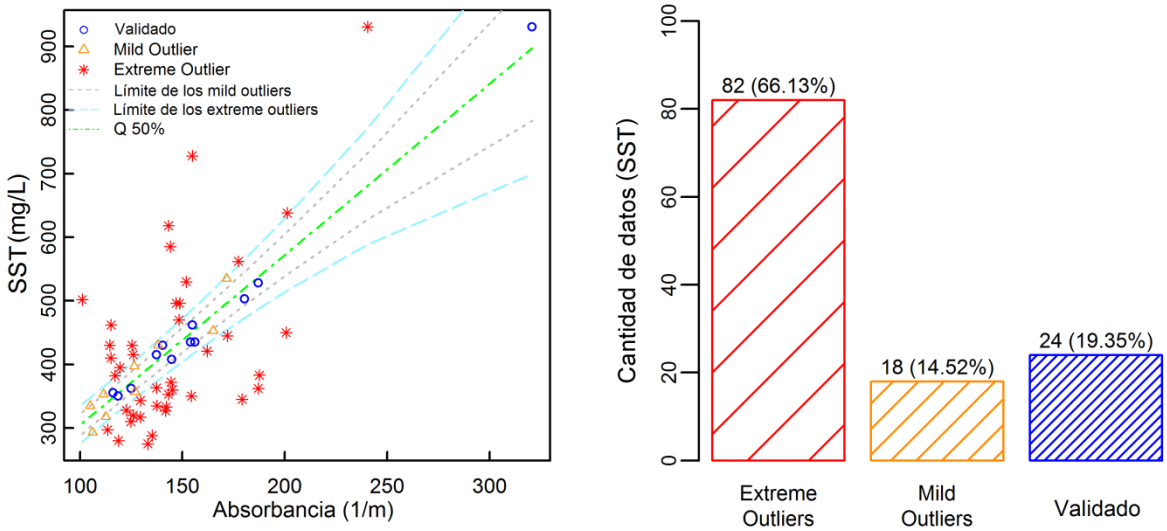


Figura 47- (Izq.) Detección de los datos *mild outliers*, *extreme outliers* y validados; (Der.) cantidad y porcentaje de los datos detectados en los conjuntos de datos de SST del afluente

Los autores determinaron para el caso de los SST que en general un mayor porcentaje de *Extreme outliers* (*Eo*) fueron detectados específicamente en el rango de absorbancias de 100 a 350 1/m. Con respecto a los *Mild outliers* (*Mo*) estos presentaron una mayor cantidad que los datos denominados validados (DR). Además, lograron establecer que las absorbancias más importantes relacionadas con la presencia de la DQO se encuentran tanto en la parte UV como en la Visible, ya que este determinante está relacionado con oxidación química de la materia disuelta y en suspensión. 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 1/m a 150 1/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 *Mo* (Zamora y Torres, 2013).

Por último, los autores confirmaron la mejora predictiva calibrando y validando modelos *PLS* utilizando el programa *OPP,* comparando la bondad de ajuste de los modelos cuando son entrenados con bases de datos con y sin *outliers*, los resultados se presentan a continuación:
En la Figura 48 y la Figura 49 se presentan los resultados de calibración y validación de los modelos con (*WOM*) y sin *outliers* (*WoOM*) de izquierda a derecha, para el afluente. En las ordenadas de estos gráficos se presentan los valores de las concentraciones calculadas mediante 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* (coeficiente de Nash-Sutcliffe (1970)) para los validados (denominados por los autores DR) y *outliers*, de los modelos entrenados con y sin *outliers*.

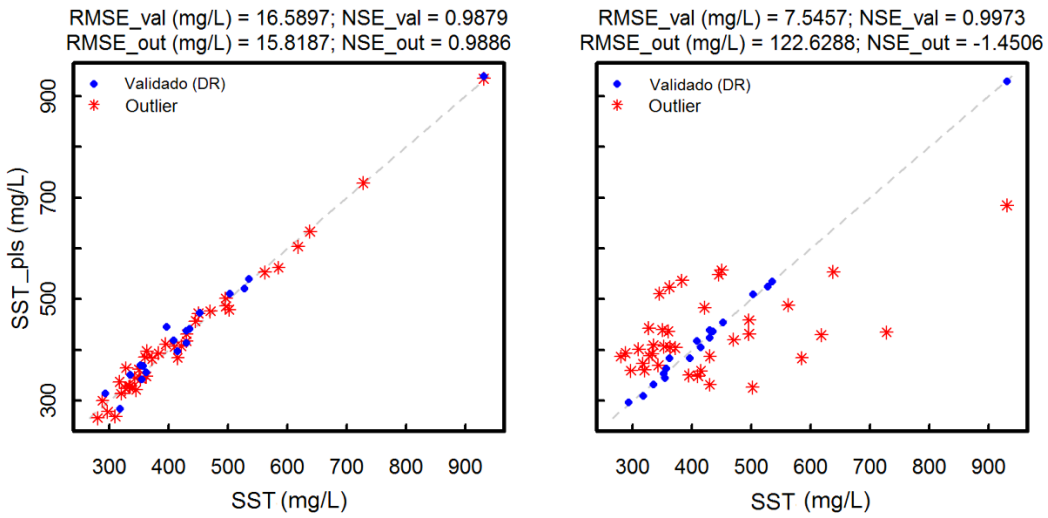


Figura 48- *RMSE* y NSC para la calibración de los modelos *PLS* con (Izq.) y sin *outliers* (Der.) para los SST (Zamora y Torres, 2013)

El nivel de ajuste alcanzado en el proceso de calibración (Figura 48) para los grupos de DR de los SST es el mejor para los resultados generados con el modelo *PLS* calibrado sin considerar *outliers*. Con respecto al grupo de datos *outliers*, se encontraron mejores resultados con el modelo con *outliers*, 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 determinantes de calidad del agua. 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 determinantes 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 son pobres. 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 incrementará el error de predicción ni afectaría la capacidad predictiva del modelo (Zamora y Torres, 2013).

Por otra parte los valores negativos del coeficiente NSC para los *outliers* en el modelo calibrado sin *outliers* denotan que los residuos de los errores generan una mayor varianza que la varianza de los datos originales.

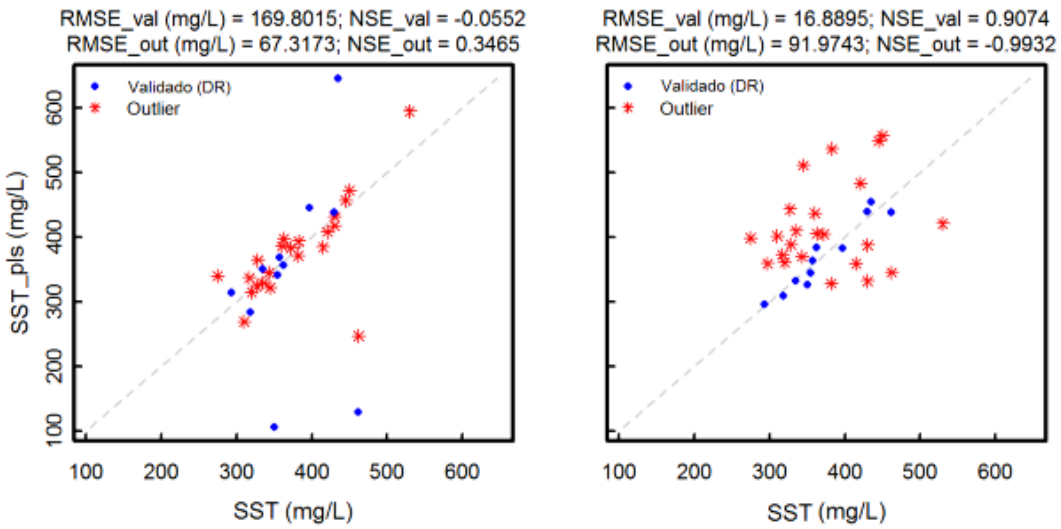


Figura 49- *RMSE* y NSC para la validación de los modelos *PLS* con y sin *outliers* (afluente)

En la validación del conjunto de DR del afluente (Figura 49), se encontró que el error en mg/L de los SST es 10 veces menor en el modelo *WoOM*, en comparación al *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*. Sin embargo, la menor diferencia del error entre ambos modelos para los datos *outliers* de SST respalda la premisa que el número de datos validados podría ser mayor, sobre todo los datos más cercanos al límite de los *Eo* como se observa en la Figura 47. 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 de *outliers* en el conjunto de datos espectros-concentraciones de SST del afluente y se corrobora con la mejora en la capacidad predictiva del modelo *PLS* para cada determinante (Zamora y Torres, 2013).

3.4.2. Múltiples escenarios *PLS:* análisis de la relación de la varianza con la bisectriz y la relación del *RMSE* local con respecto al *RMSE* global (Zamora y Torres, 2012b)

Este método para la detección de *outliers* está basado en los resultados de modelos *PLS* generados por 1000 simulaciones de Monte Carlo por medio del programa *OPP* (ver numeral 3.2), así como algunos elementos de la metodología propuesta por Tukey (1977) y evaluación del *RMSE*. Los siguientes son los elementos y pasos que conforman el método (Zamora y Torres, 2012b):

i. Calibrar 1000 modelos *PLS* con el 67 % de los datos de espectrometría y las correspondientes concentraciones de cada muestra seleccionados de forma aleatoria, y dejando el resto para la validación.
ii. Calcular los valores de concentración para los conjuntos de datos de calibración la calibración y validación utilizando cada uno de los modelos *PLS* calibrados.

iii. Se evalúa la primera condición para determinar si un par de datos es *outlier*: un conjunto espectro-concentración será declarado un *outlier* si la distancia entre el cuantil 50 % (Q_2) y $Eo = Q_{1-3} \pm 3 \times IQR$, donde Q_1 es el primer percentil, Q_3 es el tercer percentil y IQR es el rango intercuantil calculado como ($Q_3 - Q_1$) (Tukey, 1977), no intersecta la bisectriz en un gráfico de dispersión de los valores modelados versus los medidos (Figura 50-Derecha).

iv. Incluso si existe una intersección como se definió anteriormente, un dato también se considera como un *outlier* si la raíz cuadrada media del error ($RMSE_L$) (Ecuación 58) de las concentraciones estimadas por los modelos para una concentración medida es mayor que el $RMSE_G$ (Ecuación 59), el cual es obtenido de todos los datos de las 1000 ejecuciones realizadas (Figura 50-Izquierda). Esta segunda condición se propone debido a que algunos datos obtenidos de modelos regresivos pueden mostrar muy altas variabilidades en comparación con la media calculada para todo el conjunto de datos.

$$RMSE_L = \sqrt{\frac{\sum_{j=1}^{1000}\left(y_j - \hat{y}_j\right)^2}{m}}$$

Ecuación 58-

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

Ecuación 59-

donde n es el número de muestras, N es el número total de datos del conjunto de datos de calibración o validación y m es el número de veces que una muestra en las generaciones aleatorias hizo parte del conjunto de calibración o validación. Por último, y_j son los valores medidos en laboratorio de concentración y $\hat{y}_j$ las concentraciones estimadas por cada modelo *PLS* tanto para el conjunto de calibración como de validación. Por consiguiente, una pareja de datos se define *outlier* si:

$$\frac{RMSE_L}{RMSE_G} \geq 1$$

Ecuación 60-

Esta última condición indica que el valor de la concentración obtenida en laboratorio no concuerda con la calidad del agua de la muestra, y así las longitudes de onda asociadas a ella en el espectro de absorbancia no son las que físicamente o químicamente la representan.

Con base en los pasos presentados anteriormente, se muestran parte de los resultados alcanzados por los autores del método en el caso de la DQO de las muestras del afluente de la PTAR San Fernando. En la Figura 50 se puede observar la aplicación de las 1000 simulaciones de Monte Carlo, representadas en los gráficos de dispersión (Izquierda), donde el eje de ordenadas corresponde a las concentraciones equivalentes de la DQO obtenidas de los modelos de *PLS* y el eje de las abscisas corresponde a las concentraciones de referencia del análisis de laboratorio para el mismo determinante, tanto para la calibración (arriba) como para validación (abajo). En estas gráficas se puede observar la variabilidad que el valor de una concentración del determinante puede tener, e incluso la forma en que varía cuando una muestra (dato espectro-concentración) hace parte de la calibración o la validación de un modelo. Por lo tanto, la variabilidad de los errores de predicción determinará si una pareja de datos de espectro-concentración es *outlier*, tal como se define por el segundo criterio de detección (Zamora y Torres, 2012b).

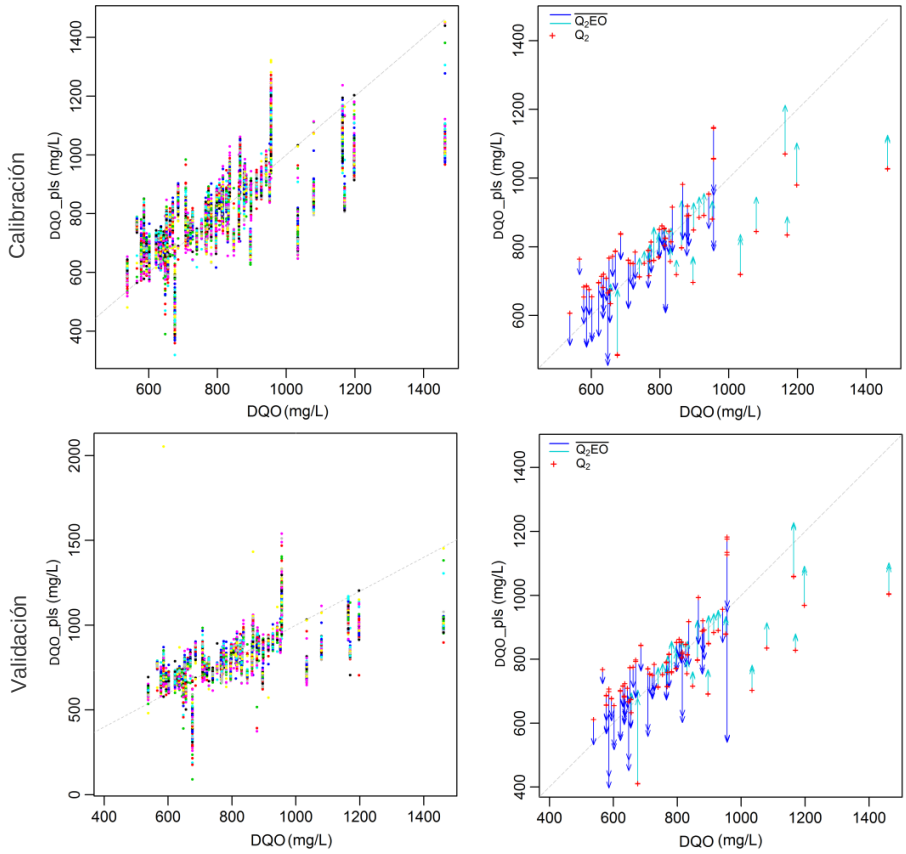


Figura 50- 1000 ejecuciones de modelos *PLS* y detección de *outliers* con el primer (Der.) y segundo (Izq.) criterio propuestos, para la DQO de las muestras del afluente del PTAR San Fernando (Zamora y Torres, 2013)

Después de ver la forma en que se determina si un par de datos es *outlier* mediante el empleo de ambas condiciones o criterios de detección, es necesario saber cuáles se clasifican como *outliers*. Por lo tanto, en la Figura 51 se muestran dos tipos de gráficos: el primero y el segundo de izquierda a derecha representan en el eje de ordenadas el valor del cuantil 50 % y en las abscisas la concentración de los valores de DQO medidos en laboratorio. En estos gráficos, se determinó cuáles eran los datos clasificados en cada condición como valores atípicos sólo en la etapa de calibración o validación (triángulo y cuadrado, respectivamente), cuáles datos fueron igualmente detectados en ambas etapas (círculo), y qué datos fueron clasificados como válidos. Por otra parte, el gráfico de la Figura 52 resume y compara cuáles muestras se clasifican como valores atípicos para cada condición y etapa.

Para el ejemplo específico de la PTAR San Fernando, se detectó para los datos de la DQO una mayor cantidad de *outliers* para la primera condición (bisectriz), pero la misma cantidad en la etapa de calibración y validación de esta condición (38 *outliers*) con respecto a la mayor cantidad detectada por la segunda condición (residuos) para la etapa de calibración (26 *outliers*). Por otra parte, dos de las muestras detectadas por la primera condición fueron diferentes tanto en la calibración [37 y 122] como en la validación [99 y 110]. Por último, el número de *outliers* detectados por la condición dos fue menor y

diferente entre etapas, puesto que en la calibración se detectaron 26 de datos, mientras que en la validación 24 (Zamora y Torres, 2012b).

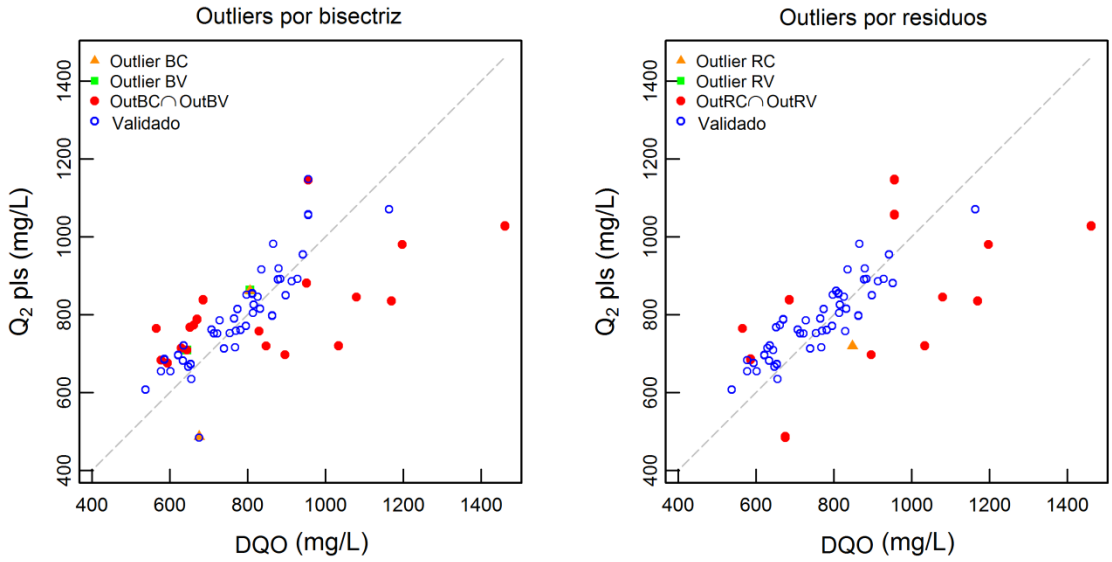


Figura 51- *Outliers* detectados por medio de los dos criterios propuestos (Zamora y Torres, 2012b)

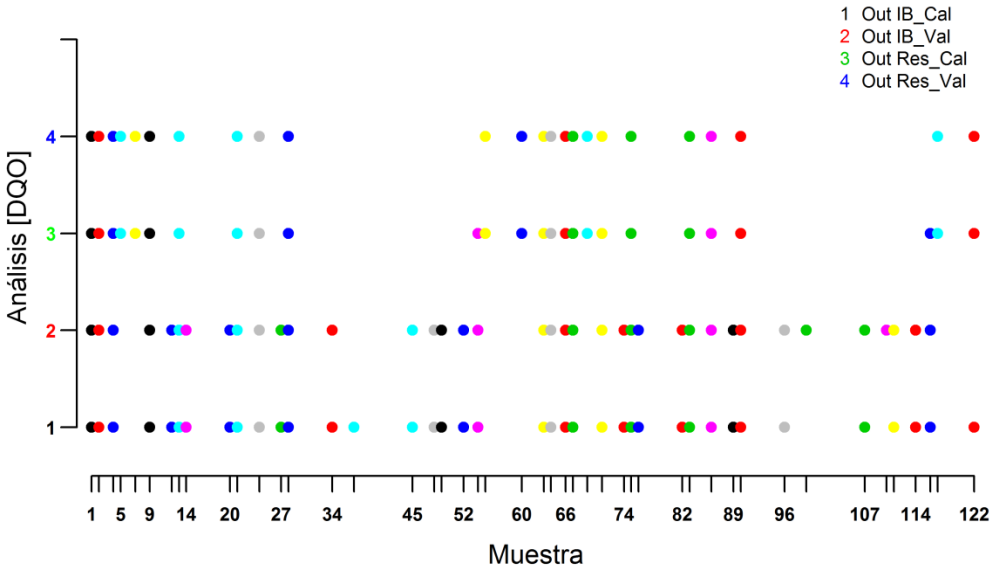


Figura 52- Resumen y comparación de los datos detectados como *outliers* por ambos criterios en el caso de la DQO (Zamora y Torres, 2012b)

Finalmente los autores entrenaron modelos *PLS* con conjuntos de datos con y sin *outliers*, para comparar el desempeño en la etapa de calibración y validación.

La Figura 53 muestra los resultados de la calibración y la validación de izquierda a derecha de los modelos con *outliers* (recuadro azul) y sin *outliers* (recuadro verde). En la ordenada de estos gráficos, se muestran los valores de las concentraciones calculados por los modelos de regresión *PLS* y en la abscisa la concentración de valores obtenidos en

laboratorio. En la parte superior de cada figura, se presentan los resultados de las métricas evaluadas r, R^2 y *RMSEP*, para cada modelo entrenado con y sin *outliers*. El nivel de ajuste alcanzado en el proceso de calibración (izquierda) para los grupos de datos validados es mejor para los resultados generados por la calibración del modelo *PLS* que aquel obtenido excluyendo los *outliers*. Por lo tanto, la aplicación del método mejoró la predictibilidad en las concentraciones de la DQO del afluente por medio de los espectros de absorbancia (Zamora y Torres, 2012b).

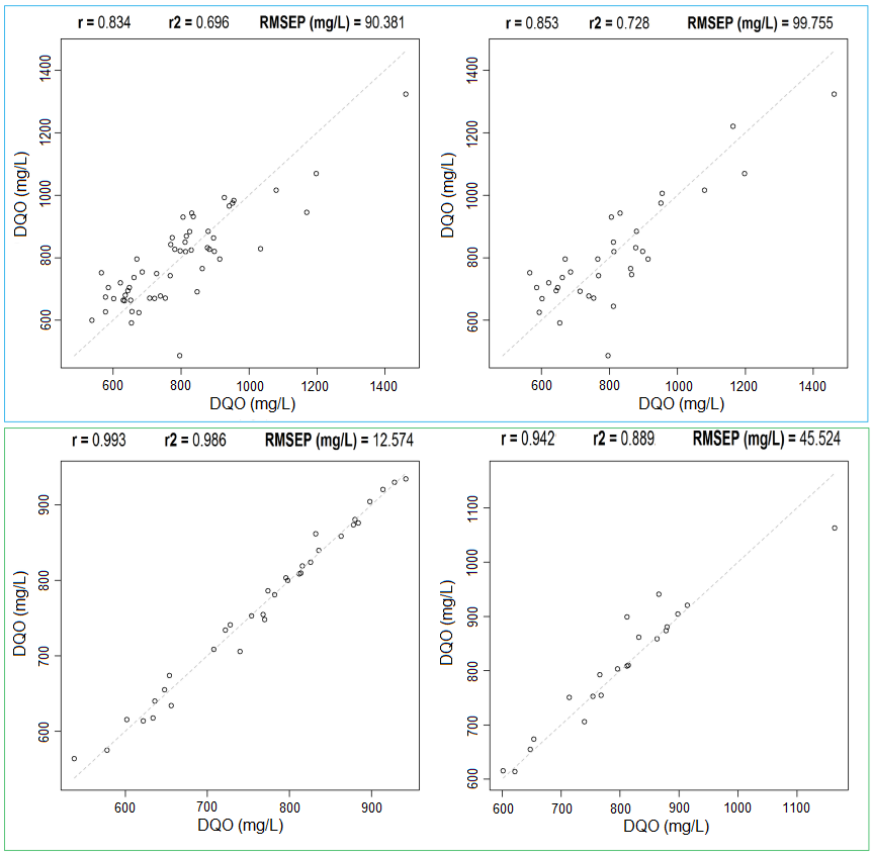


Figura 53- Los resultados de los modelos PLS calibrados para DQO con *outliers* (recuadro azul) y sin *outliers* (recuadro verde) (Der.: Calibración – Izq.: Validación)

Este último método de detección de *outliers* no será implementado para la detección de *outliers* en las bases de datos de los puntos de monitoreo PTAR de *Fontaines-sur-Saône* y EE Gibraltar, ya que los tiempos computacionales requeridos para correr 1000 PLS son muy altos, alrededor de 10 días.

3.5. ALGORITMO DE EVALUACIÓN - AEEC

Se desarrolló un algoritmo para evaluar los métodos *machine learning* RNA y SVM, y el método *PLS* empleados para estimar las concentraciones equivalentes de los determinantes SST, DQO y DQOf presentes en las muestras del afluente de la EE de Gibraltar y de la PTAR de *Fontaines-sur-Saône* a través de datos de espectrometría UV-Visible. Dicho algoritmo no solamente contempla la evaluación de los métodos regresivos de inteligencia artificial propuestos sino que se integra a un conjunto de metodologías que determinan: (i) incertidumbre de las concentraciones y espectros UV-Vis de los determinantes objetivo, (ii) detección de *outliers,* (iii) calibración y validación de los métodos regresivos, y (iv) evaluación del desempeño de los resultados de los modelos generados para cada método.

El algoritmo está conformado por cinco fases descritas a continuación (Figura 55):

Fase 1: empleando la metodología presentada en el numeral 2.4.1.1 se determina la incertidumbre de las concentraciones de los determinantes objetivo y los valores de absorbancias por cada longitud de onda del espectro UV-Vis, para lo cual se generaron simulaciones de Monte Carlo de 50000 ternas de réplicas para cada muestra.
Fase 2: suponiendo una distribución normal de los datos de concentración de cada determinante y de los valores absorbancia en cada longitud de onda de los espectros, se generan 1000 conjuntos con valores de concentraciones de cada muestra obtenidos de forma aleatoria en un rango de $\mu_{ci,k} \pm u_{ci,k}$, donde $\mu_{ci,k}$ es valor de la media y $u_{ci,j}$ es la incertidumbre compuesta de la muestra i, obtenidos del análisis de incertidumbre de la Fase anterior. De la misma forma se generan igual cantidad de conjuntos con espectros UV-Vis, evaluando el valor de absorbancia $\mu_{\lambda i,k} \pm u_{\lambda i}, k$ por cada longitud de onda (λ) y por cada muestra i.

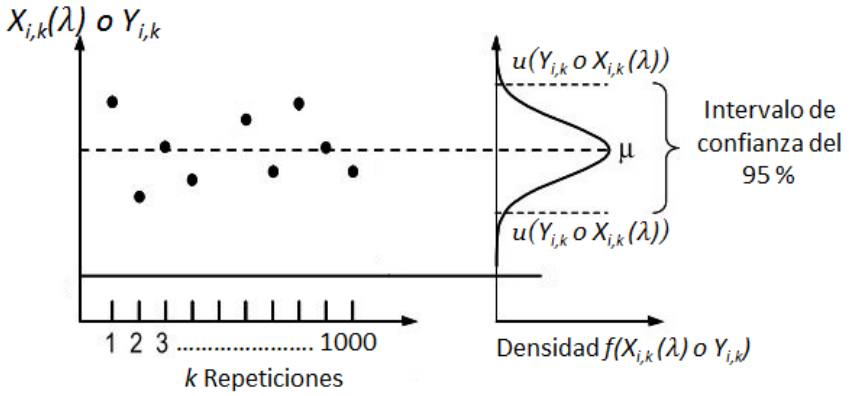


Figura 54- Generación aleatoria de valores de concentración y absorbancias en cada longitud de onda del espectro suponiendo que los datos sigue una distribución normal

Fase 3: con base en los conjuntos de datos de concentraciones Y_k y espectros de absorbancia X_k generados en la Fase anterior ($k = 1\ a\ 1000$), se realiza la detección de *outliers* de forma bivariada para cada grupo de datos $Y_k X_k$ (*e.g.* $Y_1X_1 \dots Y_6X_6 \dots Y_{1000}X_{1000}$) empleando la metodología descrita en el numeral 3.4.1.

Fase 4: se elimina de cada grupo de datos $Y_k X_k$ las muestras que fueron catalogadas como *outliers,* y con los datos restantes denominados validados (DR), se conforman de forma aleatoria las parejas $Y_{C-DR,k} X_{C-DR,k}$ para la calibración de los modelos regresivos empleando el 67 % de la información y el 33 % restante para conformar los grupos de validación $Y_{V-DR,k} X_{V-DR,k}$.

Fase 5: a partir de los grupos de datos $Y_{C-DR,k} X_{C-DR,k}$ se calibran 1000 modelos de RNA, *SVM* y *PLS*, y se validan con $Y_{V-DR,k} X_{V-DR,k}$. Los algoritmos para calibrar los modelos RNA y *SVM* son explicados en los numerales 3.5.1 y 3.5.2 respectivamente. Por otra parte, el algoritmo *OPP_*modificado fue aplicado para calibrar los modelos *PLS* (ver numeral 3.2).

Fase 6: finalmente se evalúa el desempeño por medio de las métricas de ajuste *RMSEP*, *RMSE*, coeficiente de determinación (R^2) y coeficiente de correlación de Spearman (ρ).

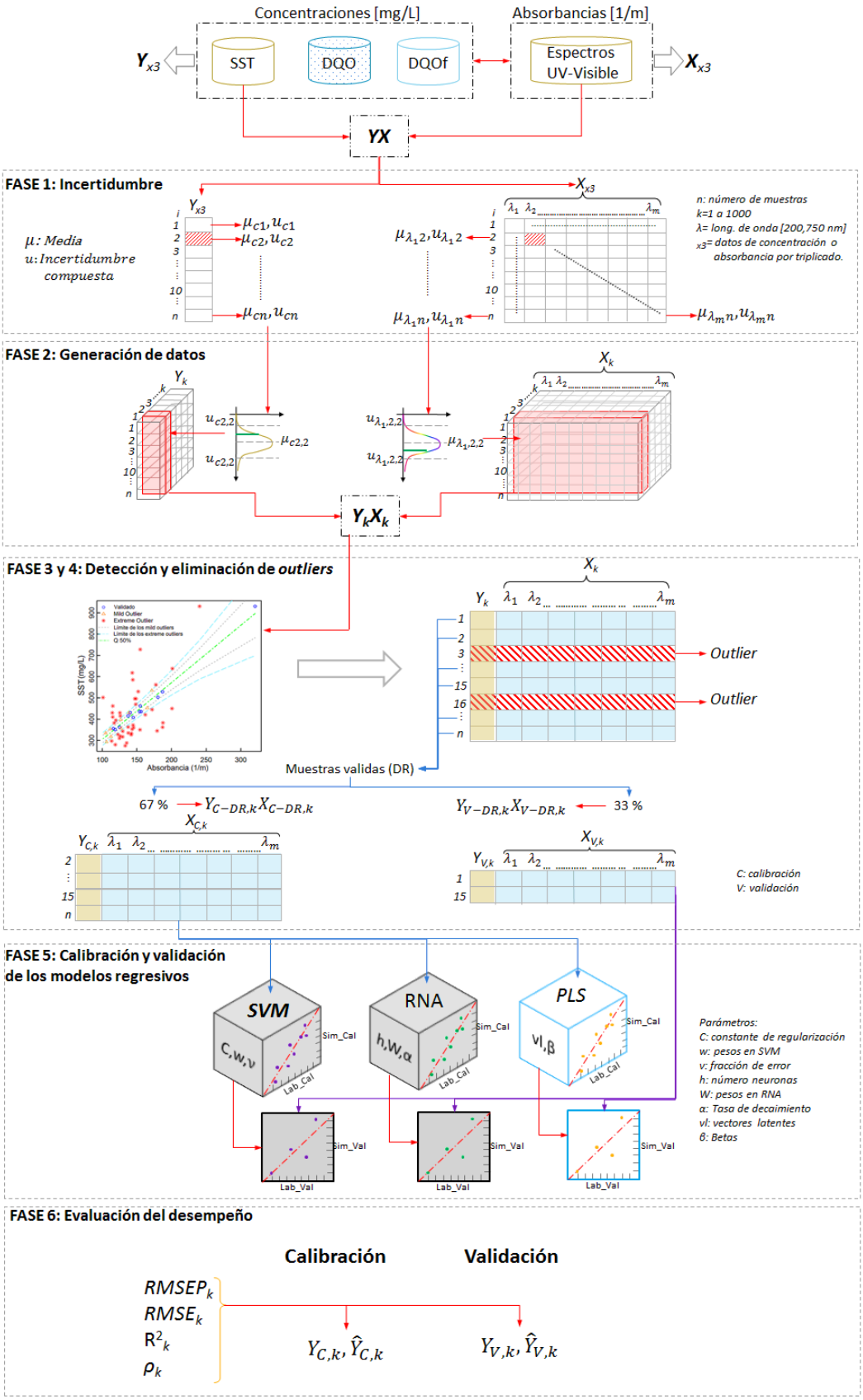


Figura 55- Algoritmo general para la evaluación de la relación espectro-concentración (Autor)

3.5.1. Algoritmo para la calibración de un modelo regresivo de RNA

La calibración del modelo de RNA tipo *feed-forward* implementado en la función *nnet* del programa R, se centró en determinar el mejor número de longitudes de onda, la tasa de decaimiento del peso (α) y el número de neuronas en la capa oculta con los cuales el modelo estimará de la mejor forma posible los valores de concentraciones equivalentes a partir de los valores de absorbancia de las longitudes de onda seleccionadas. Para este fin se desarrolló un algoritmo de evaluación, conformado por siete pasos descritos a continuación:

i. Luego de la detección de *outliers* se reescala el conjunto de datos la generación k de concentraciones $Y_{DR,k}$ entre 0.05 a 0.95, y los espectros de absorbancia $X_{DR,k}$ se reescalan entre -1 y 1 por cada longitud de onda.
ii. Los conjuntos de datos reescalados se dividen de la forma como se indicó en la Fase 2 del algoritmo presentado en el numeral 3.5.
iii. El conjunto de calibración ($Y_{C-DR,k}, X_{C-DR,k}$) es dividido a su vez en un conjunto de entrenamiento ($Y_{CE,k}, X_{CE,k}$) con el 67 % de la información de forma aleatoria y otro de prueba ($Y_{CP,k}, X_{CP,k}$) con el 33 % restante. De este tipo de conjuntos se generan 20 denominados ($[Y_E, X_E]_{z,k}$), con lo cual se pretende encontrar el mejor conjunto de variables de entrada que generen un modelo calibrado que represente un amplio universo del fenómeno (concentraciones) con bajos errores de predicción.
iv. Unificar de forma decreciente de acuerdo a la importancia de la relación (*IF*) entre cada longitud de onda de los espectros UV-Vis y las concentraciones de las muestras. Para esto se utilizó el método descrito en el numeral 3.1.
v. El primer modelo es entrenado sin neuronas (*h*) en la capa oculta, pero a medida que el proceso iterativo avanza el valor de este parámetro cambia de 1 a 10. Luego, el número de neuronas es usado para calibrar modelos de RNA, donde el número de longitudes de onda (*nl*) de la variable independiente en los datos de entrada ($[Y_E, X_E(, nl)]_{z,k}$) aumenta desde 2 al número de muestras presentes en el conjunto de entrenamiento (*m*) conforme al *IF*, y además se realiza validación cruzada tipo *leave-one-out LOO,* el cual consiste en entrenar un modelo con n-1 muestras del conjunto de calibración y validar con una sola muestra del mismo conjunto; repitiendo este procedimiento hasta que cada muestra haya sido empleada tanto para entrenar como para validar los modelos (Witten y Frank, 2005). Este proceso se repite para cada uno de los conjuntos generados en el paso anterior.

   Por ejemplo para el primer conjunto de datos, el primero modelo está conformado por cero neuronas (*h=0*) en la capa oculta, con dos longitudes de onda y omite la primera muestra i del conjunto de datos para entrenar dicho modelo $[Y_E(-i), X_E(-i, nl)]_{1,k}$, y finalmente evaluar el $RMSE_{VC}$. Cuando finaliza el proceso de validación cruzada, *h* incrementa su valor en 1 para el cual se repite el proceso anterior, y así sucesivamente hasta 10. Después que *h* llega a su máximo, el valor

de este parámetro en el siguiente ciclo vuelve a ser cero, pero la cantidad de longitudes de onda aumenta a 3. Esto se repite hasta que *nl* es igual a *m*, luego se inicia el mismo proceso para el segundo conjunto de datos.

vi. Por otra parte, en cada modelo se optimiza α por medio del algoritmo de Evolución Diferencial (ver numeral 1.4.8.2), que fue implementado por Ardia *et al*. (2011) en la función DEoptim desarrollada en R. En dicha función se deben especificar ciertos parámetros y la función objetivo. En este orden se establecen los siguientes parámetros: límite inferior (0.00001) y superior (0.999) rango en el cual probablemente se encuentra el mejor valor de α, una función regresiva de RNA que evaluará cada valor del parámetro en el proceso de optimización, número de miembros de una población (NP=10), máximo número de iteraciones (generación de poblaciones) (NI=500), y se define la convergencia cuando el porcentaje de mejora entre iteraciones está por debajo de la tolerancia T= $1x10^{-8}$ después de evaluar 25 iteraciones. La función objetivo que será optimizada es la suma de los errores cuadrados (*SSE*-Ecuación 54), la cual tendrá por fin evaluar las concentraciones equivalentes obtenidas por el modelo RNA versus las concentraciones de laboratorio, y determinar con qué valor de α se minimiza dicha función.

vii. Los valores de $RMSE_{VC}$ de cada modelo calibrado en el paso v se guardan en una matriz, y de esta última se selecciona el menor valor de la métrica. Con base en este criterio, se abstraen los valores de los parámetros de la arquitectura del modelo RNA que generó el menor $RMSE_{VC}$ ($\hat{Y}_k = nnet(nl, \alpha, h)$), y se emplea dicho modelo en la etapa de validación.

Antes de implementar el algoritmo en la evaluación de los determinantes SST, DQO y DQOf de cada caso de estudio, se realizó una prueba con las concentraciones promedio de SST y los espectros de absorbancia pertenecientes a las muestras de la PTAR de *Fontaines-sur-Saône* en tiempo seco (ver Figura 32). En este ejemplo se demuestran las limitaciones del modelo RNA para esta aplicación. Para este caso puntual se analizó un máximo de 10 longitudes de onda, y los demás parámetros como se especifican en el algoritmo.

Para ejecutar el algoritmo de calibración de RNA, se utilizaron los computadores del Centro del Alto Rendimiento Computacional-ZINE de la Universidad[4], empleando para esta tarea un total de 5 computadores cada uno con 12 procesadores. En cada equipo se ejecutaron modelos de RNA con diferente número de longitudes de onda: Equipo 1 [2], Equipo 2 [3 y 4], Equipo 3 [5 y 6], Equipo 4 [7 y 8], y Equipo 5 [9 y 10].

Los resultados obtenidos de cada modelo ejecutado fueron concatenados en una sola base de la cual se obtienen los siguientes gráficos. En la Figura 56, Figura 57 y Figura 58, se presentan los errores de estimación obtenidos de los mejores modelos calibrados para los conjuntos de datos de entrenamiento, prueba y validación respectivamente. En estas figuras se presentan en el eje de las abscisas los índices de cada uno de los conjuntos de

[4] www.zine.javeriana.edu.co

datos evaluados (1 a 20), y en las ordenadas los valores de las métricas empleadas para calcular el nivel de ajuste y la capacidad predictiva de los modelos. En el caso de los conjuntos de entrenamiento se usó el *RMSEP*, mientras que para los conjuntos de prueba y validación el *RMSE.* Aunque los valores de dichas métricas se generan a partir de los datos de concentraciones estimadas y de laboratorio reescalados, las unidades de éstos continúan en mg/L.

Es claro al observar la Figura 56 que al aumentar el número de longitudes de onda en los modelos RNA, el valor de los *RMSEP* para cada conjunto de datos en general decrece, pero tiende a ser similar para los modelos que usan 6 o más longitudes de onda.

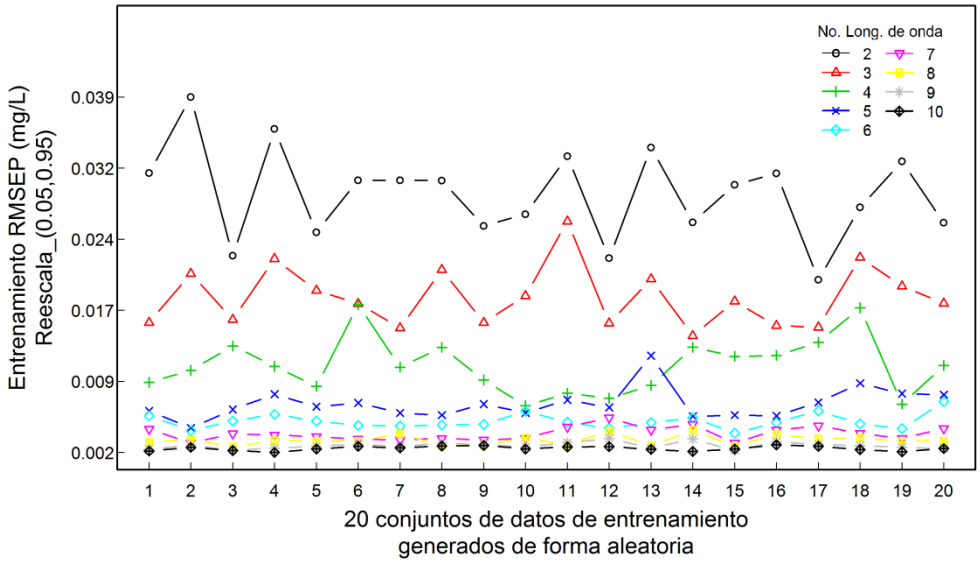


Figura 56- Menores valores de los *RMSEP* obtenidos de los modelos de RNA calibrados con diferentes conjuntos de datos y número de longitudes de onda (SST-PTAR *Fontaines-sur-Saône*, tiempo seco)

Sin embargo, al observar la Figura 57 y Figura 58 se puede establecer un comportamiento de *overfitting* generalizado para todas las arquitecturas generadas de los modelos de RNA. Por ejemplo se puede comparar los resultados obtenidos con el modelo calibrado con el conjunto de datos No. 2, el cual genera un *RMSEP* de 0.0023 mg/L para el conjunto de entrenamiento, mientras que el *RMSE* evaluado en los conjuntos de prueba y validación incrementa su valor a 0.138 mg/L y 0.152 mg/L respectivamente.

No obstante, aunque los errores obtenidos en la etapa de prueba y de validación son altos (baja capacidad predictiva del modelo *feed-forward*) se puede abstraer que para replicar los resultados obtenidos en el entrenamiento en cualquier conjunto de datos se debe establecer cuáles y cuantas longitudes de onda representan la variabilidad del fenómeno asociado a los valores de absorbancia, como se presenta, por ejemplo en el caso del conjunto de datos No. 19 para el cual se pueden presentar errores de predicción muy similares al usar 2 longitudes de onda como 10 (Figura 57 y Figura 58).

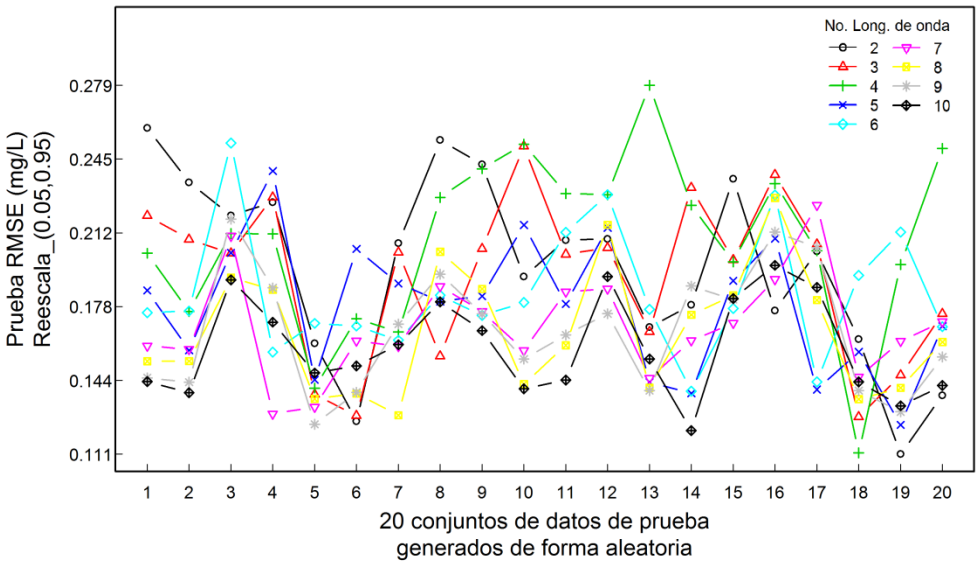


Figura 57- Resultados de la evaluación de los *RMSE* calculados para los conjuntos de datos de la etapa prueba de los mejores modelos RNA presentados en la Figura 56

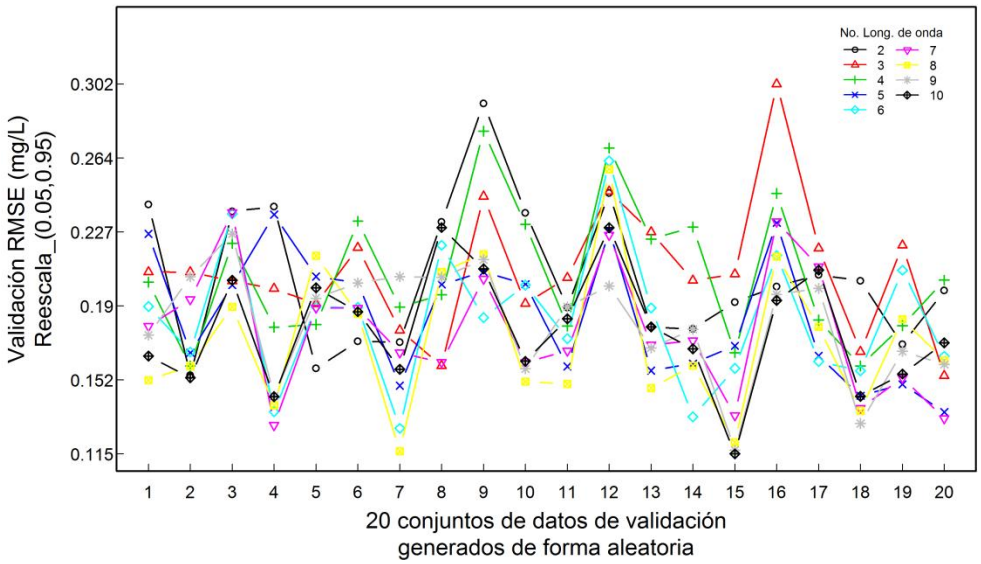


Figura 58- Resultados de la evaluación de los *RMSE* calculados para los conjuntos de datos de la etapa validación de los mejores modelos RNA presentados en la Figura 56

Por otra parte, se logró establecer que el valor de la tasa de decaimiento de pesos está directamente asociado al número de predictores en el modelo. Esta afirmación pudo establecerse gracias a los múltiples modelos desarrollados por la función de optimización.

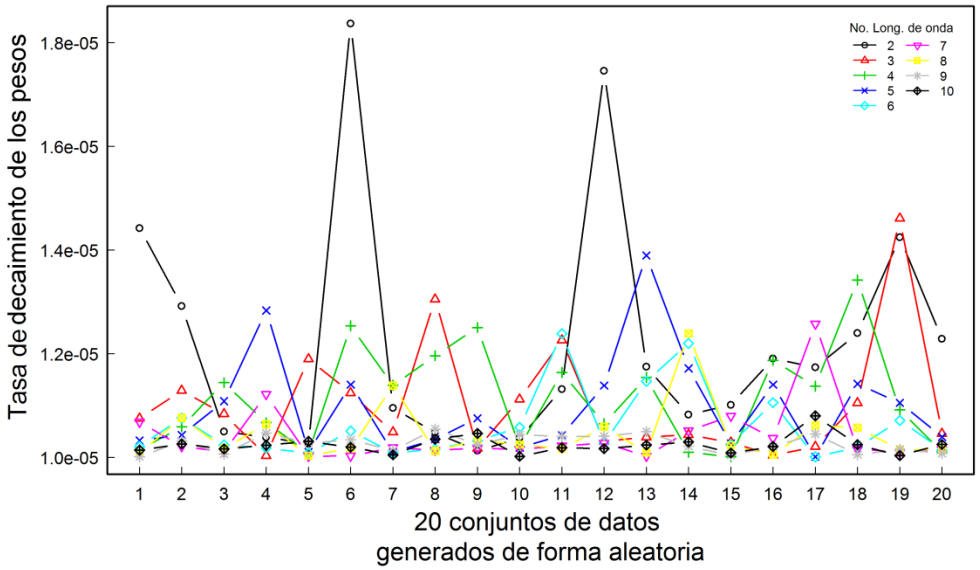


Figura 59- Valores de la tasa de decaimiento de pesos obtenidos en los procesos de optimización y calibración de los modelos RNA con diferentes números de longitudes de onda

Se realizó un ejercicio el cual consistió en seleccionar el modelo que generó el menor en cada etapa (entrenamiento, prueba y validación) como se señala en rojo en la Tabla 11, y conforme al modelo en cada etapa (*e.g.* validación) seleccionar el valor de las métricas para las otras dos etapas (*e.g.* entrenamiento y prueba).

Con este ejercicio se puede confirmar que a mayor número de longitudes de onda menor debe ser la tasa de decaimiento de pesos, y que para el fenómeno estudiado aumentar el número neuronas en la capa oculta no inciden significativamente en la calibración de los modelos. Los conjuntos de pesos generados por diferentes arquitecturas del modelo RNA *feed-forward* no permitieron replicar en los conjuntos de prueba y validación los resultados alcanzados en la etapa de entrenamiento.

Selección del mejor modelo	Entrenamiento-RMSEP (mg/L)	Validación-RMSE (mg/L)	Prueba-RMSE (mg/L)	Tasa decaimiento	No. de neuronas	No. long. de onda	Conjunto de datos
Entrenamiento	0,82	84,33	95,16	1,0237E-05	10	10	4
Validación	1,01	72,71	99,50	1,4248E-05	10	1	19
Prueba	16,97	94,96	70,88	1,0088E-05	10	10	15

Tabla 11- Mejores modelos de las etapas de entrenamiento, prueba y validación seleccionados por los menores errores de predicción

Finalmente, es importante mencionar los tiempos computacionales que tomó la calibración de los modelos conforme al número de longitudes de onda evaluadas:

Equipo 1 [2]= 15 horas
Equipo 2 [3 y 4]= 23 horas
Equipo 3 [5 y 6]= 42 horas
Equipo 4 [7 y 8]= 56 horas

Equipo 5 [9 y 10]= 81 horas

Por lo tanto, el algoritmo desarrollado para la calibración de los modelos de RNA no fue aplicado a los casos de estudio, debido al *overffiting* y a la cantidad de tiempo que tomaría evaluar 1000 modelos por cada determinante de calidad del agua objeto de estudio y un mayor número de longitudes de onda, tal como se estableció en el algoritmo AEEC.

3.5.2. Algoritmo para la calibración de un modelo regresivo *SVM-ν*

El algoritmo de calibración de los modelos regresivos *SVM-ν* (numeral 1.4.7.5) es similar al algoritmo *OPP* (numeral 3.2). La diferencia está en los parámetros que se quiere calibrar. En el caso del modelo *PLS* se pretende encontrar el número de vectores latentes y de longitudes de onda (*nl*), con los cuales se determine el mejor conjunto de betas (o pesos) que generen el menor $RMSE_{VC}$. Para el caso del modelo *SVM-ν* se evalúa la misma función objetivo, pero los parámetros a calibrar son: *nl*, la constante de regularización (C) y fracción de error (ν). La calibración de estos últimos dos parámetros no se realiza únicamente evaluando el error con validación cruzada y modificando el número de *nl*, sino que se emplea el algoritmo de optimización ED (DEoptim).

Por consiguiente, se explicará el procedimiento y los elementos tenidos en cuenta para la calibración de los parámetros del modelo *SVM-ν*:

i. Ver Fase 2 del algoritmo AEEC.
ii. Igual a *OPP* (numeral 3.2).
iii. El proceso iterativo de calibración del modelo *SVM-ν* utilizando la función *ksvm* (ver numeral 2.5) comienza con el conjunto de datos $Y_{C-DR,k}, X_{C-DR,k}(, nl(IF))$, donde la matriz de espectros de absorbancia emplea inicialmente dos longitudes de onda de acuerdo a su *IF*. Este conjunto de datos es empleado para calibrar los parámetros de modelo *SVM-ν* por medio de la función DEoptim. Después de encontrar los mejores valores de los parámetros del modelo, se da inicio al proceso de calibración con validación cruzada tipo *LOO* para determinar en ausencia de cuál muestra se genera el menor valor de $RMSE_{VC}$. En el siguiente ciclo del proceso iterativo *nl* aumenta a 3 y así sucesivamente hasta un máximo igual al número de muestras del conjunto de calibración.
iv. En la función DEoptim se deben especificar ciertos parámetros y la función objetivo. En este orden se establece los siguientes parámetros: límite inferior [0, 0.0001] y superior [100, 1] rango en el cual probablemente se encuentran los mejores valores de C y ν respectivamente, una función regresiva de *SVM-ν* (*ksvm*- ver numeral 2.5) que evaluará cada valor de los parámetros en el proceso de optimización, número de miembros de una población (NP=10), máximo número de iteraciones (generación de poblaciones) (NI=500), y se define la convergencia cuando el porcentaje de mejora entre iteraciones está por debajo de la tolerancia T= $1x10^{-8}$ después de evaluar 25 iteraciones.

Adicionalmente dentro *ksvm* se especifica la función *kernel RBF* (Ecuación 35). Esta última es función del hiperparámetro denominado a ó σ, y cuya calibración se realiza mediante la función *sigest*. Esta función estima un rango de valores del parámetro sigma que permiten generar buenos resultados cuando se utiliza *SVM*. La estimación se basa en el cuantil 0.1 y 0.9 de $|x - x_d|^2$. Básicamente cualquier valor entre esos dos límites producirá buenos resultados. La función objetivo que será optimizada es la suma de los errores cuadrados (*SSE*-Ecuación 54), la cual tendrá por fin evaluar las concentraciones equivalentes obtenidas por el modelo *SVM* versus las concentraciones de laboratorio, y determinar con qué valor de α se minimiza dicha función.

v. Los valores de $RMSE_{VC}$ de cada modelo calibrado en el paso iii se guardan en una vector, y de este último se selecciona el menor valor de la métrica. Con base en este criterio, se abstraen los valores de los parámetros de la arquitectura del modelo *SVM* que generó el menor $RMSE_{VC}$ ($\hat{Y}_k = ksvm(C, \nu, nl)$), y se emplea dicho modelo en la etapa de validación.

vi. Por último, se toma nuevamente el conjunto de datos de calibración y se evalúa su desempeño por medio de las métricas presentadas en el numeral 1.4.8.3 y de la misma forma para el conjunto de validación, y así se determina la eficiencia y robustez del modelo.

4. RESULTADOS

4.1. INCERTIDUMBRE EN LOS VALORES DE CONCENTRACIÓN

Las Figuras 76, 78 y 80 presentan los valores de concentración asociados a la incertidumbre de los valores concentración (triplicado) para cada determinante y a la presión de los aparatos empleados en su medición. En dichas figuras se presenta de arriba hacia abajo los resultados obtenidos de los determinantes SST, DQO y DQOf. En cada gráfica el eje de las ordenadas representa los valores de concentración en mg/L y en las abscisas los índices de cada una de las muestras en cada caso de estudio. Por otra parte, en cada gráfico se presenta en color azul y naranja los límites superior e inferior de la banda de confianza (gris), y en color rojo los valores de la concentración media de cada muestra.

En cuanto a los espectros UV-Vis, se presentan en las Figuras 79, 81 y 83 un espectro típico de las muestras del afluente para cada uno de los casos de estudio. Los valores de absorbancia presentados en el eje de las ordenadas de cada gráfico están asociados a la precisión del aparato de medición (sonda spectro::lyser), y en el eje de las abscisas las longitudes de onda en las cuales el espectrómetro midió la atenuación de la luz (absorbancia).

4.1.1. Incertidumbre en los valores de SST, DQO, DQOf y absorbancias espectro UV-Vis del afluente de la EEG

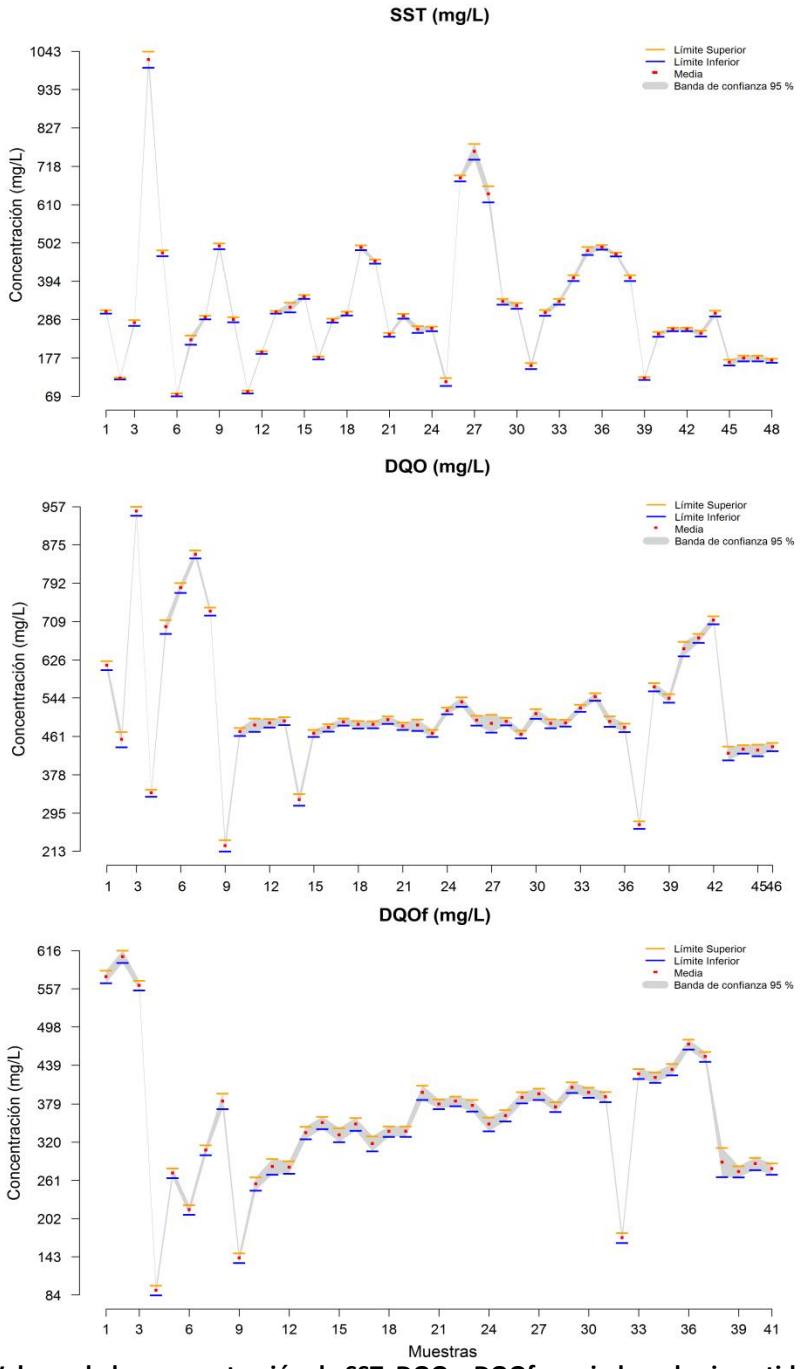


Figura 60- Valores de las concentración de SST, DQO y DQOf asociados a las incertidumbre de las concentraciones y de los instrumentos de medición – Estación Elevadora Gibraltar

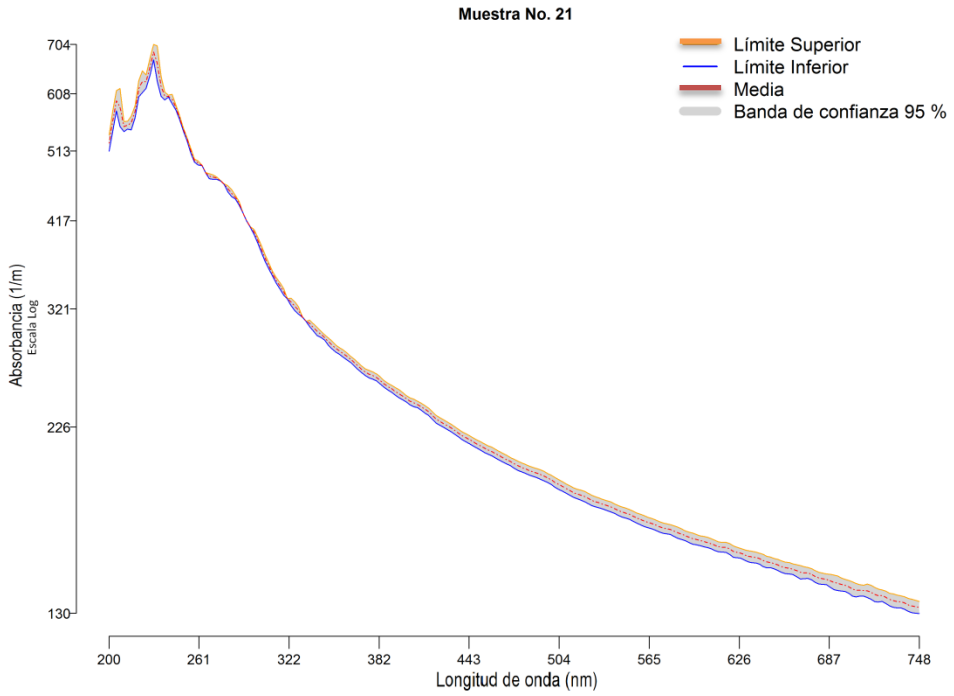


Figura 61- Espectro UV-Visible típico de las muestras del afluente de la EE de Gibraltar con valores de absorbancia asociados a la incertidumbre del aparato de medición

4.1.2. Incertidumbre en los valores de SST, DQO, DQOf y absorbancias espectro UV-Vis del afluente de la PTAR de *Fontaines-sur-Saône*

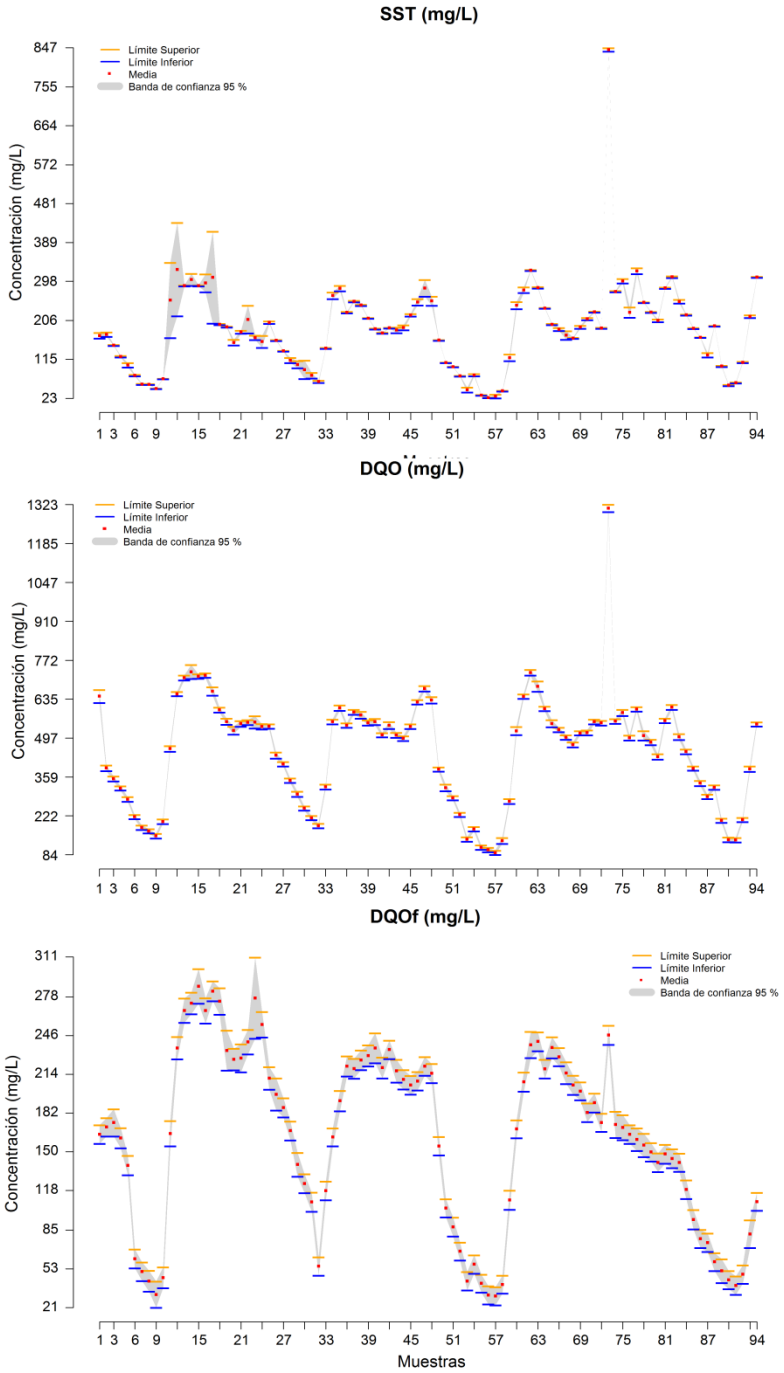


Figura 62- Valores de las concentración de SST, DQO y DQOf asociados a las incertidumbre de las concentraciones y de los instrumentos de medición – PTAR *Fontaines-sur-Saône* (Tiempo seco)

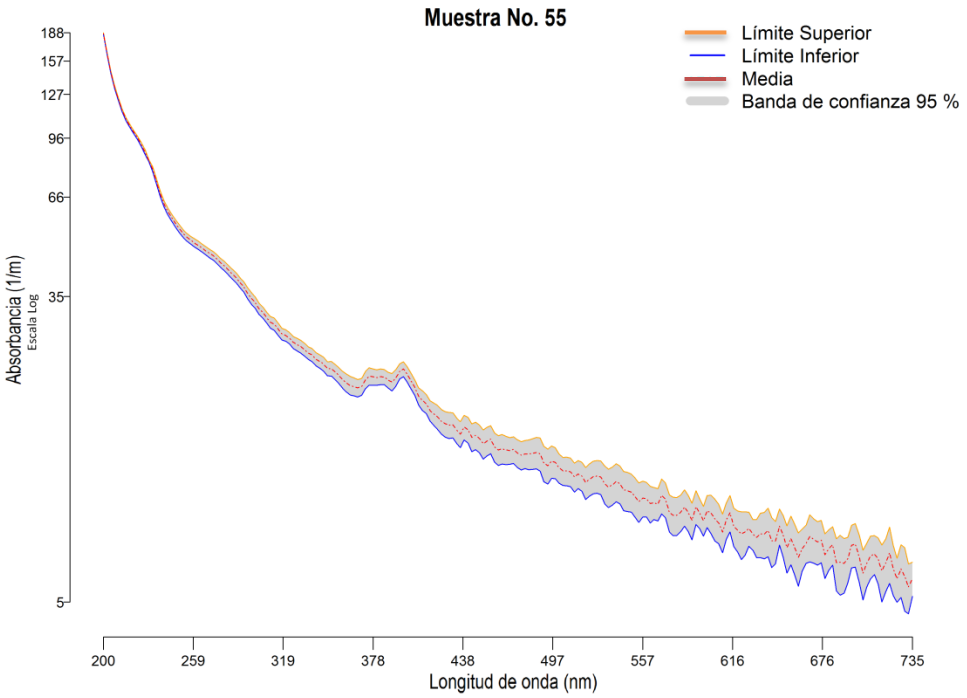


Figura 63- Espectro UV-Visible típico de las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo seco) con valores de absorbancia asociados a la incertidumbre del aparato de medición

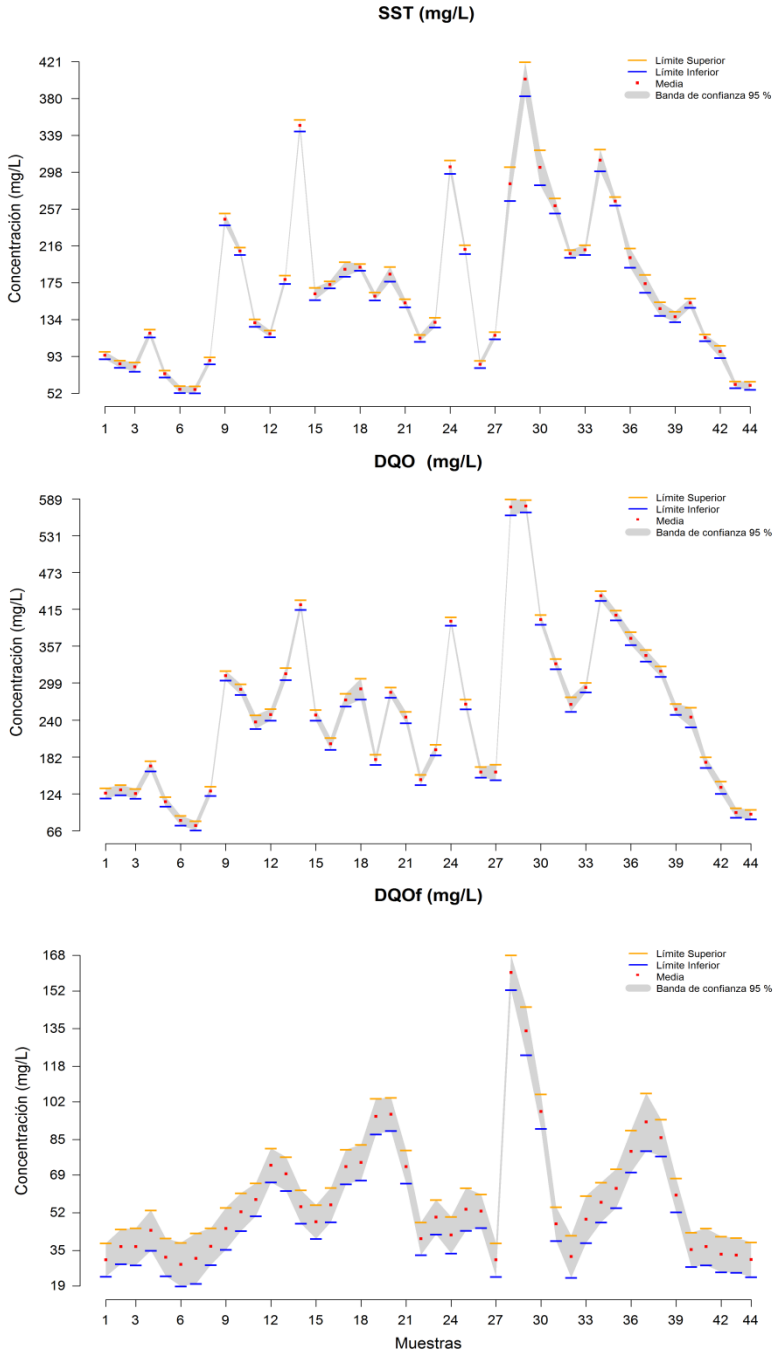


Figura 64- Valores de las concentración de SST, DQO y DQOf asociados a las incertidumbre de las concentraciones y de los instrumentos de medición – PTAR de *Fontaines-sur-Saône* (Tiempo lluvia)

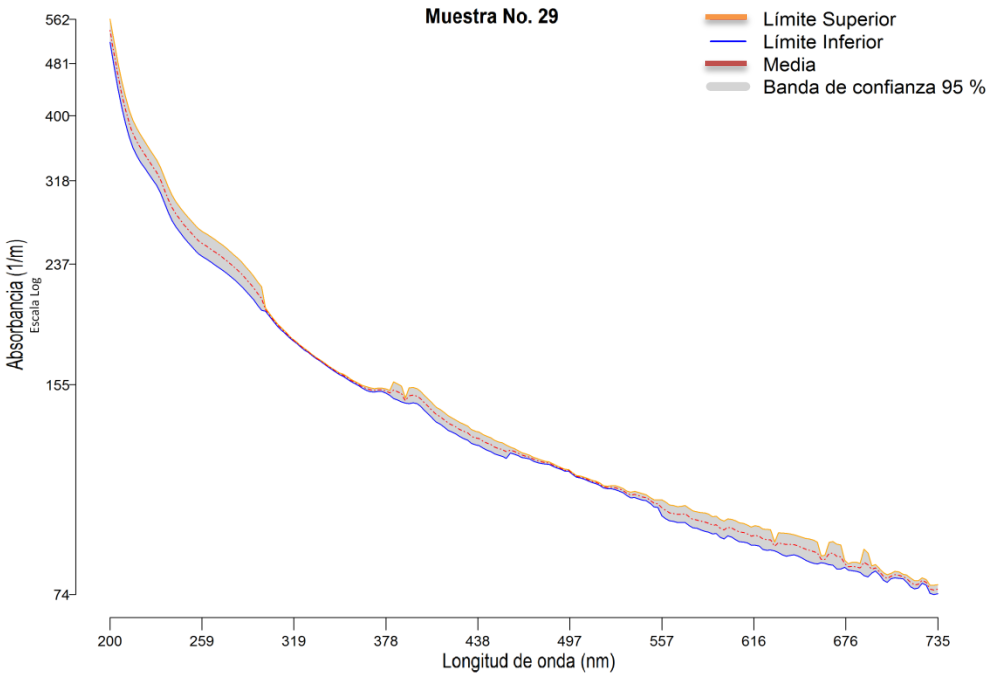


Figura 65- Espectro UV-Visible típico de las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo lluvia) con valores de absorbancia asociados a la incertidumbre del aparato de medición

4.2. ELIMINACIÓN DE *OUTLIERS*

De la aplicación de la metodología descrita en el numeral 3.4.1 se obtuvieron 6 gráficos que resumen el comportamiento de los *outliers* en las 5000 simulaciones de Monte Carlo, generadas para los valores de concentraciones y los espectros de absorbancia para cada de muestra de los casos de estudio. Por lo tanto, a continuación se presentan y describen los gráficos correspondientes a los resultados obtenidos para las concentraciones de los SST del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo seco. En cuanto a los gráficos de los resultados para los demás determinantes y casos de estudio, se pueden consultar en el ANEXO B.

La Tabla 12 muestra cuáles son las figuras similares a las descritas en el numeral 4.2.1 para los demás determinantes y casos de estudio presentadas en el ANEXO B.

Figura típica PTAR de *Fontaines-sur-Saône* (SST-tiempo Seco)	Figuras similares		
	SST	**DQO**	**DQOf**
Figura 66		PTAR de *Fontaines-sur-Saône* (tiempo seco)	
		Figura 97	Figura 102
	PTAR de *Fontaines-sur-Saône* (tiempo lluvia)		
	Figura 107	Figura 112	Figura 117
	EE Gibraltar		
	Figura 122	Figura 127	Figura 132
Figura 67		PTAR de *Fontaines-sur-Saône* (tiempo seco)	
		Figura 98	Figura 103
	PTAR de *Fontaines-sur-Saône* (tiempo lluvia)		
	Figura 108	Figura 113	Figura 118
	EE Gibraltar		
	Figura 123	Figura 128	Figura 133
Figura 68		PTAR de *Fontaines-sur-Saône* (tiempo seco)	
		Figura 99	Figura 104
	PTAR de *Fontaines-sur-Saône* (tiempo lluvia)		
	Figura 109	Figura 114	Figura 119
	EE Gibraltar		
	Figura 124	Figura 129	Figura 134
Figura 69		PTAR de *Fontaines-sur-Saône* (tiempo seco)	
		Figura 100	Figura 105
	PTAR de *Fontaines-sur-Saône* (tiempo lluvia)		
	Figura 110	Figura 115	Figura 120
	EE Gibraltar		
	Figura 125	Figura 130	Figura 135
Figura 70		PTAR de *Fontaines-sur-Saône* (tiempo seco)	
		Figura 101	Figura 106
	PTAR de *Fontaines-sur-Saône* (tiempo lluvia)		
	Figura 111	Figura 116	Figura 121
	EE Gibraltar		
	Figura 126	Figura 131	Figura 136

Tabla 12- Índices de las figuras similares a las descritas en el numeral 4.2.1 para cada caso de estudio en la detección de *outliers*

4.2.1. Muestras *Outliers* del afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo Seco) – caso SST

En la Figura 66 se presenta la frecuencia (ordenadas) que una longitud de onda (abscisas) fue la más correlacionada con la presencia de SST en el espectro UV-Vis. De esta gráfica se puede destacar que siempre las correlaciones entre los valores de absorbancia y concentraciones de los SST estuvieron presentes en el rango visible del espectro y con una mayor recurrencia se presentaron en las longitudes de onda cercanas a 717.5 nm. Sin embargo, no se presentó una longitud de onda que cuya recurrencia fuera significativamente alta en proporción al número de simulaciones generadas. Este

comportamiento se presentó en general para todos los conjuntos de datos de concentraciones de SST y de DQO en los casos de estudio. No obstante, la DQOf presentó una única longitud de onda correlacionada con la presencia de este determinante (*e.g.* Figura 97) pero diferente en los casos de estudio de las muestras en tiempo seco y lluvia de la PTAR de *Fontaines-sur-Saône* (200 nm y 200.5 nm respectivamente). Sin embargo, en el caso de la EE de Gibraltar este determinante presentó diferentes longitudes de onda de las cuales sobresale la longitud de onda a 225 nm como la más correlacionada con una recurrencia de 2750/5000.

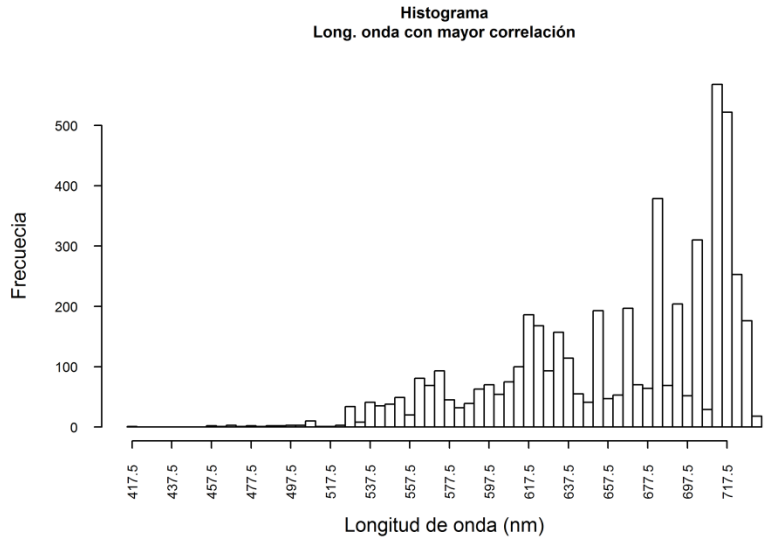


Figura 66- Longitudes de onda con mayor correlación para identificar la presencia de los SST en las 5000 simulaciones de Monte Carlo del afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo seco)

Los histogramas presentados en la Figura 67 representan los valores de absorbancias correlacionados con las concentraciones de SST en las muestras detectadas como *outliers* (izquierda) y las muestras catalogadas como validados-DR (derecha). Al comparar los valores de absorbancia entre ambos gráficos se puede establecer que el método detección no tiende a eliminar las muestras cuya relación con valores de absorbancia sean menos recurrentes, y de allí lo exigente y poco funcional que resulta detectar un *outlier* en relación a una sola longitud de onda, lo cual conlleva a sobreestimar el número de muestras catalogadas como *outliers*. Por lo tanto, la presencia de un determinante puede estar asociada a más de una longitud de onda y a partir de esto reducir los efectos generados por la sensibilidad cruzada a otros compuestos presentes en las muestras, los cuales pueden afectar el valor de absorbancia y con esto a la longitud de onda con mayor correlación a la cual se asocia la presencia de un determinante.

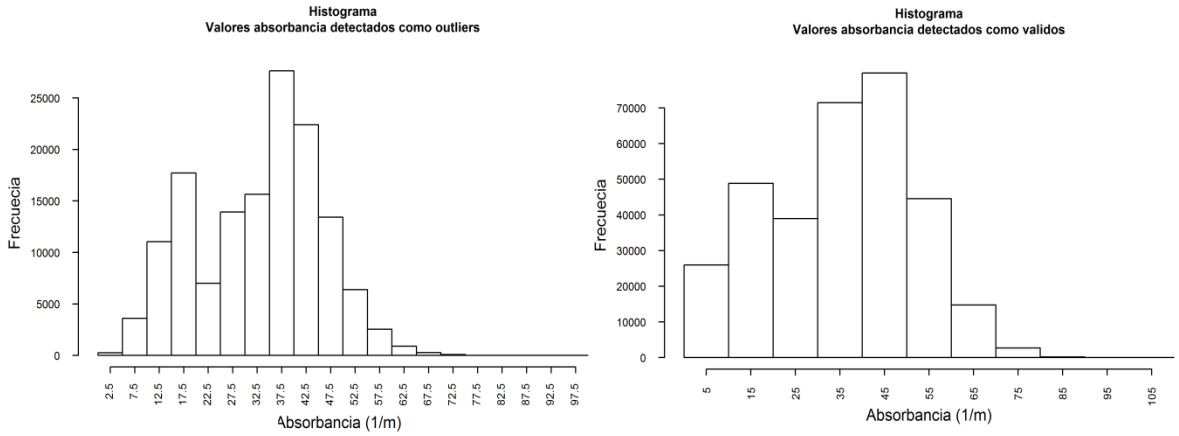


Figura 67- Histograma de los valores de absorbancia detectados como *outliers* (Der.) y de los valores establecidos como válidos (Izq.) en el conjunto de muestras del afluente de la PTAR de *Fontaines-sur-Saône* (SST-Tiempo seco)

En la Figura 68 se presenta en la parte superior un gráfico con los valores máximos y mínimos de las concentraciones evaluadas en primeros 1000 modelos generados en las simulaciones de Monte Carlo de fase 3 del algoritmo AEEC, sobre los cuales se realizó la detección de outliers. En dicho gráfico se presenta en el eje de las ordenas los valores de las concentraciones en mg/L de SST y en las abscisas los índices de las muestras. En la parte inferior de esta figura se presenta en el eje y el número de veces que una muestra fue catalogada como un outlier, señalando en color cian las muestras que 300 o más veces fueron outliers, y una gama de colores grises para aquellas que estuvieron por debajo de dicho umbral. En general los determinantes analizados en los diferentes casos de estudio tiendan a presentar un porcentaje de outliers en promedio entre el 33 % y el 47 %, pero hasta un máximo de 68 % en el caso de la DQOf, el cual es el determinante que para todos los casos de estudio presentó el mayor número de outliers (ver
Tabla 13).

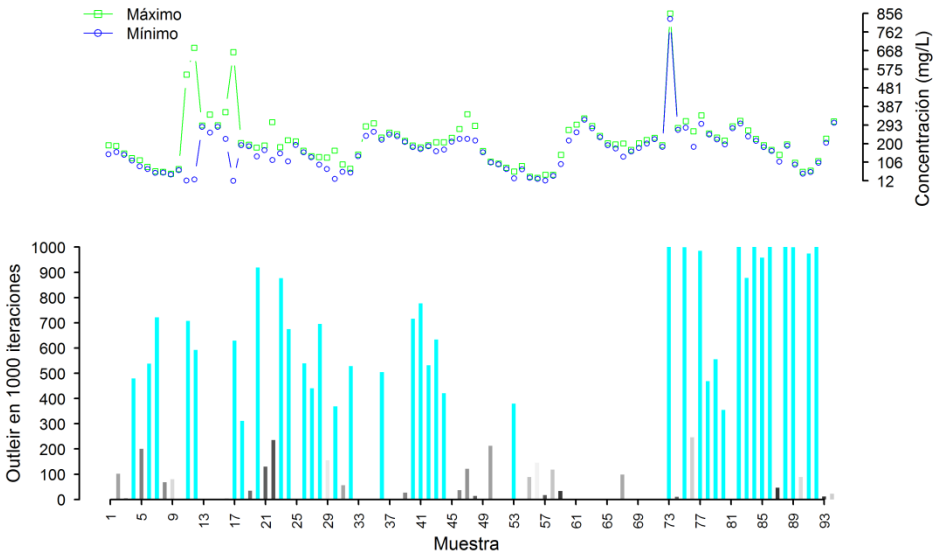


Figura 68- En la gráfica superior se presenta el valor máximo y mínimo de la concentración de los SST de cada muestra generada en 1000 simulaciones de Monte Carlo y en el gráfico inferior se presentan en color cian las muestras que en 300 o más simulaciones de 1000 fueron catalogados como *outliers*

Punto de monitoreo	Estadístico	SST (*outliers*)		DQO (*outliers*)		DQOf (*outliers*)	
		No.	%	No.	%	No.	%
PTAR de Fontaines-sur-Saône (tiempo seco) No. de muestras 94	*Máximo*	41	43,62	48	51,06	64	68,09
	Mínimo	17	18,09	24	25,53	38	40,43
	Promedio	29	30,38	35	37,23	51	54,26
PTAR de Fontaines-sur-Saône (tiempo lluvia) No. de muestras 44	*Máximo*	25	56,82	24	54,55	30	68,18
	Mínimo	9	20,45	10	22,73	10	22,73
	Promedio	17	39,52	17	38,64	20	45,45
EE de Gibraltar No. de muestras 41	*Máximo*	20	48,78	13	31,71	21	51,22
	Mínimo	15	36,59	8	19,51	14	34,15
	Promedio	18	43,71	10	24,39	17	41,46
Global	*Máximo (%)*	56,82		54,55		68,18	
	Mínimo (%)	18,09		19,51		22,73	
	Promedio (%)	37,55		33,93		47,33	

Tabla 13- Resumen general de la cantidad y porcentajes de los *outliers* detectados en los conjuntos de datos de SST, DQO y DQOf para cada caso de estudio

La siguiente la figura presenta en porcentaje (en el eje y) la cantidad de veces que una muestra fue catalogada como *outlier* con respecto a las 5000 simulaciones realizadas. Los números que aparecen en el eje de las abscisas corresponden a las muestras catalogadas como *outliers* en el 60 % o más de las simulaciones generadas (barras color gris).

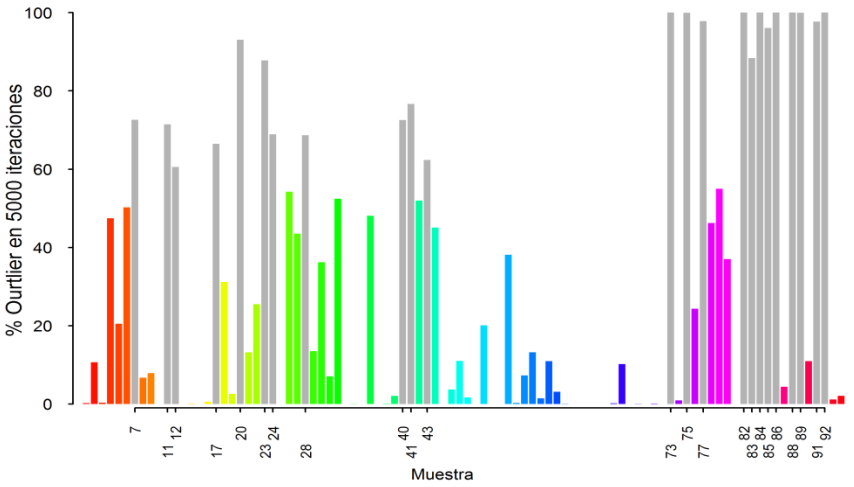


Figura 69- Porcentaje que una muestra fue catalogada como *outlier* en 5000 simulaciones de Monte Carlo en el caso de las concentraciones de SST en el afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo seco). En gris se presenta las muestras catalogas como *outliers* en el 60 % o más de las simulaciones generadas

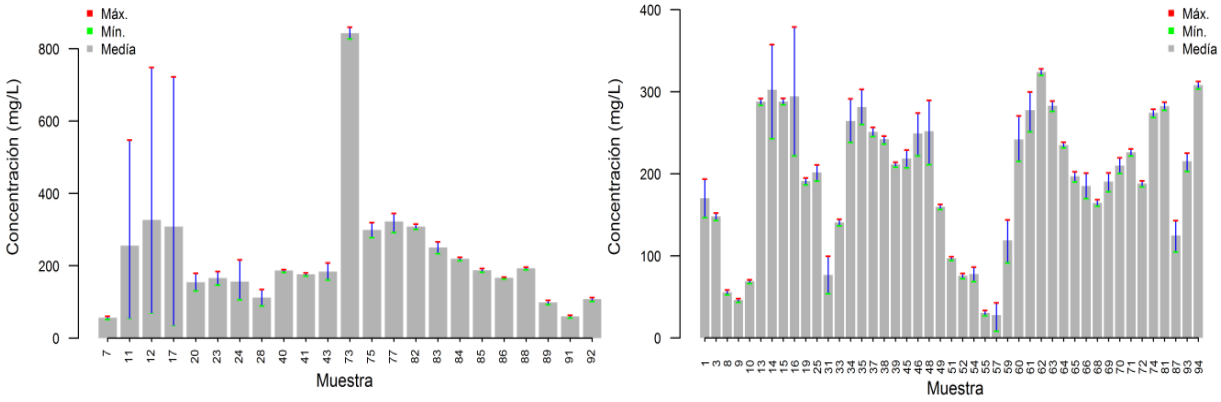


Figura 70- A la izquierda se presentan los rangos de concentración de las muestras catalogadas como *outliers* en más del 10 % de las 5000 simulaciones de Monte Carlo y a la derecha las muestras que estuvieron por debajo de dicho porcentaje y catalogadas en algunas de las simulaciones como datos validos (SST de la PTAR de *Fontaines-sur-Saône* en tiempo seco)

4.3. CALIBRACIÓN Y VALIDACIÓN DE LOS MODELOS REGRESIVOS

Luego de la detección y eliminación de los *outliers* presentes en cada una de las 5000 bases de datos generadas en la fase 2 del algoritmo AEEC, se seleccionaron 1000 para la calibración y validación de los modelos regresivos *PLS* y *SVM-ν*. Esto obedece a los tiempos computacionales que conlleva el entrenamiento de estos modelos en especial *SVM*, lo cual se evidencia en la ANEXO D.

4.3.1. Ajuste y capacidad predictiva de los modelos *PLS* y *SVM*

A partir de la división de las bases de datos (espectros-concentraciones) en conjuntos de calibración y validación realizada en la fase 4 del algoritmo AEEC, se evalúan el ajuste y la capacidad predictiva de los modelos por medio de las siguientes métricas: RMSEP, RMSE, R^2 y ρ (ver numeral 1.4.8).

Por lo tanto, se realizaron gráficos en los cuales se presentan de forma simultánea los resultados obtenidos en las etapas de calibración y validación de los modelos 1000 generados de *PLS* y *SVM* por cada determinante y caso de estudio. A continuación se presentan en la Figura 71 los resultados para los SST de la PTAR de *Fontaines-sur-Saône* en tiempo seco, obtenidos por medio del algoritmo desarrollado para la calibración de los modelos *SVM* (numeral 3.5.2). Los demás determinantes y casos de estudio son presentados en su totalidad en el ANEXO C, el cual se divide en dos: la primera parte muestra los resultados alcanzados con algoritmo *OPP*_modificado (*PLS*), y la segunda los resultados del algoritmo para *SVM*.

En la Figura 71 de izquierda a derecha se muestran los resultados obtenidos de la evaluación de las métricas: coeficiente de Spearman, *RMSE* y coeficiente de determinación para la etapa de calibración y validación respectivamente. Las ordenadas de las gráficas representan el número de cada simulación, y en los ejes de las abscisas el valor de cada métrica.

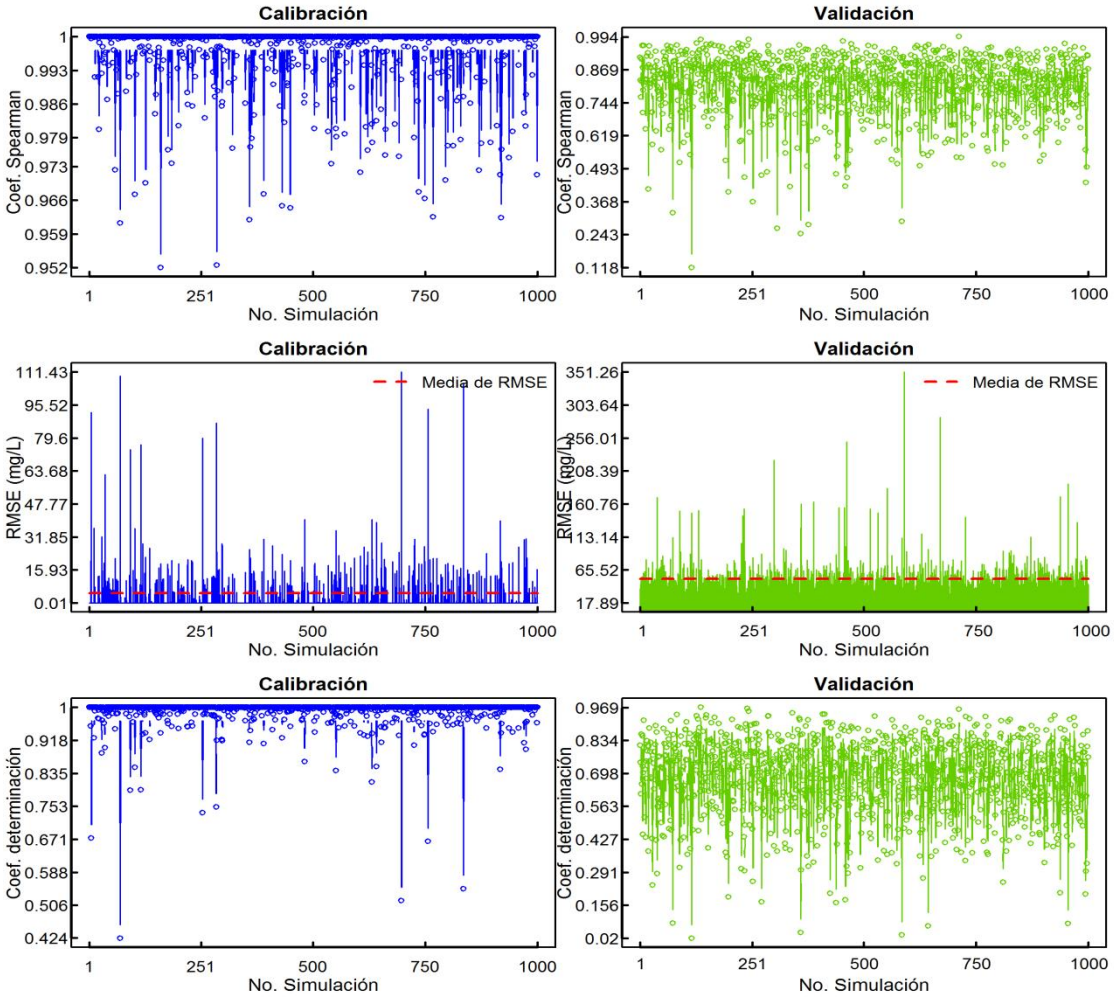


Figura 71- Evaluación del desempeño de los modelos *SVM* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de SST del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo seco

Para determinar los modelos que mejor simularan los valores de las concentraciones de los determinantes objeto de estudio, cada uno de los algoritmos de calibración de *PLS* y *SVM* empleaban la métrica *RMSEP.* Los resultados de la evaluación de dicha métrica son presentados para el caso de los SST del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo seco obtenidos con modelos *SVM*, y los resultados para los demás casos de estudio y determinantes se pueden consultar en el ANEXO C.

La Figura 72 muestra los valores de los errores de predicción en la etapa de calibración de modelo *SVM* cuando se emplea validación cruzada y esta es evaluada por medio del *RMSEP* (ordenadas). Además, en este gráfico se presenta por medio de una línea roja el promedio de los valores *RMSEP* de los 1000 modelos *SVM* calibrados.

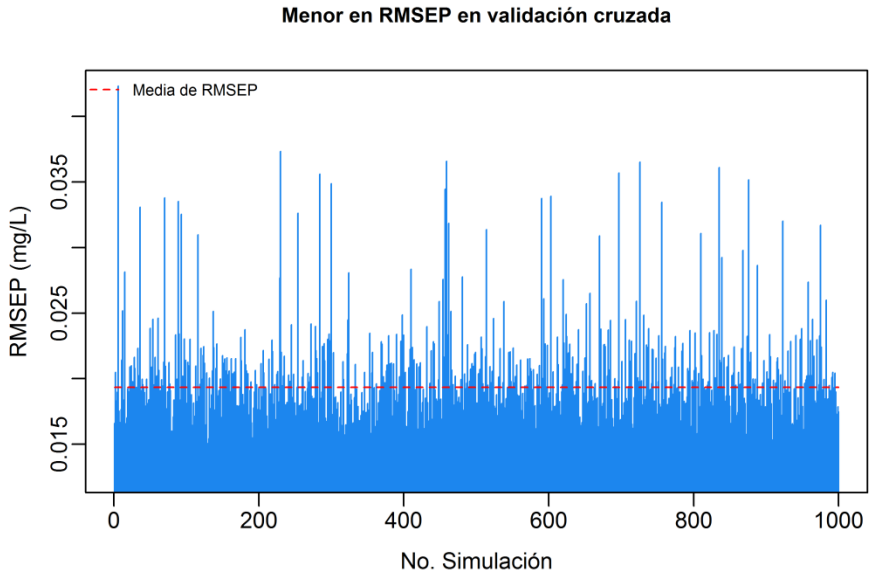


Figura 72- Valores de *RMSEP* de los modelos *SVM* calibrados para la estimación de las concentraciones equivalentes de SST del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo seco

De los 1000 modelos calibrados para la estimación de las concentraciones equivalentes de SST, DQO y DQOf utilizando la información de los espectros de absorbancia, para cada caso de estudio y empleando los algoritmos de entrenamiento (*PLS* y *SVM*), se seleccionó de cada uno de éstos el modelo con el menor *RMSEP*. A partir de esta selección se generaron gráficas de dispersión para poder evidenciar el ajuste tanto en la etapa de calibración como de validación.

Por consiguiente, se presenta a continuación por cada caso de estudio y determinante una comparación entre los gráficos de dispersión obtenidos de los resultados de los mejores modelos *PLS* y *SVM* en la etapa de calibración y validación.

Desde la Figura 73 a la Figura 81 se presenta en la parte superior e inferior los gráficos de dispersión de los mejores modelos *PLS y SVM,* y de izquierda a de derecha los resultados de los mismos en la etapa de calibración y validación respectivamente. En las figuras

indicadas se presenta en el eje de las ordenadas los valores de concentraciones equivalentes en mg/L obtenidas de los modelos *PLS* y *SVM*, y en las abscisas las concentraciones de los ensayos de laboratorio obtenidas en la fase 2 del algoritmo AEEC. Finalmente en la parte superior de cada gráfico se presentan los valores de las siguientes métricas: coeficiente de correlación de Spearman, coeficiente de determinación, *RMSEP* (calibración) y *RMSE* (validación).

Es importante recordar que los conjuntos de datos empleados por mejor modelo *PLS* y *SVM* pueden diferir en el número de muestras y en la magnitud de los valores de las concentraciones de cada determinante y los espectros de absorbancia. Esto se debe a las fases 2, 3 y 4 del algoritmo AEEC.

4.3.1.1. Concentraciones equivalentes de SST, DQO y DQOf del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

Los resultados para los SST obtenidos por el modelo *SVM* en la etapa de calibración y validación fueron mejores que los alcanzados por el modelo *PLS* en ambas etapas, como se puede observar en la Figura 73. Además, se puede evidenciar que muchos más puntos de monitoreo en la etapa de validación del modelo *SVM* se encuentran cerca a la bisectriz y en una amplio rango de valores de concentración, lo cual permite inferir que el modelo *SVM* calibrado para la predicción diferentes fenómenos de contaminación de los SST a través del espectro de absorbancias.

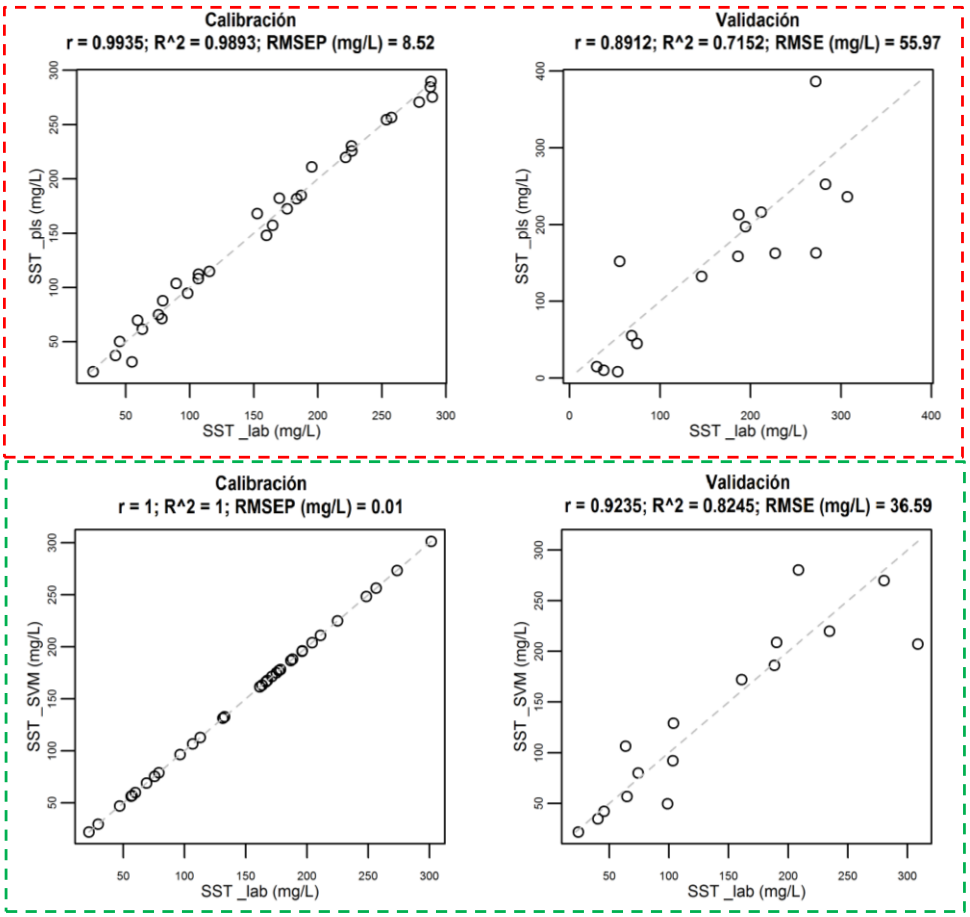


Figura 73- Comparación de las concentraciones equivalentes de SST obtenidas por el mejor model *PLS* (arriba) y *SVM* (abajo) en las etapas de calibración y validación para las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

En cuanto a los resultados para la DQO, el modelo *SVM* estima los valores de concentración de este determinante con un error mucho menor (*RMSEP*=0.03 mg/L) respecto al mejor modelo *PLS* (*RMSEP*=5.84 mg/L) en la etapa de calibración. Sin embargo, en la etapa de validación los resultados no favorecen al modelo *SVM* el cual presenta errores de predicción mayores a los modelos *PLS*. No obstante, para ambos modelos los valores de los coeficientes de Spearman y determinación están por encima de 0.9 en la validación, lo que permite determinar que el modelo *PLS* estimará el 93.4 % de los valores de concentraciones de DQO del afluente de la PTAR en tiempo seco con una alta confiabilidad, y en el caso del modelo *SVM* representar la variabilidad del determinante en un 91.6 %.

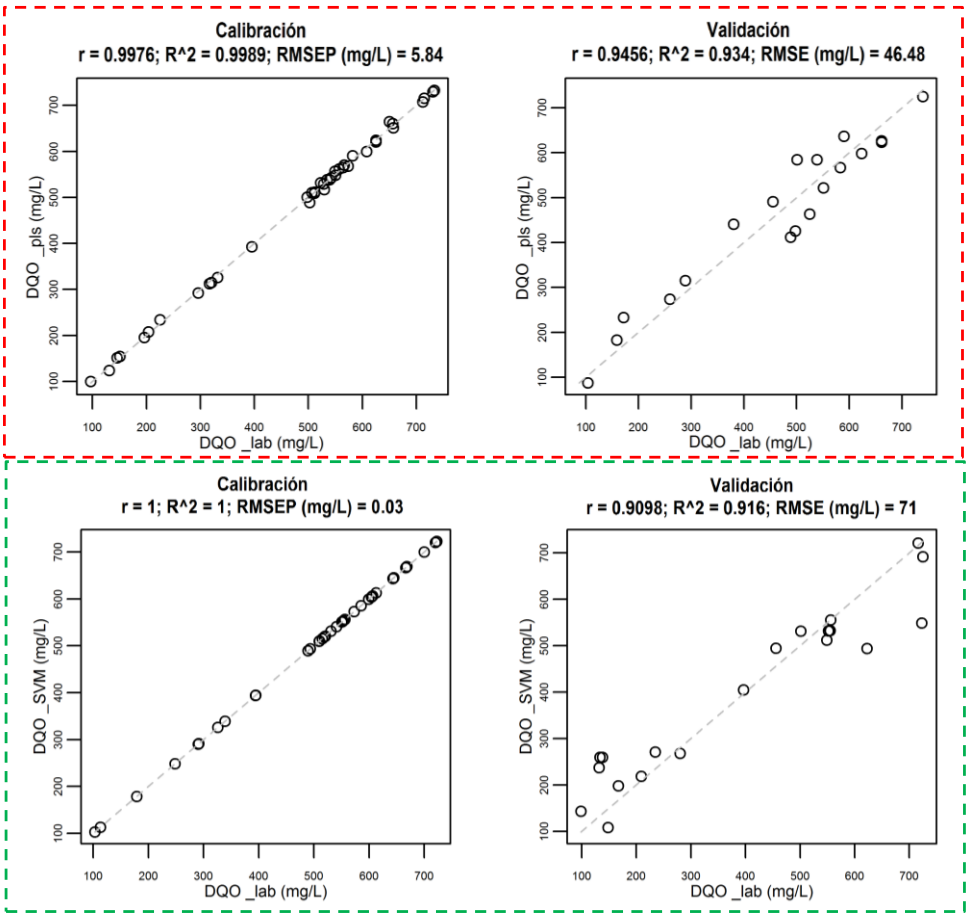


Figura 74- Comparación de las concentraciones equivalentes de DQO obtenidas por el mejor model *PLS* (arriba) y *SVM* (abajo) en las etapas de calibración y validación para las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

Finalmente la DQOf del afluente de la PTAR en tiempo seco fue estimada con un menor error de predicción por el modelo *SVM* en la calibración (*RMSEP*= 0.01 mg/L) y en la validación (*RMSE*= 7.91 mg/L) con respecto al modelo *PLS*. Sin embargo, el modelo *PLS* explica la varianza de las concentraciones de este determinante en un menor porcentaje (0.67 %) en comparación con el 1.67 % del modelo *SVM* en las etapas de validación, tal como se presenta en la Figura 75. Entonces, ambos modelos permiten estimar diferentes rangos de valores de concentración de forma confiable en función de los valores de absorbancia de las longitudes de onda seleccionadas como las mejores variables predictoras.

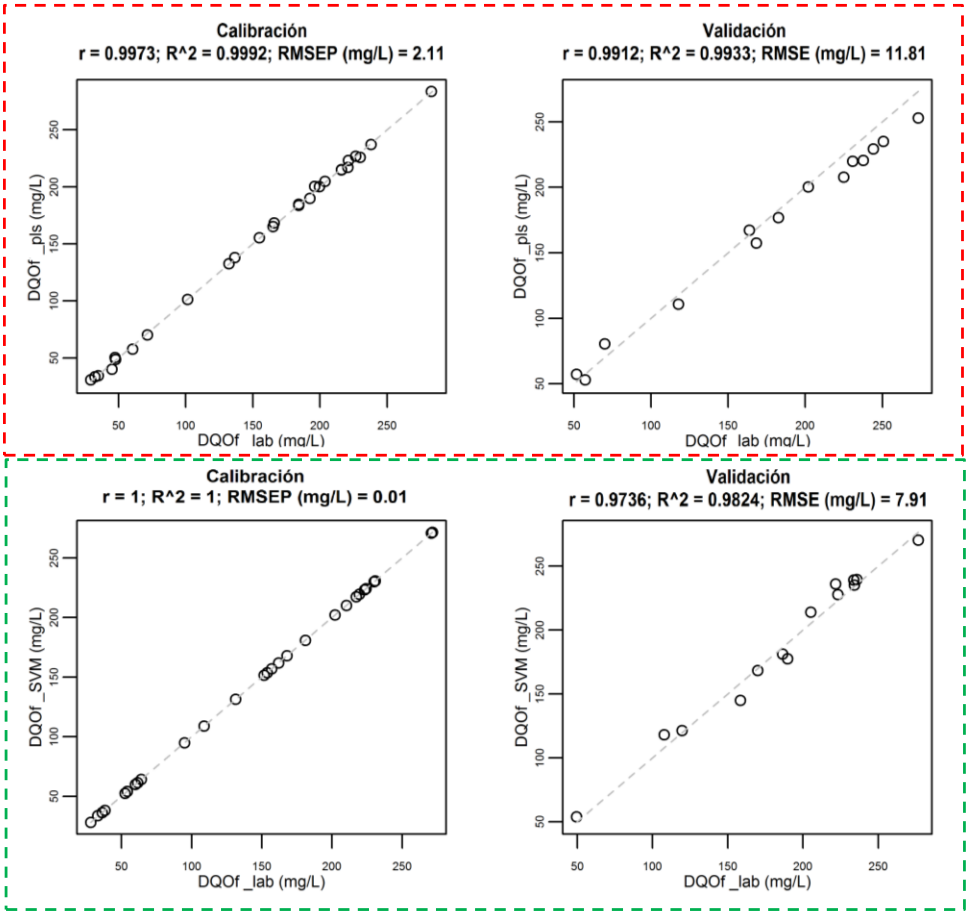


Figura 75- Comparación de las concentraciones equivalentes de DQOf obtenidas por el mejor model *PLS* (arriba) y *SVM* (abajo) en las etapas de calibración y validación para las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

4.3.1.2. Concentraciones equivalentes de SST, DQO y DQOf del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia)

Los valores de las concentraciones equivalentes de SST para las muestras del afluente de la PTAR en tiempo de lluvia, fueron estimadas con un menor error en la etapa de calibración por el modelo *SVM*, pero en la validación el error se incrementó significativamente a 67.54 mg/L, y con esto la capacidad de representar la variabilidad de los SST en tiempo de lluvia por medio de los espectros de absorbancia solo es posible en 61.5 % de los datos del conjunto de validación. Por otra parte, el modelo *PLS* genera resultados satisfactorios en la etapa de calibración, y mucho mejores que el modelo *SVM* en la etapa de validación, en cuya etapa el modelo *PLS* calibrado permite determinar el 95.84 % de la variabilidad del determinante en el afluente del PTAR en tiempo lluvia (ver Figura 76).

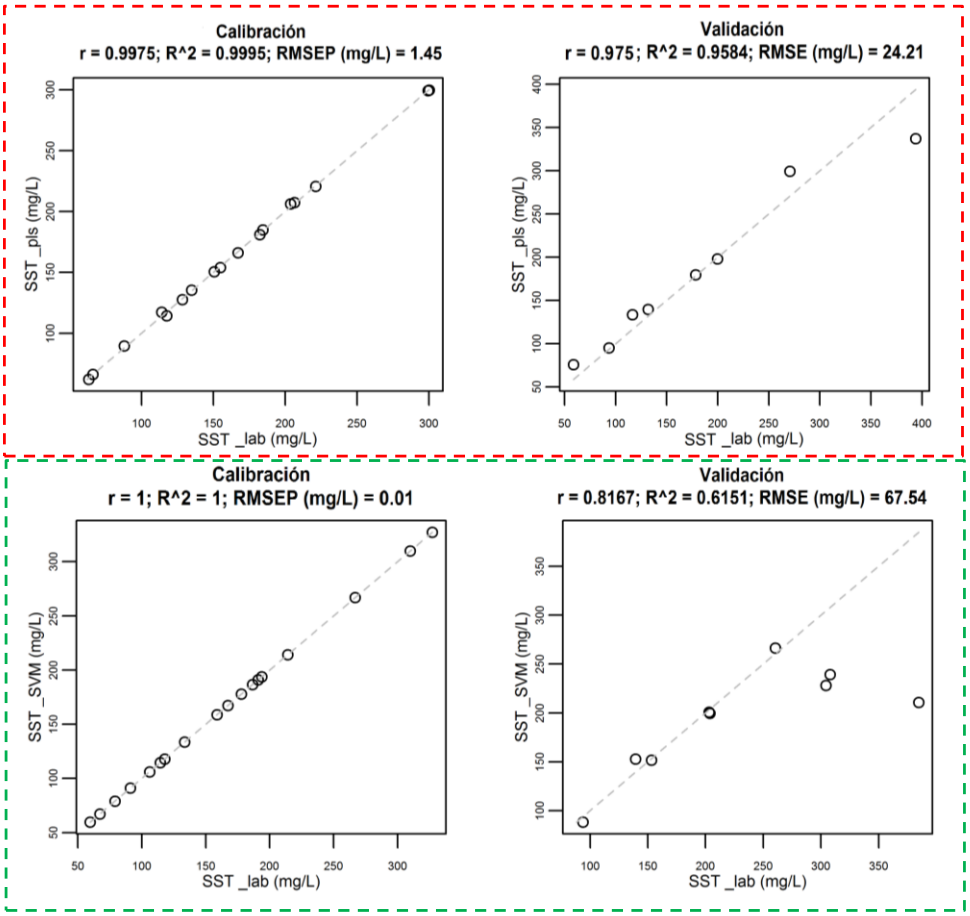


Figura 76- Comparación de los concentraciones equivalentes de SST obtenidas por el mejor model *PLS* (arriba) y *SVM* (abajo) en las etapas de calibración y validación para las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia)

En el caso de la DQO en tiempo de lluvia, ambos modelos presentaron resultados satisfactorios en las concentraciones equivalentes estimadas, lo cual se reflejó en los valores de los coeficientes de Spearman y determinación. Pero en el caso de la métrica *RMSEP* se obtuvieron mejores resultados con el modelo *SVM* (0.01 mg/L) con respecto al modelo *PLS* (3.9 mg/L). Por otro lado, en la etapa de validación, el modelo *PLS* generó un mejor comportamiento en general para todas las métricas evaluadas, pero principalmente el *RMSE* fue menor en 14.22 mg/L con respecto al valor obtenido por el modelo *SVM* en la misma etapa, tal como se puede observar en la Figura 77.

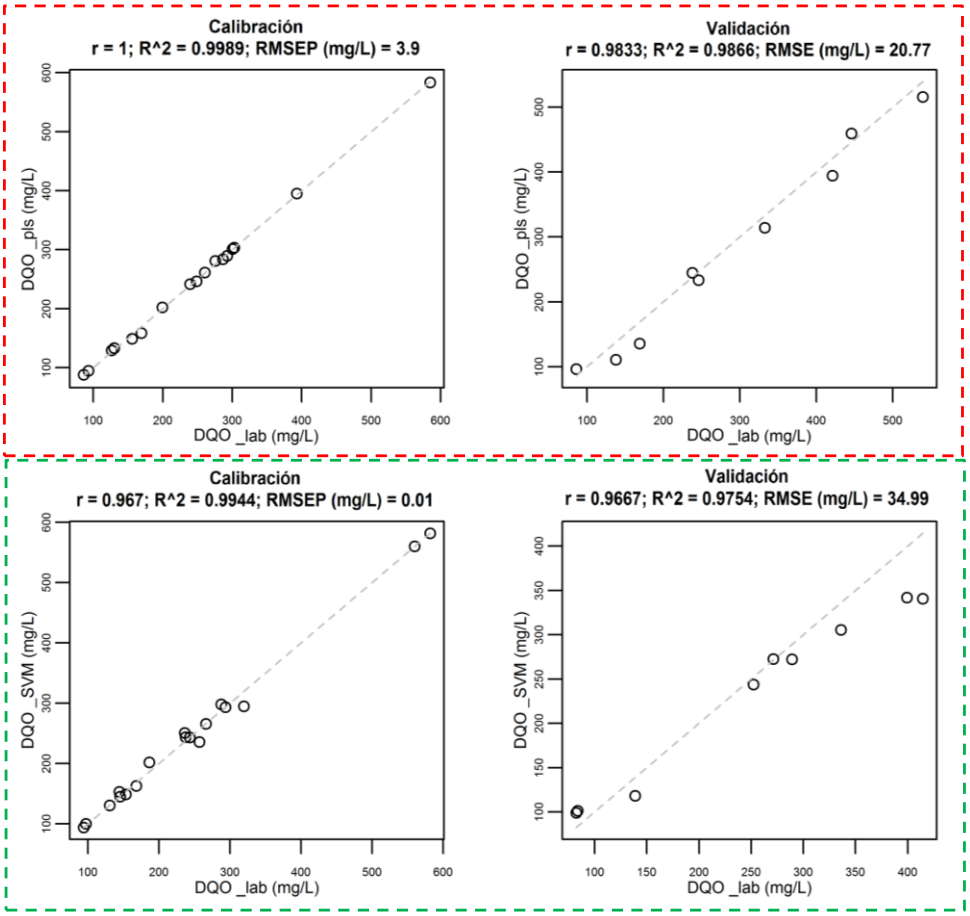


Figura 77- Comparación de las concentraciones equivalentes de DQO obtenidas por el mejor model *PLS* (arriba) y *SVM* (abajo) en las etapas de calibración y validación para las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia)

Por último, la DQOf fue el único determinante de calidad del agua de la muestras recolectas en tiempo de lluvia que por medio del modelo *SVM* no genero resultados satisfactorios en la etapa de validación para estimar concentraciones de este determinante superiores a 80 mg/L, evidenciando *overfitting* en la etapa calibración, y con esto suponer limitaciones del modelo en la estimación de ciertos determinantes asociado a que los parámetros C y ν, y σ de la función kernel no representan variabilidad del determinante, debido a que en el proceso de optimización los valores encontrados obedecen a un óptimo local y no global generado porque las variables independientes seleccionadas (longitudes de onda) no presentan un grado de dependencia entre ellas y ante esto los parámetros calibrados y aplicados a nuevos conjuntos de datos maximizaran el error de estimación. No obstante, el modelo *PLS* genera mejores resultados para las concentraciones superiores al valor mencionado en la etapa de validación, incluso permite determinar el 97.01 % de la variabilidad de las concentraciones a partir de los espectros de absorbancia (ver Figura 78).

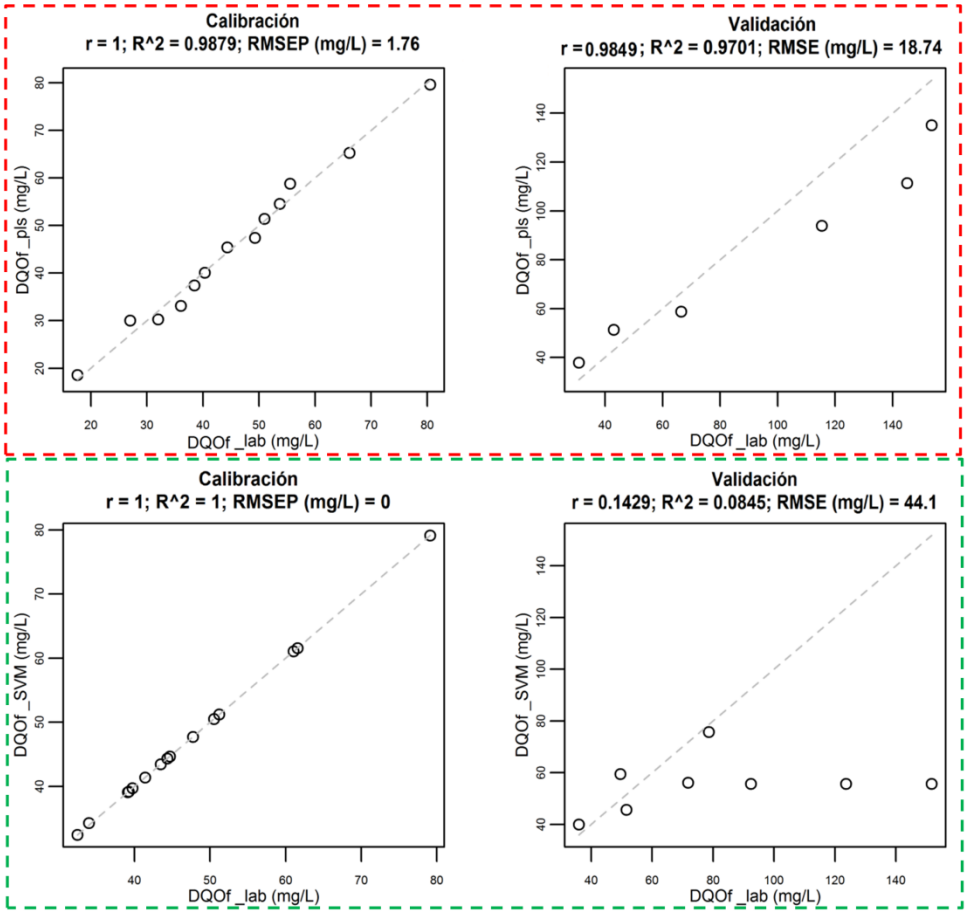


Figura 78- Comparación de las concentraciones equivalentes de DQOf obtenidas por el mejor model *PLS* (arriba) y *SVM* (abajo) en las etapas de calibración y validación para las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia)

4.3.1.3. Concentraciones equivalentes de SST, DQO y DQOf del afluente de la EE de Gibraltar

Los resultados presentados en la Figura 79 permiten determinar que el modelo *PLS* presentó mejores resultados en la estimación de las concentraciones equivalentes de SST de las muestras del afluente de la estación elevadora de Gibraltar que el modelo *SVM* en la etapa de calibración. Sin embargo, ambos modelos generaron resultados pobres en la etapa de validación, como lo evidencian los altos valores de la métrica *RMSE*, y en los valores próximos a cero del coeficiente de determinación, lo cual indica que los predictores no representan el fenómeno y la media de los datos es mejor predictor que los modelos. Una de las de principales causas que podría explicar la baja capacidad predictiva y nivel de ajuste alcanzado por ambos modelos es el paso de luz (5 mm) de la sonda spectro::lyser empleada en la medición del espectro de absorbancias, ya que es muy probable que la atenuación del haz luz resultante en una mayor trayectoria, generada en parte por la moléculas grandes en dispersión, sea de tal magnitud que las absorbancias asociadas a esta no reflejen realmente la presencia de los determinantes. Otra razón y la más probable como mencionan Winkler *et al*. (2008) es la calidad de los resultados de las concentraciones obtenidas en laboratorio, debido a problemas experimentales.

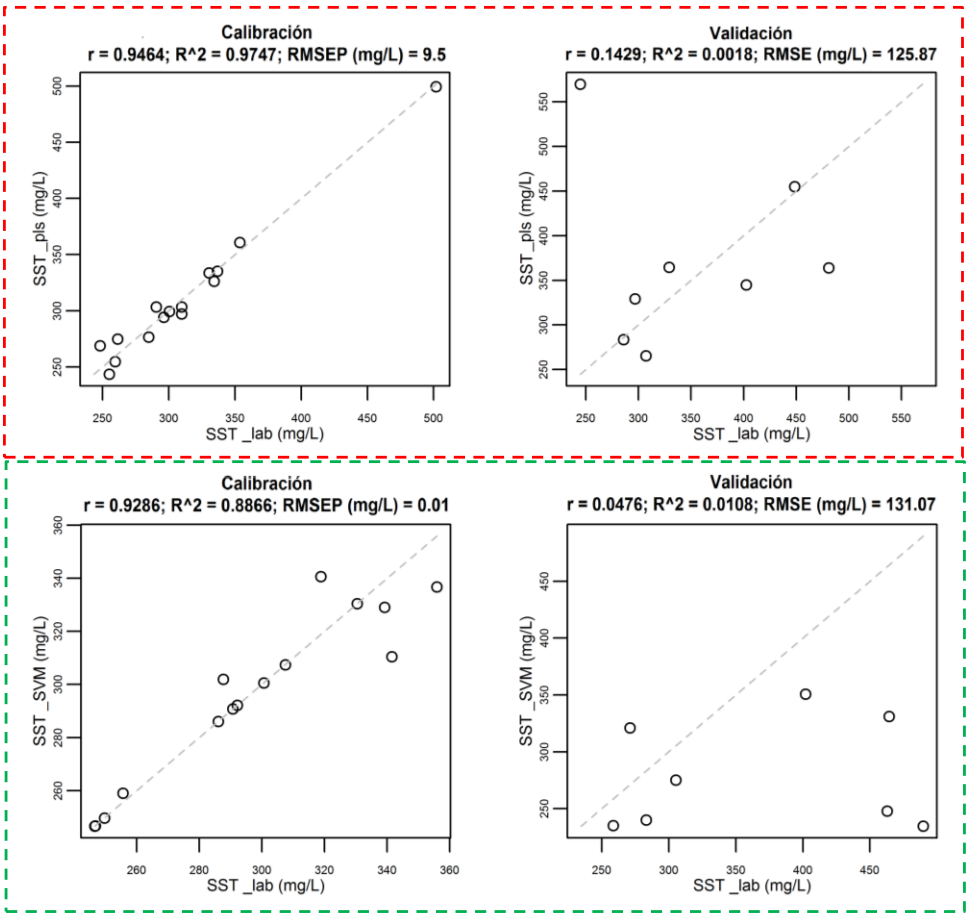


Figura 79- Comparación de las concentraciones equivalentes de SST obtenidas por el mejor model *PLS* (arriba) y *SVM* (abajo)en las etapas de calibración y validación para las muestras del afluente de la EE de Gibraltar

Al igual que los SST, las concentraciones equivalentes de la DQO (Figura 80) estimadas por el mejor modelo *PLS* fueron satisfactorias en la etapa de calibración, pero en la validación se presentó una gran dispersión, como se puede observar en la Figura 80. Además, el coeficiente de determinación cuantificado para dicha dispersión presentó valores cercano a cero, por ende la capacidad del modelo de representar la variabilidad del determinante es nula. En cuanto al modelo SVM, éste presentó en general los mejores resultados para el determinante DQO en la etapa de calibración. Sin embargo, el alcance de estos resultados claramente obedece a un *overfitting*, ya que en la etapa de validación las concentraciones estimadas permanecen bajo la bisectriz y sus valores son muy similares entre sí, pero difieren significativamente de los datos de laboratorio.

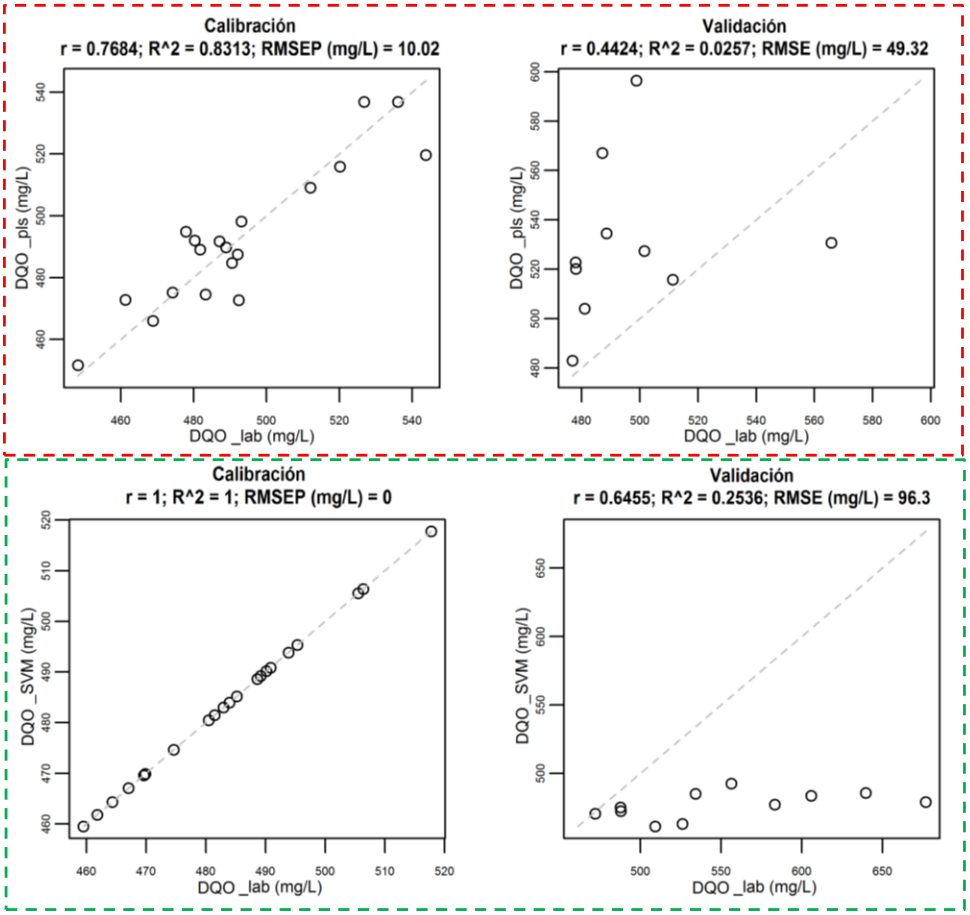


Figura 80- Comparación de las concentraciones equivalentes de DQO obtenidas por el mejor model *PLS* (arriba) y *SVM* (abajo) en las etapas de calibración y validación para las muestras del afluente de la EE de Gibraltar

De los determinantes analizados del afluente de la EE de Gibraltar es la DQOf, el único que no presenta valores cercanos a cero en el R^2 de la etapa de validación del modelo *PLS,* y explica la variabilidad de la concentraciones en el 50.9 % de los puntos de la etapa analizada. En cuanto al modelo SVM, éste presentó en general los mejores resultados para el determinante DQOf en la etapa de calibración. Sin embargo, el alcance de estos resultados claramente obedece a un *overfitting*, ya que en la etapa de validación las concentraciones estimadas permanecen por encima de la bisectriz y sus valores son muy similares entre sí, pero difieren significativamente de los datos de laboratorio (ver Figura 81).

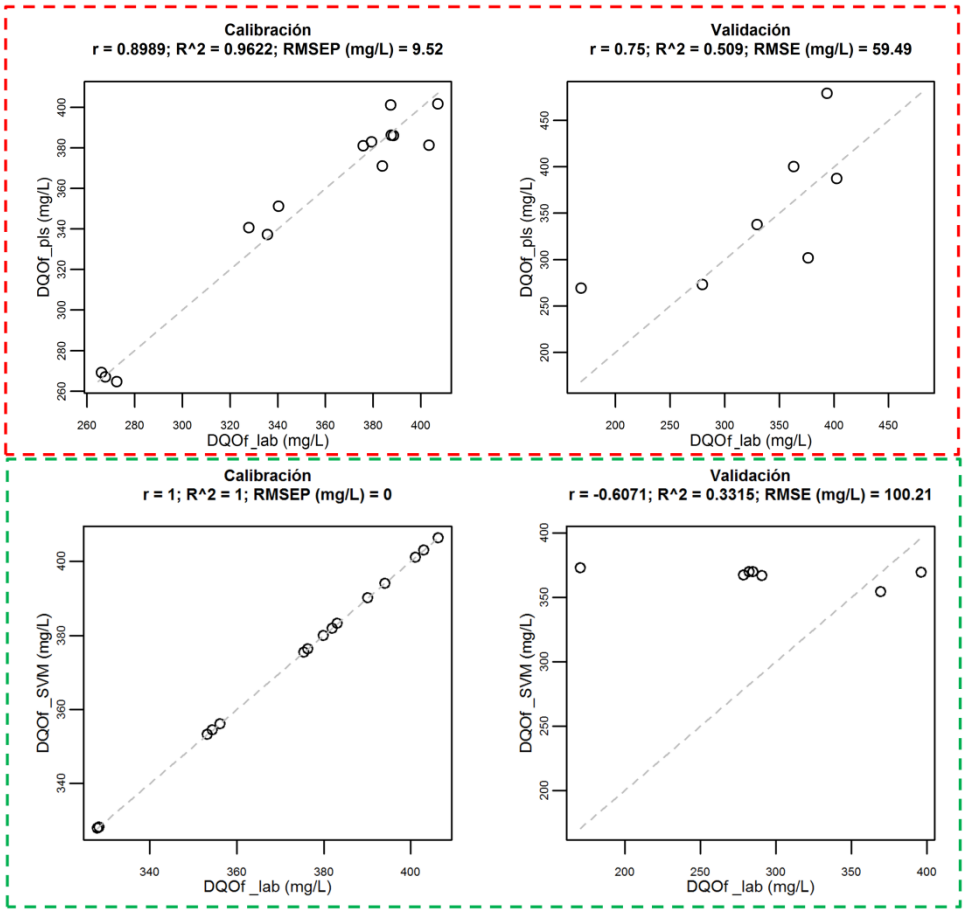


Figura 81- Comparación de las concentraciones equivalentes de DQOf obtenidas por el mejor model *PLS* (arriba) y *SVM* (abajo) en las etapas de calibración y validación para las muestras del afluente de la EE de Gibraltar

4.3.2. Arquitectura de los modelos calibrados *PLS* y SVM

Con el fin de explicar el comportamiento de las variables predictoras usadas por cada modelo, y asociar los cambios en el espectro de absorbancia a la presencia y concentración de un determinante, se realizó un análisis de la frecuencia de las longitudes de onda y valores de absorbancia del espectro UV-Vis empleadas por cada uno de los 1000 modelos de *PLS* y *SVM.*

Por lo tanto, desde la Figura 82 a la Figura 90 se presentan principalmente dos tipos de gráficos por cada modelo (*PLS*-recuadro rojo y *SVM*-recuadro verde). Los gráficos 1 y 3 de arriba hacia abajo en cada figura presentan la frecuencia (eje *y*) con la cual una longitud de onda (eje *x*) fue seleccionada para calibrar un modelo *PLS* o *SVM* respectivamente. Además, en estos gráficos se puede observar (barras en color rojo) las longitudes de onda empleadas por el mejor modelo *PLS* o *SVM* para obtener el menor *RMSEP* en la etapa de calibración. En cuanto a los gráficos 2 y 4 en mismo orden, se presentan los valores de absorbancia (eje *y*) máximo, mínimo y promedio de las longitudes de onda mostradas por los gráficos 1 y 3 para cada modelo, y señalando de estas últimas las seleccionadas por el mejor modelo *PLS* y *SVM*.

A continuación se realiza el análisis de dichos gráficos comparando el comportamiento entre puntos de monitoreo y modelos por cada determinante.

4.3.2.1. Parsimonia de los modelos calibrados para estimar las concentraciones de los SST

En primera instancia el rasgo más característico entre las Figura 82 a la Figura 84 es la diferencia significativa en la amplia gama de longitudes de onda con las cuales fueron calibrados los modelos *SVM*, con respecto a las empleadas por los modelos *PLS*. Incluso, se evidencia que los mejores modelos *SVM* son menos parsimoniosos en cada caso de estudio, por ejemplo en caso de la PTAR en tiempo seco el modelo *PLS* describe la variabilidad del determinante en función de dos longitudes de onda en comparación con 30 usadas por el modelo *SVM*; no obstante mejores resultados fueron obtenidos con este último modelo (Figura 73).

Por otra parte, las mayores frecuencias en ambos modelos y entre las muestras en tiempo seco y lluvia del afluente de la PTAR se presentaron en el rango Visible. Sin embargo, algunos modelos *SVM* emplearon longitudes de onda en el rango UV con una mayor frecuencia respecto a los modelos *PLS* que presentaron este fenómeno. Por ende, este comportamiento permite evidenciar que la relación espectros-concentraciones de las muestras de la EE de Gibraltar (Figura 84) no reflejan realmente la interacción entre el haz de luz y las partículas en suspensión, e incluso presentan mayores frecuencias en rango UV donde los determinantes que normalmente se identifican en este rango están disueltos (ver Figura 3).

Se evidenció una mayor variabilidad en la magnitud de los valores de absorbancia empleados para la calibración de los modelos *PLS* en el rango UV de los espectros de la muestras de la PTAR en tiempo seco y lluvia, con respecto a los usados por los modelos *SVM*. Además, se determinó que las concentraciones de los SST en las muestras en tiempo de lluvia de la PTAR presentan una recurrencia de la relación absorbancia-concentración con una mayor frecuencia en el rango visible de los espectros, debido a que la dilución del determinante por efecto de los incrementos en el caudal del afluente reducen su

concentración y por lo tanto la interacción entre el haz de luz y el material en suspensión es menor y con estos las pérdidas por dispersión disminuyen.

Contrario a lo presentado en el párrafo anterior, los valores de absorbancia medidos de las muestras del afluente de la EE usados para la calibración modelos *PLS* y *SVM*, presentaron una mayor variabilidad en su magnitud y recurrencia en su relación con el determinante en el rango UV. Incluso la diferencia en rango en el cual el mejor model *PLS* y *SVM* seleccionan las longitudes de onda que representarían la variabilidad del fenómeno son diferentes, tal como se puede observar en la Figura 84, pero en ambos casos se obtienen valores de concentraciones equivalentes de SST con errores significativamente altos y coeficientes de determinación que revelan que las variables predictoras no simulan el comportamiento del determinante (ver Figura 79).

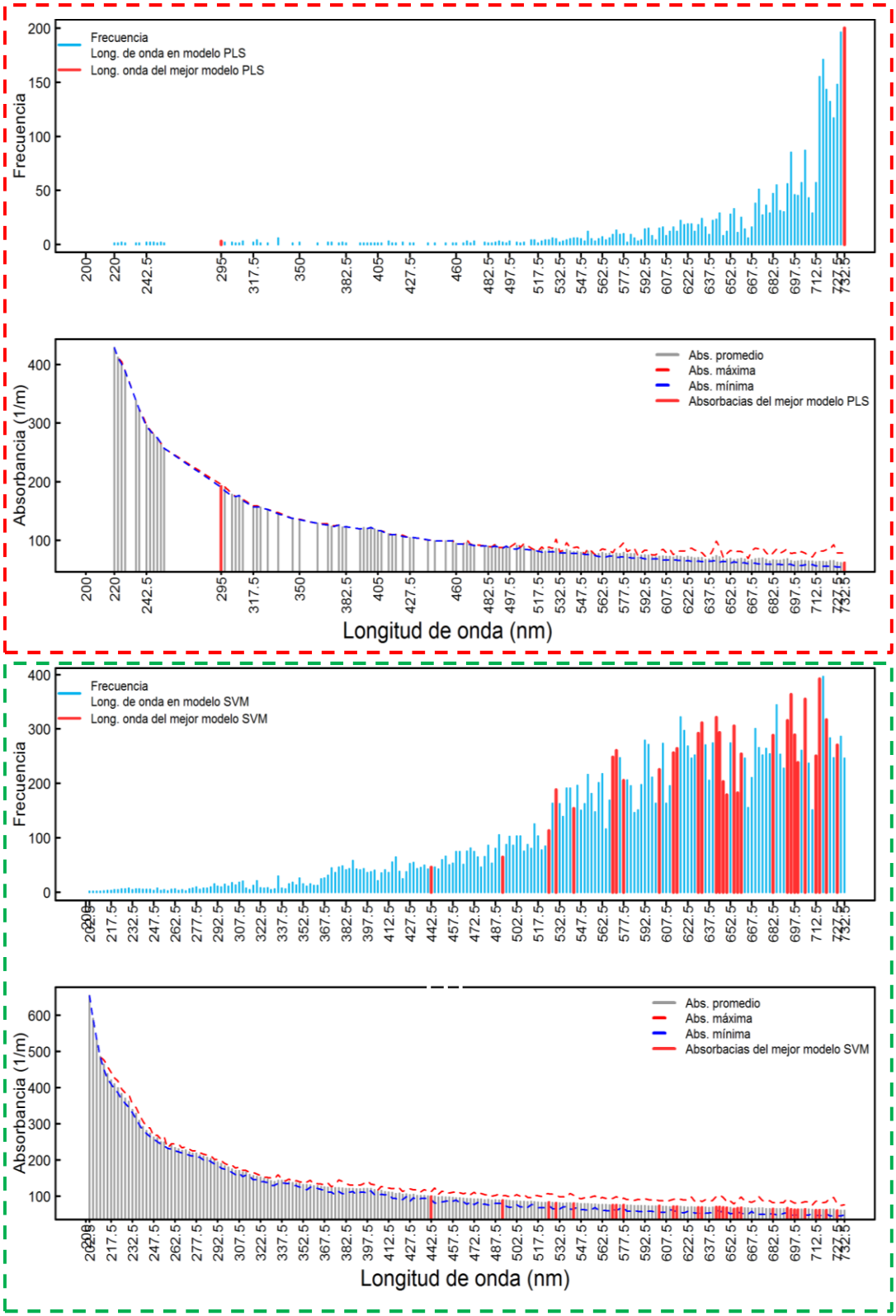


Figura 82- Parsimonia de los modelos *PLS* (recuadro rojo) y *SVM* (recuadro verde): frecuencia de las longitudes de onda y sus valores de absorbancia utilizadas en la calibración de los 1000 modelos del determinante SST del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

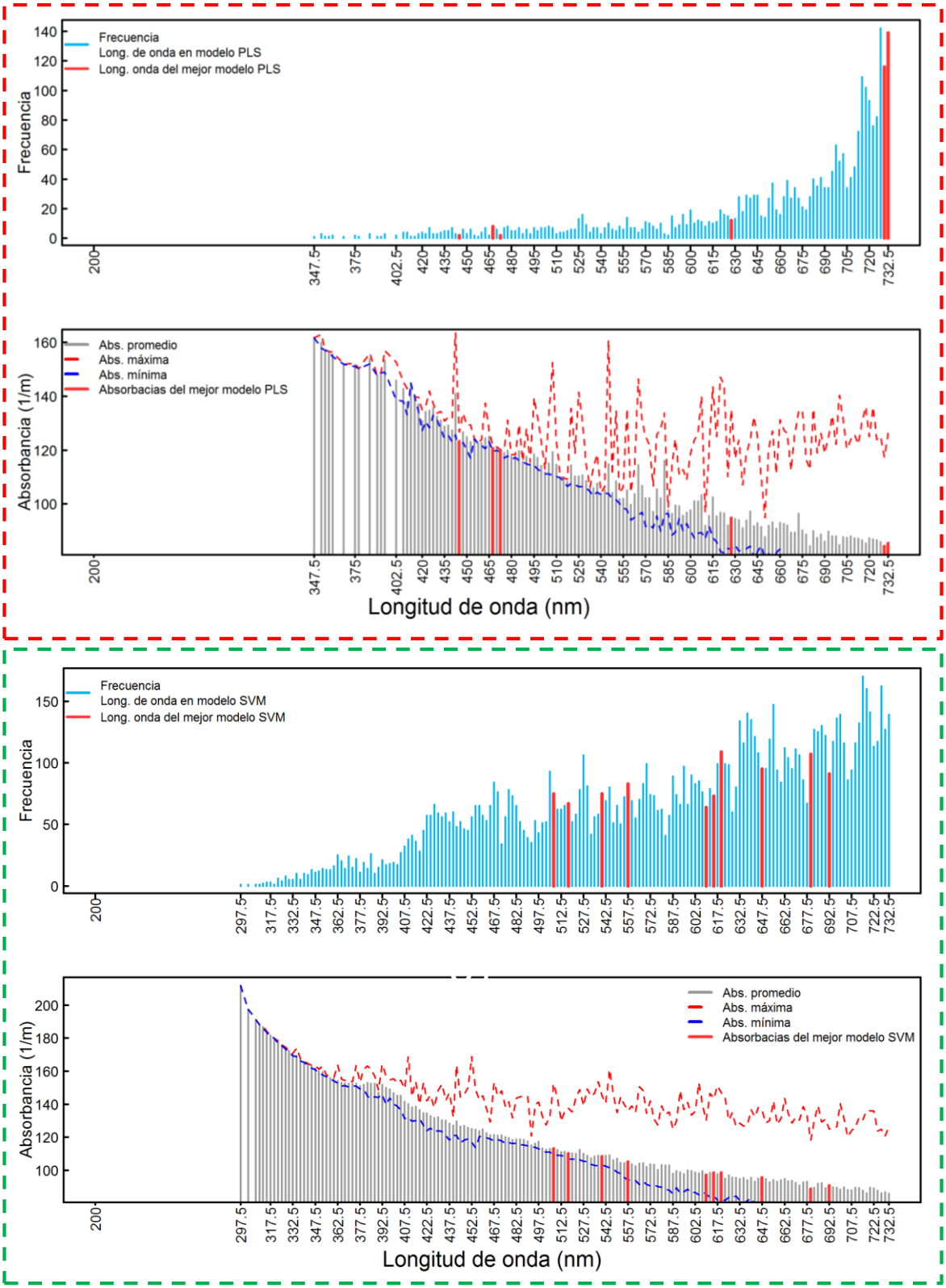


Figura 83- Parsimonia de los modelos *PLS* (recuadro rojo) y *SVM* (recuadro verde): frecuencia de las longitudes de onda y sus valores de absorbancia utilizadas en la calibración de los 1000 modelos del determinante SST del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia)

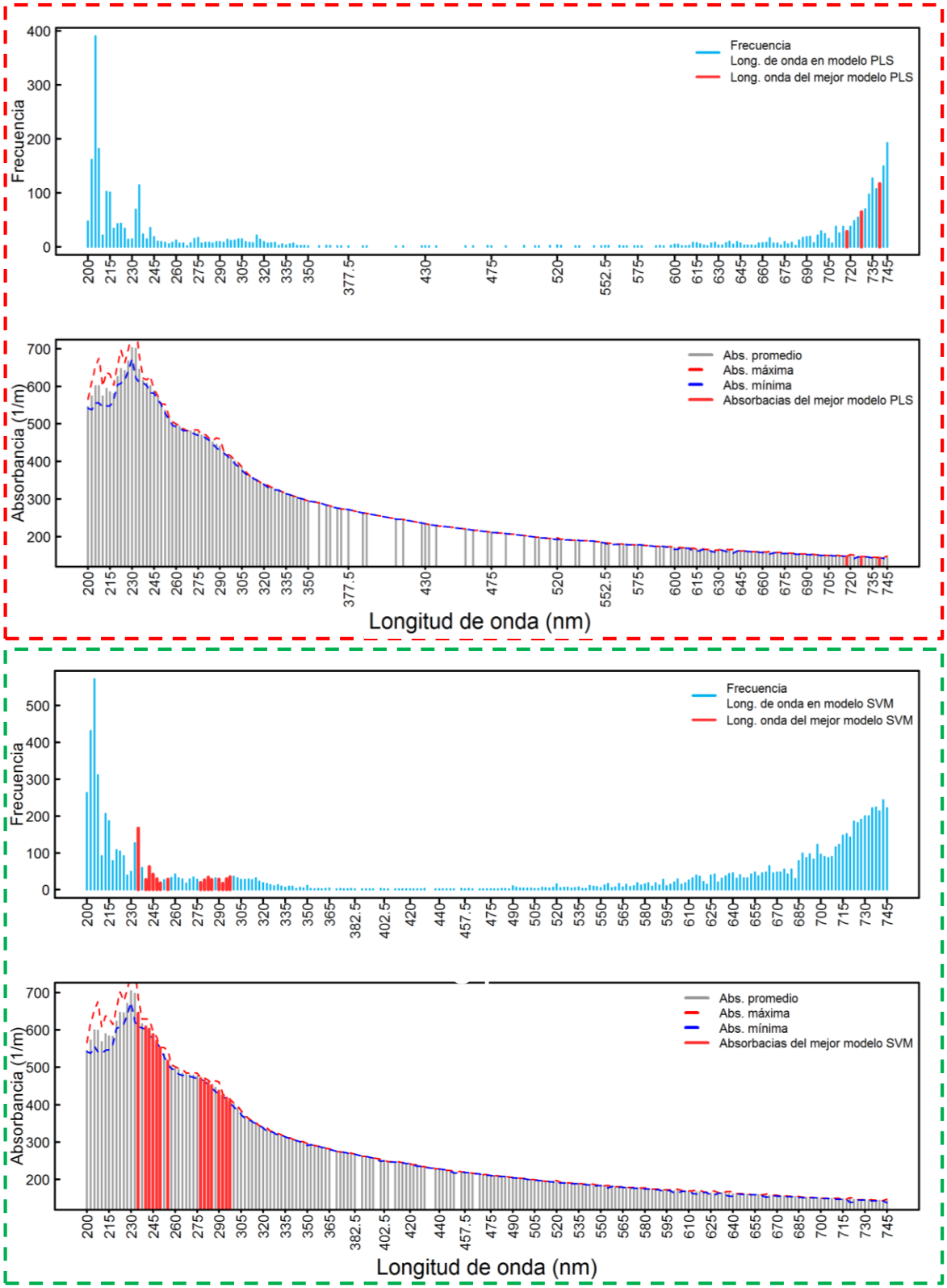


Figura 84- Parsimonia de los modelos *PLS* (recuadro rojo) y *SVM* (recuadro verde): frecuencia de las longitudes de onda y sus valores de absorbancia utilizadas en la calibración de los 1000 modelos del determinante SST del afluente de la EE de Gibraltar

4.3.2.2. Parsimonia de los modelos calibrados para estimar las concentraciones de la DQO

El número de longitudes de onda que explican la variabilidad de la concentración de la DQO continúa siendo mayor con el modelo *SVM* en general para todos los casos de estudio. No obstante, por medio de los modelos *SVM* y *PLS* se estableció que la presencia del determinante está asociada tanto al rango UV como Visible en el caso de las muestras del afluente de la PTAR en tiempo seco (Figura 85), y por ende una forma de encontrar menores errores en las estimaciones en un modelo regresivo estará ligada al uso de las principales longitudes de onda en ambos rangos del espectro para este caso. Por otra parte, las longitudes de onda seleccionadas por el mejor modelo *SVM* coinciden con las de mayor frecuencia, usadas por la mayoría de los 1000 modelos ejecutados, lo cual significa que muchos más modelos *SVM* representan la variabilidad de las concentraciones del determinante con errores de estimación similares al mejor modelo. Lo anterior contrasta con las longitudes de onda seleccionadas por los modelos *PLS*, cuya mayor frecuencia es más de cuatro veces menor a la mayor del modelo *SVM.* Lo anterior implica que un menor número de modelos *PLS* y en un rango más amplio de cantidad de longitudes de onda pueden representar la variabilidad del determinante.

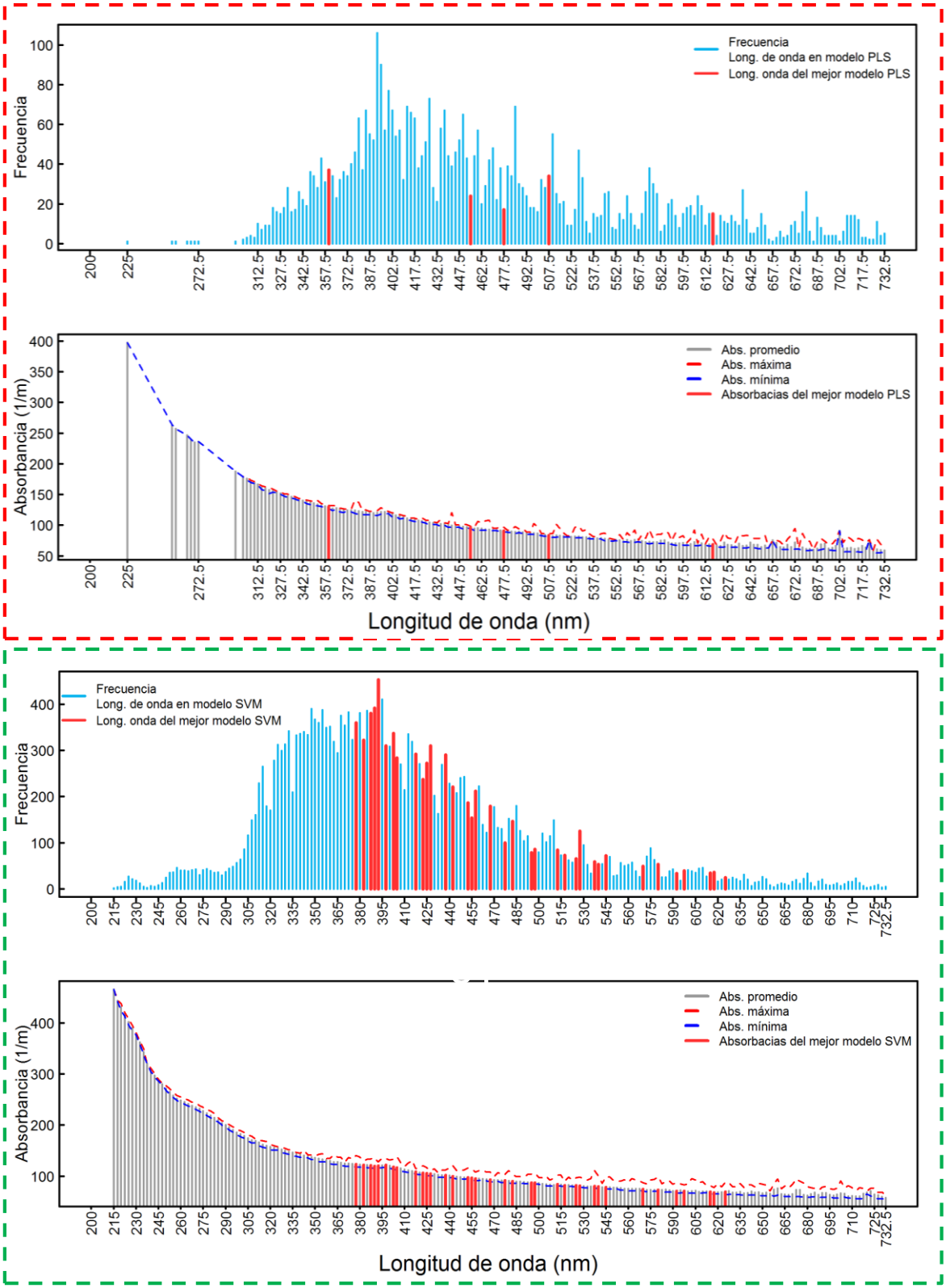


Figura 85- Parsimonia de los modelos *PLS* (recuadro rojo) y *SVM* (recuadro verde): frecuencia de las longitudes de onda y sus valores de absorbancia utilizadas en la calibración de los 1000 modelos del determinante DQO del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

En el caso de la DQO de las muestras del afluente de la PTAR en tiempo de lluvia y de la EE, se detectó que las principales variables predictoras del determinante se encuentran en el rango UV de los espectros para ambos modelos (Figura 86 y Figura 87). Sin embargo, en el caso de la EE ciertos modelos del *SVM* seleccionaron algunas longitudes de onda en el rango visible, las cuales presentan una baja frecuencia, y entre dichos modelos se encuentra el mejor, lo cual de cierta forma ratifica la relación de la DQO con interacción del haz de luz con el material soluble y particulado presente en la muestras, pero que en el caso de EE está limitado probablemente por la calidad de los datos de laboratorio y por el paso de luz de la sonda spectro::lyser empleada para la medición del espectro UV-Vis (ver numeral 4.3.1.3).

Finalmente, un rango más cerrado de longitudes de onda y con mayores frecuencias en el rango UV del espectro son usadas para calibrar ambos modelos en el caso de las muestras del afluente de la PTAR en tiempo de lluvia, lo cual permite inferir que la mayoría de sustancias susceptibles de ser oxidadas están disueltas en la muestra y no en suspensión (Figura 86).

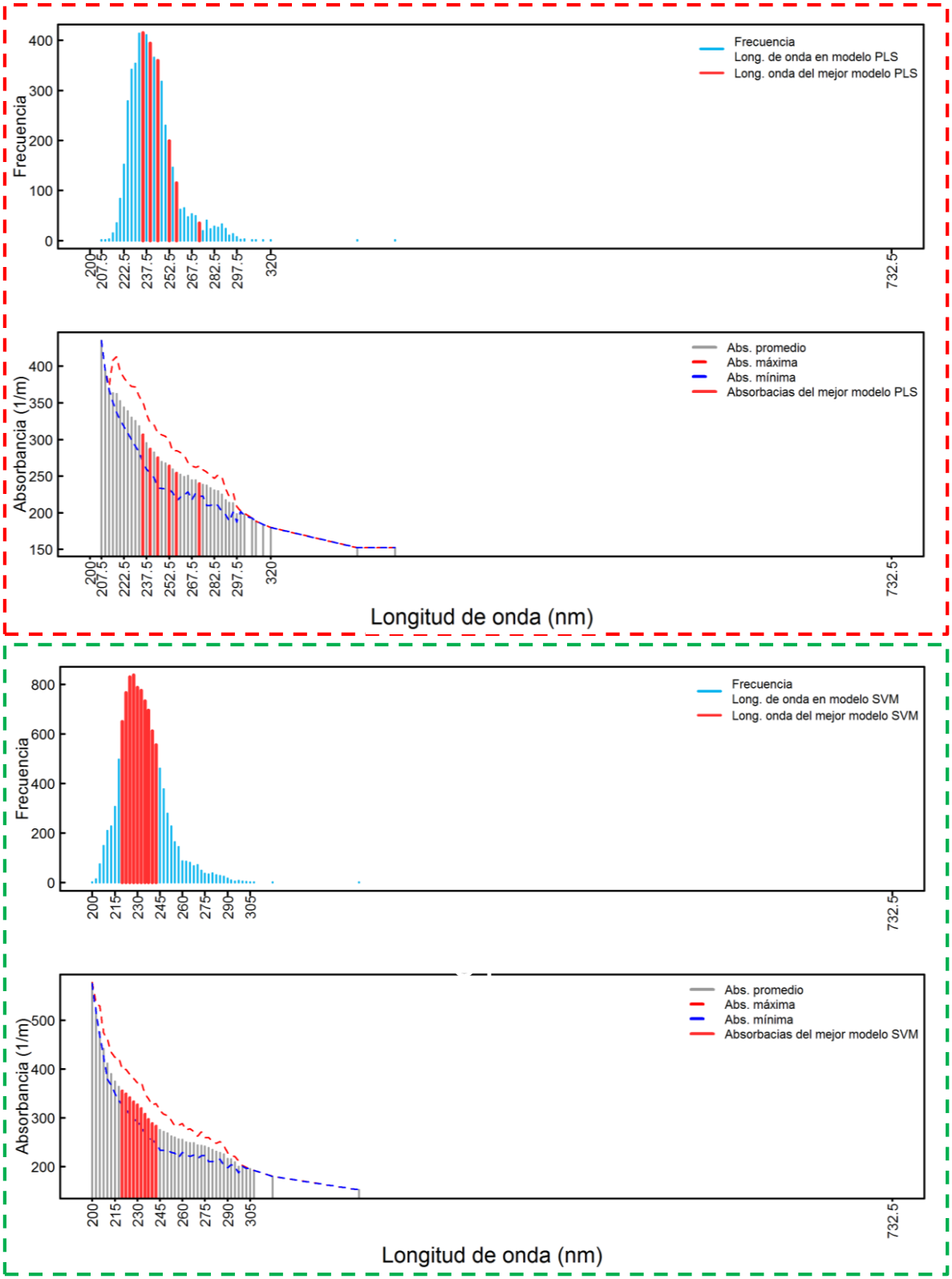


Figura 86- Parsimonia de los modelos *PLS* (recuadro rojo) y *SVM* (recuadro verde): frecuencia de las longitudes de onda y sus valores de absorbancia utilizadas en la calibración de los 1000 modelos del determinante DQO del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia)

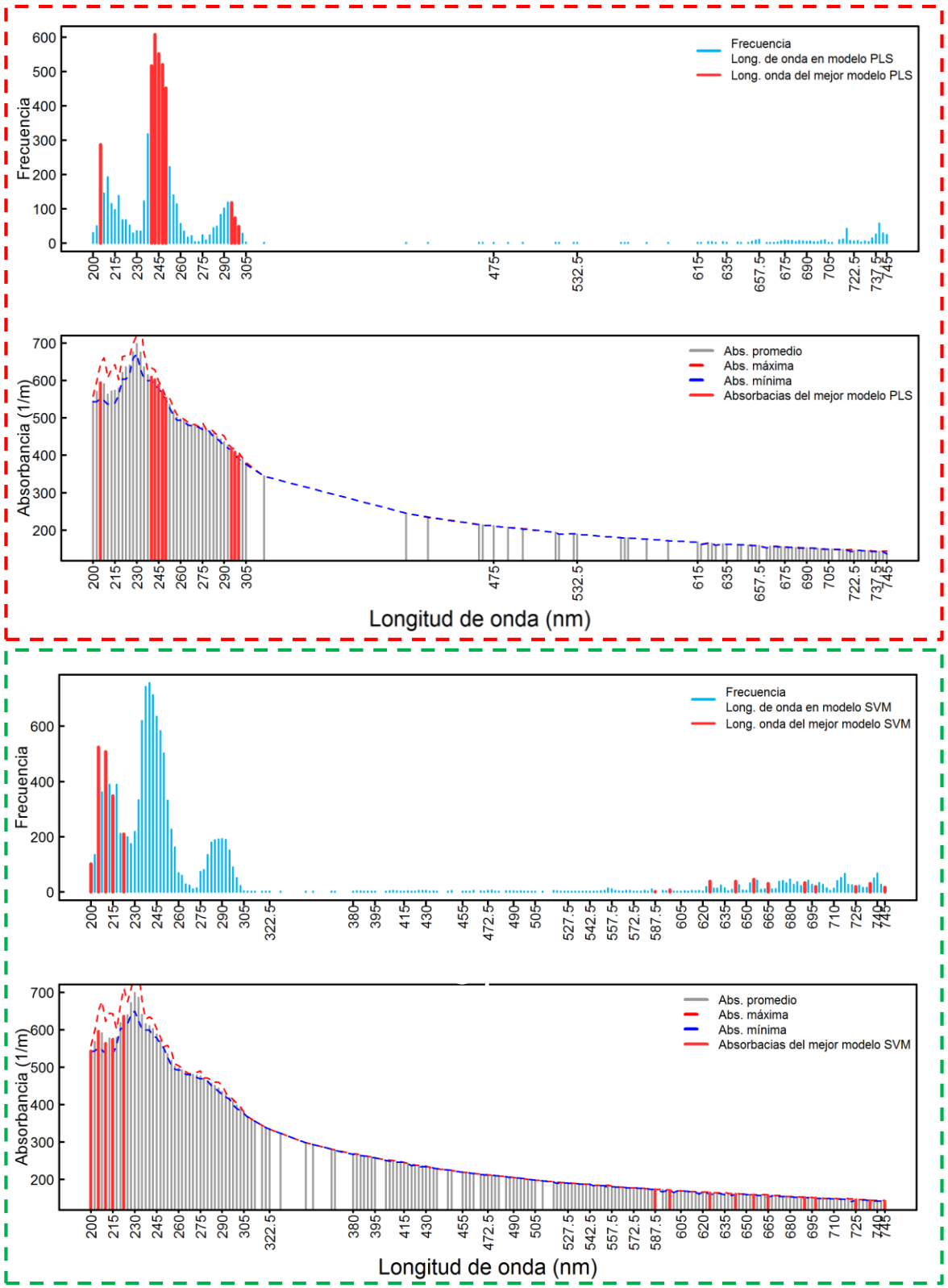


Figura 87- Parsimonia de los modelos *PLS* (recuadro rojo) y *SVM* (recuadro verde): frecuencia de las longitudes de onda y sus valores de absorbancia utilizadas en la calibración de los 1000 modelos del determinante DQO del afluente de la EE de Gibraltar

4.3.2.3. Parsimonia de los modelos calibrados para estimar las concentraciones de la DQOf

La DQOf, en general fue el determinante que para las muestras del afluente de la PTAR en tiempo seco y lluvia presentó mayores frecuencias en las longitudes de onda usadas por los modelos *PLS* y *SVM*, y que ambos casos las más importantes son las mismas. Lo cual permite inferir que es posible caracterizar el comportamiento de este determinante en el afluente de la PTAR en función de la relación de un grupo específico de longitudes de onda (absorbancias) y alguna variable hidrometereológica (Figura 88 y Figura 89).

Al igual que los otros determinantes, fueron los modelos *SVM* los que usaron en general un mayor de variables predictoras en comparación con los modelos *PLS* en los tres casos.

Por otra parte, al observar la Figura 88 es evidente que algunos modelos tanto PLS como SVM utilizaron longitudes de onda en rango visible con una frecuencia muy baja, y por lo tanto se puede suponer que dentro del conjunto de datos sin *outliers* empleados para calibrar los modelos *PLS* y *SVM*, aún existe una pareja de datos espectro-concentración que puede ser catalogado como *outlier* o que es posible que valores de absorbancia dentro de un espectro sean *outliers*.

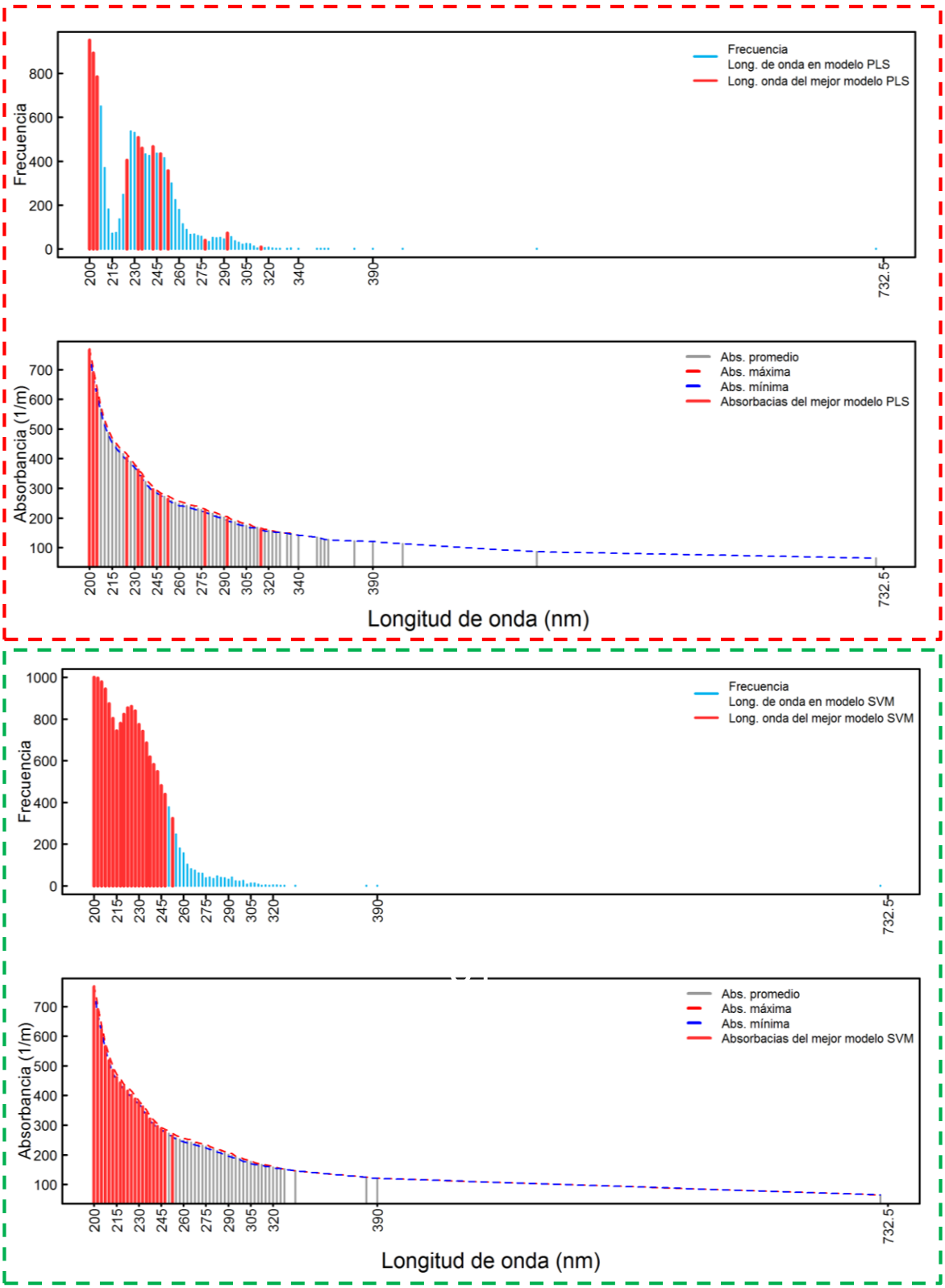


Figura 88- Parsimonia de los modelos *PLS* (recuadro rojo) y *SVM* (recuadro verde): frecuencia de las longitudes de onda y sus valores de absorbancia utilizadas en la calibración de los 1000 modelos del determinante DQOf del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

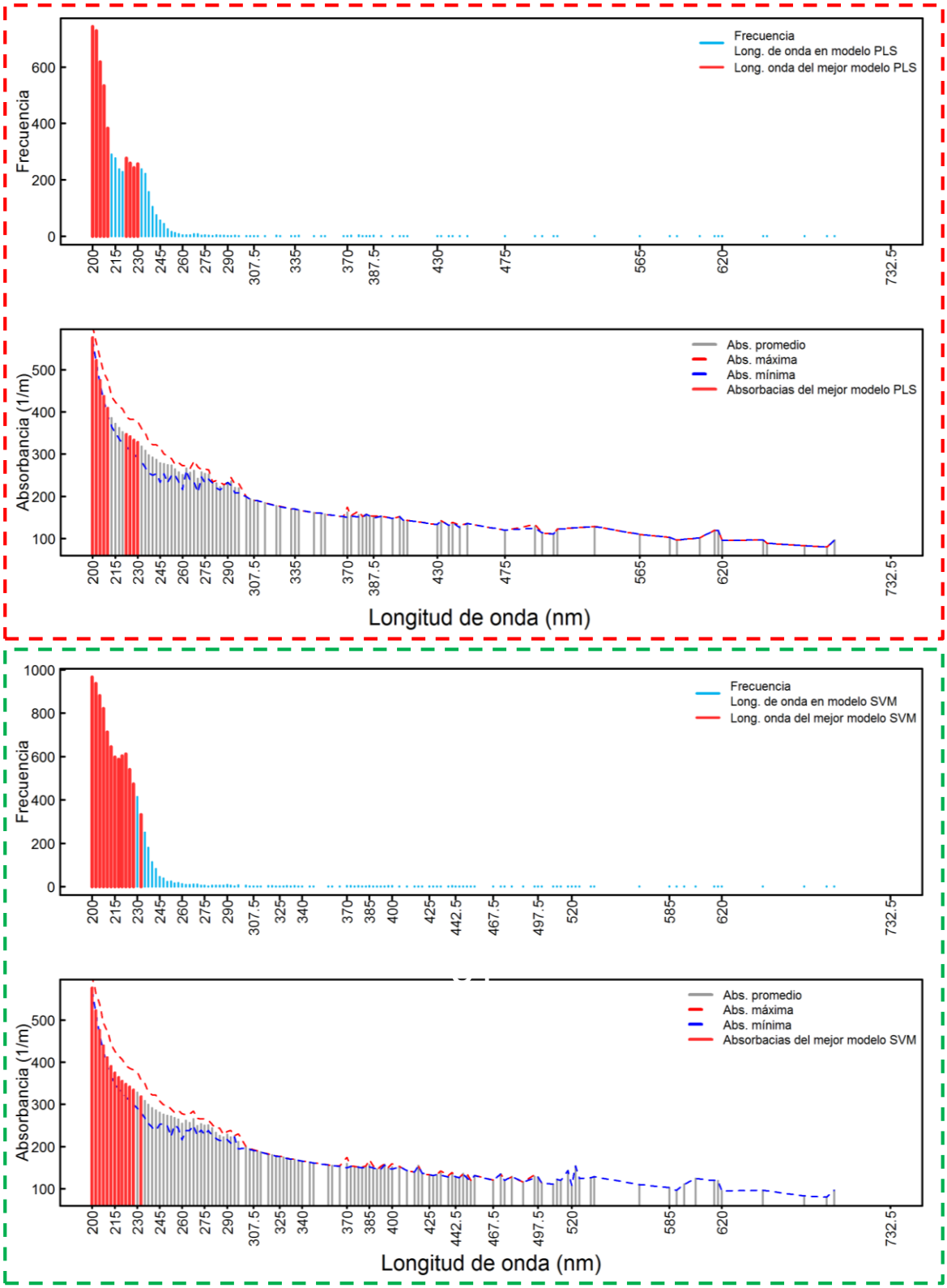


Figura 89- Parsimonia de los modelos *PLS* (recuadro rojo) y *SVM* (recuadro verde): frecuencia de las longitudes de onda y sus valores de absorbancia utilizadas en la calibración de los 1000 modelos del determinante DQOf del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia)

Por último, aunque las longitudes de onda usadas para calibrar los modelos *PLS* y *SVM* están en el rango UV en el caso de las concentraciones de DQOf de las muestras de la EE, su frecuencia con respecto a los otros casos de estudio es menor e incluso algunos modelos seleccionan longitudes de onda con mayores frecuencias a las obtenidas en el rango UV, cuya relación con el determinante no refleja su variabilidad con ningún modelo, tal como lo validan los resultados de las regresiones presentadas en la Figura 81.

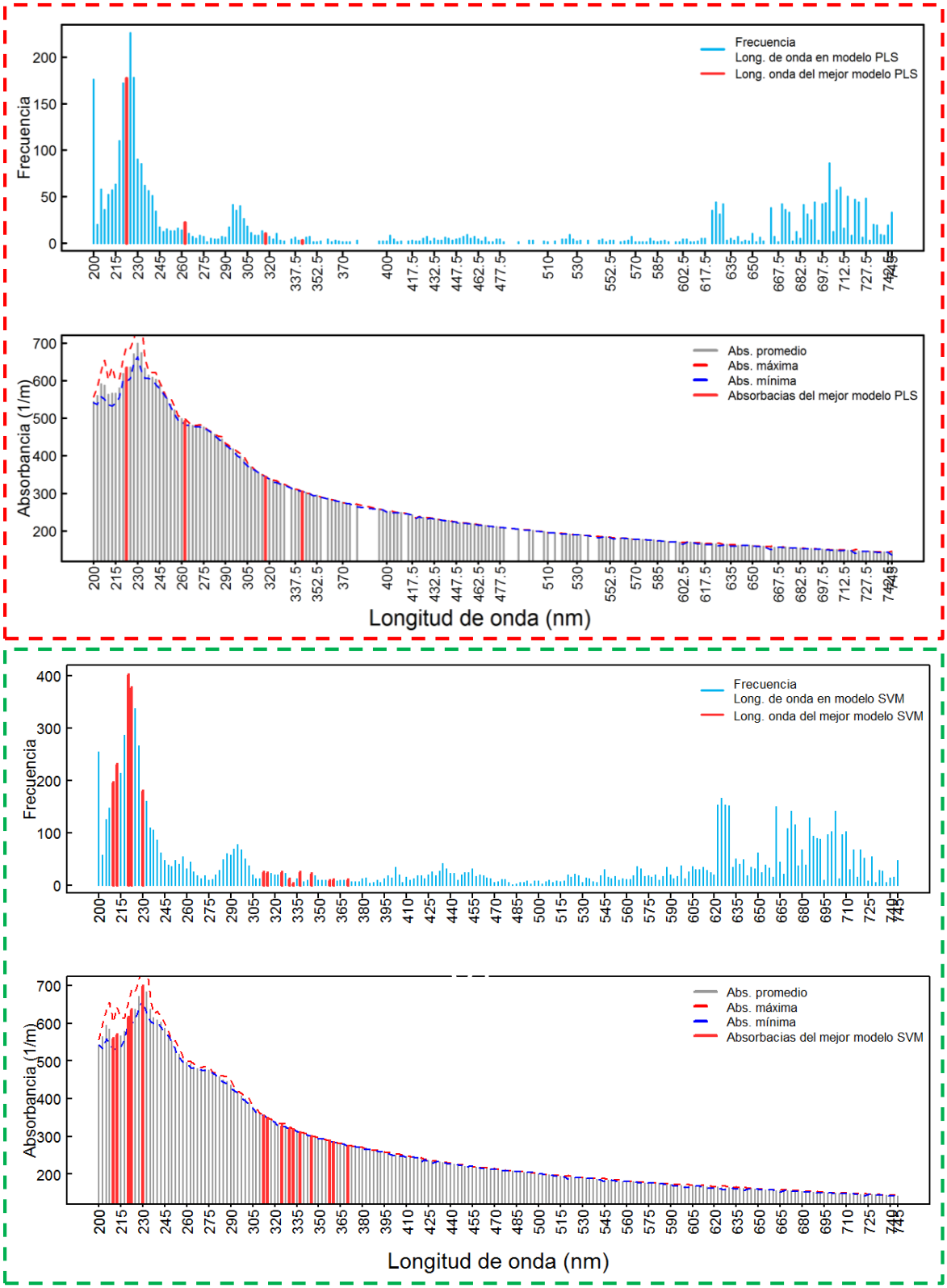


Figura 90- Parsimonia de los modelos *PLS* (recuadro rojo) y *SVM* (recuadro verde): frecuencia de las longitudes de onda y sus valores de absorbancia utilizadas en la calibración de los 1000 modelos del determinante DQOf del afluente de la EE de Gibraltar

5. CONCLUSIONES

El presente estudio pretendió contribuir en el desarrollo de dos metodologías basadas en métodos *machine learning* para estimar concentraciones de los determinantes SST, DQO y DQOf en continuo asociadas a las aguas de sistemas de saneamiento urbano mediante datos de espectrometría UV-Visible *in situ*. Luego, información de concentraciones obtenidas en ensayos de laboratorio y espectros de absorbancia en el rango UV-Vis de muestras puntuales de afluentes de cuatro puntos de monitoreo fue utilizada: PTAR San Fernando en Medellín-Colombia, PTAR de *Fontaines-sur-Saône* (tiempo seco y lluvia) en *Grand Lyon*-Francia y Estación Elevadora de Gibraltar en Bogotá-Colombia.

Se desarrollaron algoritmos basados en dos herramientas de inteligencia artificial de aprendizaje supervisado: *Support Vector Machine* (*SVM*) y Redes Neuronales Artificiales tipo *feed-forward* (RNA), donde se calibraron las funciones regresivas de estos modelos por medio de los espectros de absorbancia (variable independiente) y concentraciones de determinantes (variable dependiente) y los resultados obtenidos fueron comparados sistemáticamente con aquellos con el algoritmo quimiométrico *OPP* modificado (Torres y Bertrand-Krajewski, 2008), cuyos cambiaron estuvieron enfocados en los siguientes aspectos: (i) se rescribió el código en la plataforma *R* (R Development Core Team, 2013) con los siguientes cambios, (ii) se utilizó el paquete, (iii) el algoritmo *PLS* utilizado es *Wide Kernel,* (iv) El número óptimo de variables latentes se determina por medio de validación cruzada tipo *Jackknife* o *Leave One Out*, y (v) la selección y clasificación de forma decreciente con respecto a la relevancia de las variables independientes se determinó a través de método *ZATO* (Zamora y Torres, 2013), el cual obtiene la recurrencia y nivel afinidad entre cada variable independiente y las dependientes dentro del conjunto de datos de analizados. En general el modelo *SVM* genero resultados satisfactorios, incluso en algunos casos con un grado de error menor (SST y DQOf) al obtenido por medio de los modelos *PLS,* principalmente cuando el conjunto de datos de calibración y validación es mayor, tal como se presentó en el caso de las muestras en tiempo seco del afluente de la PTAR de *Fontaines-sur-Saône*. Lo anterior incidió en que los parámetros de la arquitectura de los modelos *SVM* sean más robustos y permitan la representación de la variabilidad de las concentraciones para diferentes comportamientos del espectro de absorbancia. Sin embargo, en las muestras para tiempo de lluvia de este punto de monitoreo fueron mejores los resultados alcanzados por los modelos *PLS* para los determinantes evaluados. No obstante, en el caso de la DQO en tiempo de lluvia, el modelo *SVM* otorgó valores de los coeficientes de determinación en la etapa de validación superiores a 0.97. En el caso de los datos de la EE de Gibraltar se obtuvieron resultados poco satisfactorios donde el *overfitting* fue el fenómeno que se mantuvo presente en la calibración de los diferentes conjuntos, generando probablemente por la baja calidad de los datos de laboratorio y el paso de luz de la sonda de medición del espectro de absorbancia. Por consiguiente, *SVM* se convierte en una herramienta alternativa para la evaluación quimiométrica de los espectros UV-Vis en conjuntos de datos con una cantidad que representen la variabilidad

del fenómeno a estimar a través de los valores absorbancias en diferentes longitudes de onda mejor que el modelo convencional *PLS.* Lo anterior implica que este tipo de herramientas permitirá aumentar la confianza en los resultados de las concentraciones equivalentes obtenidas *in situ*, en continuo y en tiempo real en diferentes sistemas de saneamiento urbano por medio de la medición del espectro de absorbancia característicos de las aguas monitoreadas.

En cuanto al algoritmo desarrollado para la evaluación del modelo RNA implementado, se determinó que la arquitectura (entradas, capa oculta-(No. de neuronas), salidas) de este tipo de modelo de red neuronal no es conveniente para la estimación de las concentraciones equivalentes de los determinantes de estudio, ya que no permite representar su variabilidad en función de ningún número de predictores (No. longitudes de onda) con un grado de error satisfactorio. No obstante, se pudo confirmar que a mayor número de longitudes de onda menor debe ser la tasa de decaimiento de pesos, y que para el fenómeno estudiado aumentar el número neuronas en la capa oculta no incide significativamente en la calibración de los modelos. Los conjuntos de pesos generados por diferentes arquitecturas del modelo RNA *feed-forward* no permitieron replicar en los conjuntos de prueba y validación los resultados alcanzados en la etapa de entrenamiento.

Por otra parte, no solamente el nivel de desempeño es responsabilidad de los modelos regresivos implementados, ya que se demostró por medio del método *ZATO* (Zamora y Torres, 2013) que determinar la recurrencia de una longitud de onda a ocupar cierta posición en 220 posiciones posibles, permite establecer su nivel de importancia y afinidad con las concentraciones del determinante objetivo. Además, el método permito evaluar la calidad de los datos, ya que una mayor dispersión o polvo de las recurrencias en diferentes niveles de importancia de muchas longitudes de onda está relacionada con la baja afinidad entre las concentraciones obtenidas en laboratorio y los espectros UV-Vis relacionados, y por ende la presencia de *outliers*. Esto implica que este método de preprocesamiento sea capaz de consolidar la variables predictoras más relevantes, lo cual conduce a que los modelos regresivos sean más parsimoniosos y cuyos tiempos computacionales empleados en su calibración sean menores. Además, la presencia y magnitud de un determinante están asociadas a los valores de absorbancia de las longitudes de onda seleccionadas por el método. Con dicha información se pueden vincular los cambios en la matriz de determinantes tanto en número de compuestos como en magnitud, y a partir de esto generar estados de alerta, alarma, reglas de control y mejoras en los procesos de tratamiento en los diferentes sistemas de saneamiento urbano.

En cuanto a la detección de *outliers* en las bases de datos conformadas por parejas de datos de espectro-concentración, se logró de establecer la ventaja de catalogar una pareja de datos como un posible *outlier* en las mejoras de los resultados de los modelos regresivos usados en este estudio. Sin embargo, el método desarrollado e implementado en el algoritmo AEEC tiene limitaciones que repercuten negativamente, si bien no en los modelos regresivos sí en los datos y la cantidad de estos que son seleccionados como

posibles *outliers*, ya que se pudo establecer que el método detección no tiende a eliminar las muestras cuya relación con valores de absorbancia sean menos recurrentes, y de allí lo exigente y poco funcional que resulta detectar un *outlier* en relación a una sola longitud de onda, lo cual conlleva a sobreestimar el número de muestras catalogadas como *outliers*. Por lo tanto, esto ratifica que la presencia de un determinante puede estar asociada a más de una longitud de onda y a partir de esto reducir los efectos generados por la sensibilidad cruzada a otros compuestos presentes en las muestras, los cuales pueden afectar el valor de absorbancia y con esto la longitud de onda con mayor correlación a la cual se asocia la presencia de un determinante. Esto implica que probablemente después de la eliminación de *outliers*, las parejas de datos catalogadas como datos válidos no representen la variabilidad de las concentraciones de un determinante cuando éste es cuantificado en función de los espectros de absorbancia, ya que se asocia la presencia del determinante a una longitud onda. Por ejemplo en el caso de la DQO, esta tenderá a estar asociada al rango UV por la presencia de sustancias disueltas susceptibles a ser oxidadas y de igual forma a las sustancias en suspensión presentes en el rango visible.

El número de longitudes de onda que explican la variabilidad de la concentración de un determinante resulta mayor en todos los casos de estudio con el modelo *SVM*. No obstante, por medio de los modelos *SVM* y *PLS* se estableció a qué rango o rangos del espectro está asociada la presencia de los determinantes objeto de estudio de la misma forma y de allí las longitudes de onda que podrán generar calibraciones satisfactorias cuyos resultados se propaguen de la misma forma al conjunto de validación. Por otra parte, las longitudes de onda seleccionadas por el mejor modelo *SVM* coinciden con las de mayor frecuencia usadas por la mayoría de los 1000 modelos ejecutados, lo cual significa que muchos más modelos *SVM* representan la variabilidad de las concentraciones del determinante con errores de estimación similares al mejor modelo. Lo anterior contrasta con las longitudes de onda seleccionadas por los modelos *PLS*, cuya mayor frecuencia es más de cuatro veces menor a la mayor del modelo *SVM*, lo cual implica que un menor número de modelos *PLS* y en un rango más amplio de cantidad de longitudes de onda puedan representar la variabilidad del determinante.

6. PERSPECTIVAS

La medición y análisis de espectros de absorbancia en el rango UV-Visible en muestras de aguas residuales resulta una alternativa interesante para detectar y cuantificar la presencia de determinantes *in situ* y en continuo. Por lo tanto, es importante continuar explorando diferentes métodos *machine learning* (aprendizaje supervisado y por refuerzo) que permitan ampliar el número de modelos capaces de estimar la variabilidad de las concentraciones de un conjunto de determinantes en una matriz de compuestos de una forma bivariada o multivariada en función de las absorbancias del espectro.

Por otra parte, se recomienda emplear no solamente métodos *machine learning* de aprendizaje supervisado sino también investigar los no supervisados y por refuerzo (*e.g. cluster*, mapas autoorganizados, *Q-Learning, etc.*) para realizar la explotación de la información espectral. Por ejemplo, en el caso de *Q-Learning* se podrían emplear en recalibración en continuo y en tiempo real de los pesos o parámetros de un modelo regresivo, ya que este método de aprendizaje por refuerzo permitiría encontrar una política de acción-selección óptima para cualquier proceso (finito) de decisión de Markov. Uno de los puntos fuertes del *Q-Learning* es que éste es capaz de comparar la utilidad esperada de las acciones disponibles sin necesidad de un modelo del fenómeno analizado.

Ningún método es realmente bueno si no se alimenta con datos de calidad y representativos de la dinámica de los flujos contaminantes. Por lo tanto, se recomienda que los métodos de análisis clásicos de laboratorio sean confiables o realmente estandarizados para que los valores de concentración de los determinantes cuantificados representen una realidad y no generen un sesgo en los modelos.

Se recomienda explorar con diferentes métodos la posible relación entre las condiciones hidrometereológicas que afectan la cantidad y la calidad de los hidrosistemas de saneamiento urbano con las formas espectrales y los cambios de magnitud en las absorbancias en el espectro.

Por otra parte, detectar cambios en la matriz de compuesto únicamente con el comportamiento del espectro UV-Vis permitiría generar estados de alerta o alarma que puedan conducir a reducir las perturbaciones que afectan, por ejemplo, los sistemas de tratamiento y con esto robustecer las reglas de control. Incluso, únicamente caracterizando la calidad del agua por medio de espectrometría UV-Vis podría contribuir a reducir la propagación de los errores asociados a las prácticas de laboratorio y a la incertidumbre compuesta ligada a la variabilidad espacial en un instante de tiempo de la concentración de un determinante y a la precisión de los instrumentos que son empleados para su detección y cuantificación en laboratorio.

Es importante en futuras investigaciones explorar otros modelos de redes neuronales de arquitecturas más complejas, tales como *multiperceptron* o *backpropagation* que permitan establecer si modelos con diferentes arquitecturas y estrategias de entrenamiento son aplicables a la estimación de concentraciones en función del espectro absorbancias en diferentes bases de datos (*e.g.* tiempo lluvia, seco; rangos de concentraciones entre otros) y con esto validar los resultados alcanzados por Zamora *et al*. (2010).

Por otra parte, se deben explorar diferentes métodos de detección de *outliers* multivariados que permitan establecer la coherencia y afinidad entre una pareja de datos espectro-concentración dentro de un universo de valores de estas variables, tal como el método presentado en el numeral 3.4.2 o como la distancia de Mahalanobis, distancia Euclidiana entre otros, ya que una detección bivariada limita a un único valor de absorbancia la presencia de un determinante de calidad del agua, cuando, por ejemplo en el caso de la DQO, éste está relacionado con las sustancias o compuestos en dilución (principalmente rango UV) y suspensión (rango visible). Incluso, se podría llegar a pensar en no eliminar siempre una pareja de datos, sino detectar y eliminar longitudes de onda cuyas absorbancias no se asocien a la presencia y magnitud del determinante y que puedan generar efectos negativos, por ejemplo en los modelos regresivos.

Es importante continuar desarrollando y aplicando rutinas de análisis de datos tipo incertidumbre-*outliers*-calibración para consolidar modelos que represente la dinámica de un fenómeno de una forma parsimoniosa y eficiente (tiempo computacional), enmarcadas en un porceso de optimización, lo cual permitiría consolidar líneas claras en el comportamiento de la relación espectro / calidad del agua en un hidrosistema.

Finalmente, el uso de tecnologías de medición *in situ*, en continuo y en tiempo cuasi real debe ser implementado, masificado y regulado por las autoridades ambientales del territorio nacional, con miras a incrementar el conocimiento del comportamiento de los flujos de determinantes en diferentes hidrosistemas (ríos, sistemas de saneamiento urbano, acuíferos, *etc.*) así incrementar la información de la calidad del agua (series de tiempo), y con esto robustecer ejercicios de modelación, toma de decisiones y estrategias de control.

7. BIBLIOGRAFÍA

Abdi, H. (2003). Partial least squares (PLS) regression. In: Lewis-Beck, M., Bryman, A.&Futing, T. (eds) Encyclopedia of Social Sciences Research Methods. Sage, Thousand Oaks, CA (USA), pp. 792–795.

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

Allan IJ, Vrana B, Greenwood R, Mills GA, Roig B y Gonzalez C. (2006). A "toolbox" for biological and chemical monitoring requirements for the European Union's Water Framework Directive. *Talenta*; 69, 302-322.

Anta, J., Pena, E., Suarez, J., y Cagiao, J. (2006). A BMP selection process based on the granulometry of runoff solids in a separate urban catchment. Water SA, 32(3) : 419 – 428.

Ardia, D., Mullen, K., Ulrich, J., y Peterson, B. (2011) Deoptim: An r package for global optimization by differential evolution. Journal of Statistical, (2006).

Basak, D., Pal, S., y Patranabis, D. (2007) Support vector regression. Neural Information Processing, 11(10), 203–224.

Bates, D. M. y Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*, Wiley.

Baurès, E, Berho, C., Pouet, M. F., y Thomas, O (2004) In situ UV monitoring of wastewater: a response to sample aging. Water science and technology, 49(1), 47–52.

Baurès, Estelle, Hélias, E., Junqua, G., y Thomas, Olivier (2007) Fast characterization of non domestic load in urban wastewater networks by UV spectrophotometry. Journal of environmental monitoring : JEM, 9(9), 959–65.

Becouze, C., Wiest, L., Baudot, R., Bertrand-Krajewski, Jean-Luc, y Cren-Olivé, C. (2011) Optimisation of pressurised liquid extraction for the ultra-trace quantification of 20 priority substances from the European Water Framework Directive in atmospheric particles by GC-MS and LC-FLD-MS/MS. Analytica chimica acta, 693(1-2), 47–53.

Bergh, S.-G. (1996). Diagnosis Problems in Wastewater Settling. Licentiate's thesis, IEA Lund Institute of Technology. Lund, Sweden. Pág. 118.

Bertrand-Krajewski, J.-L. (2006). Cours d'hydrologie urbaine. Partie 7. Les polluants des rejets urbains par temps de pluie. 54 p. Disponible sur http://jlbkpro.free.fr/teachingmaterial/oshu3-07-polluants-des-rutp.pdf, [Consulta : 2 enero 2012].

Bester, K. and Schäfer, D. (2009) Activated soil filters (bio filters) for the elimination of xenobiotics (micro-pollutants) from storm- and waste waters. Water research, 43(10), 2639–46.

Bourgeois, W., Burgess, J. E., y Stuetz, R. M. (2001) On-line monitoring of wastewater quality : a review. Chemical Technology, 348(July 2000).

Bousquet, O., Rätsch, G., y Luxburg, U. (2004) Advanced Lectures on Machine Learning LNAI 3176,

Brombach, H., Weiss, G. y Fuchs, S. (2005) A new database on urban runoff pollution: Comparison of separate and combined sewer systems. Water Science and Technology 51, 119-128.

Burton, G.A., y Pitt, R.E., 2002. Stormwater effects Handbook. A toolbox for Watershed Managers, Scientists, and Engineers. Lewis Publishers.

Butler D., y Karunaratne S. H. P.G. (1995). The suspended solids trap efficiency of the roadside gully pot. Water Resources, 29(2) : 719 – 729.

Butler, D. y Davies, J. (2011) Urban drainage (T. & Francis, ed.), London, England. Cherkassky, V. and Ma, Y. (2004) Practical selection of SVM parameters and noise estimation for SVM regression. Neural networks, 17(1), 113–26.

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

Chocat, B. (1997). Encyclopédie de l’hydrologie urbaine et de l’assainissement. Paris (France) :Technique et Documentation, 1136 p.

Chocat, B., Bertrand-Krajewski J.-L. y Barraud S. (2007). Eaux pluviales urbaines et rejets urbains par temps de pluie. Paris (France) : Les techniques de l’ingénieur, article W 8600, août 2007, 17 p. + annexes.

Clair, N., Parkin, G., y Perry, L. (2003) Chemistry for Environmental Engineering and Science, New York, USA.

Colin, F. y Quevauviller, PP (1997) The contribution of advanced technologies monitoring of water quality (ELSEVIER, ed.), Nancy, France.

Da Silva, A. M. (2008) Universidade do Minho Ana Maria da Silva Paulo Monitoring of Biological Wastewater Treatment Processes using Indirect Spectroscopic Techniques Co-Orientadora. pp. 155.

Dawson, C. W., Abrahart, R. J., y See, L. M. (2007) HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environmental Modelling & Software, 22(7), 1034–1052.

Degrémont, (2005). Mémento technique de l'eau - 10ème édition. Degrémont-Suez ed. Vol. 2, Rueil-Malmaison: Lavoisier SAS. p. 1717.

Demuth H., Beale M., Hagan M. (2009). Neural Nework Toolbox - User’s Guide. Mathworks, Inc.

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

Doyen, L. (1992). Etude de la relation entre MES et absorbance sur le collecteur EP Albert Camus (Neuilly sur Marne) et sur le collecteur unitaire d’Enghien à Leclerc (Epinay-sur-Seine). Rosny-sous-Bois (France) : Direction de l’Eau et de l’Assainissement de la Seine-Saint-Denis, rapport, septembre 1992, 51 p.

EEUU-EPA (1983) Results of the Nationwide Urban Runoff Program: Volume 1 - Final report, p. 186, US Environmental Protection Agency, Water Planning Division, Washington DC, USA.

Ellis, J.B. (1989). Urban Discharges and Receiving Water Quality. Pergamon Press, Oxford, UK, p.

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

Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P. (1996) “Knowledge discovery and data mining: Towards a unifying framework” in Discovery and Data Mining. Portland, OR, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, 82–88.

Field, J. (1987). Paramétros operativos del manto de lodos serobicos de flujo ascendente [en línea]. Tesis. Universidad Agricola de Wageningen, Holanda http://www.bvsde. paho.org /bvsacd/cd68/000866/000866b.pdf. [Consulta: 20 marzo 2013]

Fleischmann, N., Langergraber, G., y Weingartner, A [en línea]. (2001). On-line and in-situ measurement of turbidity and COD in wastewater using UV/VIS spectrometry. s-can.asia. http://www.s-can.asia/medialibrary/publications/p_2001_06.pdf. [Consultado: 20 noviembre 2012].

Fleischmann, N., Staubmann, K., y Langergraber, G. (2002) Management of sensible water uses with real-time measurements. Water science and technology, 46(3), 33–40.

Fletcher, T. D., Andrieu, H., y Hamel, P. (2013) Understanding, management and modelling of urban hydrology and its consequences for receiving waters: A state of the art. Advances in Water Resources, 51, 261–279.

Fogelman, S., Zhao, H., y Blumenstein, M. (2006) A rapid analytical method for predicting the oxygen demand of wastewater. Analytical and bioanalytical chemistry, 386(6), 1773–9.

Gamerith, V. (2011). Tesis doctoral. High resolution online data in sewer water quality modelling. pp. 236.

Ghanty, P., Paul, S., y Pal, N. (2009) NEUROSVM: An Architecture to Reduce the Effect of the Choice of Kernel on the Performance of SVM. The Journal of Machine Learning Research, 10, 591–622.

Gonzalez, C., Greenwood, R., y Quevauviller, P. P. (Eds.). (2009). Rapid chemical and biological techniques for water monitoring (Vol. 23). Wiley.

Grange, D. (1994). La mesure en continu des eaux de ruissellement en système unitaire: un outil pour le choix d'une stratégie de lutte. La Houille Blanche, 1/2 : 39 - 41.

Grange, D., y Pescheux, F. (1986). Utilisation de la spectrophotométrie d'absorption dans l'ultraviolet et le visible pour le contrôle simple en continu du fonctionnement des stations d'épuration. Trappes (France) : LROP Laboratoire Régional de l'Ouest Parisien, compte-rendude synthèse N° FAER A.51.03.5.

Gromaire, M.C., Kafi-Benyahia, M., Gasperi, J., Sadd, M., Moilleron, R. y Chebbo, G. (2008). Settling velocity of particulate pollutants from combined sewer wet weather discharges. Water Science and Technology, 58(12): 2453 – 2465.

Gujer, W. (2008) Systems analysis for water technology, Zurich, Switzerland, Springer.

HMV Ingenieros (2007). Predimensionamiento planta de tratamiento de aguas residuales canoas. Producto 2: Metodología de muestreo. Informe Inédito. EAAB. Informe de consultoría Contrato: 1-02-26100-806-2006. p. 68.

Hochedlinger, M. (2005). Tesis doctoral Assessment of combined sewer overflow emissions. pp, 219.

Holland, J. H., (1975). Adaptation in Natural Articial Systems. University of Michigan Press, Ann Arbor.

Huber, E. y Frost, M. (1998) Light scattering by small particles. J. Wat. SRT – Aqua, 47: 87–94.

Iglewicz B., and Hoaglin D. (1993). How to detect and handle outliers. ASQC Quality Press.

ISO, International Organization for Standardization, (1995); Guide to the expression of uncertainty in measurement, Geneva: ISO, p. 101.

JKwok., T. (2001). Linear Dependency between and the Input Noise in –Support Vector Regression, in: G. Dorffner, H. Bishof, y K. Hornik (Eds): ICANN 2001, LNCS 2130 pp. 405-410.

Kafi-Benyahia, M. (2006). Variabilité spatiale des caractéristiques et des origines des polluants de temps de pluie dans le réseau d'assainissement unitaire parisien. Tesis doctoral, Ecole Nationale des Ponts et Chaussées, Paris, France, 502 p.

Karatzoglou, A., Smola, A., Hornik, K., y Zeileis, A. (2004) kernlab-an S4 package for kernel methods in R, Viena, Austria.

Korostynska, O., Mason, A., y Al-Shamma'a, A. (2012) Monitoring of Nitrates and Phosphates in Wastewater: Current Technologies and Further Challenges. International Journal on Smart …, 5(1), 149–176.

Langergraber, G., Fleischmann, N, y 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., y Weingartner, A. (2004a) Monitoring of a paper mill wastewater treatment plant using UV/VIS spectroscopy. Water science and technology, 49(1), 9–14.

Langergraber, G., Gupta, J. K., Pressl, a, Hofstaedter, F., Lettl, W., Weingartner, A., y Fleischmann, N. (2004b). On-line monitoring for control of a pilot-scale sequencing batch reactor using a submersible UV/VIS spectrometer. Water science and technology. Vol. 50(10), 73–80.

Lepot, M. (2012) Mesurage en continu des flux polluants en MES et DCO en réseau d'assainissement. , 1–253.

Lorenz, U., Fleischmann, Nikolaus, y 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.

Lynggaard-jensen, A. (1999) Trends in monitoring of waste water systems. Talanta, 50(4), 707–16.

Maestre, A. y Pitt, R. (2005) The National Stormwater Quality database, Version 1.1 - A Compilation and Analysis of NPDES Stormwater Monitoring Information, p. 447, U.S. EPA Office of Water, Washington D.C., USA.

Marchandise, P., Legendre, J.P., y Lafont, R. (1978). Méthode de mesure en continu de la pollution des eaux usées – Asservissement des ajouts de réactifs à la charge de pollution sur stations d'épuration physico-chimique. La Technique de l'Eau et de l'Assainissement, 383 : 11 -18.

Maribas, A., Laurent, N., Battaglia, P., Do Carmo Lourenço da Silva, M., Pons, M.-N., and Loison, B. (2008) Monitoring of rain events with a submersible UV/VIS spectrophotometer. Water science and technology : a journal of the International Association on Water Pollution Research, 57(10), 1587–93.

Massachusetts Institute of Technology (MIT) [en línea]. (2002). Artificial Neural Networks. (http://ocw.mit.edu/NR/rdonlyres/Sloan-School-of-Management/ 15-062Data-MiningSpring2003/650A194A-828C-4990-98CE-7EB966628437/0/NeuralNet2002.pdf). [Consulta: 15 febrero 2013].

Metcalf y Eddy. (1991). Wastewater Engineering: Treatment, Disposal, and Reuse. McGraw-Hill Inc, Singapore, 1 134 p. 198.

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

Michalski, R., Carbonell, J., Mitchell, T. (1983). *Machine Learning – an artificial intelligence approach*. Morgan Kaufmann Publishers, USA.

Miranda, J. (2003) [en línea]. "EVALUACIÓN DE LA INCERTIDUMBRE EN DATOS EXPERIMENTALES." p. 43. http://depa.fquim.unam.mx/amyd/archivero/eval_incert_6905.pdf [Consulta: 10 enero 2012].

Mitchell, M. (1998). An Introduction to Genetic Algorithms. The MIT Press.

Mrkva, M. (1975) Automatic UV control system for relative evaluation of organic water pollution. Water Res., 9: 587–589.

Mujeriego, R., Bravo, J. M. y Feliu, M. T. (1982). Recreational in coastal waters: Public health implications. Vier Journee Etud. Pollutions, Cannes Centre Internationale D'exploration Scientifique de la Mer: 585-594.

Mullen, K., Ardia, D., y Gil, D. (2011) Deoptim: An r package for global optimization by differential evolution. Journal of Statistical …, 40(6).

Muttamara, S. (1996) Wastewater characteristics. Resources, Conservation and Recycling, 16, 145–159.

Nash, J. E. y Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models, 1, A discussion of principles". Journal of Hydrology., Vol. 10. pp. 282-290.

Nilsson, N. (1996) Introduction to machine learning. An early draft of a proposed textbook.

Novotny, N., y Olem, H. (1994). Water Quality: Prevention, Identification and Management of Diffuse Pollution. Van Nostrand Reinhold, New York (USA), 1 054 p.

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

Ojeda, C. B. and Rojas, F. S. (2012) Recent applications in derivative ultraviolet/visible absorption spectrophotometry: 2009–2011. A review. Microchemical Journal, 106, 1–16.

Olabe, X. (1998). REDES NEURONALES ARTIFICIALES Y SUS APLICACIONES, Bilbao, Portugal.

Olsson, G. (2007) Automation Development in Water and Wastewater Systems. Benchmarking, 12(5), 197–200.

Olsson, G., y Newell, B. (1999). "Wastewater Treatment Systems, Modelling, Diagnosis and Control," IWA Publishing, London.

Organización de las Naciones Unidas-ONU. (2010). Programa de ONU-Agua para la Promoción y la Comunicación en el marco del Decenio. Conferencia- Gestión Sostenible del Agua en las Ciudades: la implicación de las partes interesadas para un cambio y una acción eficaces. Zaragoza, España. 13 al 17 de diciembre de 2010.

Ort, C., Lawrence, M. G., Reungoat, J., y Mueller, J. F. (2010) Sampling for PPCPs in wastewater systems: comparison of different sampling modes and optimization strategies. Environmental science & technology, 44(16), 6289–96.

Pescod, M.B. (1992). Wastewater Treatment and Use in Agriculture. FAO Irrigation and Drainage Paper 47, FAO, Rome, 125 pp.

Pitt, R. y Amy, G. (1973) Toxic Materials Analysis of Street Surface Contaminants, US Environmental Protection Agency, Washington D.C., USA.

Platt, L. (1998). "Fast training of SVM using sequential optimization". In: B. Scholkopf, B. Burges and A.J. Smola (Eds.), Advances in Kernel Methods- Support Vector Learning. MIT Press, pp. 185-208. Cambridge.

Preetha, S. y Radha, V. (2011) A STUDY ON OUTLIER DETECTION METHODS FOR CATEGORICAL. International Journal Development Research, 6.

Price, KV, Storn, RM, y Lampinen, J. (2005) Differential evolution a practical approach to global optimization, Berlin, Germany, Springer.

Quevauviller, P., Thomas, O., y Van Der Beken, A. (2006) "Wastewater Quality Monitoring and Treatment" in Brussels, Belgium, John Wiley & Sons, Ltd., 397.

R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rännar, S., Lindgren, F., Geladi, P. y 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., Langergraber, G., y Siegrist, H. (2006) Uncertainties of spectral in situ measurements in wastewater using different calibration approaches. Water Science & Technology, 53(12), 187.

Rieger, L., Langergraber, G., Thomann, M., Fleischmann, N, y Siegrist, H. (2004) Spectral in-situ analysis of NO2, NO3, COD, DOC and TSS in the effluent of a WWTP. Water science and technology, 50(11), 143–52.

Ruban G., Marchandise P., y Scrivener O. (1993). Pollution measurement accuracy using real time sensors and wastewater sample analysis. Water Science and Technology, 28(11-12): 67 – 78.

Ruxton, G.D., (2006); The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test, Behavioral Ecology, 17(4), 688-690.

Sarraguça, M. C., Paulo, A., Alves, M. M., Dias, A. M. a, Lopes, J. a, y Ferreira, E. C. (2009) Quantitative monitoring of an activated sludge reactor using on-line UV-visible and near-infrared spectroscopy. Analytical and bioanalytical chemistry, 395(4), 1159–66.

Sartor, J. D., Boyd, G.B. y Agardy, F.J. (1974) Water Pollution Aspects of Street Surface Contaminants. Journal WPCF 46(3), 458-467.

Seo, S. (2006) A review and comparison of methods for detecting outliers in univariate data sets.

Simpson, B., France, D., y Lewis, B. (2013) Operating Procedure Wastewater Sampling, Athens, Georgia.

Smola, A., Murata, N., Schölkopf, B. y Muller. K. (1998). Asymptotically optimal choice of support vector machines, Proc. ICANN.

Sporea D. G., Sporea R. A., (2005). Setup for the in situ monitoring of the irradiation-induced effects in optical fibers in the ultraviolet-visible optical range. Review of Scientific Instruments. Vol. 76: 113110.

Staubmann K., Langergraber G., Fleischmann N. (2001). UV/VIS spectroscopy for the monitoring of testfilters. In: IWA (ed.), Preprints of the 2nd IWA World Water Congress, 15–19 October 2001, Berlin, Germany.

Storn, R. y Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 341–359.

Sutherland-Stacey, L., Corrie, S., Neethling, A., Johnson, I., Gutierrez, O., Dexter, R., Yuan, Z., Keller, J., y Hamilton, G. (2008) Continuous measurement of dissolved sulfide in sewer systems. Water science and technology. Vol. 57(3), 375.

Tenenhaus, M. (1998). La re´ gression PLS, the´ orie et pratique. Technip, Paris (France), p. 254.

Thomas O., Theraulaz F., Agnel C., Suryani S., (1996). Advanced UV examination of wastewater. Environmental Technology. Vol. 17: 251-261.

Thomas, O y Burgess, C. (2007) UV-visible Spectrophotometry of Water and Wastewater (O Thomas and C. Burgess, eds.), Amsterdam, Netherlands, Elsevier.

Thomas, O, Theraulaz, F., Agnel, C., y Suryani, S. (1996) Advanced UV Examination of Wastewater. Environmental Technology, 17(3), 251–261.

Thomas, Olivier, Baurès, Estelle, y Pouet, M. (2005). UV Spectrophotometry as a Non-parametric Measurement of Water and Wastewater Quality Variability. 40(1), 51–58.

Threllfall, J.L., Hyde, J., y Crabtree, R.W. (1991). Sewer quality archive data analysis. Foundation for Water Research, Rep. No. FR0203. London, UK, Foundattion of Water Research, 31 p.

Tobias, R. (1995). An Introduction to Partial Least Squares Regression. Proceedings of the Twentieth Annual SAS Users Group International Conference. SAS Institute Inc, Cary, NC (USA), pp. 1250–1257.

Torres, A. (2011). Metodología para la Estimación de Incertidumbres Asociadas a Concentraciones de Sólidos Suspendidos Totales Mediante Métodos de Generación Aleatoria Resumen. Instituto Tecnológico Metropolitano, 8(26), 181–200.

Torres, A. y 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), 581–8..

V. Cherkassky y F. Mulier, Learning from Data: Concepts, Theory, and Methods. (John Wiley & Sons, 1998)

Vaillant, S., Pouet, M. F., y Thomas, O (1999) Methodology for the characterisation of heterogeneous fractions in wastewater. Talanta, 50(4), 729–36.

Vanrolleghem, P. y Lee, D. S. (2003) On-line monitoring equipment for wastewater treatment processes: state of the art. Water science and technology : a journal of the International Association on Water Pollution Research, 47(2), 1–34.

Vapnik, V.N. (1998). “Statistical Learning Theory”. Wiley. New York, USA.

Varmuza, K. y Filzmoser, P. (2009) Introducction to multivariate statistical analysis in chemometrics, Boca Raton, FL, Taylor & Francis - CRC Press.

Velásquez, J., Olaya, Y., y Franco, C. (2009) Predicción de series temporales usando máquinas de vectores de soporte. Ingeniare. Revista chilena de ingeniería, 18(1), 64–75.

Venables, W. N., Ripley, B. D., y Venables, W. N. (1994). Modern applied statistics with S-PLUS (Vol. 250). New York: Springer-verlag.

Villalba, M. M., McKeegan, K. J., Vaughan, D. H., Cardosi, M. F., y Davis, J. (2009) Bioelectroanalytical determination of phosphate: A review. Journal of Molecular Catalysis B: Enzymatic, 59(1-3), 1–8.

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. In: IWA (ed.), Preprints of the 3rd IWA International World Water Congress, 7–12 April 2002, Melbourne, Australia.

Winkler, S., Rieger, L., Saracevic, E., Pressl, a, y Gruber, G. (2004) Application of ion-sensitive sensors in water quality monitoring. Water science and technology : a journal of the International Association on Water Pollution Research, 50(11), 105–14.

Winkler, S., Saracevic, E., Bertrand-Krajewski, J.-L., y Torres, A. (2008). Benefits, limitations and uncertainty of in situ spectrometry. Water science and technology, 57(10), 1651–8.

Witten I., Frank E. (2005). Data Mining – Practical Machine Learning Tools and Techniques.

Witten, I. y Frank, E. (2005) Data Mining: Practical machine learning tools and techniques, San Francisco, USA, Morgan Kaufman Publishers.

Zamora, D. y Torres A. (2013). Detection method in outliers of spectrometry UV-Visible databases: preliminary phase for calibration models applied to regressive real-time monitoring of water quality. Conference Latin American and Caribbean Consortium of Engineering Institutions. Cancum, Mexico. 14-16 Agosto 2013.

Zamora, D. y Torres A. (2012b). Method for outliers detection: a tool to assess the consistency between laboratory data and UV-visible absorbance spectra in wastewater samples. New Developments in IT & Water conference. Amsterdam, The Netherlands. 4-6 of November 2012. Actualmente evaluado por la revista internacional: Water Science and Technology.

Zamora, D. y Torres A. (2012a). Proposal for recurrence, level of importance and quality detection of UV-Vis spectra and target pollutant dataset. New Developments in IT & Water conference. Amsterdam, Netherlands. 4-6 of November 2012 (poster y artículo).

Zamora, D., Rossman, T., y Torres, A. (2011.) Redes neuronales aplicadas a la espectrometría UV-Visible. XIX Seminario Nacional de Hidráulica e Hidrología y el I Foro Nacional sobre la Seguridad de Embalses. Bogotá D.C., 24 y 25 de marzo.

ANEXOS

ANEXO A

Espectros UV-Vis con valores medios y mínimos de absorbancia para las muestra de la PTAR (época seca y lluvia), y del afluente de la estación elevadora de Gibraltar.

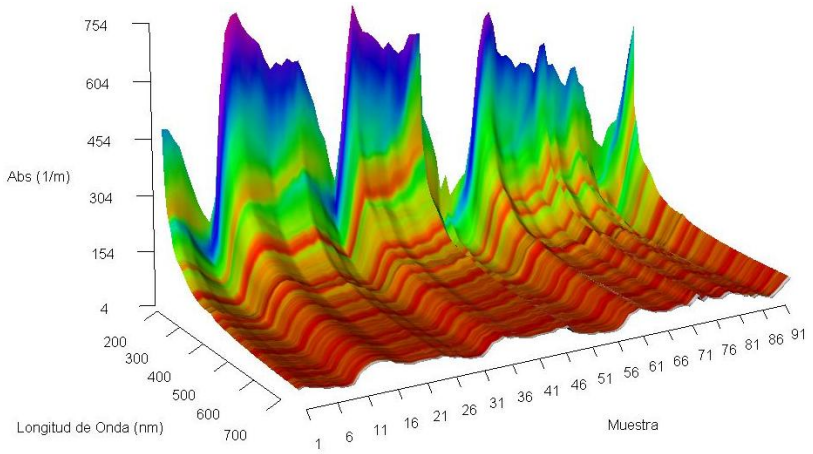


Figura 91- Espectros UV-Vis con los valores medios de absorbancia del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

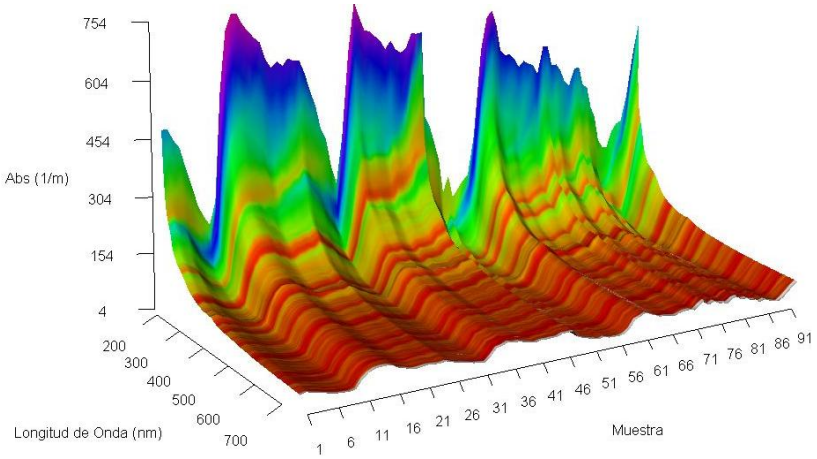


Figura 92- Espectros UV-Vis con los valores mínimos de absorbancia del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

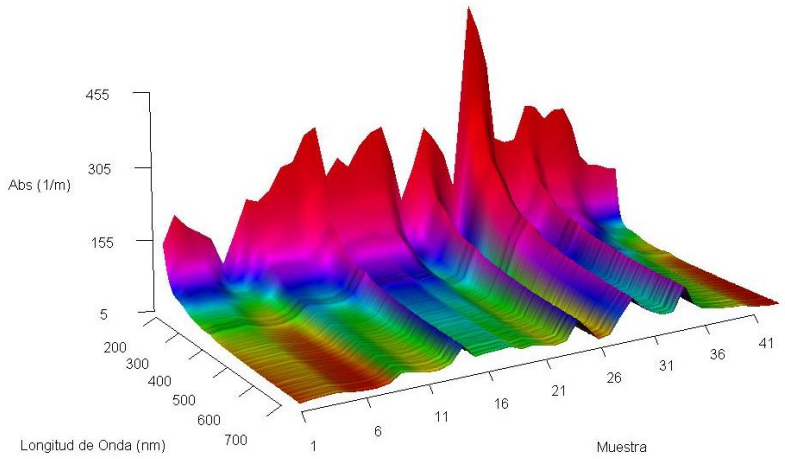


Figura 93- Espectros UV-Vis con los valores mínimos de absorbancia del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia)

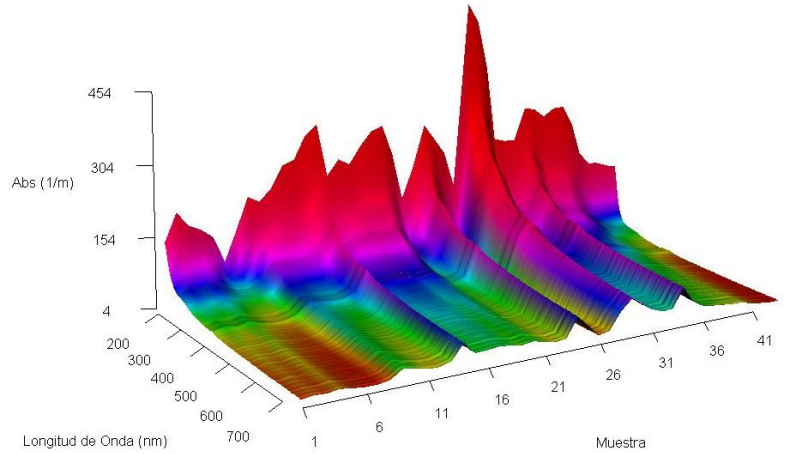


Figura 94- Espectros UV-Vis con los valores medios de absorbancia del afluente de la PTAR de *Fontaines-sur-Saône* (Epoca lluvia)

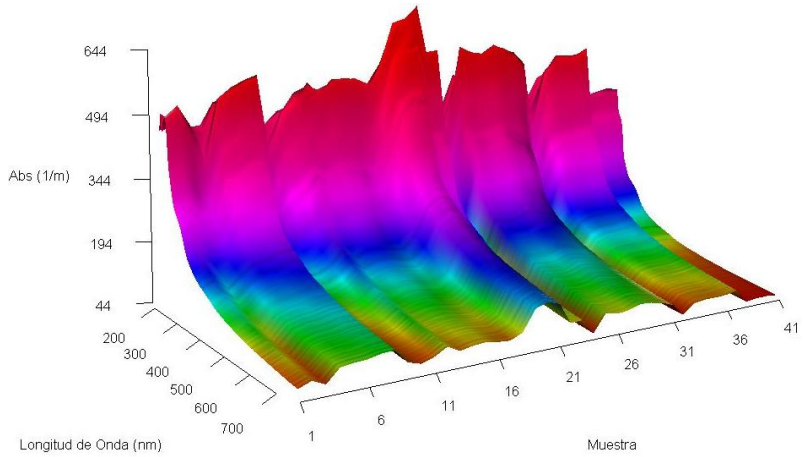


Figura 95- Espectros UV-Vis con los valores medios de absorbancia del afluente de la estación elevadora de Gibraltar

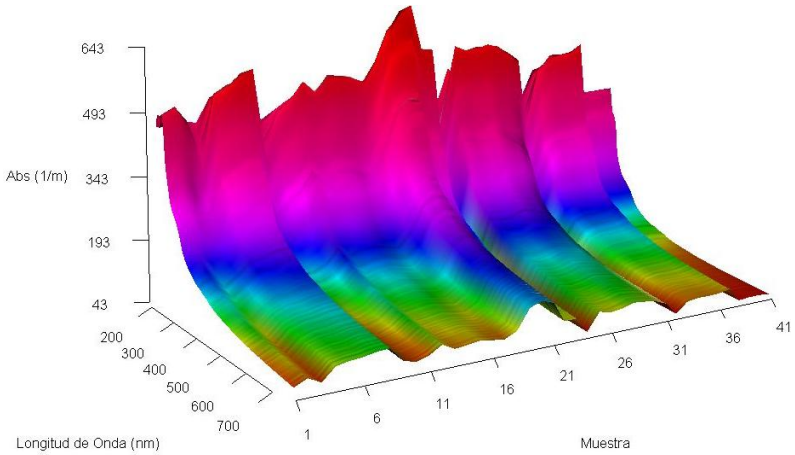


Figura 96- Espectros UV-Vis con los valores mínimos de absorbancia del afluente de la estación elevadora de Gibraltar

ANEXO B

PTAR de *Fontaines-sur-Saône* (Tiempo Seco)

Outliers en conjunto de datos de DQO

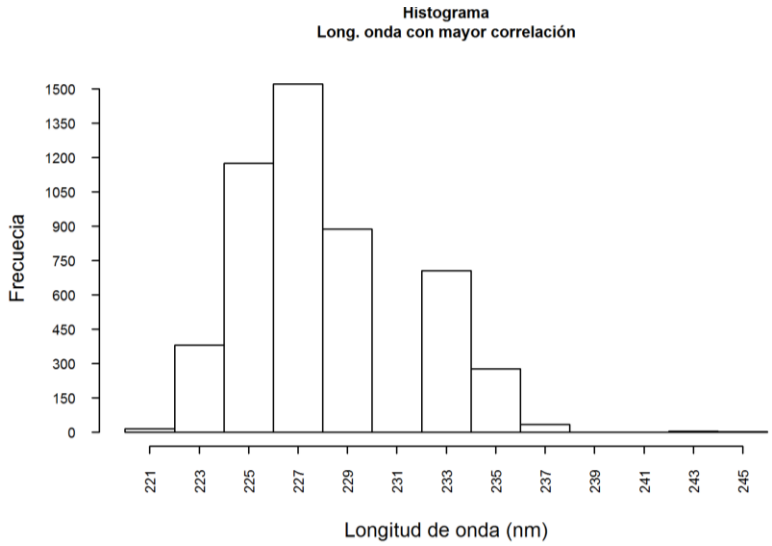


Figura 97- Longitudes de onda con mayor correlación para identificar la presencia de la DQO en las 5000 simulaciones de Monte Carlo del afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo seco)

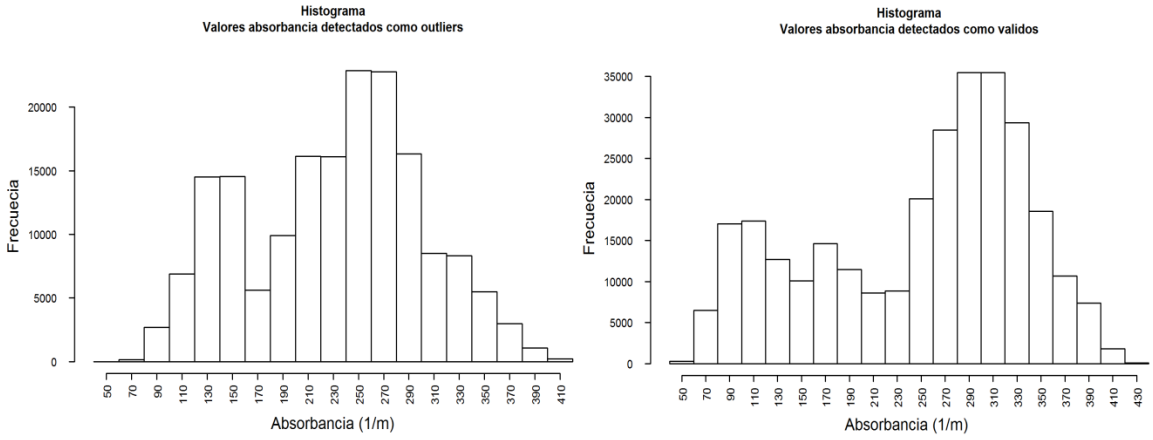


Figura 98- Histograma de los valores de absorbancia detectados como *outliers* (Der.) y de los valores establecidos como válidos (Izq.) en el conjunto de muestras del afluente de la PTAR de *Fontaines-sur-Saône* (DQO-Tiempo seco)

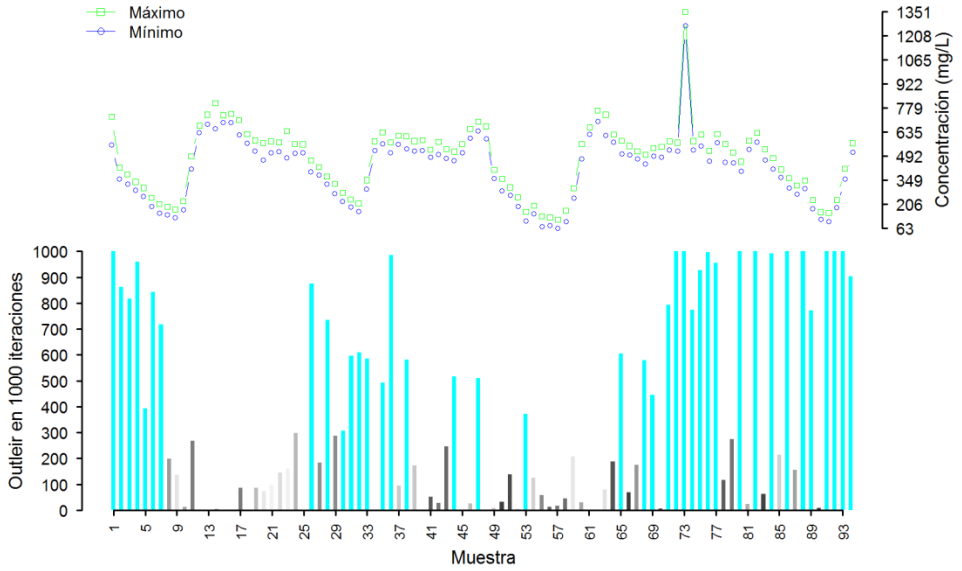


Figura 99- En la gráfica superior se presenta el valor máximo y mínimo de la concentración de la DQO de cada muestra generada en 1000 simulaciones de Monte Carlo y en el gráfico inferior se presentan en color cian las muestras que en 300 o más simulaciones de 1000 fueron catalogados como *outliers*

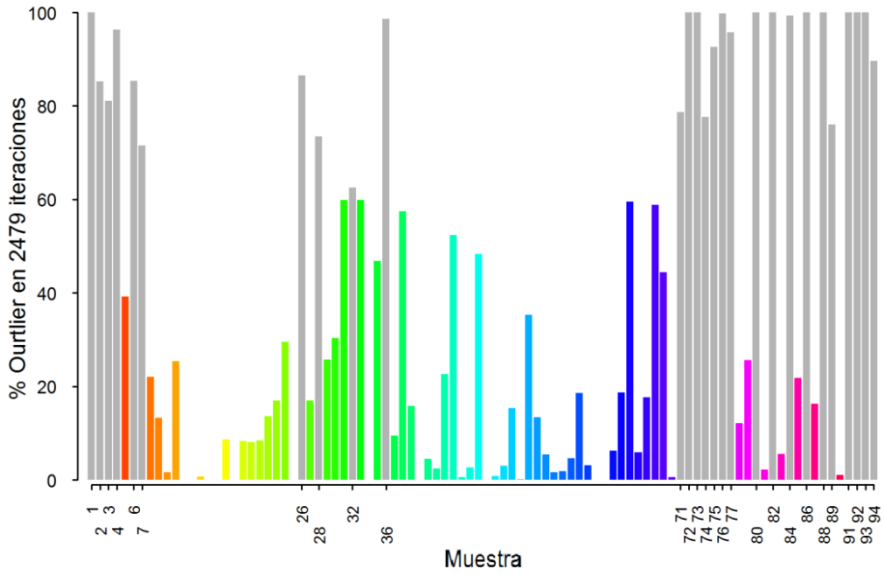


Figura 100- Porcentaje que una muestra fue catalogada como *outlier* en 5000 simulaciones de Monte Carlo en el caso de las concentraciones de la DQO en el afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo seco). En gris se presenta las muestras catalogas como *outliers* en 60 % o más de las simulaciones generadas

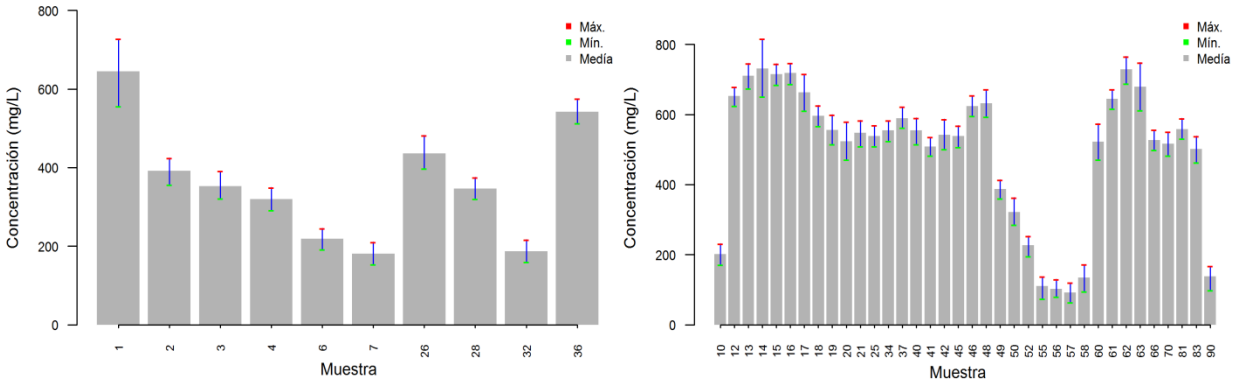


Figura 101- A la izquierda se presentan los rangos de concentración de las muestras catalogadas como *outliers* en más del 10 % de las 5000 simulaciones de Monte Carlo y a la derecha las muestras que estuvieron por debajo de dicho porcentaje catalogadas como datos validos (DQO de la PTAR de *Fontaines-sur-Saône* en tiempo seco)

Outliers en conjunto de datos de DQOf

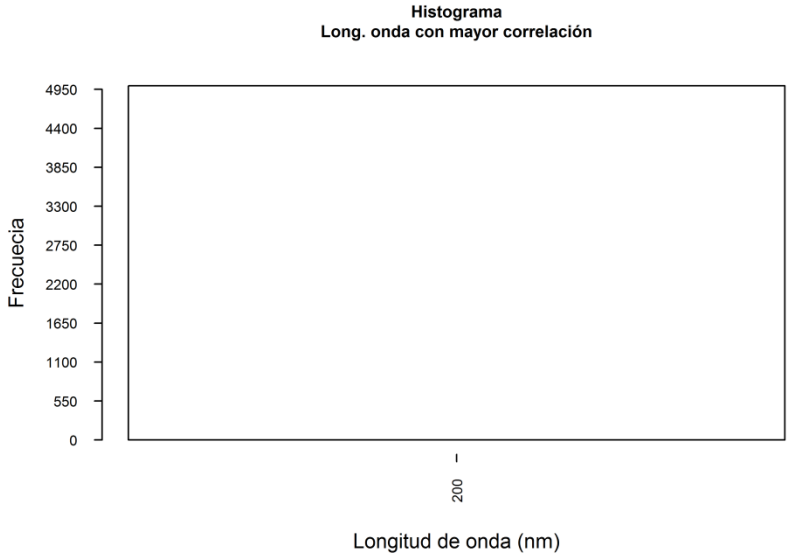


Figura 102- Longitudes de onda con mayor correlación para identificar la presencia de la DQOf en las 5000 simulaciones de Monte Carlo del afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo seco)

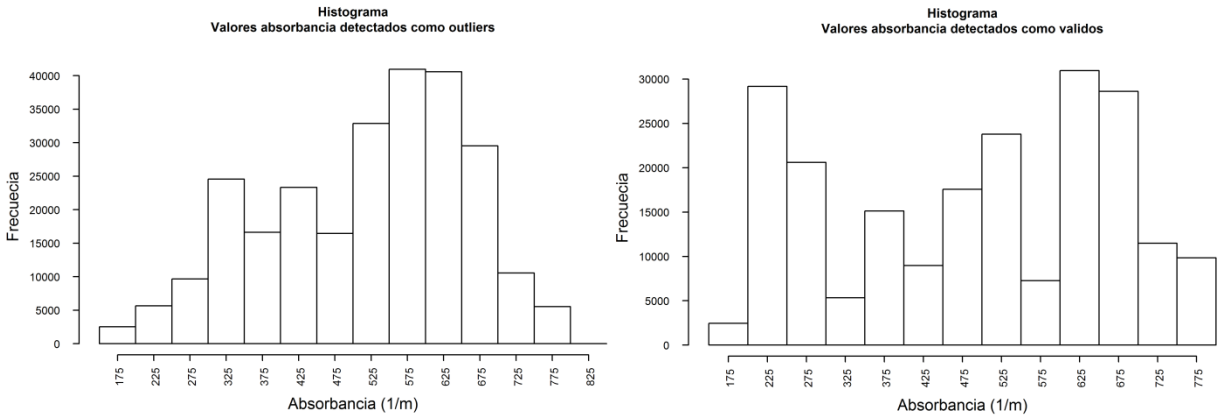


Figura 103- Histograma de los valores de absorbancia detectados como *outliers* (Der.) y de los valores establecidos como válidos (Izq.) en el conjunto de muestras del afluente de la PTAR de *Fontaines-sur-Saône* (DQOf-Tiempo seco)

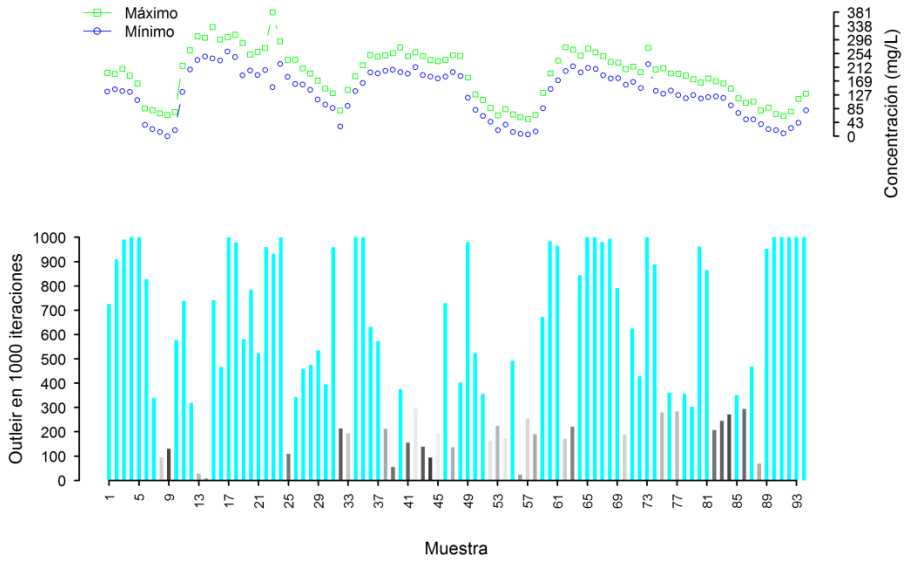


Figura 104- En la gráfica superior se presenta el valor máximo y mínimo de la concentración de la DQOf de cada muestra generada en 1000 simulaciones de Monte Carlo y en el gráfico inferior se presentan en color cian las muestras que en 300 o más simulaciones de 1000 fueron catalogados como *outliers*

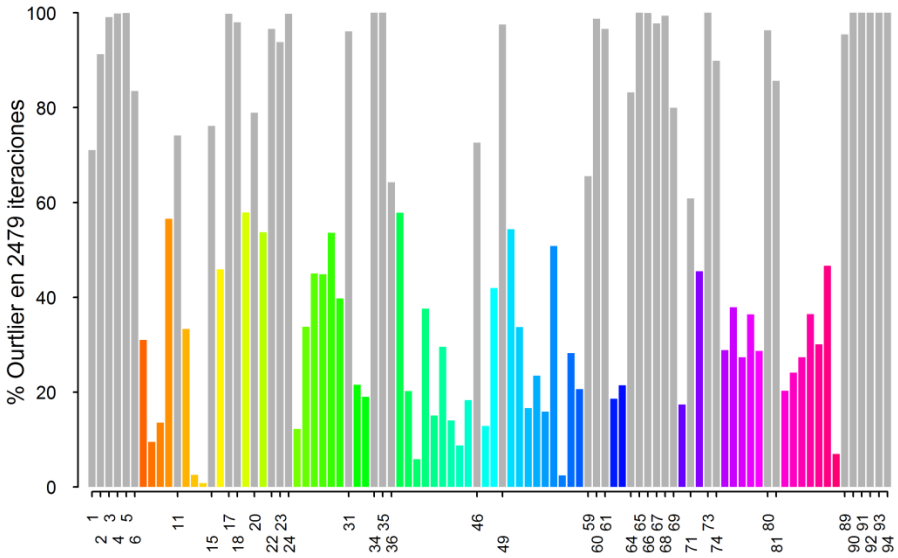


Figura 105- Porcentaje que una muestra fue catalogada como *outlier* en 5000 simulaciones de Monte Carlo en el caso de las concentraciones de la DQOf en el afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo seco). En gris se presenta las muestras catalogas como *outliers* en 60 % o más de las simulaciones generadas

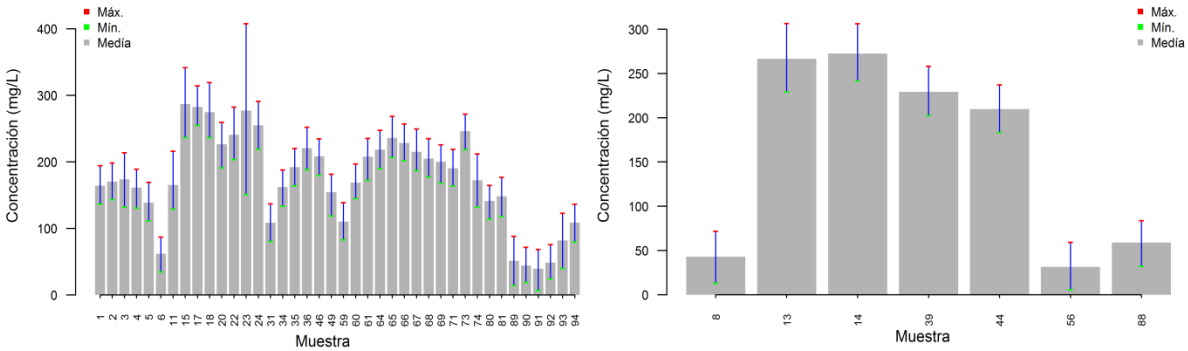


Figura 106- A la izquierda se presentan los rangos de concentración de las muestras catalogadas como *outliers* en más del 10 % de las 5000 simulaciones de Monte Carlo y a la derecha las muestras que estuvieron por debajo de dicho porcentaje catalogadas como datos validos (DQOf de la PTAR de *Fontaines-sur-Saône* en tiempo seco)

PTAR de *Fontaines-sur-Saône* (Tiempo Lluvia)

Outliers en conjunto de datos de SST

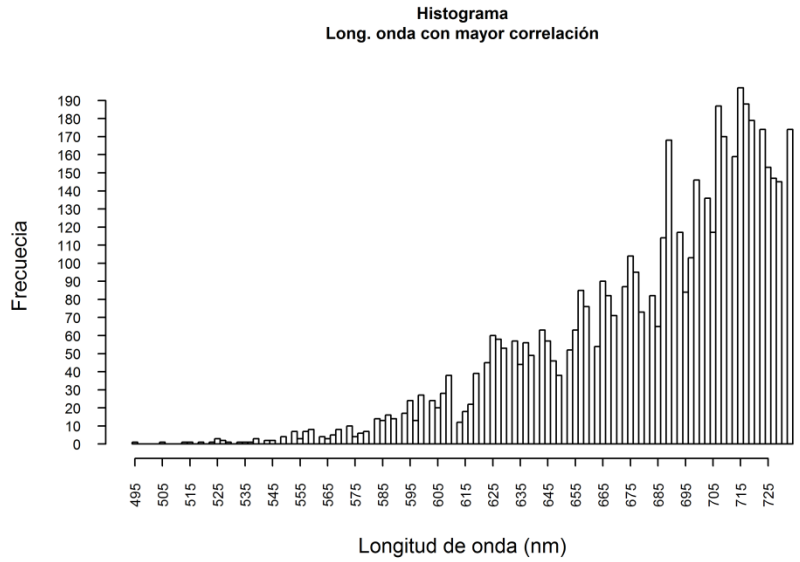


Figura 107- Longitudes de onda con mayor correlación para identificar la presencia de la SST en las 5000 simulaciones de Monte Carlo del afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo lluvia)

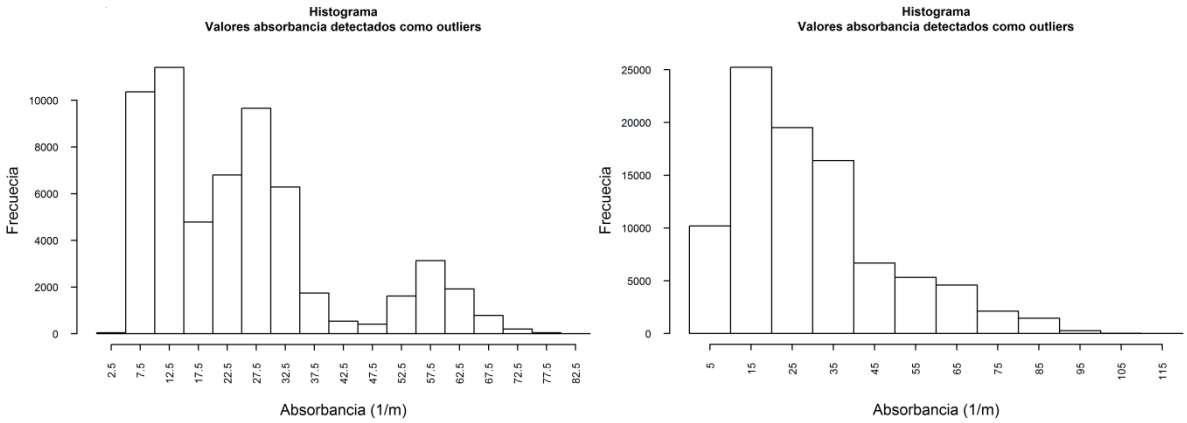


Figura 108- Histograma de los valores de absorbancia detectados como *outliers* (Der.) y de los valores establecidos como válidos (Izq.) en el conjunto de muestras del afluente de la PTAR de *Fontaines-sur-Saône* (SST-Tiempo lluvia)

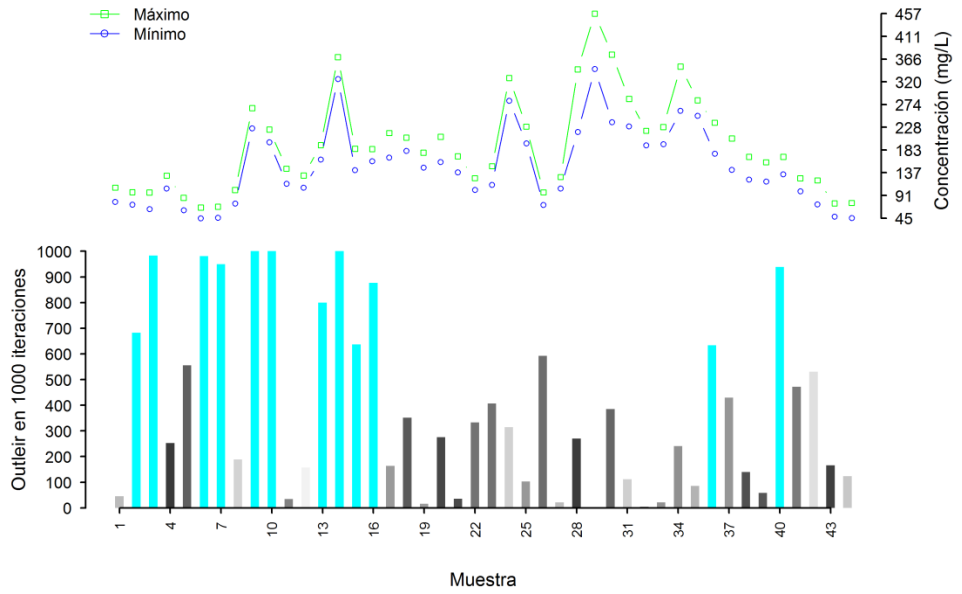


Figura 109- En la gráfica superior se presenta el valor máximo y mínimo de la concentración de la SST de cada muestra generada en 1000 simulaciones de Monte Carlo y en el gráfico inferior se presentan en color cian las muestras que en 300 o más simulaciones de 1000 fueron catalogados como *outliers*

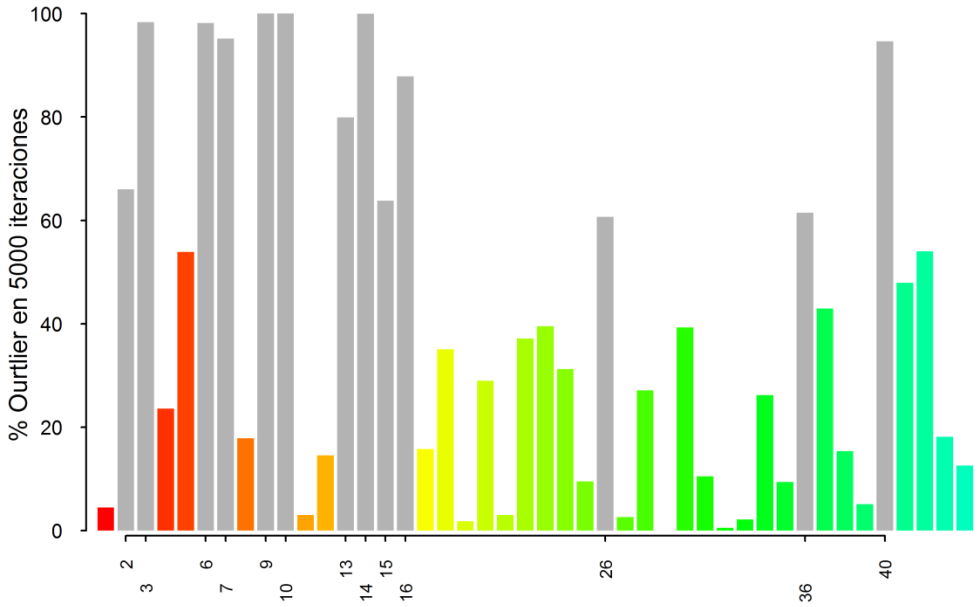


Figura 110- Porcentaje que una muestra fue catalogada como *outlier* en 5000 simulaciones de Monte Carlo en el caso de las concentraciones de la SST en el afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia). En gris se presenta las muestras catalogas como *outliers* en 60 % o más de las simulaciones generadas

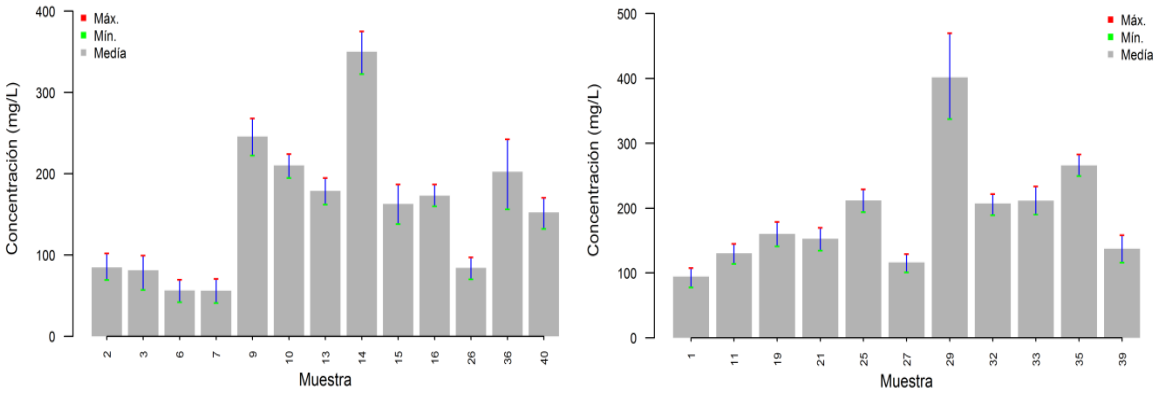


Figura 111- A la izquierda se presentan los rangos de concentración de las muestras catalogadas como *outliers* en más del 10 % de las 5000 simulaciones de Monte Carlo y a la derecha las muestras que estuvieron por debajo de dicho porcentaje catalogadas como datos validos (SST de la PTAR de *Fontaines-sur-Saône* en tiempo lluvia)

Outliers en conjunto de datos de DQO

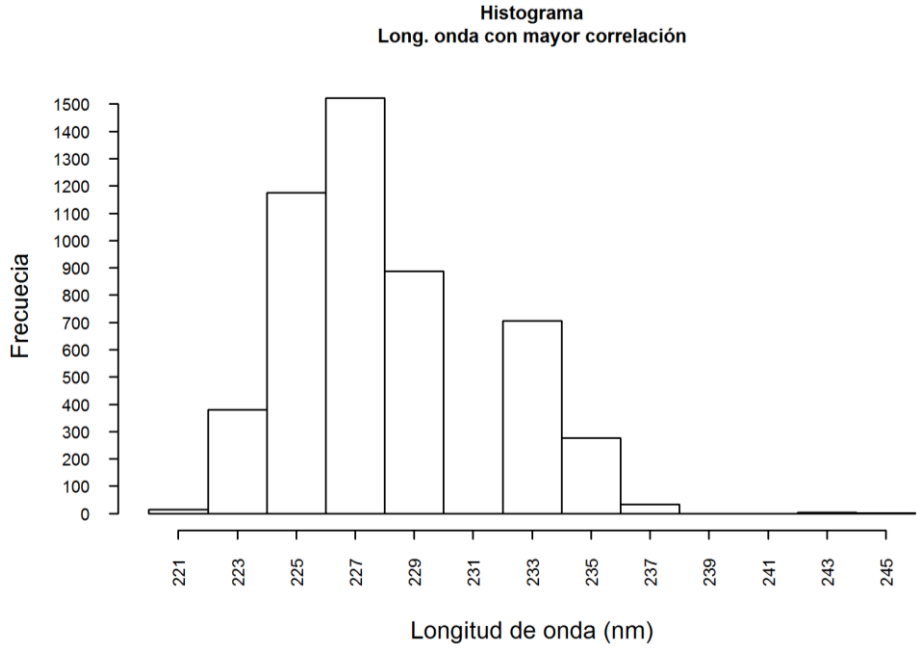


Figura 112- Longitudes de onda con mayor correlación para identificar la presencia de la DQO en las 5000 simulaciones de Monte Carlo del afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo lluvia)

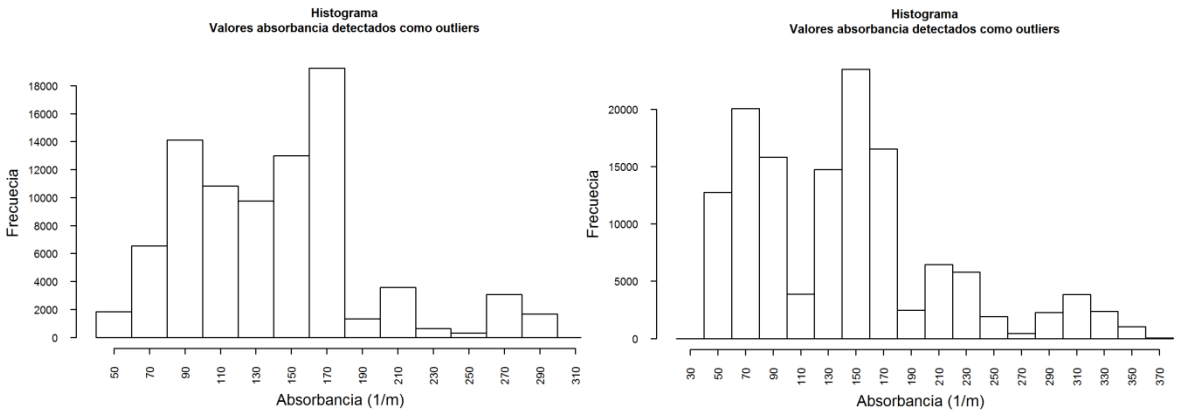


Figura 113- Histograma de los valores de absorbancia detectados como *outliers* (Der.) y de los valores establecidos como valores validos (Izq.) en el conjunto de muestras del afluente de la PTAR de *Fontaines-sur-Saône* (DQO-Tiempo lluvia)

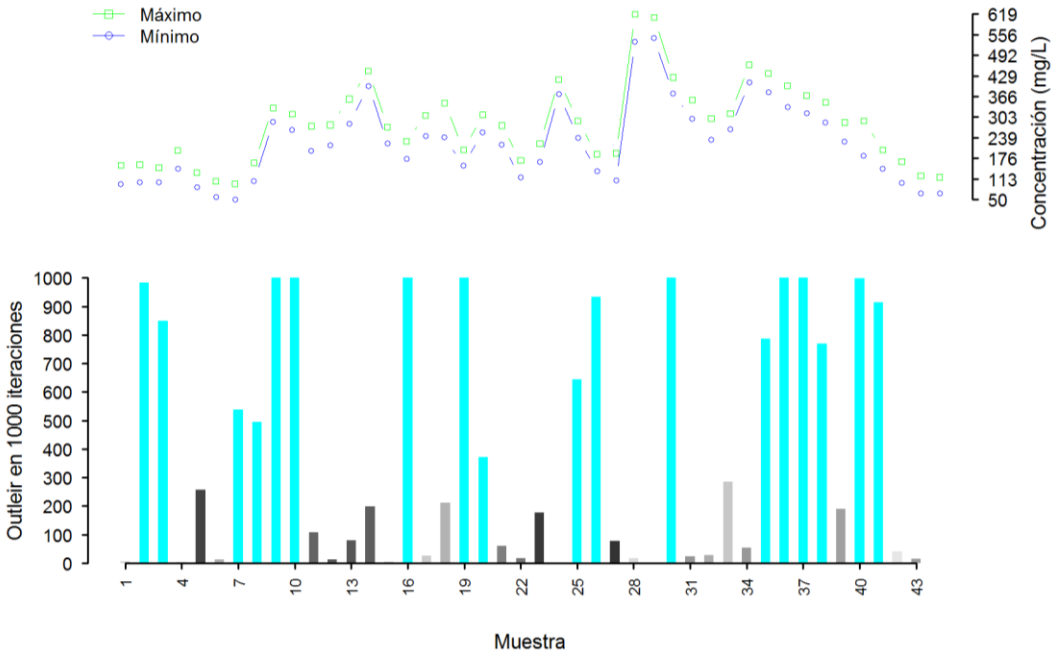


Figura 114- En la gráfica superior se presenta el valor máximo y mínimo de la concentración de la DQO de cada muestra generada en 1000 simulaciones de Monte Carlo y en el gráfico inferior se presentan en color cian las muestras que en 300 o más simulaciones de 1000 fueron catalogados como *outliers*

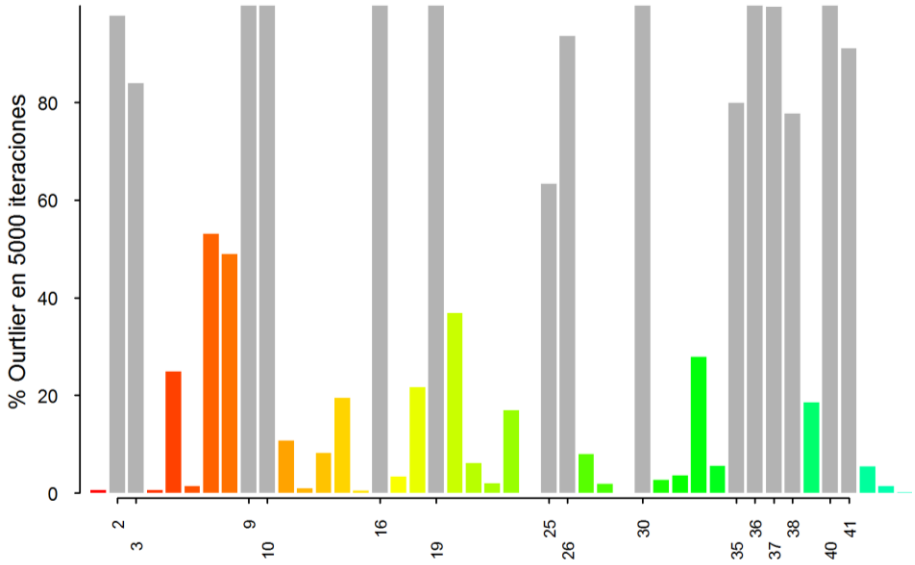


Figura 115- Porcentaje que una muestra fue catalogada como *outlier* en 5000 simulaciones de Monte Carlo en el caso de las concentraciones de la DQO en el afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia). En gris se presenta las muestras catalogas como *outliers* en 60 % o más de las simulaciones generadas

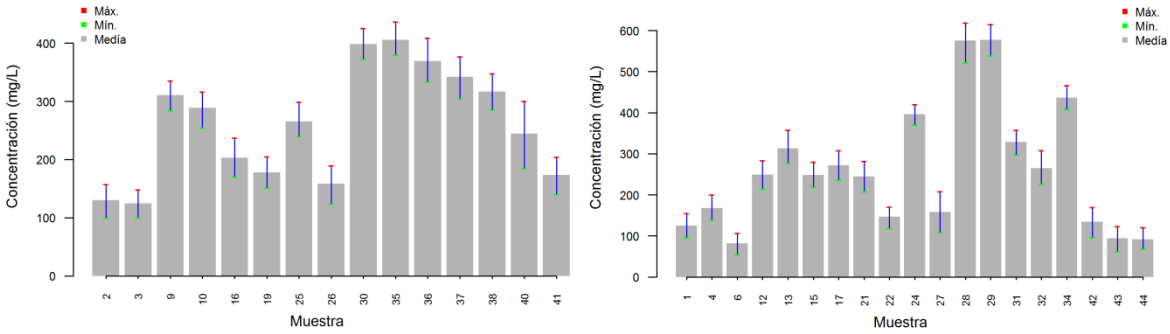


Figura 116- A la izquierda se presentan los rangos de concentración de las muestras catalogadas como *outliers* en más del 10 % de las 5000 simulaciones de Monte Carlo y a la derecha las muestras que estuvieron por debajo de dicho porcentaje catalogadas como datos validos (DQO de la PTAR de *Fontaines-sur-Saône* en tiempo lluvia)

Outliers en conjunto de datos de DQOf

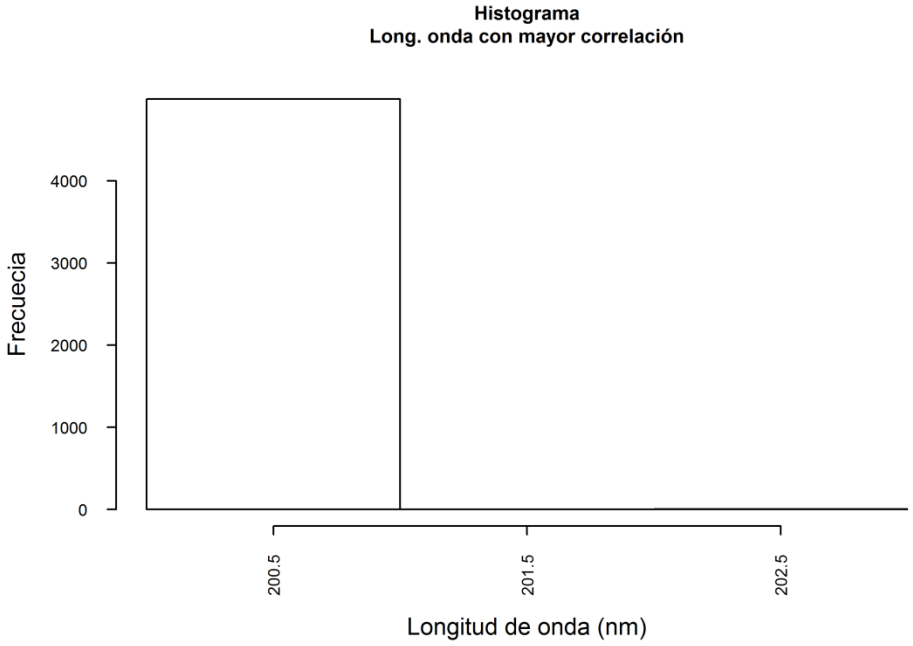


Figura 117- Longitudes de onda con mayor correlación para identificar la presencia de la DQOf en las 5000 simulaciones de Monte Carlo del afluente de la PTAR de *Fontaines-sur-Saône* (Tiempo lluvia)

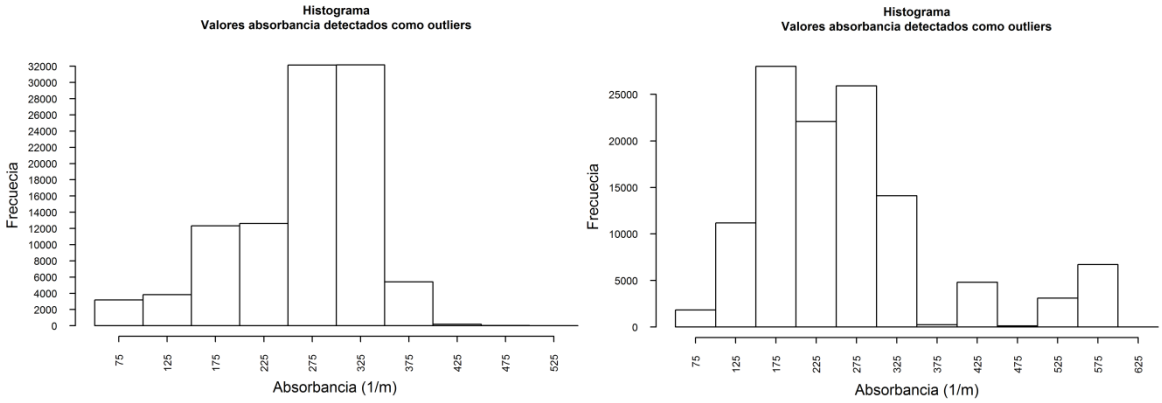


Figura 118- Histograma de los valores de absorbancia detectados como *outliers* (Der.) y de los valores establecidos como valores validos (Izq.) en el conjunto de muestras del afluente de la PTAR de *Fontaines-sur-Saône* (DQOf-Tiempo lluvia)

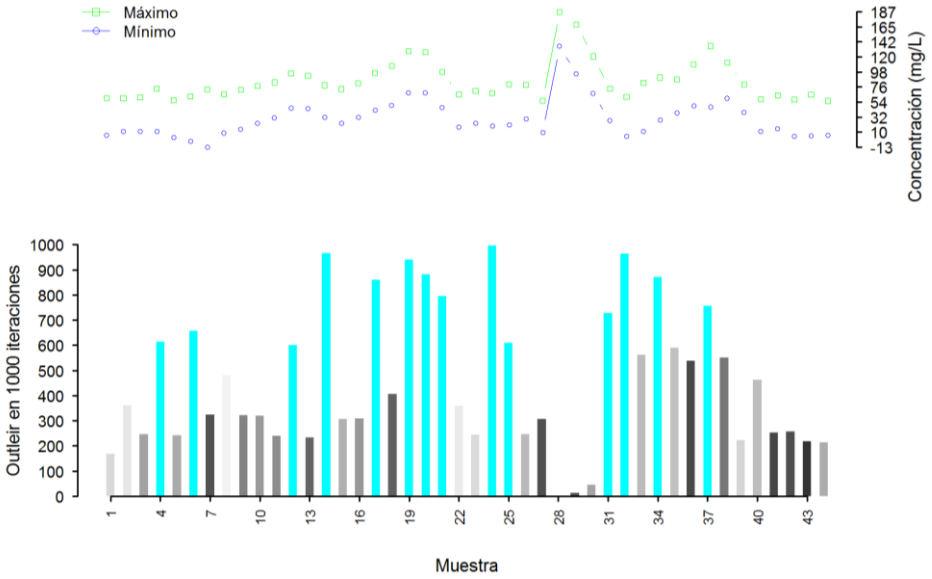


Figura 119- En la gráfica superior se presenta el valor máximo y mínimo de la concentración de la DQO de cada muestra generada en 1000 simulaciones de Monte Carlo y en el gráfico inferior se presentan en color cian las muestras que en 300 o más simulaciones de 1000 fueron catalogados como *outliers*

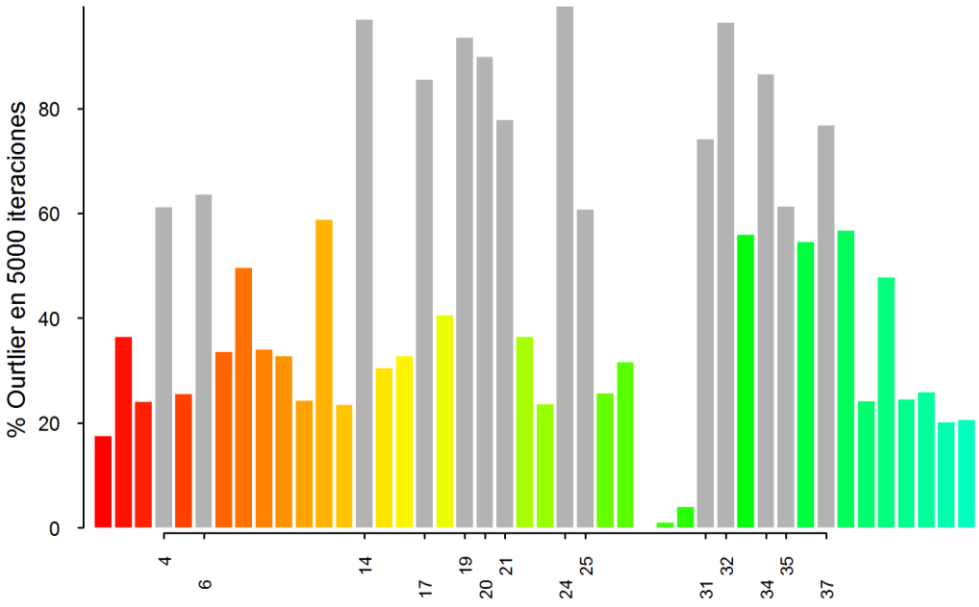


Figura 120- Porcentaje que una muestra fue catalogada como *outlier* en 5000 simulaciones de Monte Carlo en el caso de las concentraciones de la DQOf en el afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia). En gris se presenta las muestras catalogas como *outliers* en 60 % o más de las simulaciones generadas

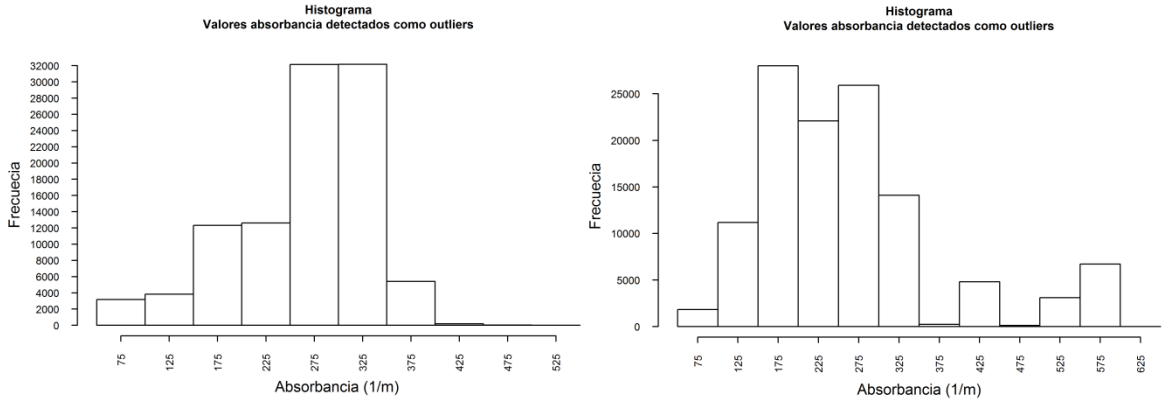


Figura 121- A la izquierda se presentan los rangos de concentración de las muestras catalogadas como *outliers* en más del 10 % de las 5000 simulaciones de Monte Carlo y a la derecha las muestras que estuvieron por debajo de dicho porcentaje catalogadas como datos validos (DQOf de la PTAR de *Fontaines-sur-Saône* en tiempo lluvia)

Outliers en conjunto de datos de los SST

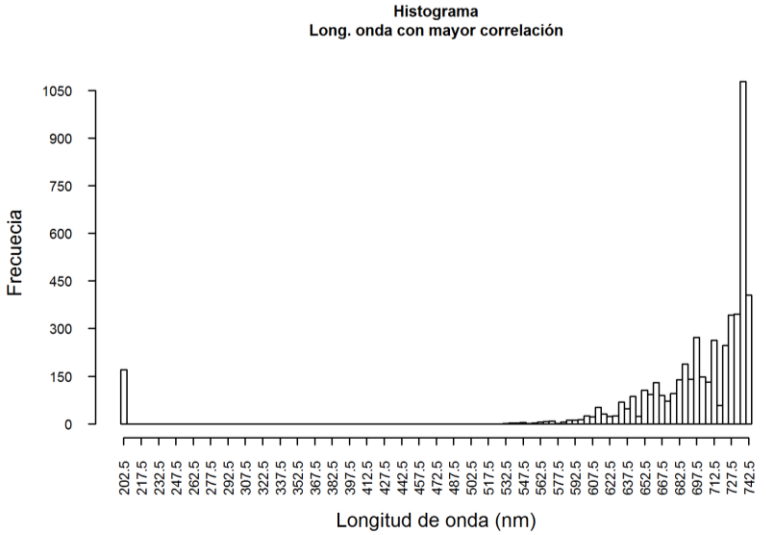


Figura 122- Longitudes de onda con mayor correlación para identificar la presencia de los SST en las 5000 simulaciones de Monte Carlo del afluente de la EE de Gibraltar

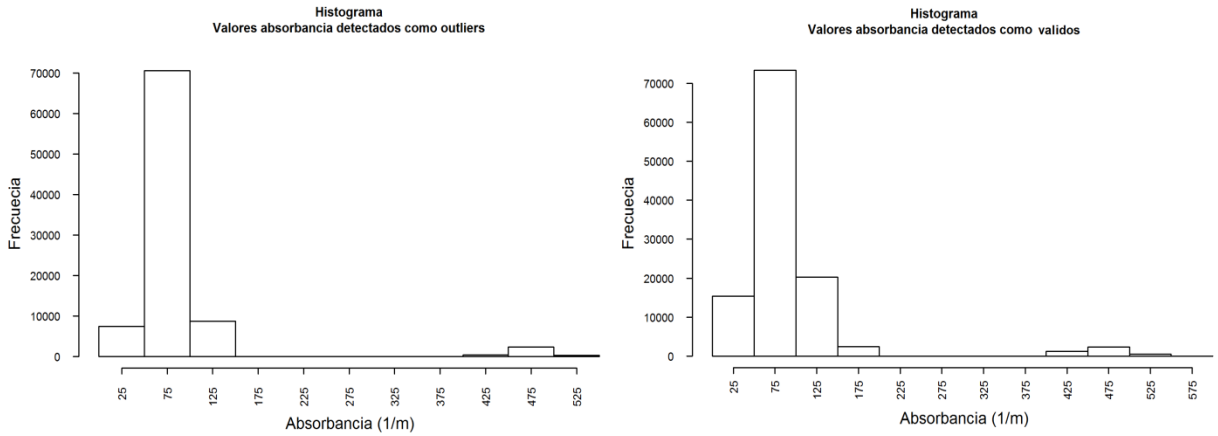


Figura 123- Histograma de los valores de absorbancia detectados como *outliers* (Der.) y de los valores establecidos como válidos (Izq.) en el conjunto de muestras del afluente de la EE de Gibraltar

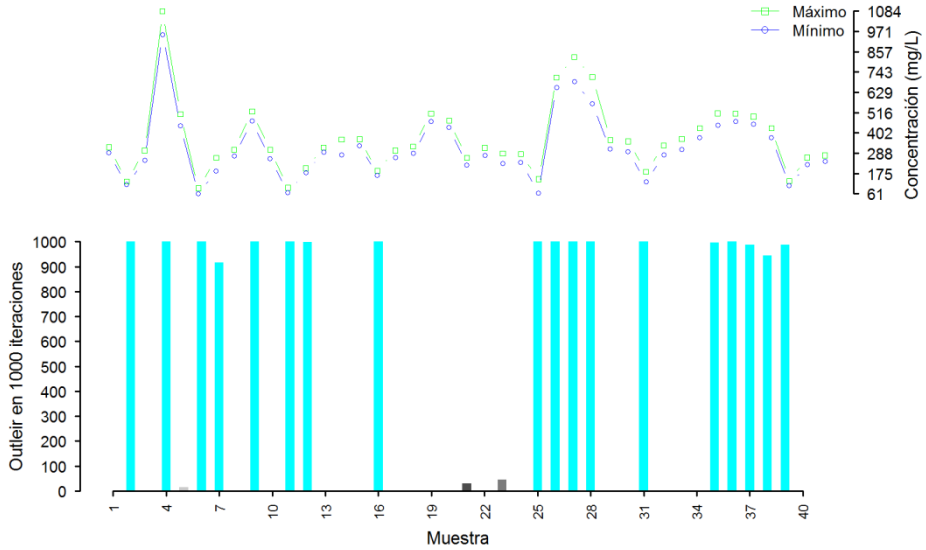


Figura 124- En la gráfica superior se presenta el valor máximo y mínimo de la concentración de los SST de cada muestra generada en 1000 simulaciones de Monte Carlo y en el gráfico inferior se presentan en color cian las muestras que en 300 o más simulaciones de 1000 fueron catalogados como *outliers* (EE de Gibraltar)

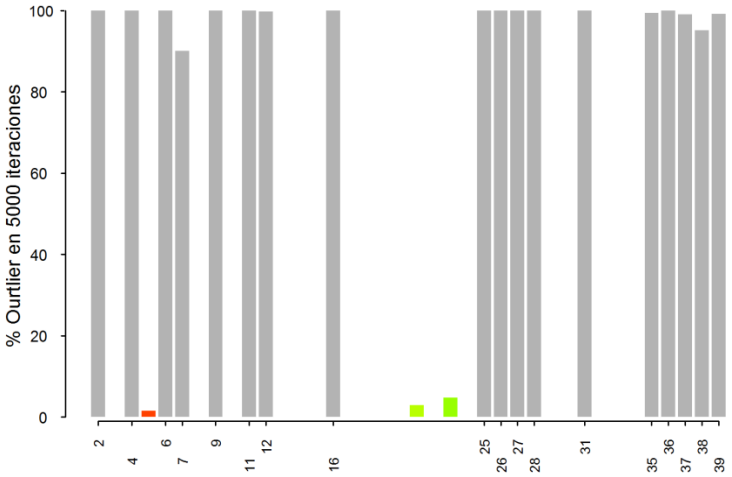


Figura 125- Porcentaje que una muestra fue catalogada como *outlier* en 5000 simulaciones de Monte Carlo en el caso de las concentraciones de SST en el afluente de la EE de Gibraltar. En gris se presenta las muestras catalogas como *outliers* en 60 % o más de las simulaciones generadas

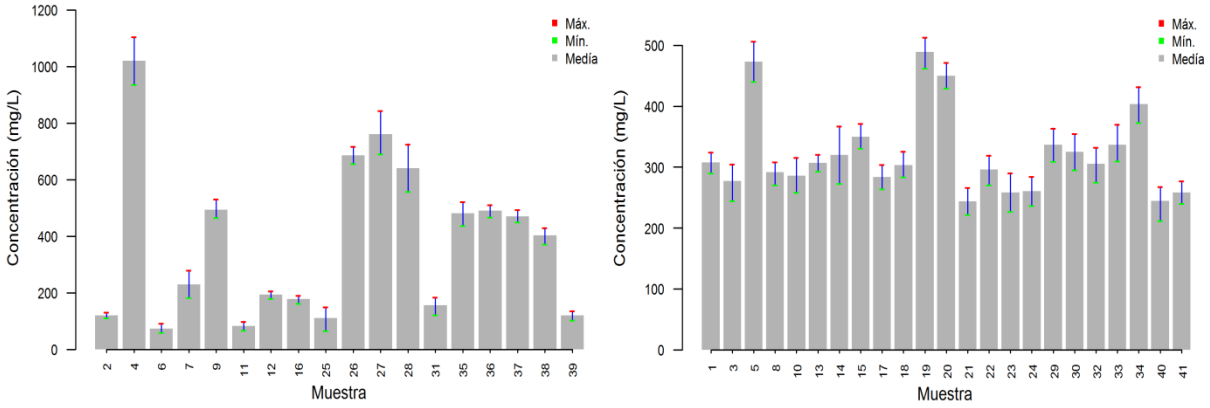


Figura 126- A la izquierda se presentan los rangos de concentración de las muestras catalogadas como *outliers* en más del 10 % de las 5000 simulaciones de Monte Carlo y a la derecha las muestras que estuvieron por debajo de dicho porcentaje catalogadas como datos validos (EE de Gibraltar)

Outliers en conjunto de datos de DQO

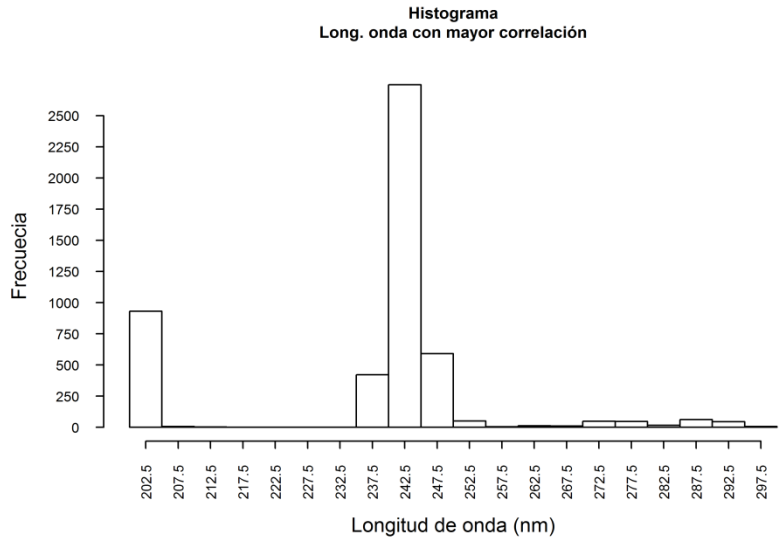


Figura 127- Longitudes de onda con mayor correlación para identificar la presencia de la DQO en las 5000 simulaciones de Monte Carlo del afluente de la EE de Gibraltar

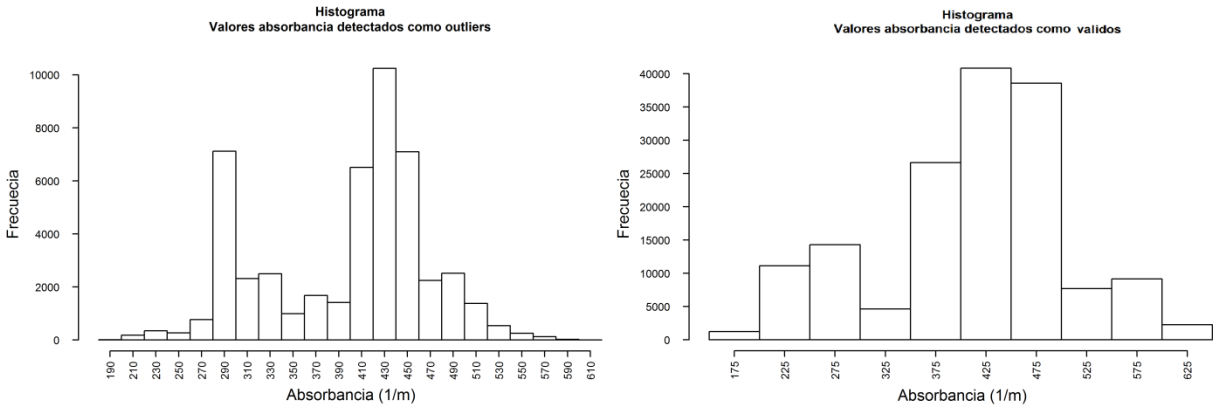


Figura 128- Histograma de los valores de absorbancia detectados como *outliers* (Der.) y de los valores establecidos como válidos (Izq.) en el conjunto de muestras del afluente de la EE de Gibraltar

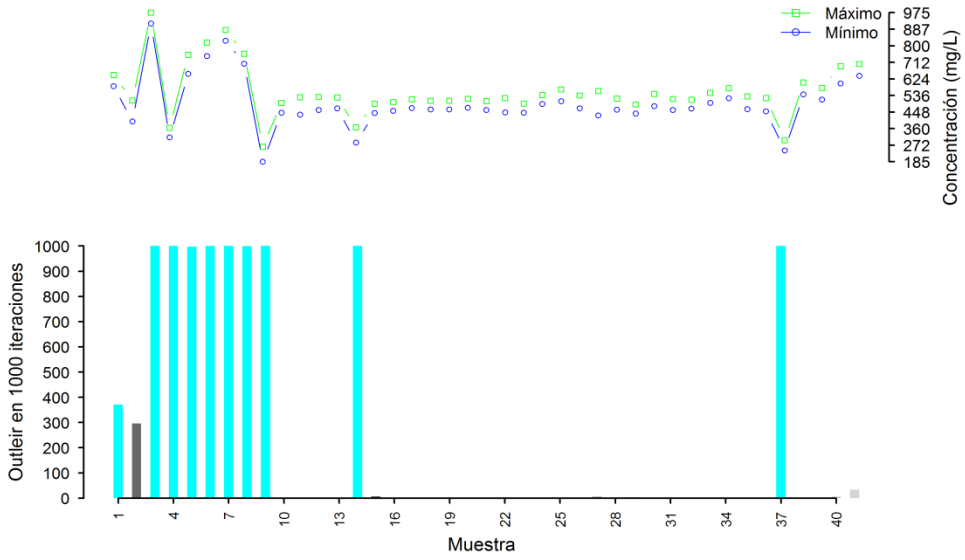


Figura 129- En la gráfica superior se presenta el valor máximo y mínimo de la concentración de la DQO de cada muestra generada en 1000 simulaciones de Monte Carlo y en el gráfico inferior se presentan en color cian las muestras que en 300 o más simulaciones de 1000 fueron catalogados como *outliers* (EE de Gibraltar)

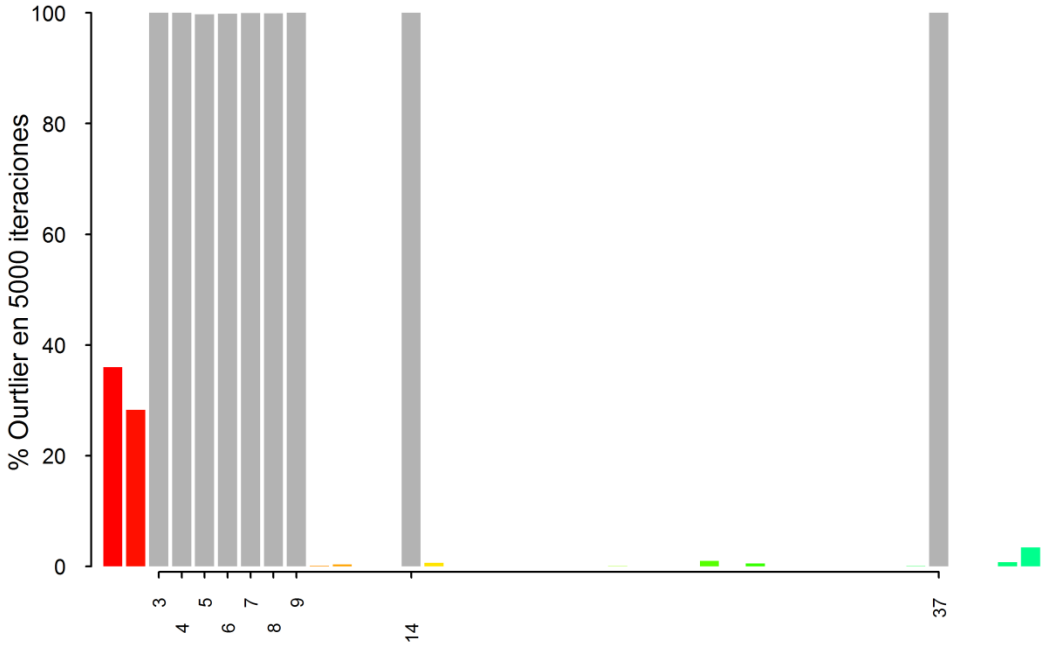


Figura 130- Porcentaje que una muestra fue catalogada como *outlier* en 5000 simulaciones de Monte Carlo en el caso de las concentraciones de la DQO en el afluente de la EE de Gibraltar. En gris se presenta las muestras catalogas como *outliers* en 60 % o más de las simulaciones generadas

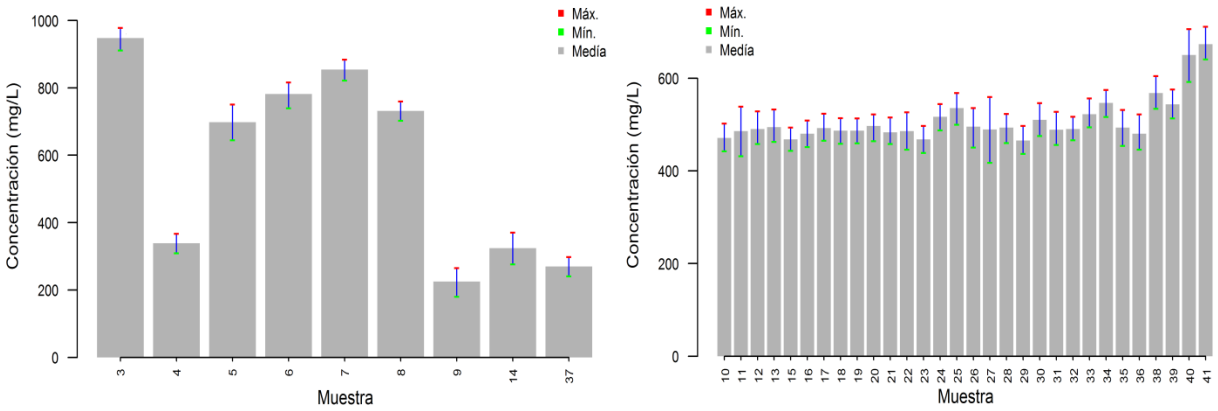


Figura 131- A la izquierda se presentan los rangos de concentración de las muestras catalogadas como *outliers* en más del 10 % de las 5000 simulaciones de Monte Carlo y a la derecha las muestras que estuvieron por debajo de dicho porcentaje catalogadas como datos validos (DQO-EE de Gibraltar)

Outliers en conjunto de datos de DQOf

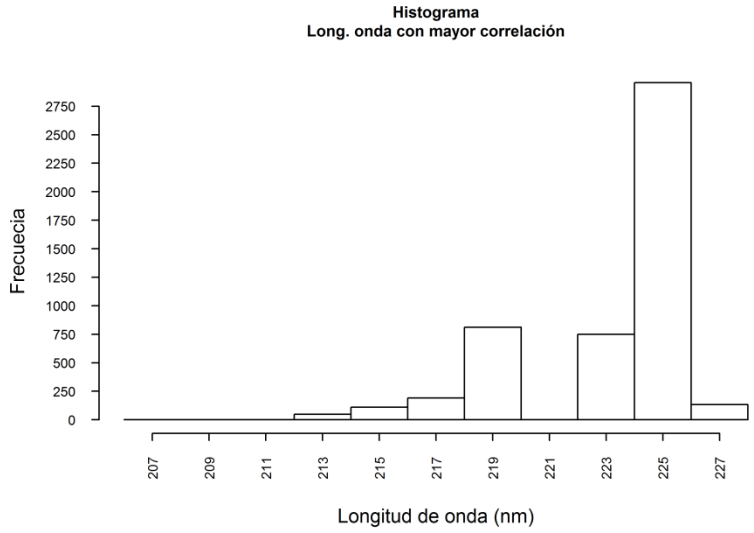


Figura 132- Longitudes de onda con mayor correlación para identificar la presencia de los DQOf en las 5000 simulaciones de Monte Carlo del afluente de la EE de Gibraltar (DQOf)

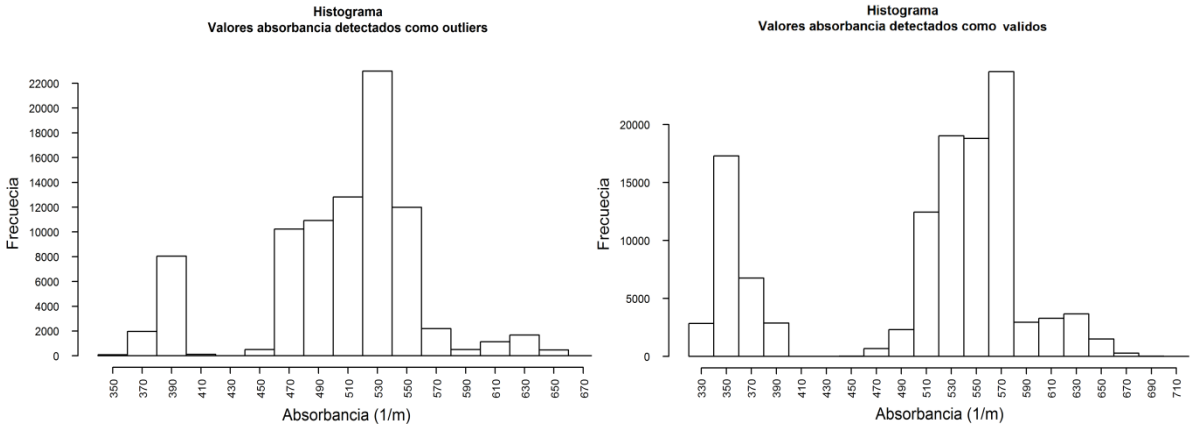


Figura 133- Histograma de los valores de absorbancia detectados como *outliers* (Der.) y de los valores establecidos como válidos (Izq.) en el conjunto de muestras del afluente de la EE de Gibraltar

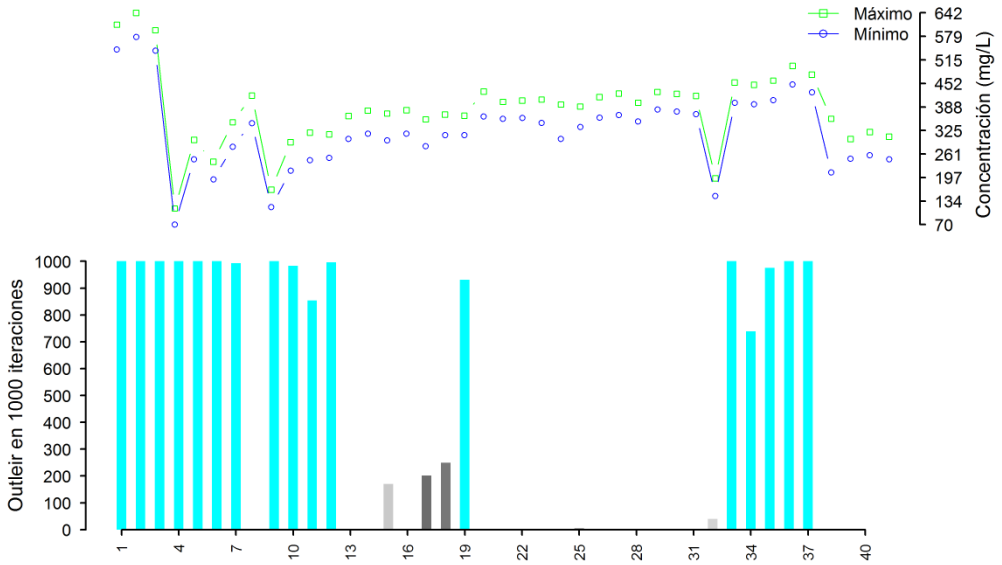


Figura 134- En la gráfica superior se presenta el valor máximo y mínimo de la concentración de los DQOf de cada muestra generada en 1000 simulaciones de Monte Carlo y en el gráfico inferior se presentan en color cian las muestras que en 300 o más simulaciones de 1000 fueron catalogados como *outliers* (EE de Gibraltar)

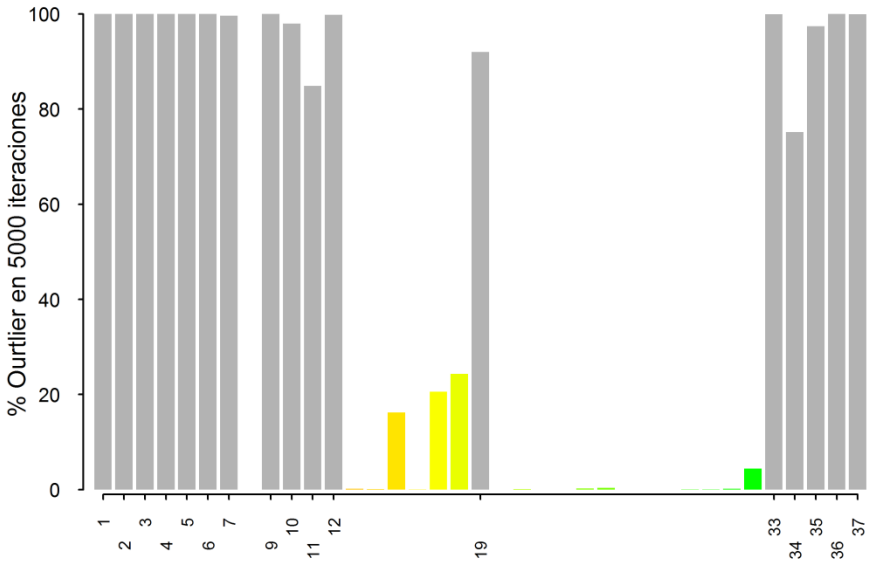


Figura 135- Porcentaje que una muestra fue catalogada como *outlier* en 5000 simulaciones de Monte Carlo en el caso de las concentraciones de DQOf en el afluente de la EE de Gibraltar. En gris se presenta las muestras catalogas como *outliers* en 60 % o más de las simulaciones generadas

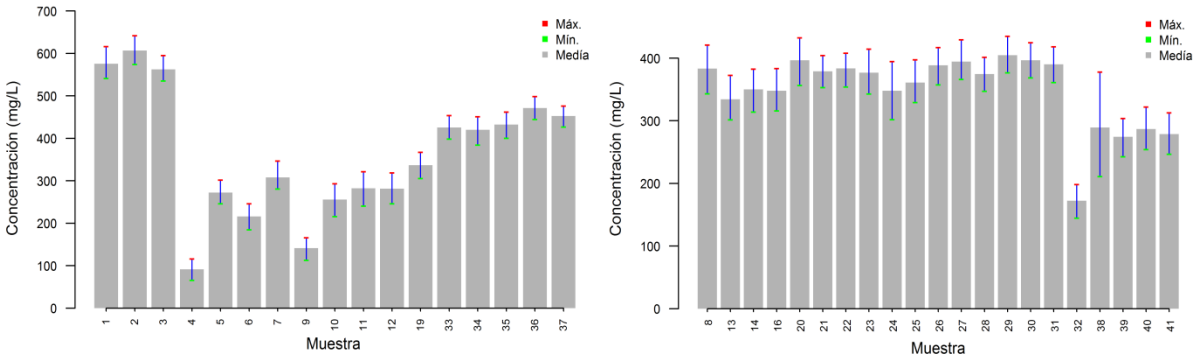


Figura 136- A la izquierda se presentan los rangos de concentración de las muestras catalogadas como *outliers* en más del 10 % de las 5000 simulaciones de Monte Carlo y a la derecha las muestras que estuvieron por debajo de dicho porcentaje catalogadas como datos validos (DQO-EE de Gibraltar)

Resultados de las 1000 ejecuciones de los modelos *PLS* en el caso de las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

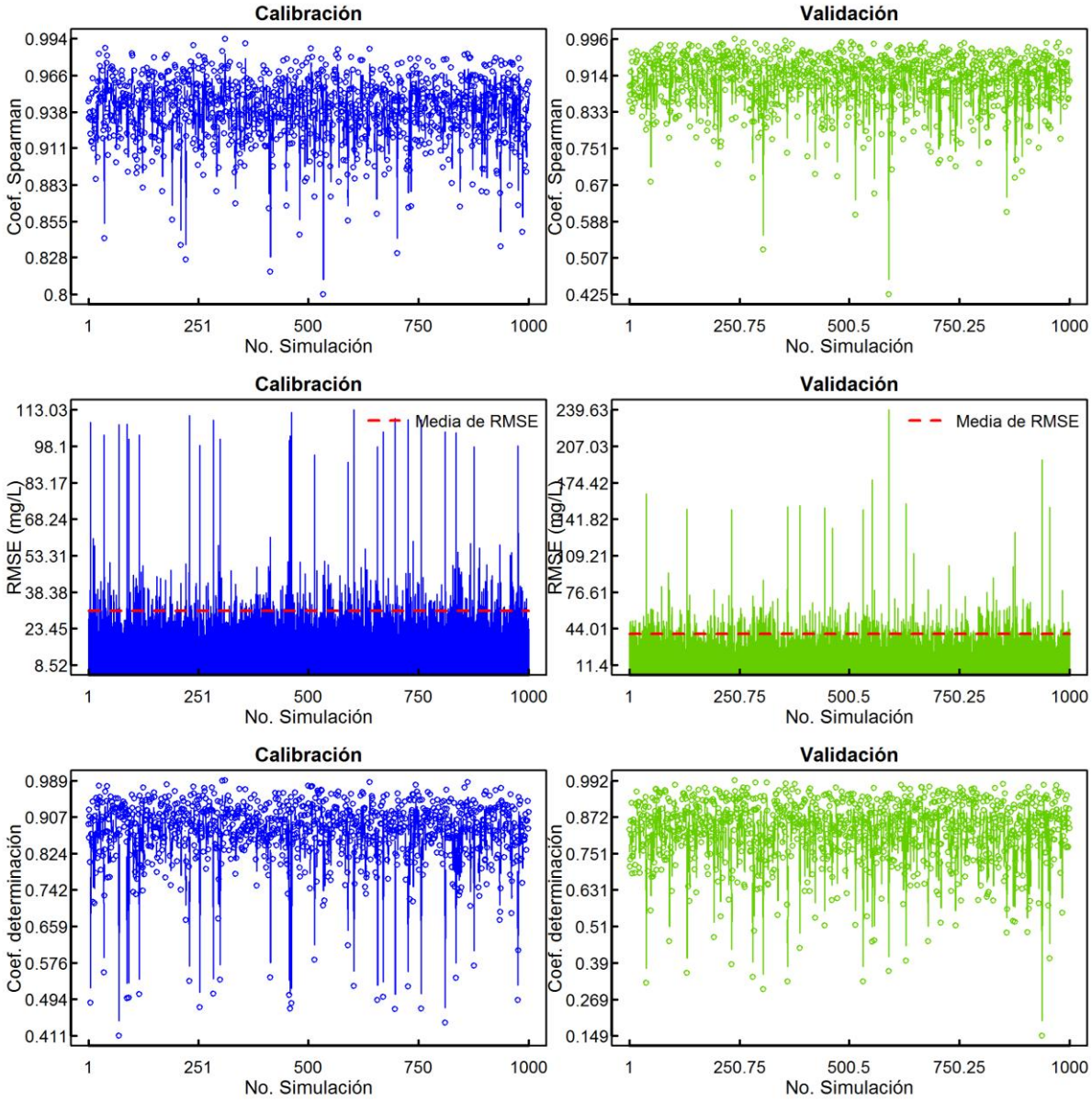


Figura 137- Evaluación del desempeño de los modelos *PLS* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de SST del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo seco

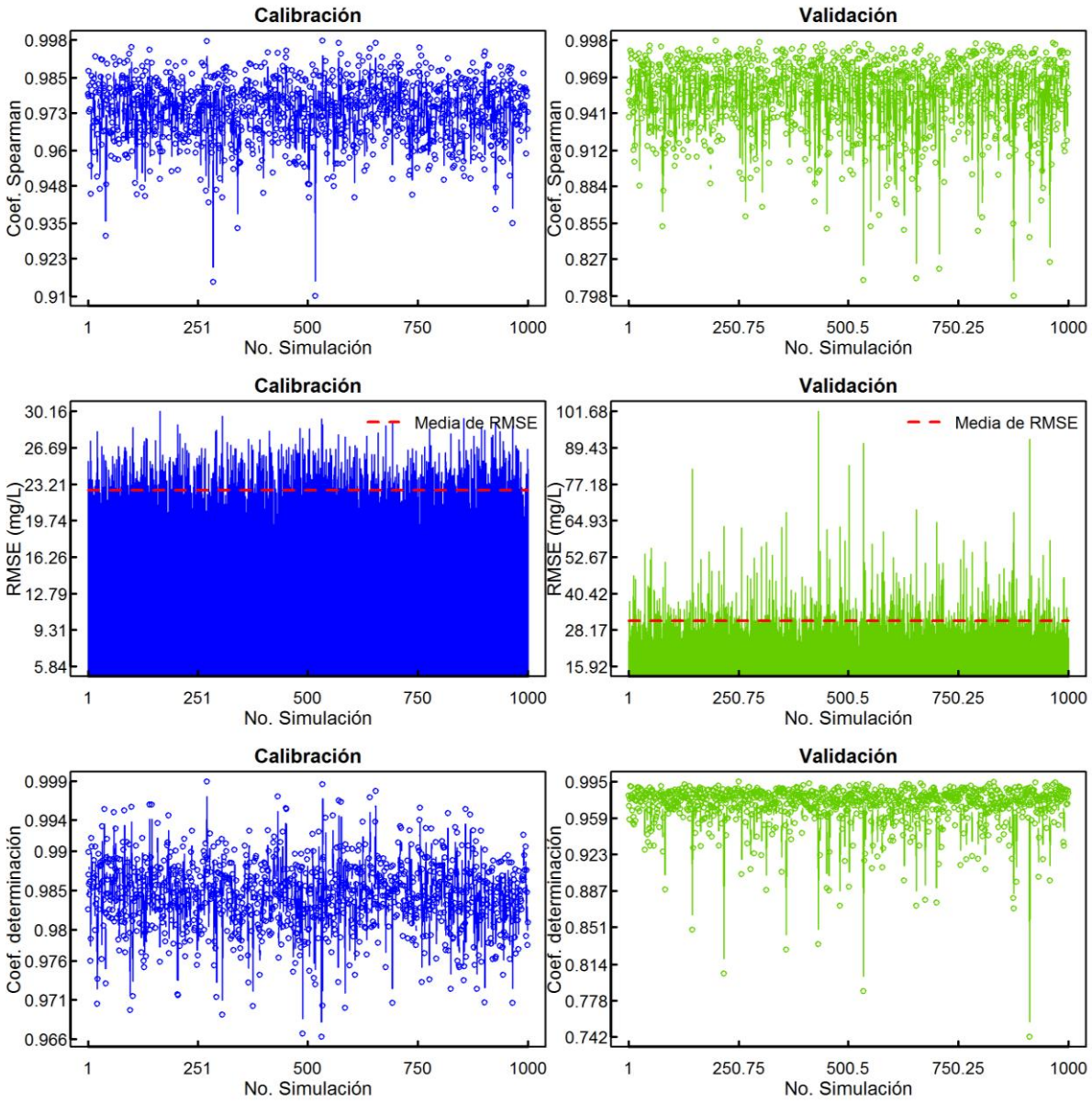


Figura 138- Evaluación del desempeño de los modelos *PLS* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQO del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo seco

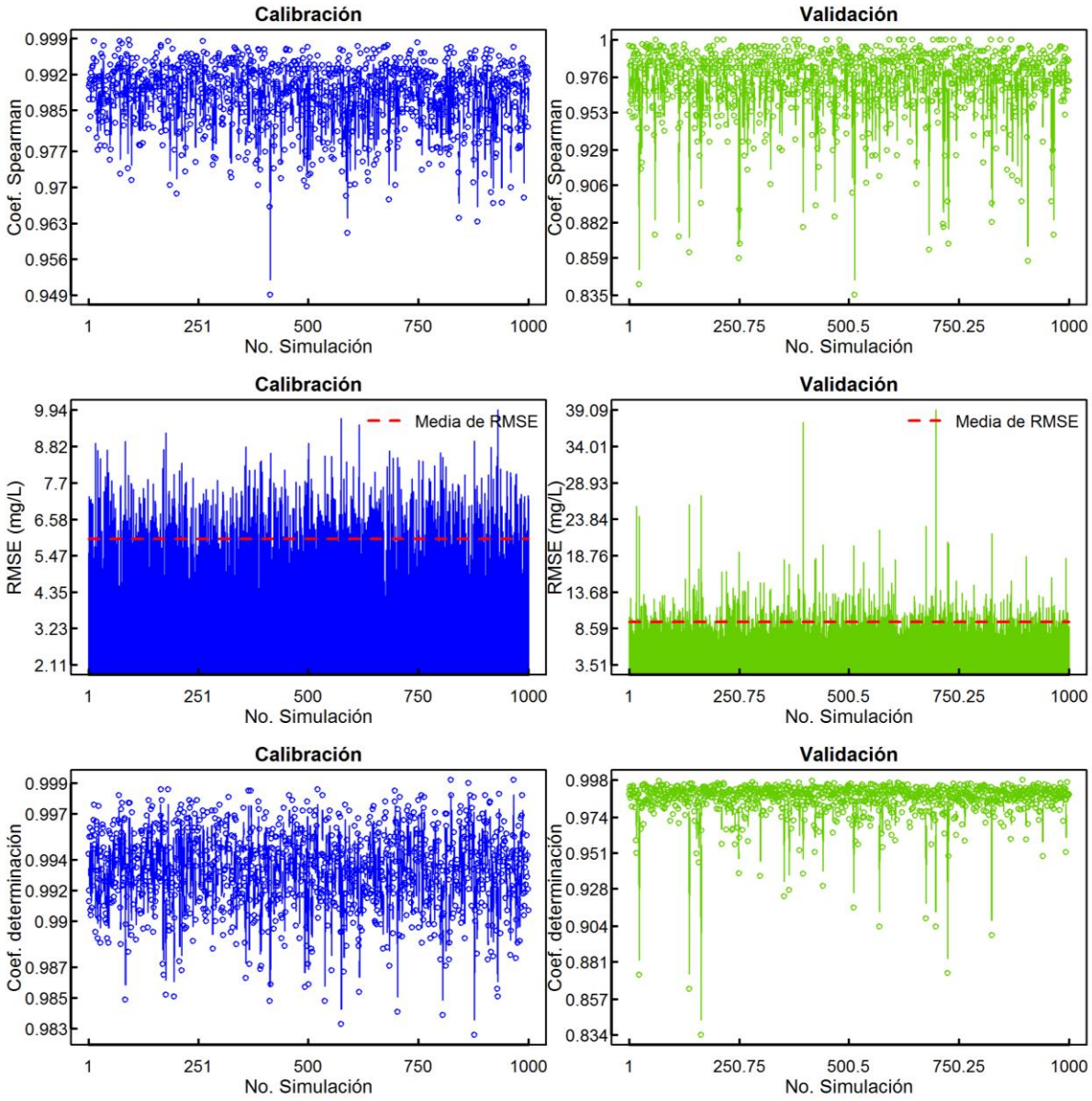


Figura 139- Evaluación del desempeño de los modelos *PLS* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQOf del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo seco

Resultados de las 1000 ejecuciones de los modelos *SVM* en el caso de las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo seco)

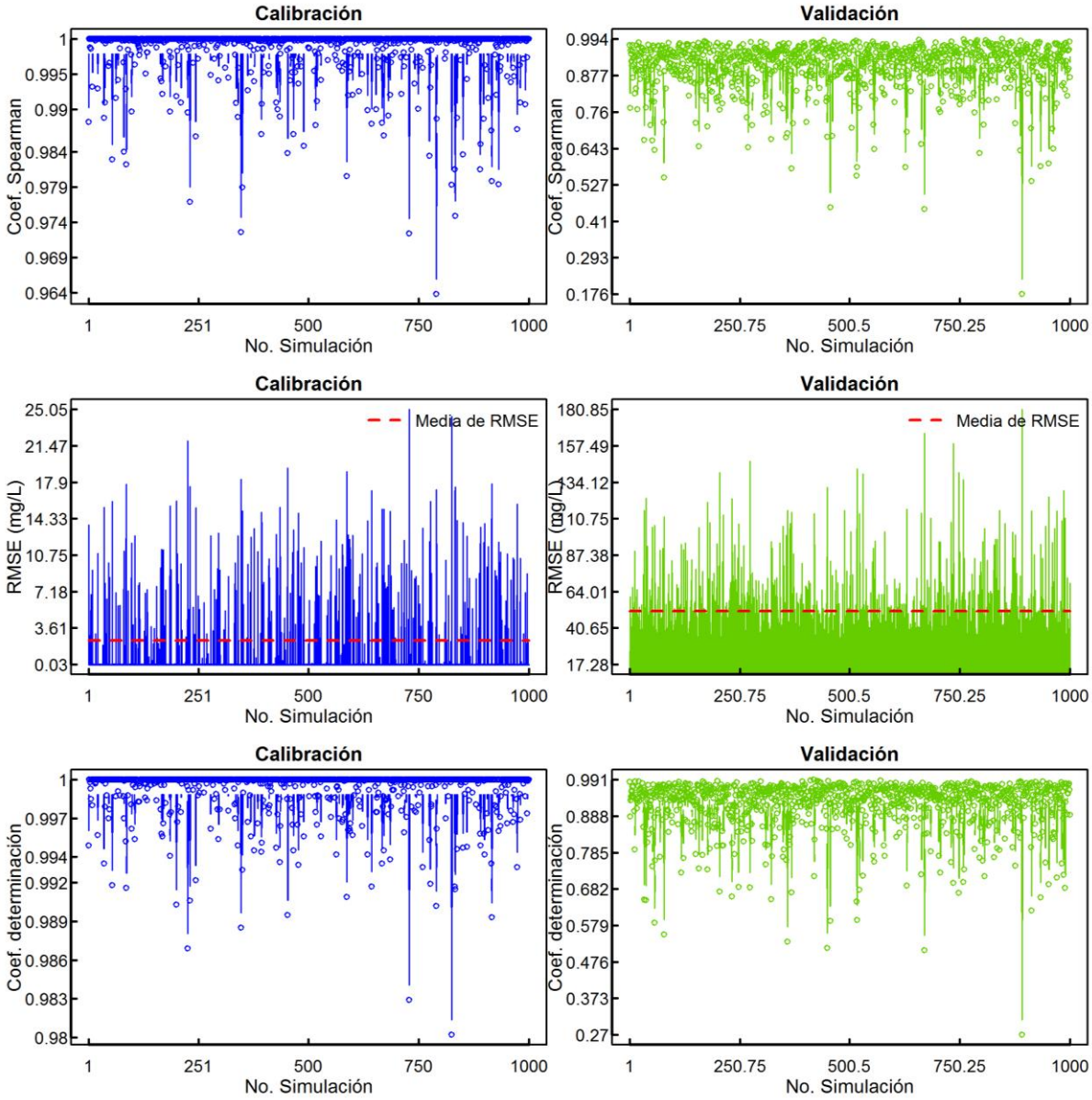


Figura 140- Evaluación del desempeño de los modelos *SVM* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQO del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo seco

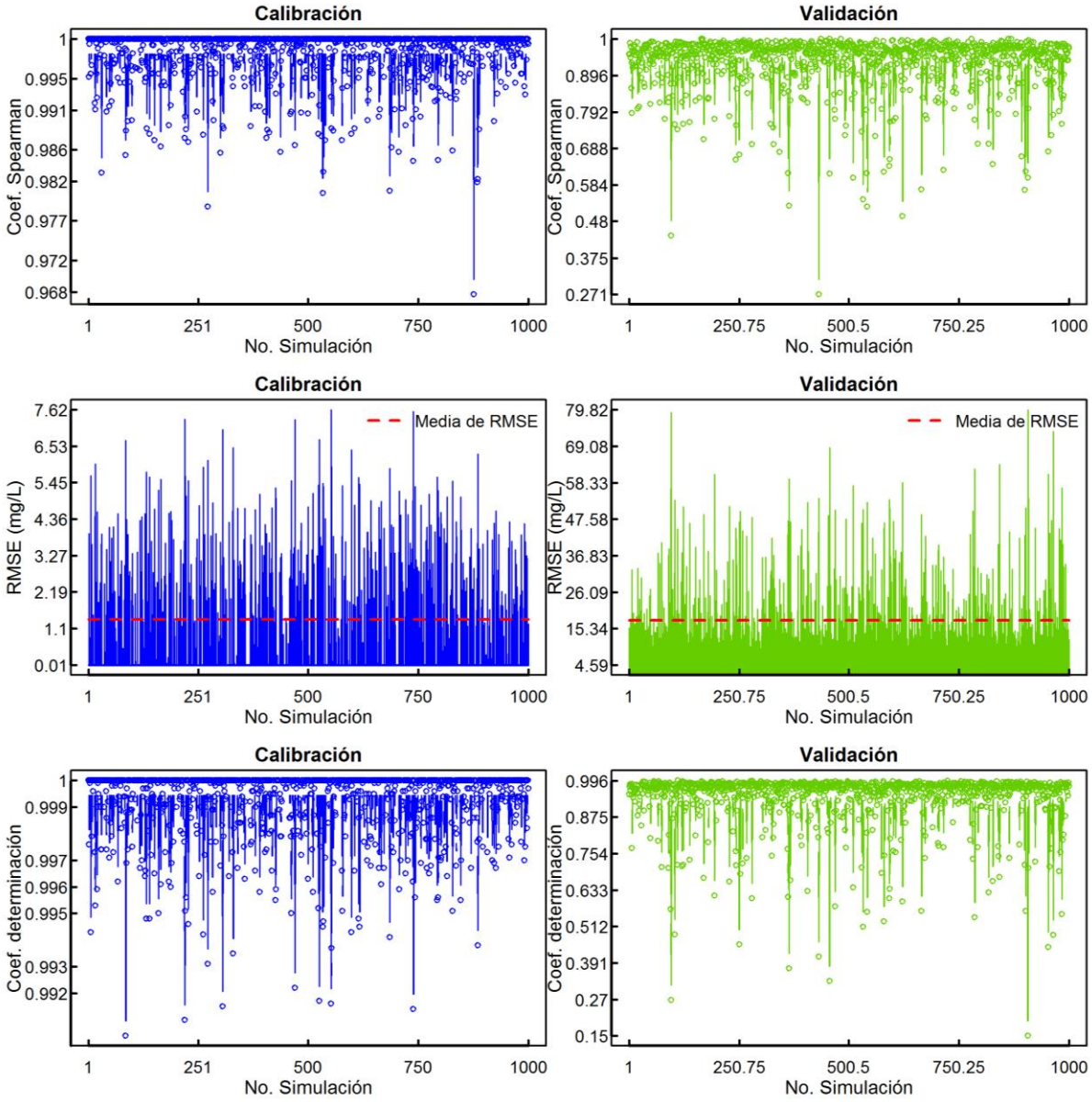


Figura 141- Evaluación del desempeño de los modelos *SVM* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQOf del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo seco

Resultados de las 1000 ejecuciones de los modelos *PLS* en el caso de las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia)

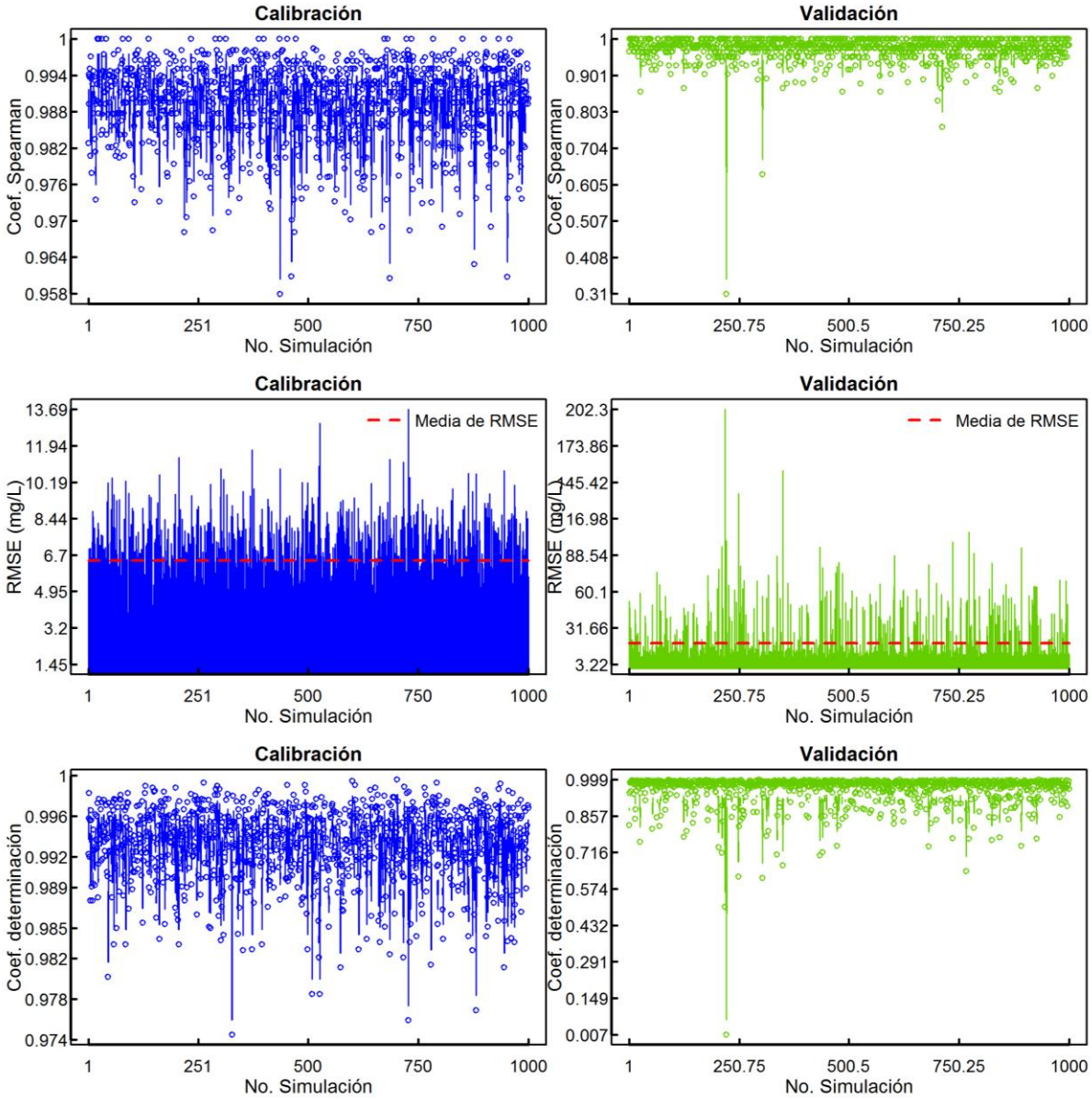


Figura 142- Evaluación del desempeño de los modelos *PLS* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de SST del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo lluvia

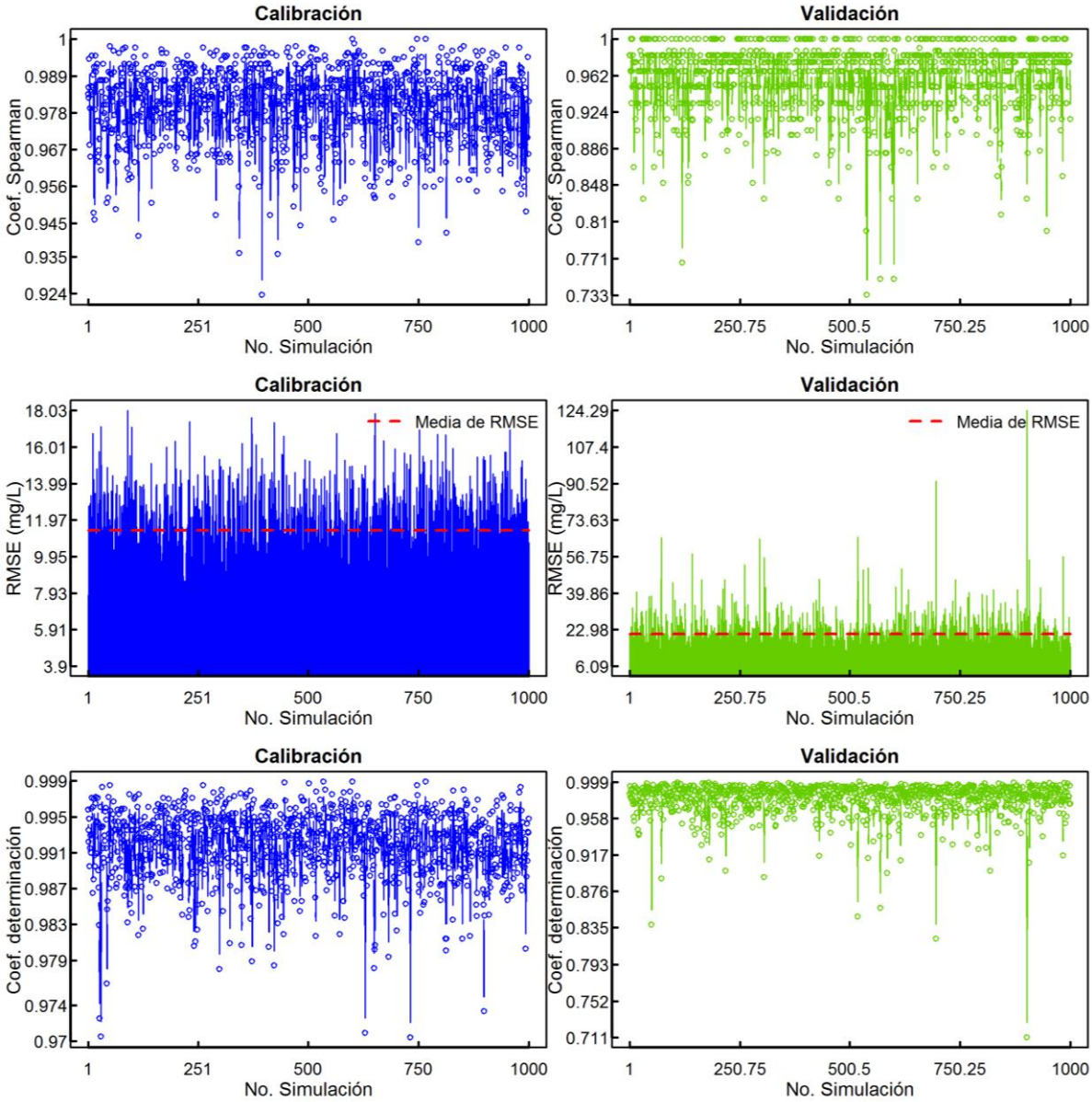


Figura 143- Evaluación del desempeño de los modelos *PLS* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQO del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo lluvia

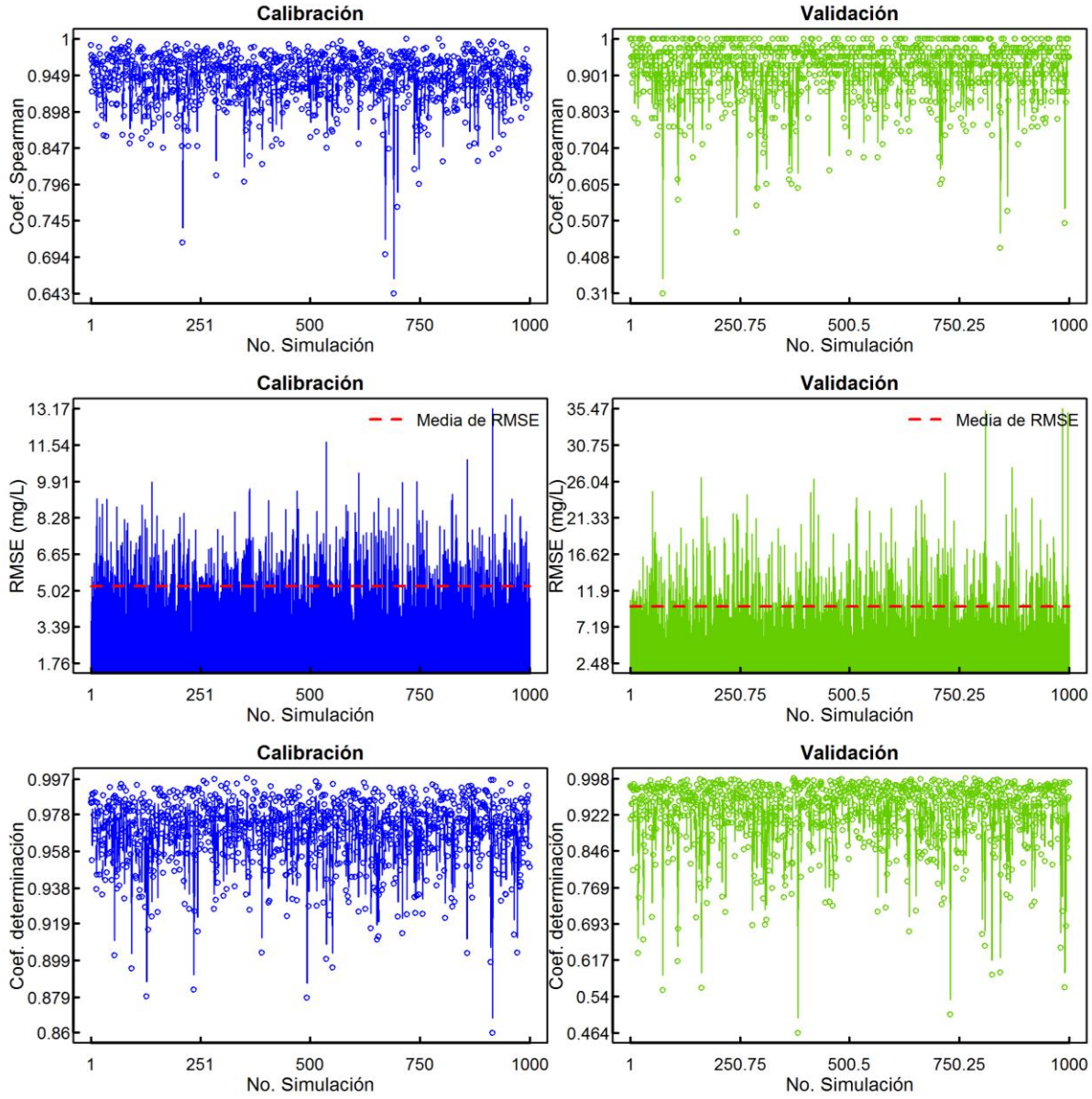


Figura 144- Evaluación del desempeño de los modelos *PLS* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQOf del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo lluvia

Resultados de las 1000 ejecuciones de los modelos *SVM* en el caso de las muestras del afluente de la PTAR de *Fontaines-sur-Saône* (tiempo lluvia)

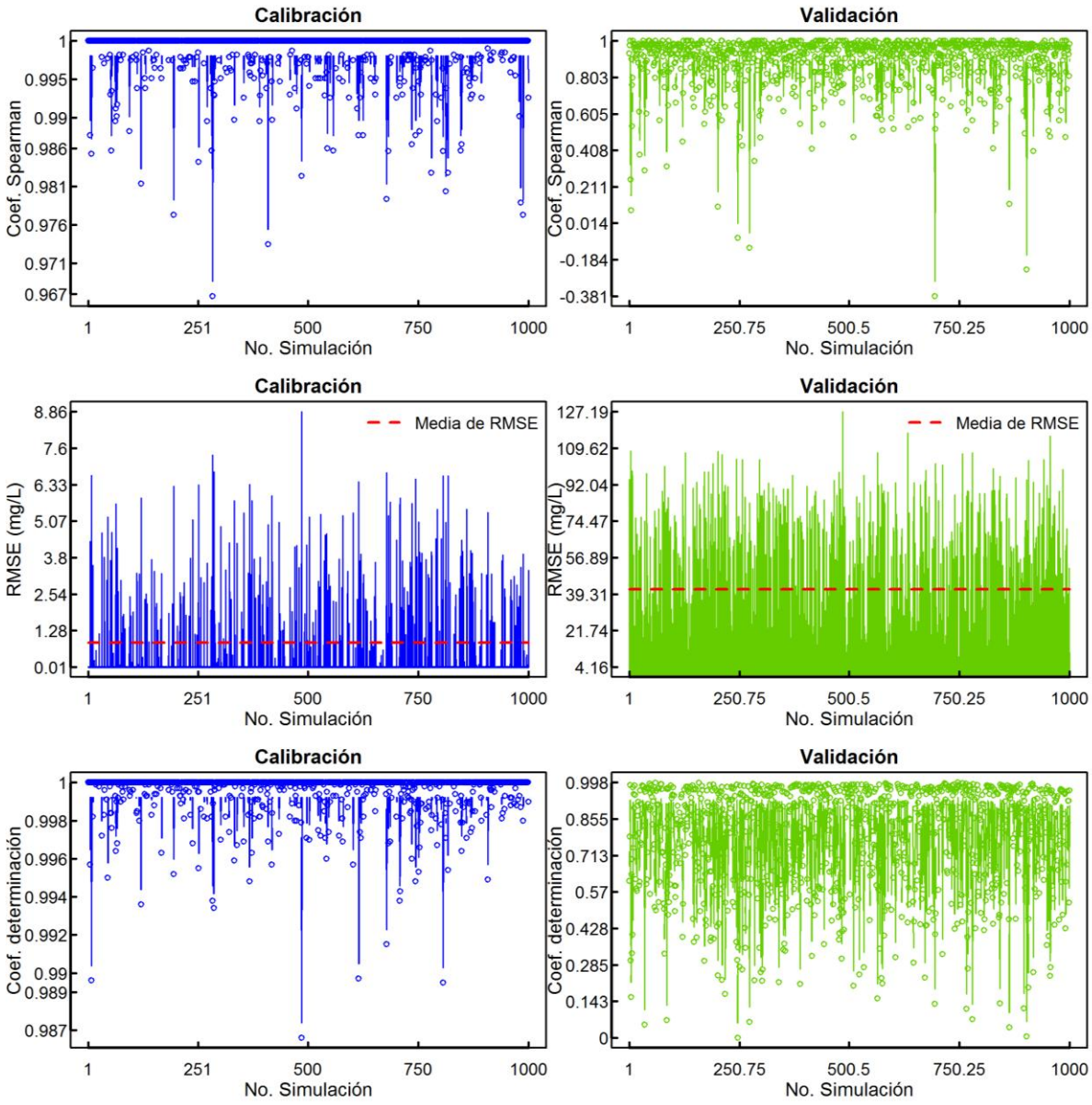


Figura 145- Evaluación del desempeño de los modelos *SVM* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de SST del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo lluvia

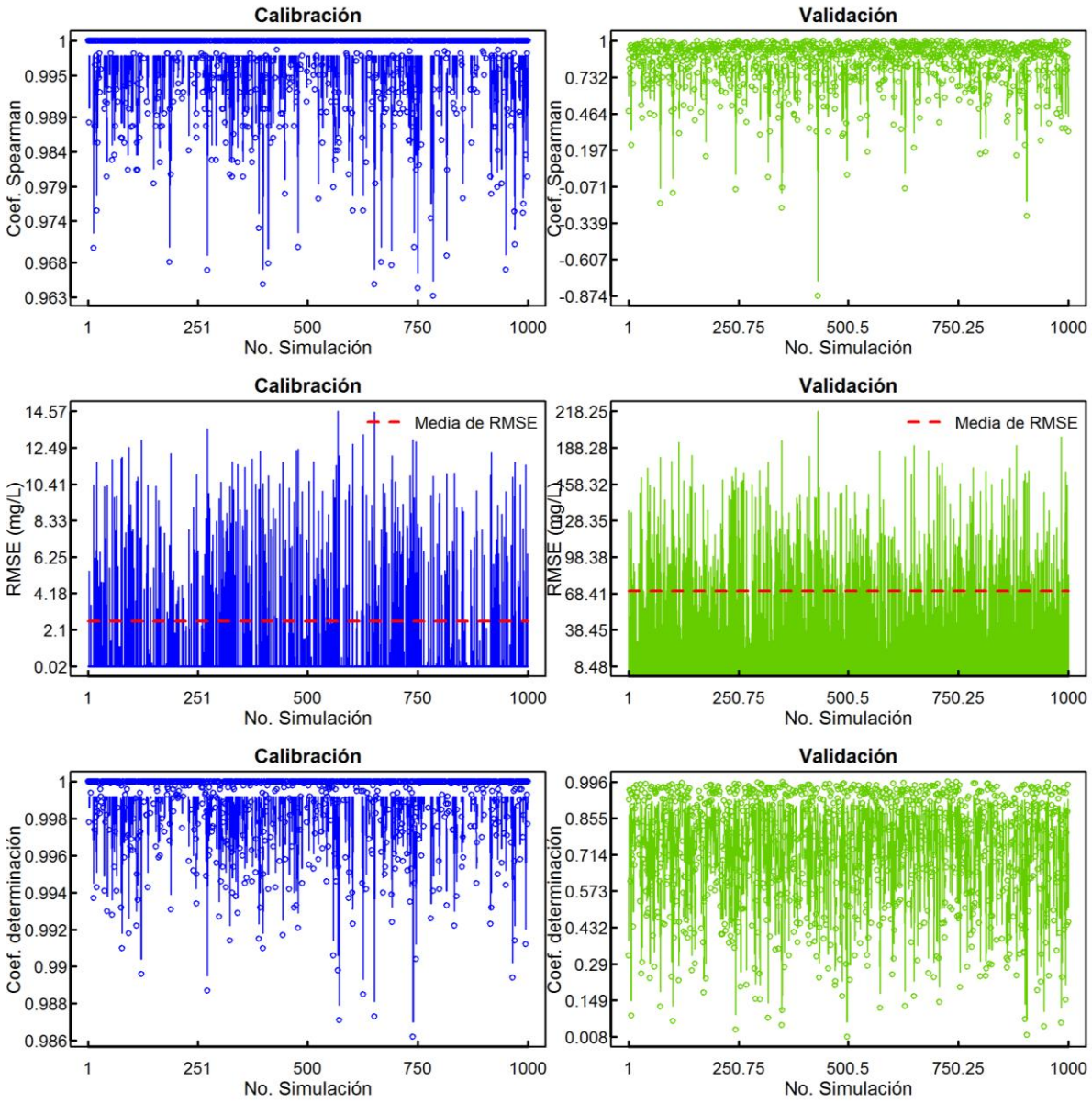


Figura 146- Evaluación del desempeño de los modelos *SVM* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQO del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo lluvia

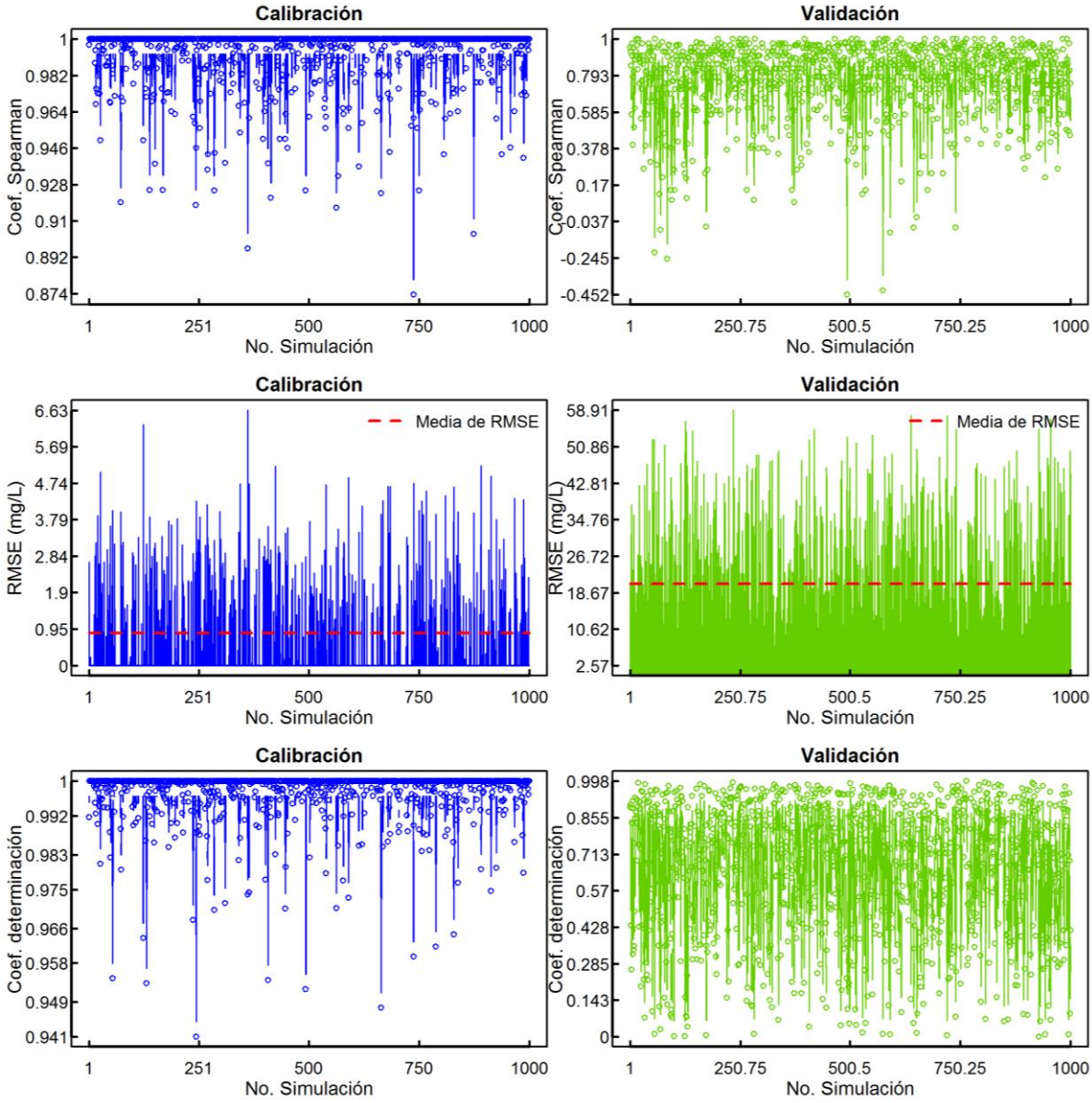


Figura 147- Evaluación del desempeño de los modelos *SVM* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQOf del afluente de la PTAR de *Fontaines-sur-Saône* en tiempo lluvia

Resultados de las 1000 ejecuciones de los modelos *PLS* en el caso de las muestras del afluente de la EE de Gibraltar

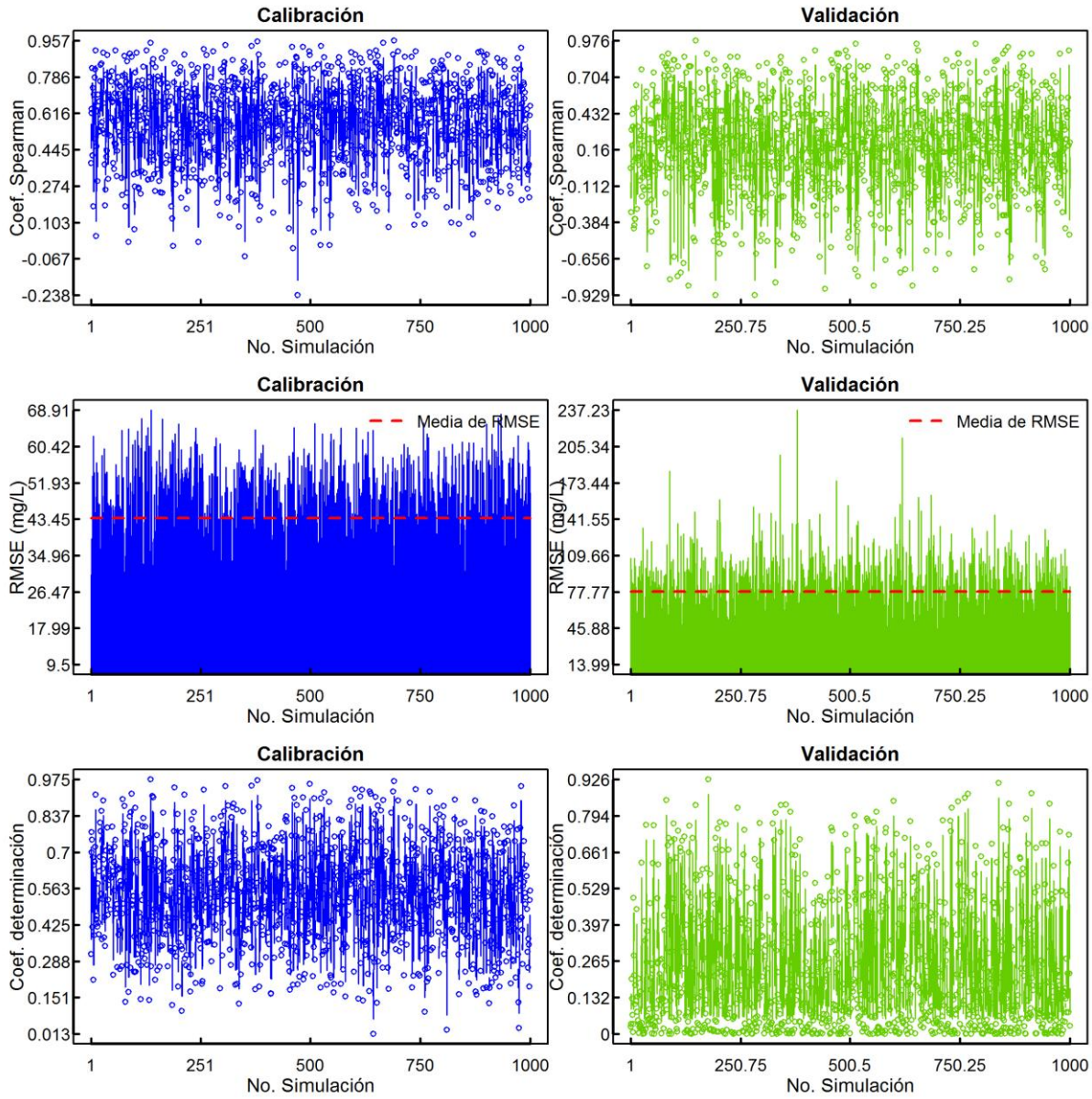


Figura 148- Evaluación del desempeño de los modelos *PLS* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de SST del afluente de la EE de Gibraltar

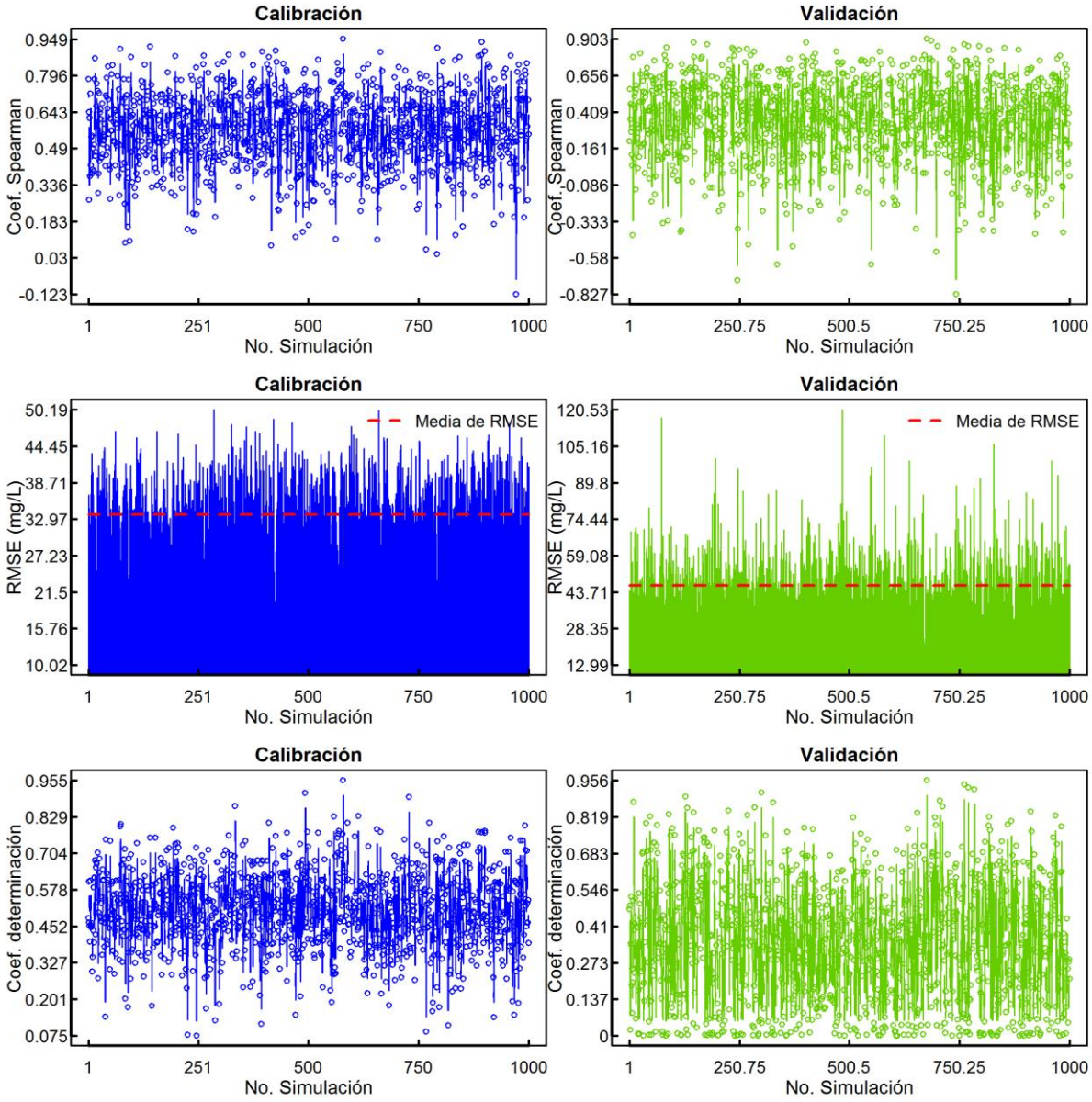


Figura 149- Evaluación del desempeño de los modelos *PLS* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQO del afluente de la EE de Gibraltar

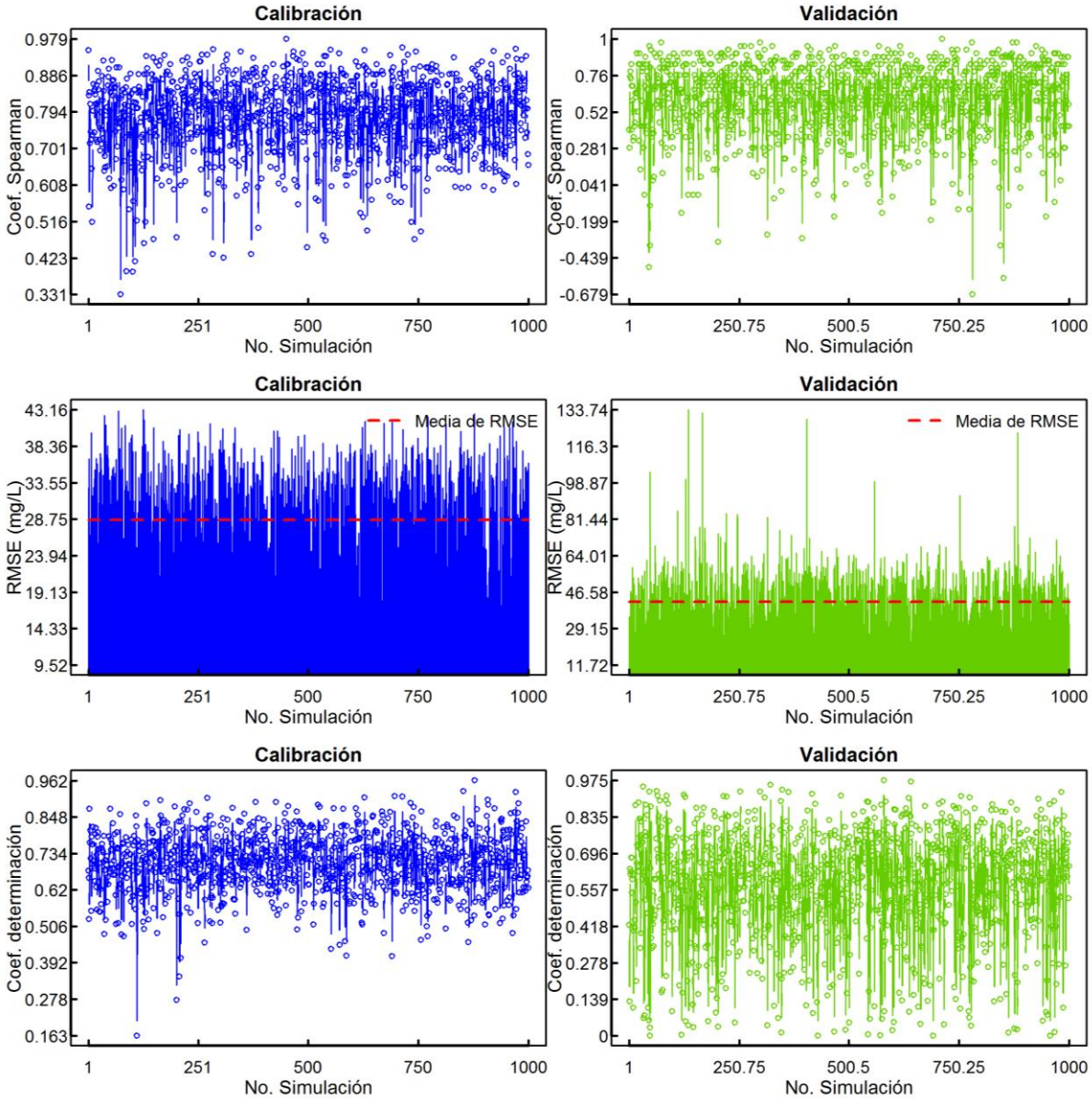


Figura 150- Evaluación del desempeño de los modelos *PLS* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQOf del afluente de la EE de Gibraltar

Resultados de las 1000 ejecuciones de los modelos *SVM* en el caso de las muestras del afluente de la EE de Gibraltar

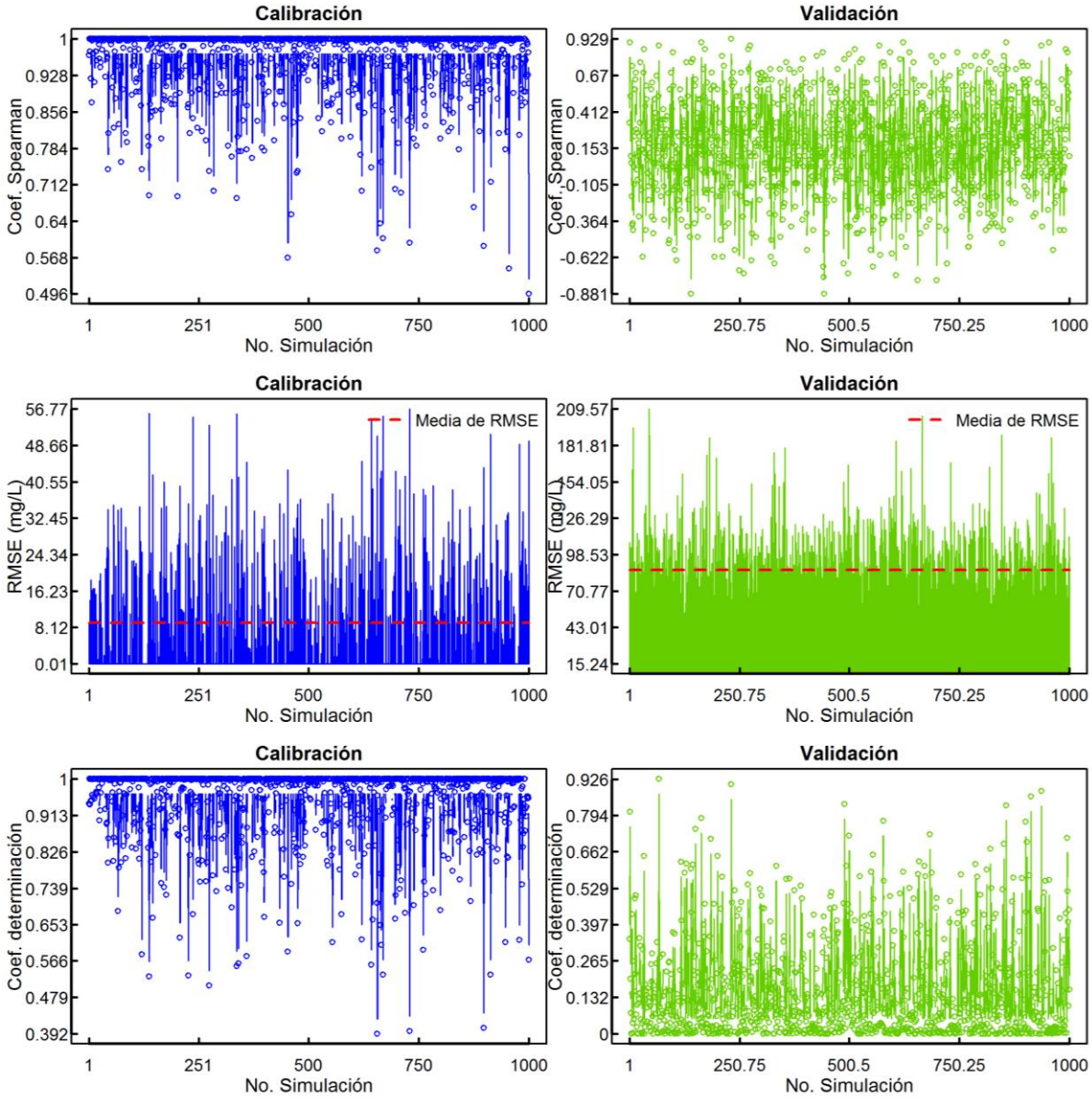


Figura 151- Evaluación del desempeño de los modelos *SVM* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de SST del afluente de la EE de Gibraltar

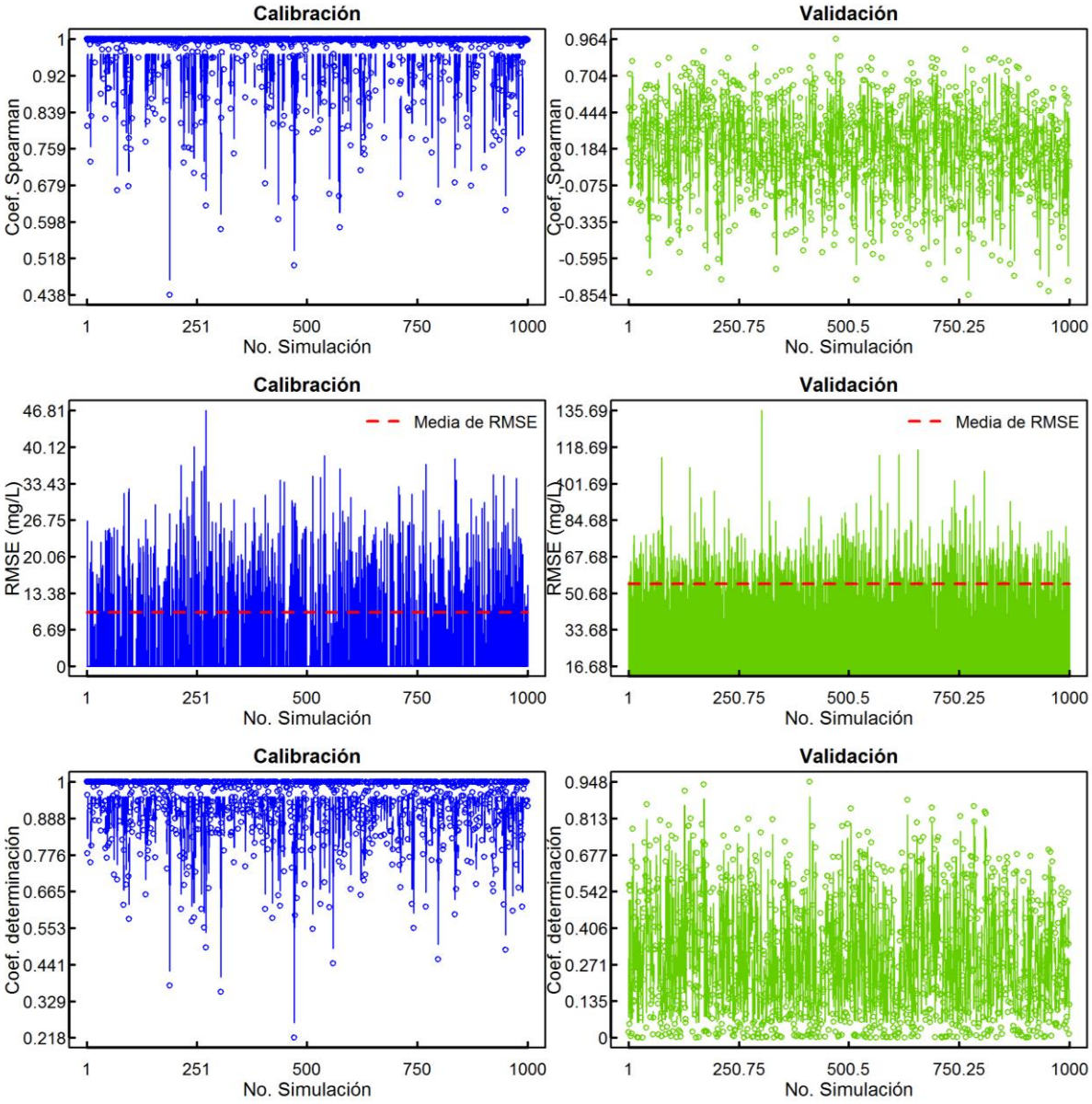


Figura 152- Evaluación del desempeño de los modelos *SVM* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQO del afluente de la EE de Gibraltar

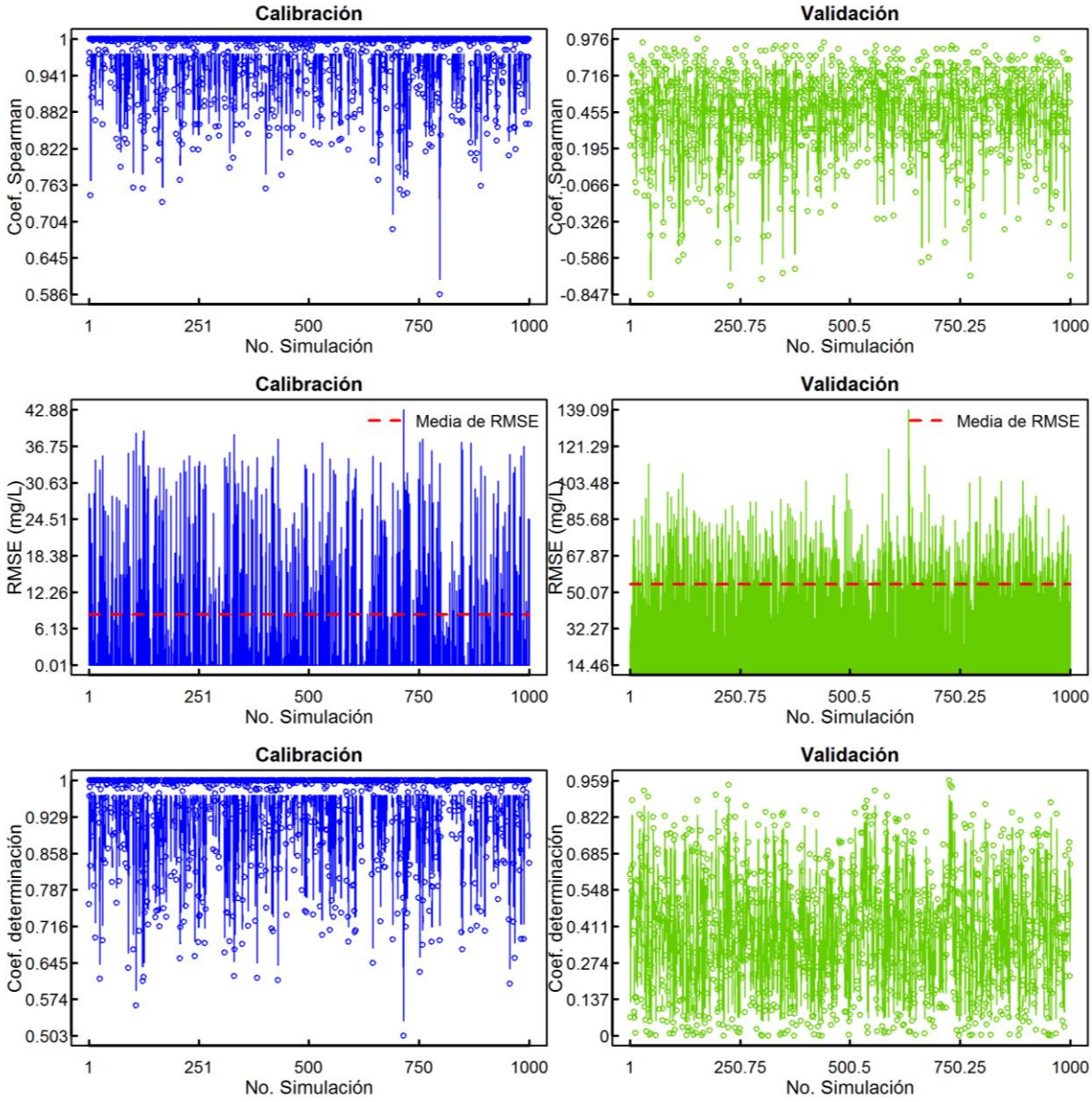


Figura 153- Evaluación del desempeño de los modelos *SVM* en la etapa de calibración y validación en el caso de la estimación de las concentraciones equivalentes de DQOf del afluente de la EE de Gibraltar

ANEXO D

Desde la Figura 154 a la Figura 159 se presenta los resultados *RMSEP* versus el tiempo de computo en segundos (izquierda) y el número de longitudes de onda (derecha) empleadas en la calibración de los modelos *PLS* y *SVM* en cada una 1000 ejecuciones por cada determinante (SST, DQO y DQOf) y caso de estudio.

En dichos gráficos se presenta de arriba abajo los resultados obtenidos para los SST (recuadro verde), la DQO (recuadro naranja) y DQOf (recuadro azul); y para cada determinante se presenta en ambas gráficas en el eje de las ordenadas los valores de *en* mg/L. Por otra parte, en los gráficos a la derecha de cada figura se presenta en las abscisas el tiempo de cómputo que tomo la calibración de cada modelo, y en los gráficos a la izquierda en las abscisas el número de longitudes de onda empleadas en la calibración de cada modelo.

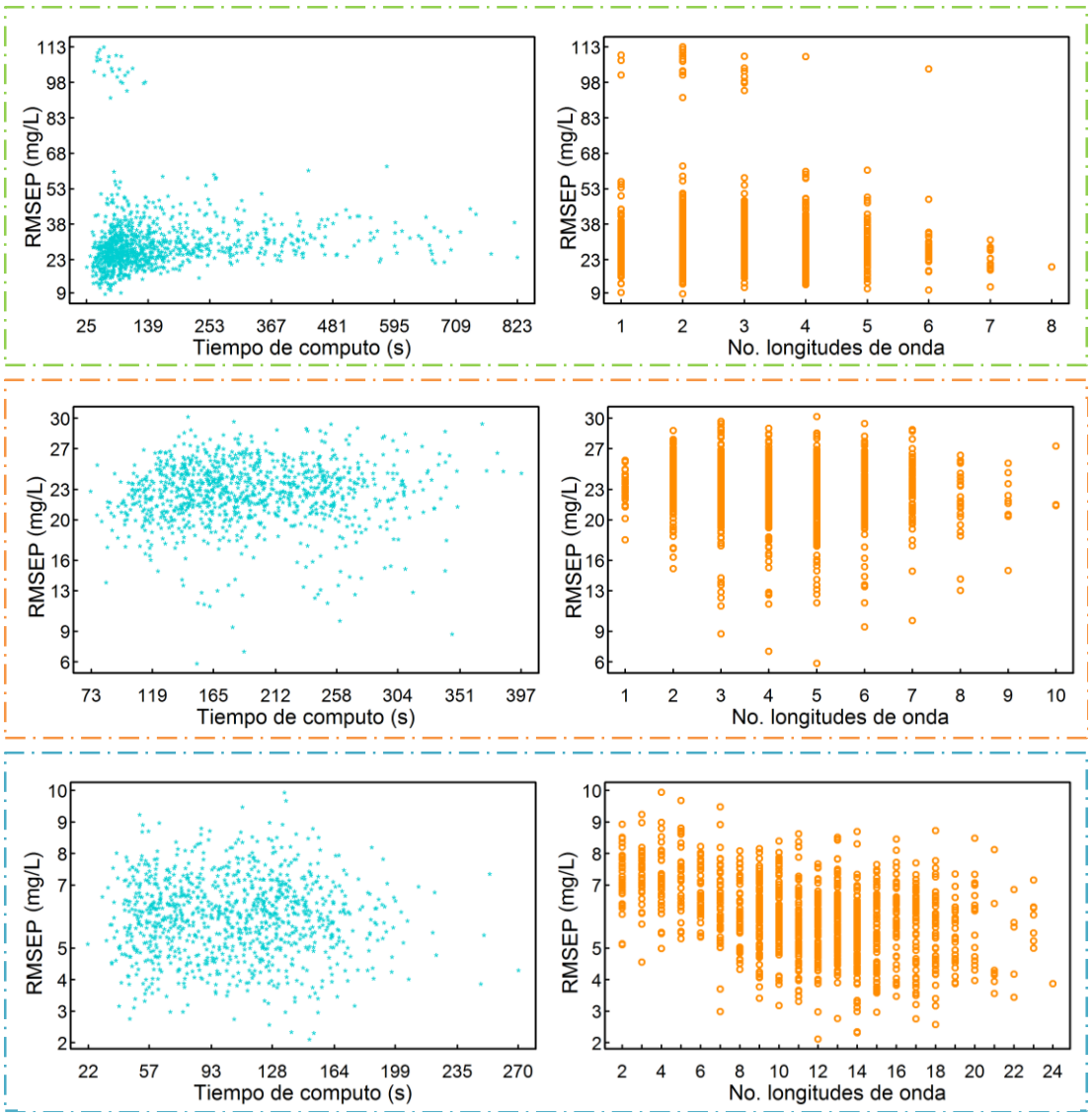


Figura 154- *RMSEP* versus tiempo de computo (Izq.) y número de longitudes de onda empleados en la calibración de los modelos *PLS* para la estimación de las concentraciones equivalentes de los determinantes de SST, DQO y DQOF de las muestras del afluente de la PTAR *Fontaines-sur-Saône* (tiempo seco)

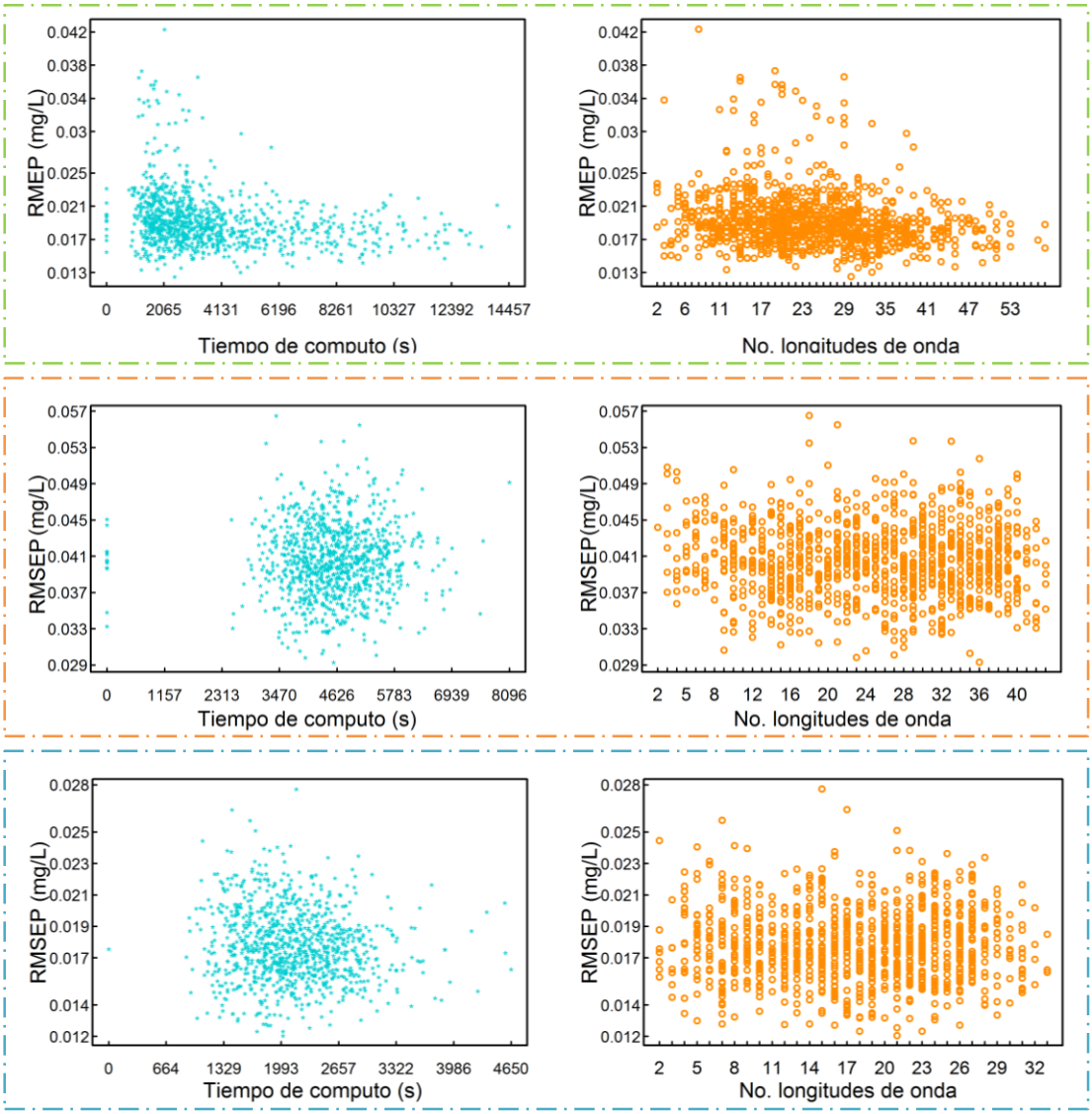


Figura 155- ***RMSEP*** **versus tiempo de computo (Izq.) y número de longitudes de onda empleados en la calibración de los modelos** ***SVM*** **para la estimación de las concentraciones equivalentes de los determinantes de SST, DQO y DQOF de las muestras del afluente de la PTAR** ***Fontaines-sur-Saône*** **(tiempo seco)**

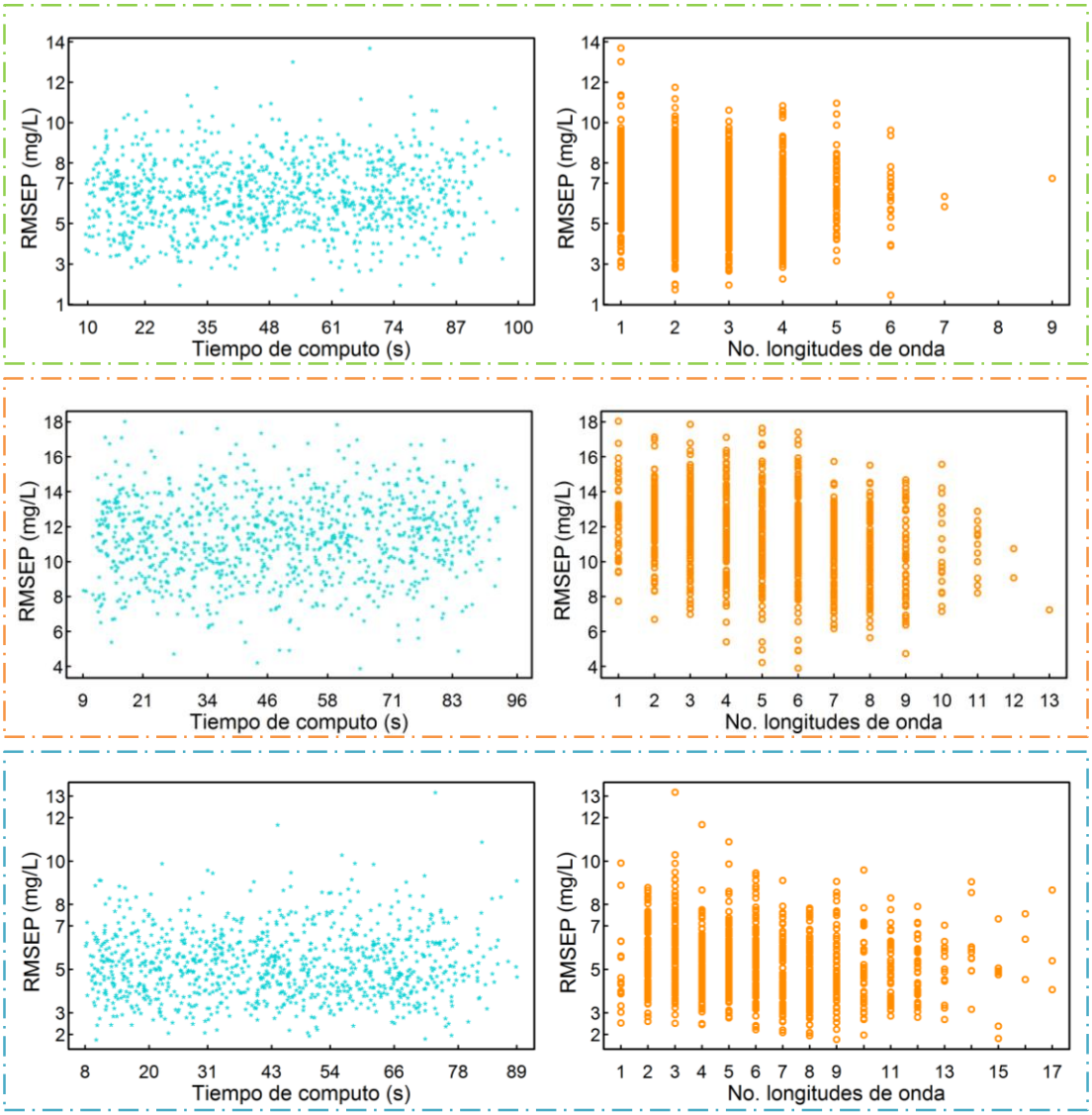


Figura 156- *RMSEP* versus tiempo de computo (Izq.) y número de longitudes de onda empleados en la calibración de los modelos *PLS* para la estimación de las concentraciones equivalentes de los determinantes de SST, DQO y DQOF de las muestras del afluente de la PTAR *Fontaines-sur-Saône* (tiempo lluvia)

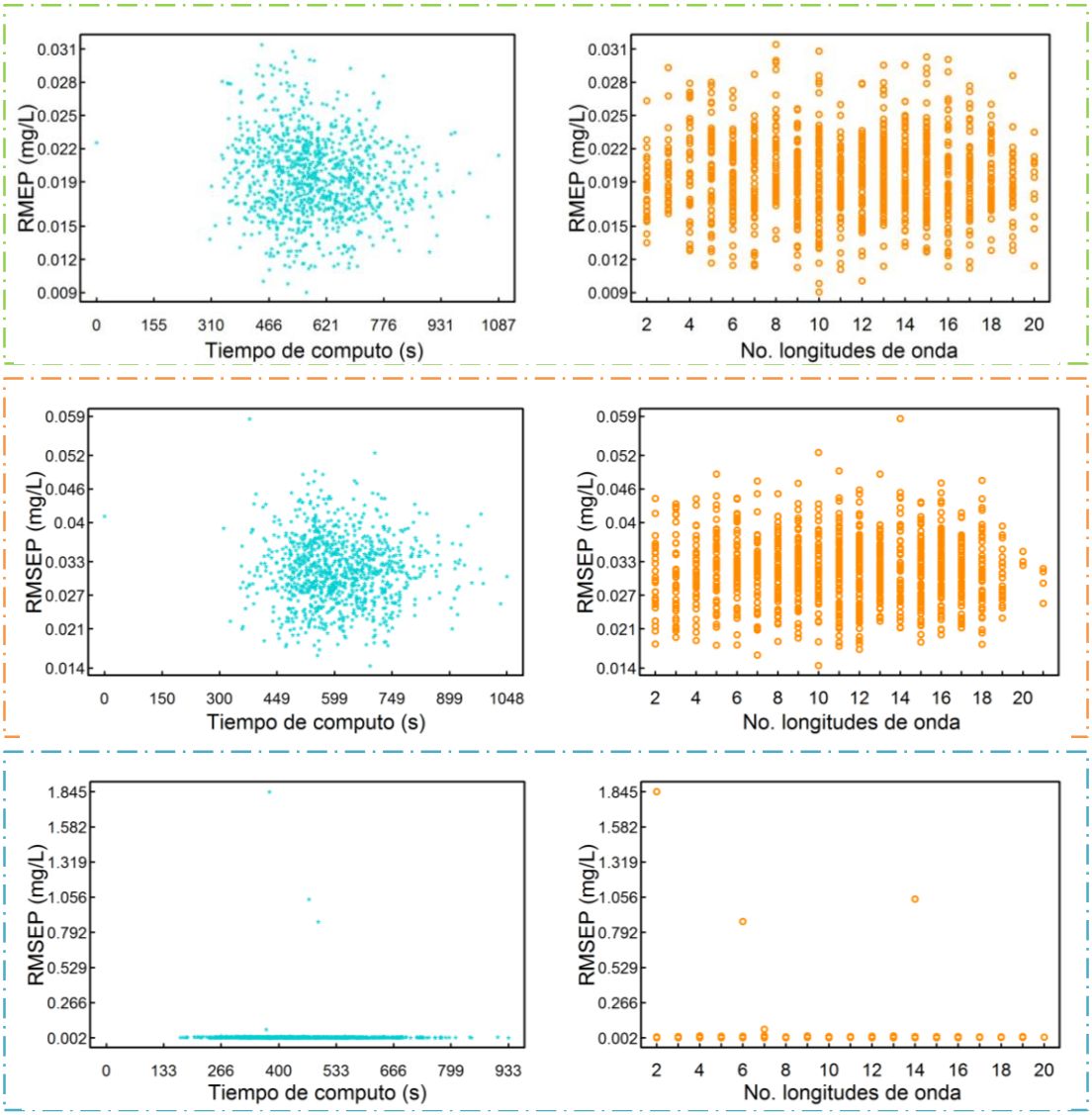


Figura 157- *RMSEP* versus tiempo de computo (Izq.) y número de longitudes de onda empleados en la calibración de los modelos *SVM* para la estimación de las concentraciones equivalentes de los determinantes de SST, DQO y DQOF de las muestras del afluente de la PTAR *Fontaines-sur-Saône* (tiempo lluvia)

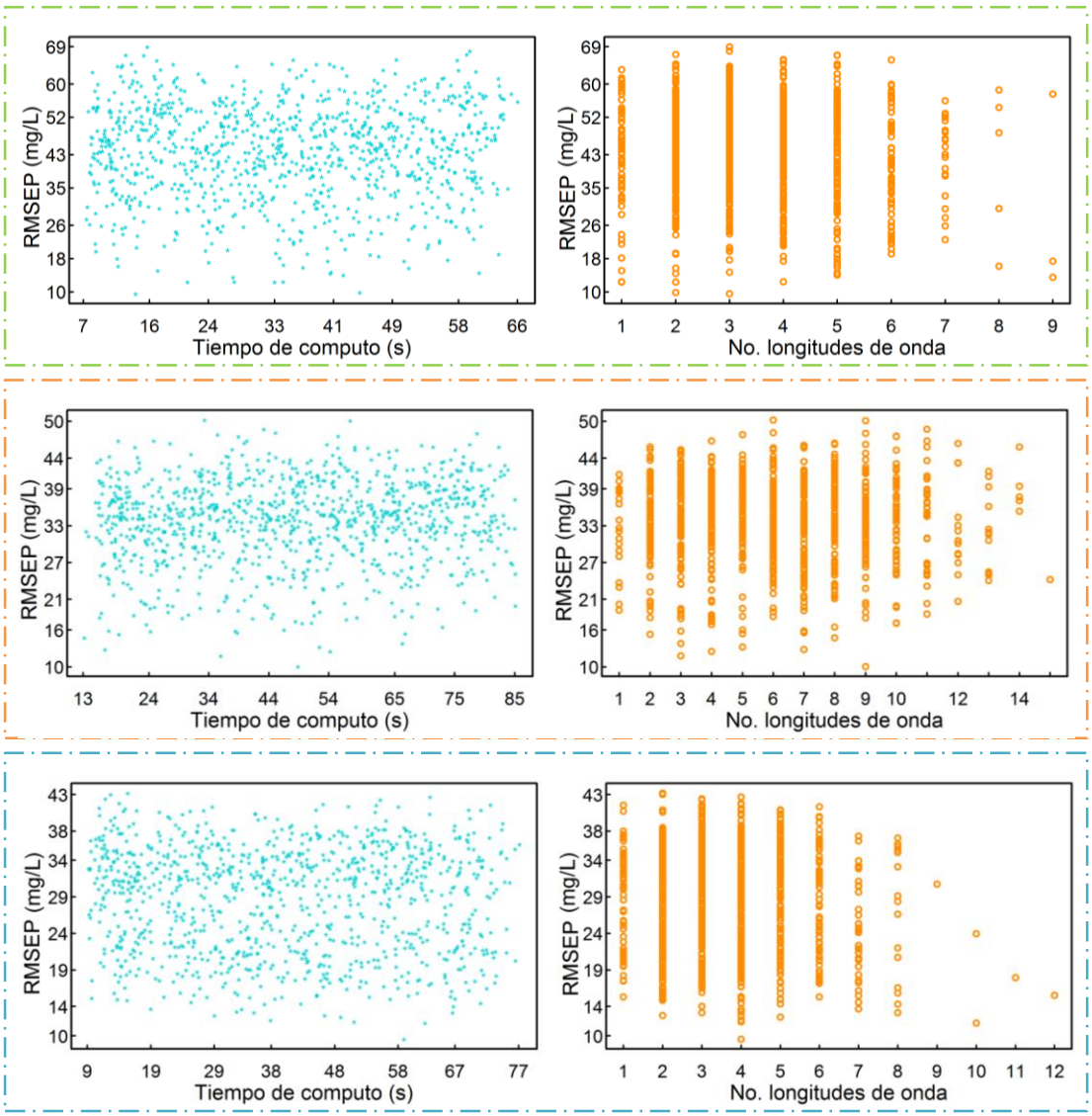


Figura 158- *RMSEP* versus tiempo de computo (Izq.) y número de longitudes de onda empleados en la calibración de los modelos *PLS* para la estimación de las concentraciones equivalentes de los determinantes de SST, DQO y DQOF de las muestras del afluente de la EE de Gibraltar

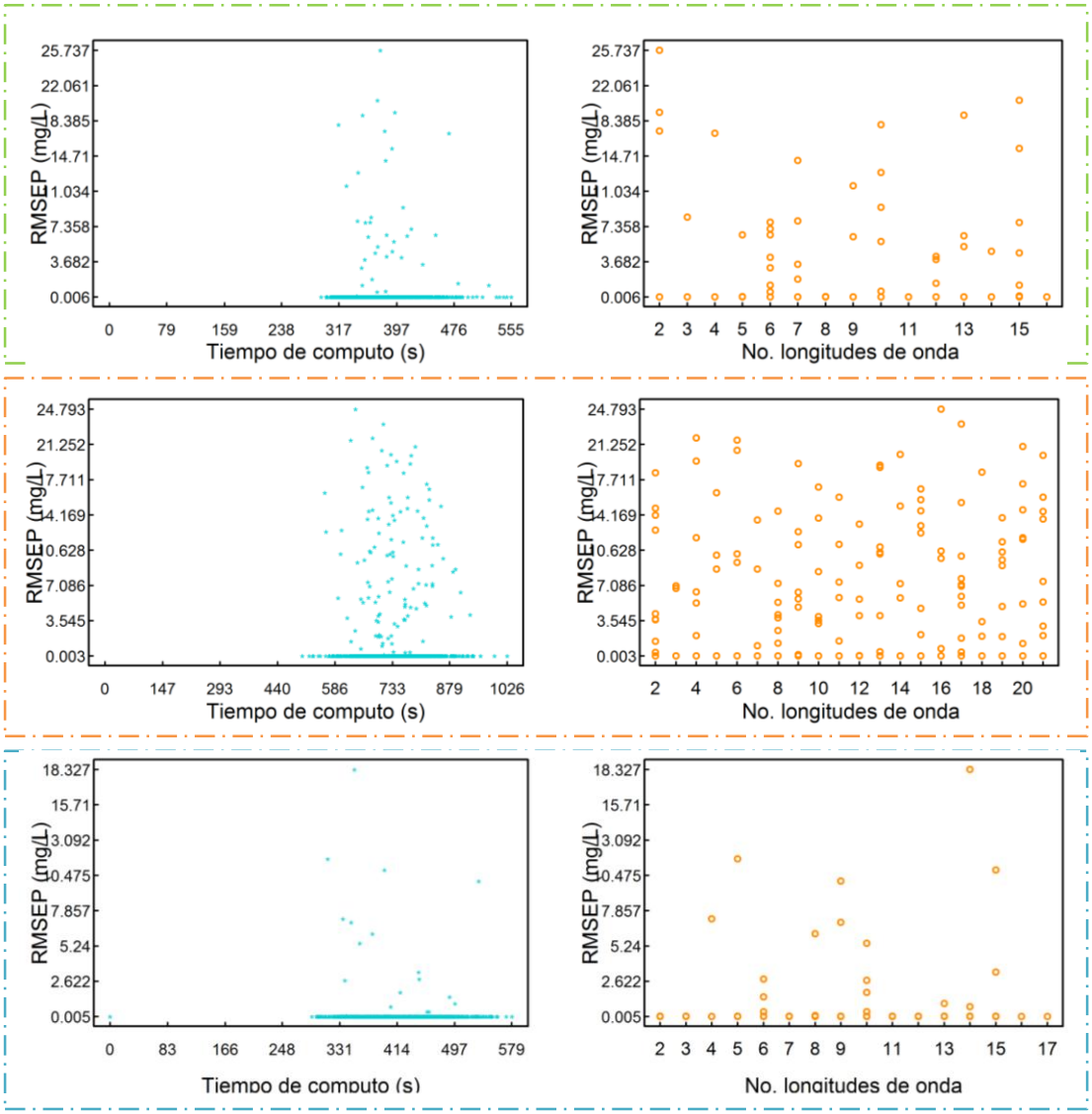


Figura 159- *RMSEP* versus tiempo de computo (Izq.) y número de longitudes de onda empleados en la calibración de los modelos *SVM* para la estimación de las concentraciones equivalentes de los determinantes de SST, DQO y DQOF de las muestras del afluente de la EE de Gibraltar

ANEXO E-1

7.1. MUESTREO Y ANÁLISIS DE LABORATORIO

El muestreo es una operación clave para el monitoreo de aguas residuales. Cualquiera que sea el objetivo del muestreo, por ejemplo dar cumplimiento a una regulación, la eficiencia de un tratamiento, el impacto de una descarga, éste es el primer paso del procedimiento clásico analítico antes del análisis de laboratorio; el cual también se lleva a cabo para validar las mediciones *in situ* u *on line*. Aunque el uso de dispositivos en línea está aumentando, la gran mayoría de las mediciones de calidad de las aguas residuales se lleva a cabo en el laboratorio, después del muestreo. Por lo tanto, antes de considerar los métodos de análisis de laboratorio para el control de calidad de aguas residuales, con base en los procedimientos estándar o técnicas alternativas, la etapa de muestreo debe ser considerada debido a su importancia como fuente de posibles errores (Quevauviller *et al*., 2006; Gonzalez *et al*., 2009).

Con el objetivo de conseguir un volumen representativo de efluentes, el muestreo tiene que enfrentarse a una serie de problemas específicos relacionados con las características de aguas residuales. Por lo tanto, el muestreo de las aguas residuales es difícil, considerando la heterogeneidad y la variabilidad de los efluentes, y además, la evolución de las muestras durante el transporte del lugar de muestreo al laboratorio, relacionados con el envejecimiento de la muestra. Incluso, en el estudio realizado por Ort *et al*. (2010) se demostró que la incertidumbre de muestreo puede variar desde "no significativa" a "mucho mayor que la incertidumbre debida a análisis químico", el cual depende del sitio y de los compuestos específicos presentes en las muestras, así como la precisión del método analítico usado (Quevauviller *et al*., 2006; Ort *et al*., 2010).

7.1.1. Características de una muestra

7.1.1.1. Heterogeneidad

Todos los tipos de aguas residuales se caracterizan por la heterogeneidad de su composición. Un agua residual se compone de agua, llevando una gran cantidad de sólidos en suspensión y sustancias disueltas que no estaban presentes originalmente (los contaminantes). Entonces los diferentes tipos de aguas residuales dependerán de su naturaleza, tanto en su concentración como en su composición y caudal (Bourgeois *et al*., 2001; Olsson, 2007).

El tipo de agua residual urbana más frecuente es la mezcla entre las aguas residuales domésticas y las industriales. La composición de las aguas residuales municipales es bien conocida y no varía mucho de un lugar a otro. Las composiciones típicas de las aguas residuales urbanas han sido reportados por la literatura (Muttamara, 1996). No obstante, los valores de concentración para la mayoría de determinantes caracterizados en una

muestra tienden a disminuir en el caso de un alcantarillado combinado (efecto de dilución por lluvia) o a incrementarse, en función de la proporción y la naturaleza de las aguas residuales industriales (derrames de sustancias tóxicas y picos de caudal) recogidos en el área urbana (Olsson y Newell, 1999; Quevauviller *et al*., 2006).

Por lo tanto, la heterogeneidad se relaciona con la diversidad de la naturaleza de los compuestos solubles, incrementa cuando se consideran contaminantes emergentes, pero también lo hace la distribución de fracciones no solubles: coloides, supra-coloides y partículas en suspensiones sedimentables (Vaillant *et al*., 1999; Quevauviller *et al*., 2006).

7.1.1.2. Variabilidad

Las características del agua residual cambian con el tiempo, no sólo en términos de caudal, sino también en términos de composición y concentración, cambiando a lo largo del sistema de alcantarillado bajo la influencia de factores físicos, físico-químicos y biológicos, los cuales varían en diferentes escalas de tiempo: horario, diario, semanal y estacional (Fletcher *et al*., 2013).

- *Factores Físicos:* El primer factor físico es la tasa de cambio del caudal en el caso de una mezcla o de una descarga, que juega un papel en la concentración o dilución de la concentración de los determinantes. Este problema se hace más evidente en las redes de alcantarillado combinado durante los eventos de lluvia. Al inicio del evento, los materiales particulados en las carreteras, tejados y zonas de estacionamiento y cambio de aceite, *etc.*, pueden llegar a la alcantarilla (sobre todo después de un período de tiempo largo y seco) generando picos en la carga contaminante. Luego, después del lavado, el fenómeno principal sigue siendo la dilución (reducción de la carga) (Thomas *et al*., 2005; Quevauviller *et al*., 2006).

  Por otra parte, los efectos de las aguas pluviales en las alcantarillas combinadas varían con las características de las alcantarillas (longitud, diámetro, *etc.*) y la topografía (pendiente) que conduce a la igualación de las cargas en el caso de pequeños caudales y grandes volúmenes. Por último, la variación de la temperatura, en general incrementos, conducen al aumento de la cinética en las reacciones biológicas y físico-químicas (biodegradación, reacciones químicas), principalmente por el aumento de las constantes de equilibrio, pero también por el aumento de solubilidad de algunos compuestos orgánicos (por ejemplo, la solubilidad del benceno en el agua aumenta 20% hasta 1900 mg/l, entre 10° C y 30° C) (Quevauviller *et al*., 2006; Fletcher *et al*., 2013).

- *Factores físico-químicos:* El primer factor es la variación de pH (potencial de iones de Hidrógeno) responsable de la modificación de las reacciones ácido-básicas. Otro factor físico-químico es el potencial *redox* E_H, el cual representa los procesos de oxidación y/o reducción de las sustancias presentes en el agua. Una disminución de la E_H puede dar condiciones sépticas (por ejemplo, EH ≤ 40 miliVoltios para un

pH = 7) y conducen a la producción de olores fétidos y a la presencia de sulfuros (Degrémont, 2005).

- *Factores Biológicos:* La degradación de las sustancias orgánicas en las redes de alcantarillado se debe principalmente a las reacciones biológicas anaeróbicas. Incluso si la concentración de oxígeno disuelto es muy baja (<2 mg/L), algunos procesos aeróbicos pueden ocurrir de la misma forma que en una planta de tratamiento biológico (Quevauviller *et al*., 2006).

7.1.1.3. Envejecimiento de las muestras

Al igual que dentro del alcantarillado, la composición de las aguas residuales puede variar muy rápidamente en la muestra (Baurès *et al*., 2004). Este fenómeno es conocido como el envejecimiento de la muestra, y se produce principalmente bajo la influencia de tres factores:

- El agua residual es un medio heterogéneo que se encuentra agitado en el alcantarillado. Sin embargo, los sólidos en suspensión se asientan rápidamente en el frasco de muestreo modificando la distribución del tamaño de la fracción particulada por efecto de floculación, adsorción, *etc.* (ver Figura 160).

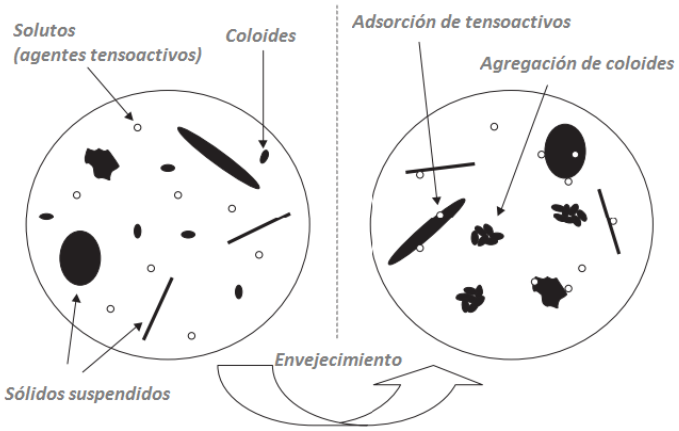


Figura 160- Fenómenos de agregación/adsorción durante los procesos de envejecimiento (adaptado de Baurès, *et al*., 2004)

- El segundo factor es de naturaleza química, ya que reacciones de reducción, la formación de complejos, la modificación de los equilibrios ácido-base, etc., se producen cuando el agotamiento del oxígeno disuelto conduce a condiciones anaeróbicas y a la variación del potencial redox y el pH. Por ejemplo, la adsorción de tensoactivos en sólidos en suspensión es responsable que, en las aguas residuales en bruto o tratadas, se genere agregación de la fracción coloidal y, por lo tanto se incremente la presencia de sólidos en suspensión (Baurès *et al*., 2004).

- El tercer factor es probablemente el más importante con el efecto de la biodegradación por los microorganismos presentes en las aguas residuales. La principal consecuencia es la degradación de la materia orgánica, en condiciones aerobias o anaerobias, como se puede observar en la Figura 161 (Quevauviller *et al*., 2006).

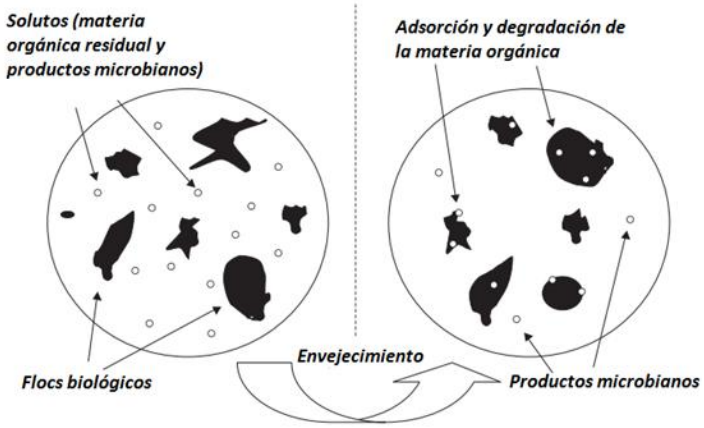


Figura 161- Degradación/adsorción durante los procesos de envejecimiento (adaptado de Baurès, *et al*., 2004)

7.1.2. Técnicas de muestreo y tipos de muestra

La toma de muestras en aguas residuales se realiza generalmente por uno de dos métodos: muestreo manual (puntual o compuesto) o muestreo automático (secuencial o compuesto). El primer método es simple, de bajo costo y se utiliza ampliamente, mientras que el segundo es mejor para monitorear la relevancia de un determinante, teniendo en cuenta la heterogeneidad y la variabilidad de las aguas residuales (Quevauviller *et al*., 2006).

La elección de un método de muestreo se relaciona con el objetivo del muestreo, los requisitos reglamentarios, la medición de la eficiencia de tratamiento, la gestión de alcantarillado, o en algunos casos simplemente por conocimiento (Muttamara, 1996).

Luego, existen tres tipos de muestra que pueden ser recolectadas por cada una de las técnicas de muestreo: puntual, compuesta o integrada.

7.1.2.1. Muestra puntual o simple.

Este tipo de muestra proporciona información sobre la calidad en un punto y momento dado, para luego analizar de ella los componentes de interés (ver Figura 162). Es útil para la detección de la fluctuación en la composición y la descarga de compuestos o sustancias, especialmente en las investigaciones de efluentes y aguas residuales industriales durante eventos de lluvia (Muttamara, 1996; Metcalf y Eddy, 2003). Dicha información se vuelve aún más relevante si a la toma de la muestra la acompaña la medición de caudal en el

mismo instante de tiempo. Sin embargo, esta muestra no es representativa del comportamiento del agua residual en un periodo de 24 horas, por ejemplo.

Figura 162- Muestra puntual en Shivajinagar-Bangalore, India (www.pulitzercenter.org)

Por otra parte, la fiabilidad de la medición y el análisis llevado a cabo a partir de una muestra puntual tomada al azar se limita a la composición de las aguas residuales para un punto de control e instante dado. El valor de un muestreo instantáneo es único y debe analizarse muy cuidadosamente, ya que puede ocurrir que se tome la muestra en un instante de concentración mínima o en un instante de concentración máxima, o que algún componente se presente en un mínimo cuando el resto se encuentre en sus valores máximos, o viceversa (WEF, 1996). Sin embargo, incluso si el procedimiento de toma de muestras puntuales parece ser simple, existen varias recomendaciones que deben hacerse para garantizar la fiabilidad de la medición:

- Uso de frascos limpios y adaptados de acuerdo al ensayo a realizar.
- Elegir un sitio de muestreo con una sección homogénea que prevenga la variabilidad de la calidad de las aguas residuales (especialmente en la medición caudal).
- Realizar los respectivos ensayos de laboratorio en el menor tiempo posible.
- Preservar y conservar las muestras en función del ensayo a realizar.
- Siempre tener en cuenta las condiciones de temperatura y clima durante el muestreo.

7.1.2.2. Muestra compuesta

Una muestra compuesta proporciona una estimación de la composición promedio del agua residual muestreada durante el período de tiempo que comprende el muestreo (con frecuencias variables como minutos, horas, días). Cada muestra puede ser tomada de forma automática o manual, como se puede ver en la Figura 163. Este tipo de muestras se aplica, por ejemplo, en el seguimiento de vertidos industriales cuya calidad puede variar mucho a lo largo de una jornada de trabajo (Simpson *et al*., 2013).

Figura 163- Toma de una muestra compuesta (isovolumétrica) en el punto Doña Juan del río Tunjuelo (RCHB-SDA, 2012)

Existen varios métodos para realizar la composición, los cuales se basan en el tiempo o en la proporcionalidad con respecto al flujo. La elección de un esquema de muestreo compuesto proporcional al flujo o al tiempo depende de la variabilidad del flujo de las aguas residuales o la concentración de determinantes, la disponibilidad del equipo y del lugar de muestreo. En general, un muestreo compuesto en función del tiempo es aceptable (e.g. tomar una muestra cada 15 minutos durante dos horas y al finalizar se combinan las muestras puntuales) (Quevauviller *et al*., 2006).

Una muestra compuesta en el tiempo consiste en la toma de muestras puntuales de igual volumen recogidas en intervalos de tiempo constantes en un mismo contenedor. Por otra parte, una muestra compuesta proporcional de flujo puede ser recogida usando uno de dos métodos. El primer método consiste en recoger un volumen constante de la muestra a diferentes intervalos de tiempo proporcional al caudal del agua residual. Para el otro método, la muestra se recoge mediante la variación del volumen de cada toma proporcional al caudal, manteniendo el mismo intervalo tiempo entre tomas (Simpson *et al*., 2013).

7.1.2.3. Muestra integrada

Una muestra integrada es similar a una muestra compuesta, se diferencian en que la integrada está conformada por varias muestras puntuales tomadas espacialmente en lugares diferentes, pero en un mismo instante de tiempo, y la compuesta está conformada por muestras puntuales en diferentes instantes, pero de un mismo lugar. Sin embargo, realizar la integrada en un mismo instante de tiempo no siempre es posible, sobre todo en colectores de aguas residuales de un ancho considerable o en el caso de los ríos, y por lo tanto se establece que la muestra integra debe ser realizada en el menor tiempo posible.

7.1.2.4. Muestreo Automático

Cada uno de los tipos de muestreo mostrados desde el numeral 7.1.2.1 al 7.1.2.3 pueden ser realizadas manualmente. Sin embargo, hoy día existen aparatos que realizan el muestreo de forma automática, y que facilitan esta actividad principalmente en los casos en los cuales la calidad del agua requiera de control permanente debido a su variabilidad, por ejemplo el afluente de una PTAR (Gonzalez *et al*., 2009).

Un muestreo automático principalmente se puede realizar utilizando el modo secuencial, por evento, de tipo puntual o compuesto y este último en función del tiempo o del volumen.

El modo secuencial completo realiza un muestreo en intervalos de tiempo regulares de un volumen dado recogido en recipiente. Después de tomar una muestra, el sistema de distribución se mueve dentro de la toma de muestras con el fin de llenar el siguiente recipiente, es decir, varios frascos se colocan en la toma de muestras (por lo general 24 ó 12), correspondientes a las muestras horarias o bihorarias (ver Figura 164). Este tipo de muestreo puede ser combinado con el muestreo compuesto, cuando se necesita una frecuencia de muestreo más alta, con la colección de submuestras de igual volumen a intervalos de tiempo regulares en los recipientes para cada hora del muestreo (*e.g.* 200 ml cada 15 min durante una hora en el recipiente X) (Quevauviller *et al*., 2006).

Figura 164- Estación de monitoreo de una red pluvial con muestreo automático en Greenville, Carolina del Sur-E.E.U.U. (www.apwa.net)

Las muestras proporcionales de flujo pueden ser recogidos directamente con un muestreador automático, cuando éste es conectado a un dispositivo de medición de flujo compatible. Un muestreador automático también puede ser utilizado para recoger muestras puntuales, de acuerdo a la programación que se realice en el muestreador (*e.g.* incrementos o disminución del caudal por fuera de una banda de confianza). Esto se conoce como muestreo por evento. No solamente las mediciones de flujo pueden ser utilizadas para realizar el muestreo automático, otros equipos como sondas de turbiedad, conductividad, pH, *etc.*, pueden ser empleadas para este fin (Simpson *et al*., 2013).

7.1.3. Análisis de laboratorio

Con respecto a los análisis, todos los países del mundo utilizan reconocidos métodos para la detección y cuantificación de determinantes (Quevauviller *et al*., 2006). Tradicionalmente, la calidad de las aguas residuales se define mediante la medición de los determinantes globales, tales como la demanda biológica de oxígeno (DBO), demanda química de oxígeno (DQO), carbono orgánico total (TOC), sólidos suspendidos totales (SST), *etc*. (Bourgeois *et al*., 2001). En los últimos años, los determinantes más específicos, tales como nitrógeno total, fósforo total, hidrocarburos aromáticos policíclicos, halógenos orgánicos absorbibles, *etc*., y una lista de sustancias peligrosas han aparecido, por ejemplo, en el contexto de la Directiva Marco del Agua (2000/60/CE). Sin embargo, existe una tendencia en el sentido de aceptar métodos de ensayo rápidos o instrumentación en línea.

A continuación se proporciona información básica sobre lo que es un método estándar de laboratorio para la detección de los determinantes de interés estudiados en el presente trabajo: DQO, DQOf y SST, y el método específico implementado en las muestras de los casos de estudio se presentará desde el numeral 7.2.1 al 7.2.2.

7.1.3.1. Detección de los SST

La detección de Sólidos Suspendidos Totales (SST) se relaciona con las formas de partículas de materia orgánica o mineral en el agua que no pasan a través de un filtro de 0.45 µm. Normalmente, se realiza la separación de SST por filtración o centrifugación, seguido por secado a 105° C. El método de centrifugación se utiliza cuando la filtración no es aplicable debido a un alto riesgo de obstrucción de los filtros. Los sólidos decantados corresponden a los SST que se decantan durante un tiempo convencionalmente de dos horas (Quevauviller *et al*., 2006; Gonzalez *et al*., 2009).

Las concentraciones de los SST varían ampliamente en las redes de alcantarillado. En las redes de agua de lluvia (ALL), las concentraciones pueden variar de 1 mg/L a alrededor de 4000 mg/L. La misma magnitud de variación ocurre en tiempo húmedo en los sistemas de alcantarillado combinado (SAC-H), aunque las concentraciones mínimas son ligeramente más altas. Las concentraciones en aguas residuales domésticas (ARD) varían entre 100 y 600 mg/L. Las concentraciones mínimas son generalmente un poco más altas en los sistemas combinados (en tiempo seco (SAC-S)), esto se debe a que los efectos de dilución generados por las aguas lluvias no están presentes (Lepot, 2012).

Concentraciones (mg/L)			Tipo de agua	Fuente
Mínimo	Media	Máximo		
1	190	4582	ALL	(Chocat *et al*., 2007)
200	-	800	ALL	(Chocat, 1997)
0	-	250	ALL	(Doyen, 1992)
84	-	3526	ALL	(Anta *et al.*, 2006)
50	-	590	ALL	(Anta *et al.*, 2006)
300	-	600	ARD	(Chocat *et al.*, 1997)
72	-	456	ARD	(Grange y Pescheux, 1986)
200	-	1000	ARD	(Chocat *et al.*, 1997)
50	-	700	ARD	(Marchandise *et al.*, 1978)
154	288	509	SAC-H	Grommaire *et al.*, 2008)
-	350	-	SAC-H	(Butler y Karunaratne, 1995)
53	358	2035	SAC-H	(Threllfall *et al.*, 1991)
176	425	647	SAC-H	(Chocat *et al.*, 2007)
50	-	800	SAC-H	(Grange, 1994)
176	-	2500	SAC-H	(Bertrand-Krajeswki, 2006)
74	-	874	SAC-H	(Kafi-Benyahia, 2006)
-	100	-	SAC-H	(EEUU-EPA, 1983)
21	190	2582	SAC-H	(Ellis, 1989)
67	-	101	SAC-H	(Metcalf y Eddy, 1991)
3	650	11000	SAC-H	(Novotny y Olem, 1994)
150	-	500	SAC-S	(Bertrand-Krajeswki, 2006)
40	-	250	SAC-S	(Doyen, 1992)
81	-	960	SAC-S	(Ruban *et al.*, 1993)
40	-	758	SAC-S	(Kafi- Benyahia, 2006)

Tabla 14- Valores de las concentraciones de SST observados en las redes de saneamiento (adaptado de Lepot, 2012)

7.1.3.2. Detección de la DQO

La prueba de DQO es ampliamente utilizada para medir la resistencia orgánica de los residuos domésticos e industriales, la cual a menudo reemplaza la necesidad de cuantificar la DBO como el principal determinante en las aguas residuales. La detección de la DQO se basa en el hecho de que la mayoría de los compuestos orgánicos se pueden oxidar por la acción de agentes oxidantes fuertes en condiciones ácidas (Bourgeois *et al*., 2001).

Por lo tanto, la medición de la DQO se lleva a cabo sobre la base del 'reflujo cerrado y el método de medición colorimétrico' como se describe en las normas de calidad del agua (NF T90-101/ISO 6060:1989/Método EPA 410.3) (Quevauviller *et al*., 2006). La muestra 'en blanco' y el estándar están contenidos en tubos sellados, los cuales son calentados en un horno o bloque digestor en presencia de dicromato a 150° C. Después de dos horas, se sacan los tubos del horno o del digestor, se enfrían y luego se mide la absorbancia en un espectrofotómetro UV-Visible en la longitud de onda de 600 nm. El método descrito esencialmente consiste en medir la cantidad de oxígeno requerido por las sustancias presentes en la muestra, ya que tiene en cuenta cualquier compuesto o elemento que presenta un carácter reductor (Bourgeois *et al*., 2001).

La principal ventaja de la prueba de DQO es la rapidez, en el cual se pueden obtener resultados (aproximadamente 2 horas en lugar de 5 días para la DBO_5). Sin embargo, una de las principales limitaciones de esta prueba es su incapacidad para distinguir entre la materia orgánica biodegradable y la biológicamente inerte presente en la muestra (ver numeral 7.2.2).

Por otra parte, Lepot (2012) recopila de la literatura una serie de rangos de concentración de la DQO de acuerdo al tipo de red de alcantarillado: pluvial, residual y combinada (en tiempo seco y lluvia), tal como se presentan en la Tabla 15.

Para este determinante, los rangos de variación son menos extensos a diferencia de los presentados para SST. Las menores concentraciones se presentan en los alcantarillados de aguas lluvias y los combinados en tiempo húmedo (ALL, SCA-H). Las concentraciones medias de eventos lluvia (ALL, SCA-H) están por debajo de las concentraciones medias durante tiempo seco (ARD, SCA-S). En cuanto a las concentraciones máximas observadas se tienen valores similares para los diferentes tipos de redes (500 a 1500 mg/L). Por último, es interesante la consistencia de los valores de concentración en alcantarillados combinados en tiempo seco, a diferencia de la gran variabilidad de los valores máximos reportados en tiempo de lluvia (entre 73 y 1600 mg/L) (Lepot, 2012).

Concentraciones (mg/L)				Tipo de agua	Fuente
Mínima	Mediana	Promedio	Máximo		
20	-	85	365	ALL	(Chocat *et al.*, 2007)
20	-	-	500	ALL	(Bertrand-Krajeswki, 2006)
50	-	-	1500	ALL	(Chocat *et al.*, 1997)
25	-	89	180	ALL	(Anta *et al.*, 2006)
92	-	212	388	ALL	(Anta *et al.*, 2006)
205	-	-	1284	ARD	(Grange y Pescheux, 1986)
-	-	720	-	ARD	(Chocat *et al.*, 1997)
250	-	380	530	ARD	(Chocat *et al.*, 2007)
200	-	-	1100	ARD	(Chocat *et al.*, 1997)
100	-	-	1400	ARD	(Marchandise *et al.*, 1978)
300	-	-	1000	ARD	(Bertrand-Krajeswki, 2006)
70	-	-	1600	SAC-H	(Grange, 1994)
42	-	-	900	SAC-H	(Bertrand-Krajeswki, 2006)
-	-	65	-	SAC-H	(EEUU-EPA, 1983)
20	-	85	365	SAC-H	(Ellis, 1989)
40	-	-	73	SAC-H	(Metcalf y Eddy, 1991)
108	-	-	809	SAC-H	(Kafi-Benyahia, 2006)
131	-	-	1512	SAC-S	(Ruban *et al.*, 1993)
96	-	-	1084	SAC-S	(Kafi-Benyahia, 2006)

Tabla 15- Valores de las concentraciones de la DQO observados en las redes de saneamiento (adaptado de Lepot, 2012)

7.1.3.3. Detección de la DQO filtrada

Para determinar la porción de DQO filtrada en una muestra, se puede realizar el mismo procedimiento señalado en el numeral anterior, pero la muestra antes de ser calentada y titulada con dicromato debe ser filtrada. Esta parte del proceso tiene por fin separar la mayoría de sólidos insolubles de los sólidos coloidales y solubles, y así cuantificar la cantidad de materia orgánica que los microorganismos no están consumiendo, es decir la fracción inerte coloidal y soluble (por ejemplo en algunos efluentes de tratamientos anaeróbicos esta fracción permanece inalterada (Field, 1987)).

ANEXO E-2

7.1.3.4. Clases de monitoreo de acuerdo al tipo de determinante

En el contexto del proyecto *SWIFT-WFD* (*European Union's Sixth Framework Project*, *Water Framework Directive)*-(Sexto Proyecto Marco de la Unión Europea, financiado por la Dirección General de Investigación), Allan *et al*. (2006) elaboraron un inventario de las técnicas existentes y emergentes para el monitoreo de la calidad del agua. En dicho inventario se clasifican las herramientas de monitoreo, incluyendo métodos/equipos de medición de la siguiente forma (Gonzalez *et al*., 2009):

- Características físico-químicas (por ejemplo, carbono orgánico total, pH, temperatura, nutrientes).
- Sustancias químicas prioritarias (por ejemplo, Hidrocarburos Aromáticos Policíclicos (HAP), pesticidas, metales).

- Los efectos de la presencia de los determinantes (por ejemplo, mortalidad y efectos subletales, tales como estrogenicidad, reducción de la alimentación o de la actividad locomotriz).

A continuación se describen cada una de las categorías de las herramientas de monitoreo, presentando distintos métodos y tecnologías que son consideras brevemente, haciendo énfasis en la primera categoría ya que la mayoría de instrumentos de monitoreo en línea (*on-line*) disponibles en la actualidad son usados para este tipo de determinantes (ver numeral 7.1.3.5). No obstante, la evolución de los sistemas de monitoreo en línea está orientada a la detección y cuantificación de los determinantes clasificados en las otras categorías de monitoreo.

a) Monitoreo de las características físico-químicas

Determinantes tales como el pH pueden ser medidos directamente mientras que para otros se utilizan medidas indirectas, como por ejemplo para determinar la salinidad por medio de conductividad. Nutrientes tales como amonio, nitrito, nitrato, fosfato, y más generalmente nitrógeno total y fósforo pueden ser indicadores útiles en los programas de vigilancia, ya que están involucrados en los procesos de eutrofización, y algunos (por ejemplo, nitratos) pueden contaminar el agua subterránea después de las aplicación de fertilizantes. Otros determinantes se pueden utilizar para caracterizar el grado de oxigenación de un cuerpo de agua, e incluyen oxígeno disuelto, demanda química de oxígeno, demanda bioquímica de oxígeno, condiciones redox o mediciones de respirometría. La cantidad de oxígeno disuelto, la materia orgánica en suspensión se pueden evaluar de diferentes maneras: mediante la medición de materia orgánica total, carbono orgánico total, aromaticidad, y turbidez del agua. A su vez, la presencia y niveles de materia orgánica influyen fuertemente en la DQO de una muestra de agua. Los diferentes tipos de dispositivos están disponibles en una serie de electrodos específicos, sensores ópticos, UV-Visible espectroscopia, colorimetría, quimioluminiscencia, métodos volumétricos o cromatografía de iones (Colin y PP Quevauviller, 1997; Gonzalez *et al*., 2009). En la Tabla 5 se presentan algunos determinantes y las diferentes técnicas analíticas por medio de las cuales pueden ser detectados y cuantificados.

Determinante	Electrodo especifico	Técnicas ópticas	Cromatografía Iónica	Polarografía	Titulometría
Amonio	√	UV, V, C, L	√		√
BOD					
COD	√	UV, V, L			√
Conductividad	√				
Oxígeno disuelto	√			√	
Materia orgánica		UV			
pH	√				
Fosfato		V,C,N	√		√
Redox	√				
TOC	√	IR, UV			
Total nitrógeno		UV-V			
Total fosforo		V, C, N			
Turbiedad		N, UV, IR			

L= Luminiscencia, C= Colorimetría, IR= Infrarrojo, N= Nefelometría, UV= Ultravioleta, V=Visible

Tabla 16- Técnicas analísticas disponibles (comerciales y en desarrollo) para monitoreo físico-químico Métodos analíticos para detectar la (Gonzalez *et al*., 2009)

b) Monitoreo de sustancias químicas prioritarias

En algunos continentes y países existen marcos regulatorios (*Environmental Protection Agency*-EE.UU y *Water Framework Directive*-Union Europea), que priorizan una serie de sustancias consideradas como perjudiciales o potencialmente perjudiciales. Éstas se clasifican normalmente en tres categorías principales: compuestos orgánicos no polares (por ejemplo, algunos pesticidas, y algunos productos químicos industriales, tales como los HAP), compuestos orgánicos polares (algunos pesticidas y productos farmacéuticos), y los metales pesados (por ejemplo, mercurio y cadmio). Una amplia gama de técnicas pueden emplearse para su detección, tales como cromatografía (cromatografía de gases o cromatografía líquida) vinculados a un detector sensible (por ejemplo ionización de llama, captura de electrones, espectrómetro de masas, espectrómetro de fluorescencia o un espectrómetro UV), y para los metales métodos tales como plasma acoplado inductivamente a espectrometría de masas, o espectrometría de absorción atómica. Algunos de estos métodos se basan en reacciones de formación de complejos específicos cuyos productos (generalmente de color) pueden ser cuantificados en un colorímetro o un espectrofotómetro. Otros pueden ser medidos directamente en la base de la modificación de un haz electromagnético para producir un espectro característico (por ejemplo, espectroscopia UV-Visible o Infra Rojo), o por medio de la emisión de una característica en el espectro (por ejemplo, espectroscopia de fluorescencia y luminiscencia) (Gonzalez *et al*., 2009; Quevauviller *et al*., 2006). La muestra como ejemplo algunas de las principales clases de sustancias químicas prioritarias establecidas en *WFD-EU* y los diferentes métodos que se pueden utilizar para su análisis.

	Ejemplo	Técnicas ópticas	Biosensor/ Bioensayos	Inmunoensayo	Electroquímica
Metales	Cd, Hg, Ni. Pb	√			√
PAH	Benzo pireno	√	√	√	
	Benzo fluoranteno				
Compuestos policlorados	Hexaclorobenceno		√	√	
	Pentaclorofenol				
Pesticidas	Atrazina, Diuron		√	√	
Disruptores endocrinicos	Nonilfenol	√	√	√	

Tabla 17- Sustancias prioritarias en *WFD-EU* (Gonzalez *et al*., 2009)

c) Monitoreo de los efectos de los determinantes

Las mediciones de la respuesta biológica a los niveles de los determinantes presentes en el ambiente pueden ofrecer un enfoque alternativo o complementario a las mediciones de las concentraciones de sustancias químicas individuales para la evaluación de la calidad del agua y para la identificación de las tendencias de la calidad. En general, existe una relación proporcional entre la concentración de un analito en el agua y su impacto en un organismo vivo. Esta vigilancia biológica se puede basar en los cambios medidos en el nivel de todo el organismo, los tejidos, las células o los mecanismos bioquímicos aislados. Los datos pueden ser cualitativos, semicuantitativos o cuantitativos. Estas técnicas de control biológico incluyen bioensayos, biomarcadores y sistemas biológicos de alerta temprana (sigla en inglés *BEWS*) (Gonzalez *et al*., 2009).

Los bioensayos se basan en el uso de organismos completos (incluyendo una gama de animales (vertebrados e invertebrados), levaduras, algas y bacterias) o de partes aisladas de los organismos (incluyendo células y tejidos aislados, sistemas de enzimas). Estos pueden ser usados para detectar o cuantificar los niveles de contaminantes orgánicos e inorgánicos o para medir la toxicidad general de las muestras de agua. La toxicidad es la aplicación más usada, la cual depende en parte de la biodisponibilidad de las sustancias tóxicas, y esto se ve afectado por factores tales como la presencia de materia en suspensión y la concentración de carbono orgánico disuelto. También se ve afectada por las variables fisicoquímicas tales como el pH, la temperatura, la dureza del agua, la tensión de oxígeno y la salinidad. La información sobre los límites de detección de diversas sustancias químicas y los procedimientos operativos estándar se suministra normalmente por los productores. Estos ensayos suelen depender de la utilización de muestras *in situ* y del laboratorio. Sin embargo, es difícil añadir conservantes a las muestras para el transporte, ya que en algunos casos afectarían notablemente el resultado del ensayo. Algunos ensayos pueden llevarse a cabo *in situ* o en línea mediante el despliegue de organismos en los sistemas directamente en el medio ambiente de retención, o por su inclusión en un flujo de agua bombeada desde la masa de agua que se está supervisando. Algunos ensayos utilizan una gama de especies de organismos con diferentes sensibilidades a los diferentes tipos de contaminantes con el fin de proporcionar la máxima sensibilidad en toda la gama de los contaminantes encontrados (Colin y Quevauviller, 1997; Gonzalez *et al*., 2009).

7.1.3.5. Principios de análisis instrumental y equipos de monitoreo *in situ* (*on line*) para la detección y cuantificación de determinantes en aguas residuales

Avances importantes en la teoría de control y cada vez mayores capacidades de los sensores, se han presentado en las últimas dos décadas para aguas residuales. Además, la demanda sobre la calidad del agua se hace cada vez más restricta en el aspecto regulatorio. Una percepción común es que los sensores representan el eslabón más débil de la aplicación *on-line* de control y seguimiento de aguas residuales (adaptado de Harremoës *et al*., 1993). Sin embargo, el rendimiento y la fiabilidad de muchos sensores en línea (por ejemplo, sensores de nutrientes, respirómetros) han mejorado notablemente durante la última década y hoy en día se pueden utilizar directamente en muchas de las diferentes estrategias de seguimiento y control (Jeppsson *et al*., 2002).

A continuación se presentan algunos métodos y tecnologías empleados para la detección de los principales determinantes estudiados en la calidad del agua residual:

***a) Conductividad*:**

La conductividad es una medida que determina el nivel de concentración iónica en una solución. Entre más sales, ácidos o bases estén disociados en la solución, mayor será la lectura de conductividad. En el agua, la conductividad está principalmente relacionada con iones salinos, por lo que puede tratársele como un índice de la carga de sal, en el agua residual, o de pureza, en la potable. La conductividad de un electrolito se determina con una resistencia electroquímica. En su configuración más simple, la celda de medición utiliza dos electrodos a los que se les aplica un voltaje alternante. Se mide, entonces, la corriente eléctrica que es directamente proporcional a los iones libres en el electrolito. El instrumento electrónico calcula la conductividad de la solución tomando en consideración la constante absoluta de la celda del sensor (factor geométrico de la celda de medición) (ver Figura 165) (Vanrolleghem y Lee, 2003). La conductividad se expresa, generalmente, en S/cm (o mS/cm). La escala de soluciones acuosas inicia con el agua ultra pura, cuyo valor es de 0.05 µS/cm a 25 °C. Las aguas naturales, como el agua potable o superficial, suelen estar en el rango de 100 a 1000 µS/cm. La parte superior de la escala la ocupan algunos ácidos y bases (WTW, 2012).

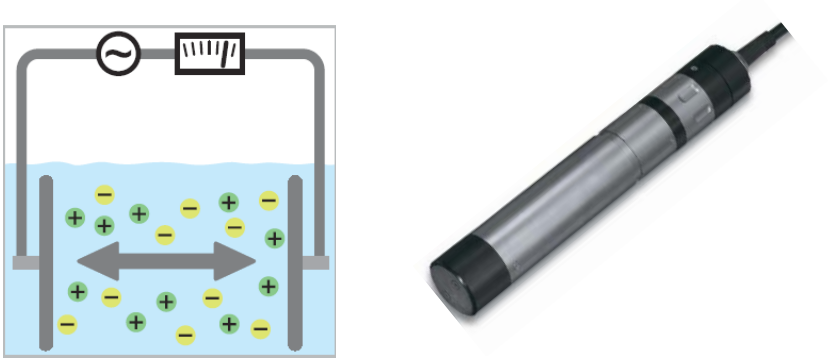


Figura 165- Ilustración del principio de medición de un conductivimetro (WTW, 2012; Endress & Hauser, 2012)

b) Turbiedad:

Típicamente, este determinante se mide con el principio de luz difusa a 90 grados según la norma EN ISO 7027, mediante una luz infrarroja con una longitud de onda de 860 nm (WTW, 2012). Claramente fuera del rango visible, este haz de luz no se ve afectado por los colores de la muestra. Si se introduce radiación óptica a un sistema de dispersión, los sólidos disueltos la transforman en otro tipo de energía y reducen su intensidad. A este efecto se le conoce como absorción. La relación entre la energía inicial y la final se define como turbidez. Los sensores para medición en línea se calibran desde la fábrica con un método de puntos múltiples. Este procedimiento es muy estable, por lo que no se necesitará re-calibrar. Se utiliza formazina, diluida a la concentración necesaria (de acuerdo al rango de medición), como estándar de calibración. Bajo este método la luz dispersa se mide a un ángulo de 90 grados, lo cual es ideal para turbidez en los rangos medio y bajo hasta 4000 UNF (Unidades Nefelométricas de Formazina) (WTW, 2012).

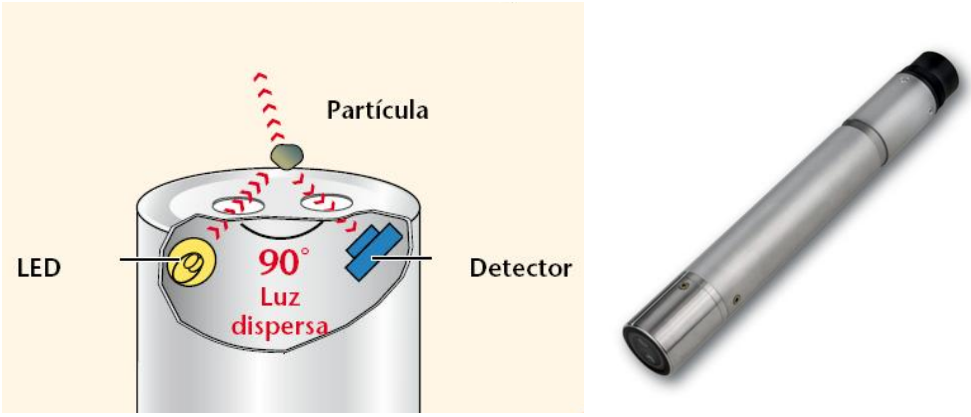


Figura 166- Ilustración del principio de luz dispersa para cuantificar la turbiedad en un medio acuoso (WTW, 2012)

d) Sólidos Suspendidos Totales:

Probablemente la variable más importante en los procesos de tratamiento de aguas residuales es la concentración de sólidos en suspensión (SS). Tres principios de medición han encontrado aplicación para este caso: medición óptica, el ultrasonido y la espectrometría de dieléctrico.

La dispersión de la luz incidente y la absorbancia por partículas en suspensión han sido métodos de estimación de la biomasa tradicional fuera de línea (Kennedy *et al*., 1992). Aunque es imposible hacer un cálculo directo (*a priori*) de la concentración de peso en seco de cualquier medida de la "densidad óptica", los sistemas pueden proporcionar estimaciones razonables siempre que se realiza una calibración regular entre los valores de concentración en laboratorio y la dispersión de la luz. Diferentes principios han sido llevados a la práctica. Generalmente, parte de la luz es absorbida, otra parte pasa a través de la muestra (transmisión), y por último, la dispersión de la luz en todas las direcciones se produce. Dicha dispersión no es homogénea con el ángulo en que inicialmente incidió la

luz: la dispersión hacia delante es la más pronunciada, mientras que la retrodispersión es el menos eficaz (Vanrolleghem y Lee, 2003).

Dependiendo de la concentración de sólidos, una o la otra medición de la luz será la más beneficiosa. Se prefieren las técnicas de dispersión para las muestras de una baja concentración de sólidos, típicamente turbidez de efluente. La retrodispersión también es bastante ventajosa en los sistemas de alto contenido en sólidos donde la absorción puede ser demasiado alta para permitir su medida. La mayoría de los sensores disponibles en el mercado utilizan una fuente que emite luz en el rango inferior al visible y/o infrarrojo cercano, que tiene la ventaja de que la mayoría de los medios de comunicación tienen una absorbancia baja en este rango (Olsson y Nielsen, 1997).

e) Demanda Química de Oxígeno-DQO:

diferentes métodos se han desarrollado, varios son el resultado de modificación de los métodos originales de laboratorio enfocados a su aplicación en línea. Los cambios más importantes se refieren al método de digestión aplicada. Meredith (1990), por ejemplo, propone no utilizar dicromato como un agente oxidante, en cambio propone el uso de peróxido de hidrógeno acoplado a la luz UV para producir ozono *in situ*. Productos químicos tales como el cromo y el ácido producen residuos peligrosos líquidos que requiere su eliminación. Dichos residuos no solamente son perjudiciales para el ambiente, sino afectan la medición de este determinante (Korenaga *et al*., 1990; Meredith, 1990). Aun no se ha desarrollado una técnica donde la muestra no requiera de un tratamiento físico o químico anteriores a su cuantificación. Sin embargo, se han desarrollado métodos que permiten reducir los análisis de laboratorio a través de la calibración de funciones matemáticas, como se estudiará en los numerales 1.3 y 2.3.

f) Demanda Biológica de Oxígeno-DBO:

Tradicionalmente, el componente biodegradable de las aguas residuales se mide por la norma, fuera de línea, el método de la demanda bioquímica de oxígeno (DBO_5). La DBO_5 es una medida de la cantidad de materia orgánica a ser consumida u oxidada en 5 días por la acción de los microorganismos. Sin embargo, la prueba de DBO_5 es inadecuada para la supervisión y el control automatizado debido al tiempo requerido para completar la prueba y la dificultad de obtener mediciones consistentemente precisas. Por lo tanto, la medición en línea de la carga DBO en las aguas residuales se basa en la estimación a corto plazo (BOD_n). Hay tres tipos de métodos BOD_n en línea que son actualmente utilizados: métodos respirométricos, sondas microbianas y espectrometría UV-Visible. A continuación se describirán los dos primeros métodos brevemente, mientras el tercero será explicado en detalle en el siguiente numeral (Vanrolleghem y Lee, 2003).

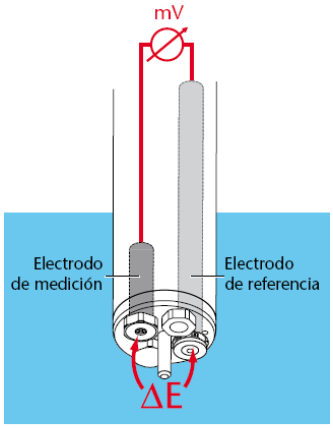
El método respirométrico utiliza respirómetros continuos en los cuales la BOD_n se calcula a partir de un respirograma resultante del balance de masa de oxígeno en la cámara de respiración, medición que se puede realizar cada 30 minutos.

La sonda microbiana se compone de células inmovilizadas, una membrana y un electrodo de oxígeno disuelto. La señal de la sonda es una medida de la actividad de las células, que es, a su vez, una medida de la concentración de sustrato en las aguas residuales (DBO). Estos sensores microbianos requieren un cuidadoso almacenamiento y mantenimiento para mantener su eficacia, ya que tienen una vida útil corta, desde unos pocos días hasta varios meses, y requieren baja temperatura de almacenamiento de 4 a 10° C. Debido a las reacciones bioquímicas que participan en la detección de DBO, su aplicación también está limitada por su especificidad con respecto al substrato, el pH, la temperatura y su susceptibilidad a los cambios químicos agresivos de la matriz de compuesto del agua residual (Vanrolleghem y Lee, 2003).

g) Nitrógeno:

El nitrógeno se encuentra en muchos compuestos y de muchas formas, por lo que se le considera el máximo "artista de cambio". En las aguas residuales domésticas se le encuentra principalmente como úrea, parcialmente convertido a amonio por el proceso de amonificación. No obstante, existen procesos de nitrificación en los cuales se oxida el nitrógeno del agua convirtiendo el nitrito a nitrato, para lo cual se requiere oxígeno. Luego de este proceso puede ocurrir la desnitrificación, en la que el nitrato (NO_3-N) se convierte en nitrógeno elemental (N_2) ante la ausencia de oxígeno (WTW, 2012).

Actualmente es común utilizar un método de medición que se enfoca en la disponibilidad de oxígeno en el agua residual y relacionarlo con la presencia de nitrógeno. Sin embargo, existen métodos para determinar de forma directa amonio, nitrato y nitrito que resultan mucho más interesantes pues permiten optimizar el control con lecturas inmediatas. Para realizar este tipo de mediciones en continuo existen sondas que utilizan Electrodos de Ion Selectivo (sigla en inglés *ISE*) basadas en el principio de medición potenciométrico (Cammann, 1979). La sonda incluye al menos dos electrodos, una referencia y un electrodo de medición. El electrodo de medición está equipado con una membrana especial, a la cual se le unen de forma reversible iones específicos. Dependiendo de la actividad de los iones medidos en el líquido, un número variable de iones se unirá al electrodo de medición. La selectividad ocasiona una tensión (ΔE) entre el electrodo de medición y el de referencia que se puede leer con la ayuda de un transformador, como se muestra en la (WTW, 2012). El potencial medido se pone en relación con la actividad del ión medido por medio de una función de calibración (Winkler *et al*., 2004).



$$\Delta E = E_{(ISE)} - E_{(Ref)}$$

Figura 167- Principio potenciómetro para la medición de nitrógeno (WTW, 2012)

Por último, con respecto al monitoreo de amonio en las redes de alcantarillado, se ha detectado que el valor de concentración de este elemento depende en un 50 % de la topografía de la red, lo cual genera que se reduzca la presencia de nitrógeno orgánico que no se hidroliza, posteriormente el amonio medido representa sólo el 50% o menos de la carga total de nitrógeno. Sobre todo en los sistemas de alcantarillado combinados durante los eventos de lluvia esta porción puede disminuir aún más. Luego, mediciones de laboratorio paralelas tienen que ser llevadas a cabo para estimar el N-total en relación con NH_4-N para las diferentes condiciones típicas en el lugar de muestreo. Por lo tanto, se reconoce que un *ISE* de este tipo en una red de alcantarillado tendrá una exactitud limitada, pero por otro lado la operación de la sonda es relativamente sencilla y proporciona un registro continuo de la concentración de amonio, que posteriormente debe ser evaluado por expertos para evidenciar el impacto que generan las condiciones del lugar sobre la medición (Winkler *et al*., 2004).

h) Fósforo:

Se considera que los compuestos del fósforo –particularmente el ortofosfato– limitan los nutrientes en las aguas tanto estancadas como en movimiento. Un incremento en su concentración, ocasionado por aguas residuales y otras fuentes, resulta en la eutrofización del agua cuyos efectos comunes son un aumento en la población de algas, lo cual se traduce en el agotamiento del oxígeno para la vida acuática en el ecosistema, y en condiciones anóxicas en las regiones más profundas. El fósforo se presenta en el agua de 3 diferentes formas naturales: i) inorgánico, orto-fosfato disuelto, ii) compuestos fosfóricos orgánicos disueltos, y iii) fósforo en partículas (unido a la biomasa o a otras partículas (Korostynska *et al*., 2012); WTW, 2012).

Normalmente el monitoreo del fósforo y sus derivados se lleva a cabo por medio de la recolección de muestras de forma manual, seguido por un proceso de filtración y su respectivo análisis en el laboratorio. Varias estrategias de detección para el fosfato incluye electrodos de iones selectivos basados en técnicas potenciométricas, voltamétricas indirectas basadas en la reacción del fosfato con diversos metales y complejos asociados,

y biosensores basados en las reacciones enzimáticas donde el fosfato actúa como un inhibidor o sustrato (Villalba *et al*., 2009). El principal método para la detección del fósforo utiliza un foto sensor que mide en una longitud de onda del espectro UV-Visible la presencia de un color específico (*e.g.* azul o amarillo) que resulta de una reacción química entre el fósforo y el reactivo especial (Figura 168). La concentración del colorante resultante indica la concentración de fósforo en la muestra. Existen dos métodos ópticos UV-Visible estándar, a saber, el método del azul de molibdeno (longitud de onda 880 nm) y el método amarillo de vanadato/molibdato (longitud de onda 380 nm) (Korostynska *et al*., 2012). Ambas se basan en la medición de ortofosfato, ya que para determinar el fósforo total es necesario realizar la digestión del fósforo ya sea orgánico disuelto o en partículas generalmente, calentando la muestra con peroxodisulfato y ácido sulfúrico. Por lo tanto, se debe utilizar una muestra no filtrada para incluir toda la materia sólida en el proceso de digestión (WTW, 2012).

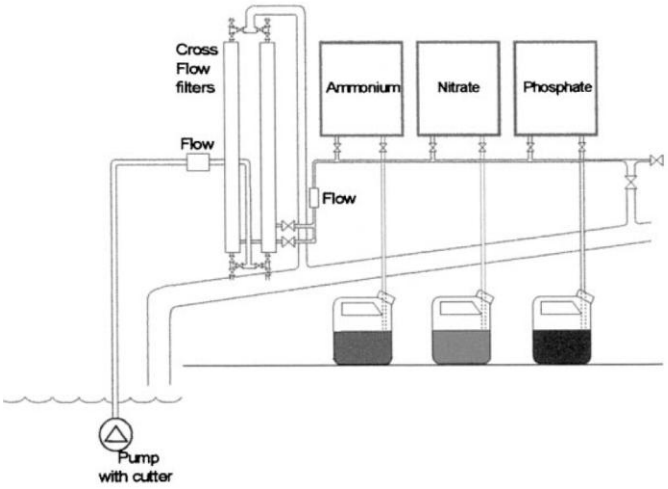


Figura 168- Sistema de muestreo y filtración de analizadores de nutrientes (Lynggaard-jensen, 1999)

Como se pudo evidenciar, la mayoría de determinantes de interés de las aguas residuales pueden ser detectados y cuantificados por diferentes técnicas análisis instrumental. Sin embargo, existen técnicas que permiten detectar y cuantificar a partir de una sola medición múltiples determinantes en un mismo instante de tiempo, sin necesidad de realizar un tratamiento físico o químico a la muestra y basadas en la interacción entre el medio y la luz: espectroscópicas. Una técnica basada en este principio es la espectrometría UV- Visible, la cual será explicada en detalle a continuación.

ANEXO E-3

7.1.4. Espectroscopía, espectrometría y espectrofotometría

Procesos espectroscópicos se basan en el hecho de que la radiación electromagnética (REM) interactúa con los átomos y las moléculas en forma discreta para producir la absorción característica o perfiles de emisión. Antes de que podamos investigar el origen de los espectros, tenemos que ver algunas de las propiedades de la REM. Nuestra propia

capacidad de percibir el color se debe a que el ojo humano actúa como un detector para REM. La propiedad de la REM que determina la gama de color que se percibe es la longitud de onda. La parte del espectro electromagnético que el ojo puede detectar es conocida como la región visible. REM puede ser simplemente representada como una onda sinusoidal. La longitud de onda, λ, es la distancia entre los picos o depresiones adyacentes. Nuestra capacidad de percibir el color depende de muchos factores. Sin embargo, el mecanismo de interacción de la REM con la materia es de gran importancia. Desde un punto de vista visual la detección es nuestra capacidad para percibir diferentes colores que depende de la óptica del proceso involucrado, por ejemplo, si la luz es absorbida o reflejada por el objeto observado. La longitud de onda, λ, de la REM puede ser expresada como una función de su frecuencia, ν, y la velocidad de la luz, c, por la Ecuación 61 (Ojeda y Rojas, 2009):

$$\nu = \frac{c}{\lambda}$$

Ecuación 61-

La Figura 169 muestra la sensibilidad relativa del ojo a la luz visible. Esta figura ilustra la importancia de la sensibilidad del detector y el rango de longitud de onda para los detectores espectroscópicos. Sin embargo, la REM se comporta como una partícula y como onda (la naturaleza dual de la luz), y la longitud de onda de dicha partícula, un fotón, está relacionada con la energía por la Ecuación 62.

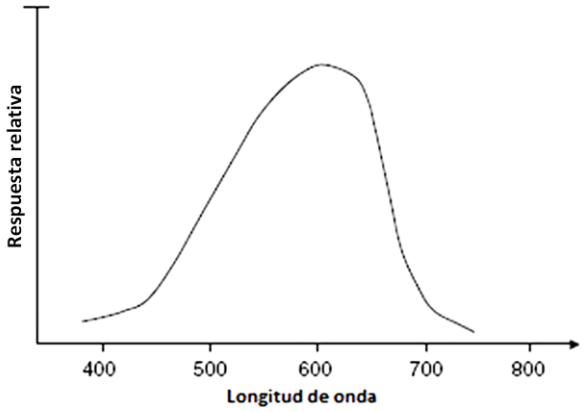


Figura 169- Sensibilidad espectral del ojo como un detector

$$E = \frac{hc}{\lambda}10^{9}$$

Ecuación 62-

donde *h* es la constante de Planck (6.63 x 10^{-34} Julios·segundo), *c* es la velocidad de la luz en el vacío (2.998 × 10^{8} m/s), *E* es la energía del fotón y *λ* es la longitud de onda en nanómetros (nm). La región visible del espectro electromagnético constituye una parte pequeña, como puede verse en la Figura 170. Es evidente que existe un enorme lapso de energías de más de 18 órdenes de magnitud. La ecuación que vincule la energía a la

longitud de onda es de importancia fundamental en la espectroscopia y se discutirá con más detalle en la siguiente sección.

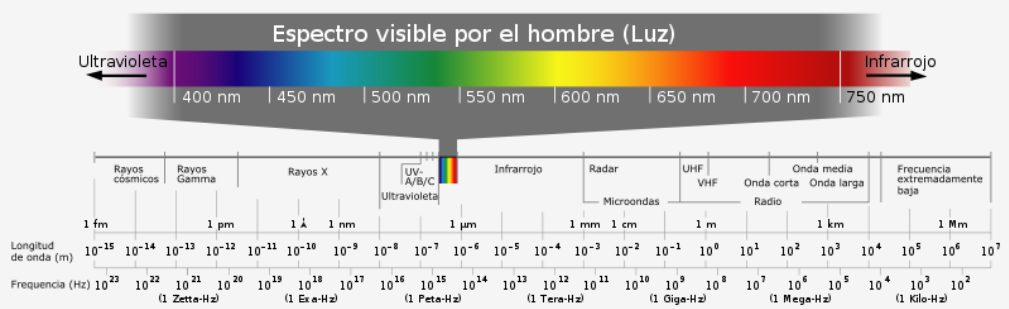


Figura 170- Espectro electromagnético - Frecuencia y longitud de onda (Horst Frank, 2006 [*on line*: www.es.wikipedia.org/wiki/Archivo:Electromagnetic_spectrum-es.svg])

7.1.4.1. Origen de un espectro: absorción de la radiación por átomos, iones y moléculas

Cuando un fotón interactúa con una nube de electrones de materia, lo hace de una manera específica y discreta. Esto es en contraste la atenuación física de la energía por un filtro que es continuo. Este proceso de absorción discreto es cuantificado y las energías asociadas a éste se relacionan con el tipo de transición involucrado (Thomas y Burgess, 2007).

El anterior proceso se puede ilustrar por medio de un cálculo simple utilizando la -. Por lo tanto, se puede suponer que un fotón de energía 8.254×10^{-19} Julios interactúa con la nube de electrones de una molécula en particular y hace que se promueva un electrón desde un estado basal (inicial) a un estado excitado. Esto se ilustra en la Figura 171, donde la diferencia en los niveles de energía moleculares, E_2 - E_1, corresponde exactamente a la energía del fotón, lo cual revela una transición electrónica en la parte ultravioleta del espectro (λ= 240 nm) (Thomas y Burgess, 2007).

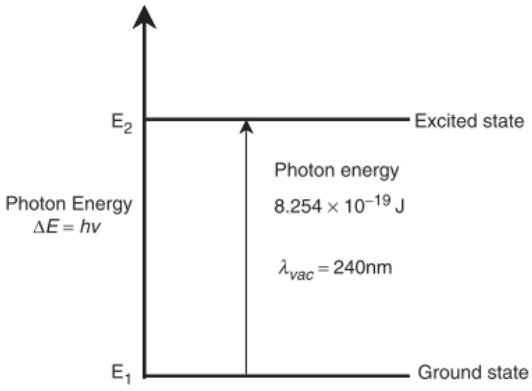


Figura 171- Captura de un fotón por un molécula (Thomas y Burgess, 2007)

Entonces la absorción en cada longitud de onda depende de la diferencia entre los niveles de energía. Algunas transiciones requieren menos energía y por lo tanto aparecen en

longitudes de onda más largas. Luego, la espectrometría UV-Vis se define como la cuantificación de las transiciones electrónicas de componentes orgánicos e inorgánicos que absorben radiación en la zona del ultravioleta o visible (Da Silva, 2008).

Sin embargo, cuando se realizan mediciones espectrales, es necesario tener en cuenta otros procesos ópticos. Esto es particularmente importante para espectrometría en soluciones. Cuando la luz incide sobre una cubeta que contiene una muestra con una sustancia de interés (soluto) en una solución (disolvente), otros procesos ópticos pueden ocurrir, tales como: transmisión, reflexión, refracción, dispersión y luminiscencia. Todos estos procesos, junto con los efectos instrumentales, se combinan para distorsionar o degradar la calidad del espectro, lo cual se debe minimizar para reducir su impacto y obtener mediciones confiables y representativas (Thomas y Burgess, 2007).

ANEXO E-4

7.1.4.2. Instrumentación UV-Visible

El instrumento utilizado en la espectrometría ultravioleta-visible se llama espectrofotómetro UV-Vis. Mide la intensidad de luz que pasa a través de una muestra (I), y la compara con la intensidad de luz antes de pasar a través de la muestra (I_0) cuantificando el valor de absorbancia.

Las partes básicas de un espectrofotómetro son una fuente de luz (a menudo una bombilla incandescente para las longitudes de onda visibles, o una lámpara de arco de deuterio en el ultravioleta) que en conjunto cubren el rango normalmente de 200 a 800 nm (dependiendo en parte del tipo de detector), un soporte para la muestra, una rejilla de difracción o monocromador para separar las diferentes longitudes de onda de la luz y un detector (ver Figura 172). El detector suele ser un fotodiodo. Los fotodiodos se usan con monocromadores, que filtran la luz de modo que una sola longitud de onda alcanza el detector con el fin de mejorar la sensibilidad de las medidas de absorbancia (Da Silva, 2008).

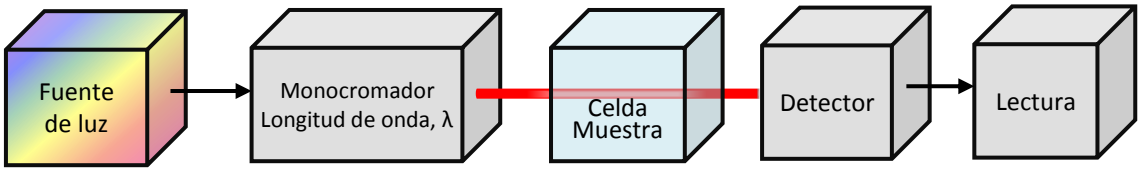


Figura 172- Partes de un espectofotometro (Adaptado de Ojeda y Rojas, 2009)

Un espectrofotómetro puede ser único o de doble haz. En un instrumento de un solo haz, toda la luz pasa a través de la célula muestra. La I_0 debe medirse retirando la muestra. En un instrumento de doble haz, la luz se divide en dos haces antes de llegar a la muestra: un haz se utiliza como referencia, y el otro haz de luz pasa a través de la muestra. Algunos instrumentos de doble haz tienen dos detectores (fotodiodos), y el haz de referencia y el de la muestra se miden al mismo tiempo (Ojeda y Rojas, 2012).

Como contenedores de muestra se utilizan cubetas que generalmente deben estar hechas de un material transparente a la radiación, como sílice o cuarzo para la región UV-Visible y vidrio o plástico para la región visible. Sin embargo, las cubetas sólo son factibles para uso *off-line.* No obstante, nuevos materiales (como la fibra óptica) conectados a sondas sumergibles pueden ser más adecuados para realizar análisis espectroscópicas *on-line*. Las fibras ópticas son, junto con los espejos y ventanas, componentes ópticos pasivos de gran interés para su uso en diferentes aplicaciones y también como enlaces de comunicación de datos ópticos (Sporea y Sporea, 2005).

Una variedad de detectores están disponibles para las mediciones de UV-Visible. Los detectores más comunes se presentan de acuerdo al rango de la región UV-Vis en el cual funcionen, (Tabla 18).

Tipo de Detector	Rango de trabajo útil (nm)
Fotodiodo de silicio	350-1100
Tubos fotomultiplicadores	160-1100
Dispositivo de carga acoplado	180-1100
Series de fotodiodos	180-1100

Tabla 18- Diferentes tipos de detectores UV-visible y rangos de trabajo útiles en nanómetros

ANEXO E-5

7.2. ENSAYOS DE LABORATORIO

7.2.1. Detección y cuantificación de los SST

Los sólidos suspendidos totales fueron cuantificados mediante ensayos estándar de laboratorio, para lo cual se siguieron los protocolos de la normatividad vigente en Colombia (TP0088-03-2007) y en Francia (NF EN 872-2005). En ambos casos el procedimiento es similar y se puede describir en seis pasos (Figura 173):

i. Agitar las muestras para evitar el asentamiento de los sólidos y garantizar la homogeneidad. Se usó un agitador magnético a 800 rpm.
ii. Pesar en la balanza el crisol junto con el disco de fibra de vidrio (filtro diámetro de 47 mm) m_{s1}, m_{s2}, m_{s3}. Luego, disponer ambos elementos en el aparato de filtración (bomba de vacío).
iii. Tomar de la muestra principal tres submuestras de volúmenes iguales (V_1, V_2, V_3)
iv. Cuando entre en funcionamiento la bomba de vacío depositar cada submuestra en cada uno de los crisoles. La bomba de vacío debe operar por un lapso de tiempo de 1 minuto.
v. Colocar en un horno a 105° C los crisoles-filtros durante 2 horas.
vi. Por último, pesar los crisoles-filtros en la balanza para obtener los pesos secos de las submuestras m_{p1}, m_{p2}, m_{p3}.

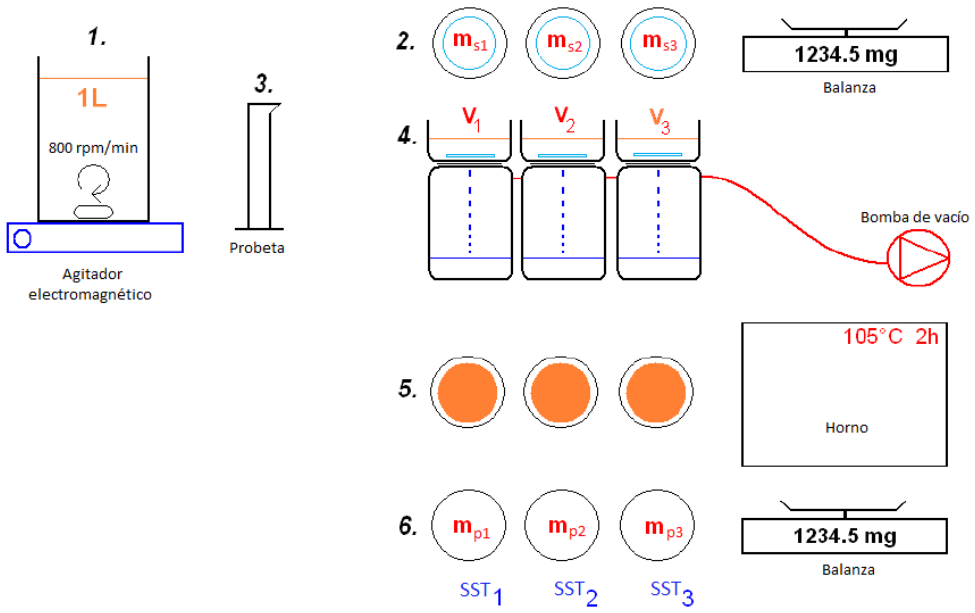


Figura 173- Principio de medición de los SST por triplicado

Para cuantificar la concentración de SST se usa la siguiente ecuación:

$$SST_i = \frac{(m_{pi} - m_{si})}{V_i}$$

Ecuación 63-

donde SST_i es la concentración de SST en la submuestra i (en mg/L), m_{si} es la masa del filtro limpio para la submuestra i (en mg), m_{pi} es la masa del filtro cargado para la submuestra i (en mg) y V_i el volumen de la submuestra i a filtrar (en L).

7.2.2. Detección y cuantificación de la DQO y de la DQOf

La demanda química de oxígeno total y filtrada fue cuantificada mediante ensayos estándar de laboratorio, para lo cual se siguieron los protocolos de la normatividad vigente en Colombia (TP0086-05-2007) y en Francia (NF T90-101-2001). En ambos casos el procedimiento es similar, y se puede describir en cinco pasos (Figura 174 y Figura 175):

i. Mezclar fuertemente la muestra por dos minutos (Figura 174) en el caso de la DQO total o preparar la bomba de vacío y un filtro de vidrio de diámetro 47 mm para filtrar la muestra en el caso de la DQO filtrada (Figura 175).
ii. Con la pipeta tomar de la muestra 2 mL y cuidosamente inyéctarlo en tubo que contiene el reactivo (H_2SO_4). Tapar herméticamente el tubo y agitar suavemente por inversión para homogenizar.
iii. Precalentar el microdigestor a 150° C. Luego colocar los tubos en el digestor y manteniendo una temperatura de 150° C incubar las muestras por dos horas.
iv. Sacar las submuestras del digestor y dejar enfriar.

v. Ajustar el espectrofotómetro a la longitud de onda de 620 nm y llevar a cero con el blanco o patrón de referencia. Colocar en cada celda las submuestras y realizar la medición del valor absorbancia en la longitud de onda indicada.

Figura 174- Principio de medición DQO por triplicado (adaptado de Lepot, 2012)

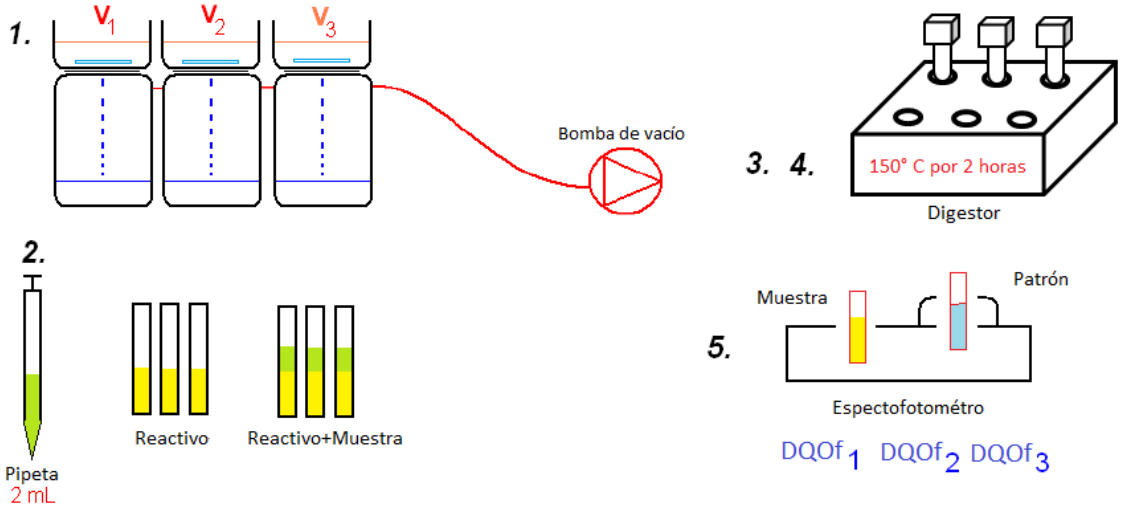


Figura 175- Principio de medición DQO filtrado por triplicado (adaptado de Lepot, 2012)

La Ecuación 64 se utiliza para cuantificar la concentración de DQO y DQOf.

$$DQO(f)_i = \frac{\left(Abs_i(620nm) - k_1\right)}{k_2}$$

Ecuación 64-

donde $DQO(f)$ es la concentración de DQO o DQOf de la submuestra i (en mg/L), $Abs_i(620\ nm)$ es el valor de absorbancia para la submuestra i en la longitud de onda 620 nm, k_1 y k_2 son los coeficientes calibrados para el aparato de medición.