Universitat de Girona

CHEMOMETRIC METHODS TO PROCESS ONLINE SPECTROMETRY FOR QUALITY MONITORING OF DIFFERENT WATER MATRICES

Mireia Plà Castellana

TESI DOCTORAL INDUSTRIAL

CHEMOMETRIC METHODS TO PROCESS ONLINE SPECTROMETRY FOR QUALITY MONITORING OF DIFFERENT WATER MATRICES

Mireia Plà Castellana

2023

TESI DOCTORAL INDUSTRIAL

CHEMOMETRIC METHODS TO PROCESS ONLINE SPECTROMETRY FOR QUALITY MONITORING OF DIFFERENT WATER MATRICES

Mireia Plà Castellana

2023

Programa de Doctorat en Ciència i Tecnologia de l'Aigua

Dirigida per

Oriol Gutiérrez Wolfgang Gernjak Jordi Raich-Montiu

Tutelada per

Joaquim Comas Matas

Memòria presentada per optar al títol de doctora per la Universitat de Girona

Certificat de Direcció de la tesi doctoral industrial

Universitat
de Girona

El Dr. Oriol Gutiérrez de l'Institut Català de Recerca de l'Aigua (ICRA),

el Dr. Wolfgang Gernjak professor d'investigació Institució Catalana de Recerca i Estudis Avançats (ICREA) a l'Institut Català de Recerca de l'Aigua (ICRA), i

el Dr. Jordi Raich-Montiu de l'empresa s::can Iberia Sistemas de Medición S.L.U.,

DECLAREM:

Que el treball titulat *Chemometric Methods to Process Online Spectrometry for Quality Monitoring of Different Water Matrices*, que presenta Mireia Plà Castellana per a l'obtenció del títol de doctor/a, ha estat realitzat sota la nostra direcció i que compleix els requisits per poder optar a Menció Internacional i industrial.

I, perquè així consti i tingui els efectes oportuns, signem aquest document.

Signat,

Oriol Gutiérrez

Wolfgang Gernjak

Jordi Raich–Montiu

Girona, a dia 28 de Setembre de 2022

Llistat d’Abreviatures

BTEX	Benzè, Toluè, Etilbenzè, Xilè
CAT	Consorci d’Aigües de Tarragona
COD	Carboni Orgànic Dissolt
COT	Carboni Orgànic Total
DBO	Demanda Biològica d’Oxigen
EB1	Estació de Bombament
EDAR	Estació Depuradora d’Aigües Residuals
ERA	Estació de Regeneració d’Aigua
ETAP	Estació de Tractament d’Aigua Potable
$PostO_3$	Post-Ozonització
$PreO_3$	Pre-Ozonització
ST	Sòlids Totals
STS	Sòlids Totals en Suspensió
THM FP	Trihalomethane Formation Potential
UV-Vis	Ultraviolat-Visible

List of Abbreviations

AI	Artificial Intelligence
AIC	Akaike's Information Criterion
ANN	Artificial Neural Networks
BAC	Biological Activated Carbon
BE	Backward Elimination
BTEX	Benzene, Toluene, Ethylbenzene, Xylene
DWTP	Drinking Water Treatment Plant
EB1	Pumping Station
FS	Forward Selection
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multivariate Linear Regression
NTU	Nephelometric Turbidity Unit
PAC	Polyaluminium chloride
PCA	Principal Component Analysis
PCR	Principal Component Regression
PostO_3	Post-Ozonization
PreO_3	Pre-Ozonization
RF	Random Forest
RMSE	Root Mean Squared Error

RSE	Relative Standard Error
SS	Stepwise Selection
SVM	Support Vector Machines
THM FP	Trihalomethane Formation Potential
TOC	Total Organic Carbon
UV-Vis	Ultraviolet-Visible
VIF	Variance Inflation Factor
WRP	Water Reclamation Plant
WWTP	Wastewater Treatment Plant

Aquesta tesi la dedico al meu arbre:
Al Francisco i la Maria Dolors, les arrels que em fan ser qui soc
A l’Antònia, el tronc que em fa anar sempre pel bon camí
Al Josep i l’Anna, les branques que m’abracen sempre que ho necessito
A l’Olívia i la Vilma, les fulles que són el meu oxigen, tan necessari

*A*graïments

La realització d'aquesta tesi doctoral no hagués estat possible sense l'ajuda de moltes persones al llarg de tot aquest temps, i els vull expressar el meu agraïment més sincer.

L'agraïment profund per tota l'ajuda incommensurable dels meus tres directors de tesi, Dr. Oriol Gutiérrez, Dr. Wolfgang Gernjak i Dr. Jordi Raich-Montiu. Moltes gràcies pel vostre temps i per tot el coneixement aportat al llarg d'aquest recorregut.

Agraeixo de tot cor als treballadors i les treballadores de l'empresa s::can Messtechnik GmbH (seu de s::can – Viena, Àustria) la seva ajuda i dedicació a la feina ben feta. En particular, li agraeixo a l'Andreas Weingartner la confiança dipositada en mi i tot el bon humor que sempre ha transmès i que el caracteritza. També m'agradaria agrair de forma particular als companys de *Solutions* i I+D durant la meva estada predoctoral a Viena: la Dra. Janelcy Alferes, el Semi Araya, el Cristian Haselberger, el Philip Worschischek i el Lukas Kornfeind, gràcies per la paciència amb el meu anglès i les bones cerveses. Evidentment, per tu Judith Benet tinc un agraïment particular, perquè va ser molt maco (i necessari) poder-te conèixer, i perquè emocionalment em vas ajudar molt, tot i que potser no n'eres conscient. Ets una persona que omples les habitacions de llum, i això espero que no ho perdis mai.

Vull agrair d'una manera especial al Dr. Jordi Raich-Montiu, gerent de la filial s::can Iberia Sistemas de Medición S.L.U., per confiar amb mi per dur a terme aquesta investigació, per totes les bones reunions i converses compartides i per l'oportunitat laboral, de coneixement i vital que ha suposat fer un Doctorat Industrial. També m'agradaria agrair profundament a les dues persones que van viure de ben a prop la meva trajectòria com a investigadora i com a tècnica d'instal·lacions de sondes de s::can Iberia. Primerament, al Roger Peris, qui va ser el meu mentor en moltes de les primeres tasques com a tècnica, a qui vaig poder preguntar tot el que necessitava, sempre molt proper i disposat a ajudar-me. Sé que en moltes ocasions em vaig fer un pèl pesada, i per això t'agraeixo tota la paciència que vas tenir. A l'Ester Tejedor li reservo un paràgraf per a ella tota sola, que se'l mereix.

A tu Ester, t'ho agraeixo tot. Haver demostrat una força incommensurable, un esperit de canvi i de superació de l'adversitat dignes d'admirar. Ser tan intel·ligent emocionalment, i haver-me obert els braços quan em sentia tan sola. Haver-me ajudat a continuar, a través de les teves paraules d'ànim diàries, sense defallir ni un moment. Haver estat sempre, SEMPRE, al meu costat, en els pitjors i els millors moments, fins i tot quan no m'ho mereixia. Aquesta tesi ha significat l'inici d'una molt bona amistat, i espero que no s'acabi per res del món. Moltes gràcies!

També m'agradaria tenir unes paraules de record i agraïment profund a totes les persones que em vau acompanyar durant aquest trajecte vital des de l'Institut Català de Recerca de l'Aigua, el Dr. Wolfgang Gernjak i el Dr. Oriol Gutiérrez, no només com a directors de tesi, si no també com a mentors i referents. M'agradaria fer una referència especial a les meves dues companyes de despatx, la Sílvia Busquets i la Dra. Inma Velo. Totes les paraules d'agraïment que tinc són poques per poder-vos transmetre com em vau ajudar durant l'inici del doctorat amb les rialles, les converses profundes, les teories conspiranòiques, els acudits de gats, enviant cartes no gaire amables al Rei, i tantes altres coses impossibles de nombrar. Als Dr. Luca Sbardella i Dr. Federico Ferrari per intentar passar desapercebuts durant tant de temps, i acabar sent persones tan properes. Al Lluis Bosch, l'Adrià Riu, el Dr. Gianluigi Buttiglieri, la Dra. Lucia Gusmaroli, la Nikoletta Tsiarta, i a tots els doctorands, doctorandes, estudiants post-doctorals i tècniques de laboratori amb qui he compartit algun àpat o cafè i amb els i les que sempre és tan amable mantenir-hi una conversa. I, evidentment, m'agradaria dedicar-li un agraïment particular i molt sincer a la Dra. Maria José Farré. Gràcies per totes les estones de laboratori en les que em vas ajudar tant, i no només a extreure els DBPs de l'aigua del CAT. Et mereixes totes i cada una de les fites que has assolit perquè hi ha molt poca gent que posi tanta dedicació i passió en la feina com ho fas tu.

M'agradaria agrair al Departament d'Informàtica, Matemàtica Aplicada i Estadística en general, i en particular al Dr. Josep Antoni Martin i al seu grup de recerca per tota l'ajuda estadística rebuda. Vau convertir un problema complicat en un problema complicat que tenia probabilitats de solucionar-se aplicant-hi temps i dedicació.

Agraeixo la col·laboració a les persones del Consorci d'Aigües de Tarragona per tota la informació cedida durant un any i escaig de diferents punts de tractament de la planta, i als tècnics Lluís Cladelles i al Carlos Pozuelo en particular per totes les bones estones passades i les converses interessants anant a dipòsits d'aigua. Amb vosaltres vaig aprendre que ser *curiós* a l´hora de fer/mantenir una instal·lació es condició *sine qua non* perquè surti bé. També, m'agradaria agrair l'ajuda rebuda per part de l'Andreu Fargas i a l'Agustí Colom per totes les dades rebudes i reunions mantingudes. Un agraïment efusiu pel Dr. Josep Ruana, per tota l'ajuda rebuda, no només en termes de tractaments de l'ETAP, sinó també en quimiometria i models matemàtics. Josep, li dec un llibre, encara! Des del CAT sempre heu sumat, i us ho agraeixo de tot cor.

Agraeixo també la col·laboració dels treballadors de l'ERA Camp de Tarragona, AITASA i Veolia Water Technologies per tot el suport i la informació cedida de l'entrada i la sortida del sistema Actiflo® durant sis mesos. En particular, m'agradaria agrair profundament tota l'ajuda rebuda per part del Dr. Joan Sanz i sobretot de l'Alfonso Garcia. Gràcies per tota la informació i el temps que em vau donar.

Vull agrair a la família que no es tria, el Francisco Pla, la Maria Dolors Benet, la Maria Antonia Pla, el Josep Pla, l'Anna Castellana, la Núria Boleda i el Xavier Farrús, totes les estones de dinar on la pregunta "quan acabaràs això que fas" era casi obligatòria, totes les vegades que, fent la cervesa al Foment, es deia "vaig explicar que feies una tesi, però no recordo de què, i vaig dir de *números*". Al final, sembla que me n'he sortit! A l'Olívia i la Vilma Pla els agraeixo el fet d'existir, ras i curt. En definitiva, a tots i totes vosaltres us vull agrair ser el Primer Arbre del Bosc, estar sempre al meu costat, a les verdes i a les madures.

També vull agrair a la família política que no es tria, la Lourdes Grané, el Josep Mujal, la Núria Mujal i l'Anna Mujal, i també a les nenes i al nen de la nova generació, ser un suport tan gran quan la situació va fer un gir cap a l'estranger. Agraeixo la forma que teniu de ser, tant properes, fent-me sentir de casa i com a casa.

També vull agrair a la família que sí que es tria, el Dr. Sergi Torramadé, la Clàudia Saltó, el Marc Torns, el Dr. Eudald Mujal, la Txell Tur i la Gabi Fuentes, persones que he tingut a prop des de fa molts i molts anys, que sou la vitamina dels divendres, i que m'heu fet costat durant aquest procés, que moltes vegades és més amarg que dolç. Agraeixo infinitament haver-vos conegut i tenir-vos al costat.

A les persones que m'han acompanyat durant aquests dos últims anys també m'agradaria mostrar-los un agraïment efusiu i sincer. Particularment a la Caroline

Sielfeld, la Dra. Ieva Sapkaite i la Dra. Lídia Fernàndez els he d'agrair que s'hagin convertit en la meva xarxa de seguretat, les persones a les que sé que puc acudir sense ni qüestionar-m´ho, si ho necessito. És en els moments més baixos on trobaràs les persones que són llum i que val la pena mantenir a prop, i a vosaltres us he d'agrair haver-me fet costat en alguns moments molt durs, i haver-me fet veure lo necessari que és desconnectar per poder tornar amb les piles carregades.

A la Bridgette Sanders per la paciència, la bona conversa i la correcció de la part en anglès d'aquesta tesi. Per tot el que he après amb ella, no només de vocabulari i pronunciació de l'anglès, sinó a confiar en mi mateixa quan parlo en una llengua que no és la meva.

I l'agraïment més sentit i sincer és per tu, Eudald, que has estat al meu costat durant tot aquest procés tan llarg i tant complicat, i que tantes vegades hagués deixat córrer. Una no sap el que significa la solitud fins que no fa una tesi doctoral i és ben cert que no ha sigut gens fàcil. Però ho he aconseguit, i m'he sentit molt recolzada i abraçada per tu cada dia de tots aquests anys i això ha sigut gràcies a la teva manera de ser, la teva paciència, i la teva estima incondicional. N'hi ha que diuen que *la vida que ens hem perdut simplement no existeix*... Quina sort no haver-me perdut la meva amb tu!

Índex general

Índex de figures

Índex de taules

Resum

La monitorització de l'aigua durant els processos de potabilització o sanejament en les Estacions de Tractament d'Aigua Potable (ETAP), Estacions Depuradores d'Aigües Residuals (EDAR) i Estacions de Regeneració d'Aigua (ERA), és un pas necessari per a observar el funcionament dels sistemes involucrats, i controlar contaminants que estan al punt de mira de les administracions. Cada vegada és més comú obtenir informació de qualitat a través de sensors i sondes instal·lats *in situ* i en continu. La present investigació ha tingut com objectiu dotar la sonda spectro::lyser® de més capacitats en la detecció i predicció de contaminants, tant en aigua potable com en aigua residual.

A través de l'espectre Ultraviolat Visible (UV–Vis), i utilitzant mètodes d'inferència estadística avançada i eines quimiomètriques, com ara els motors de selecció de variables Forward Selection (FS), Backward Elimination (BE) i Stepwise Selection (SS), i algoritmes de regressió com Multiple Linear Regression (MLR), Support Vector Machines (SVM) i Artificial Neural Networks (ANN), s'han desenvolupat diferents models matemàtics per a predir el potencial de formació de trihalometans (THM FP) durant el procés de sanejament de l'aigua potable en ETAP, la concentració d'una mescla d´hidrocarburs (toluè, m-xilè i p-xilè) en aigües residuals urbanes provinents d'influent d'EDAR, i la concentració de coagulant afegit a l'aigua residual durant el procés de coagulació-floculació en ERA.

Els resultats demostraren que el millor sistema de selecció de variables va ser SS, combinat amb una selecció manual de variables posterior usant el Variance Inflation Factor (VIF). En la modelització duta a terme per predir el THM PF s'ha conclòs que ANN (R^2 = 0,92, RMSE = 0,77) va ser l'algoritme amb més capacitats per observar les relacions no-lineals entre les longituds d'ona. Per altra banda, en la modelització desenvolupada per a predir la presència d'una mescla controlada d'hidrocarburs en aigua residual, l'algoritme de MLR va donar molt bons resultats (R^2_{MLR1} = 0,82; $RMSE_{MLR1}$ = 0,22; R^2_{MLR2} = 0,87; $RMSE_{MLR2}$ = 0,21; R^2_{MLR3} = 0,79; $RMSE_{MLR3}$ = 0,24). Per predir la concentració òptima de coagulant que s'ha d'afegir durant el procés de coagulació-floculació, es conclou que ANN és l'algoritme que va donar millors resultats (R^2 = 0,86, RSE = 0,02).

Els algoritmes desenvolupats són específics de cada punt estudiat, i poden utilitzar-se com una eina per a donar una resposta ràpida i eficaç, en cas de que els operadors de planta en tinguin necessitat.

Resumen

La monitorización del agua durante los procesos de potabilización o saneamiento en las Estaciones de Tratamiento de Agua Potable (ETAP), Estaciones Depuradoras de Aguas Residuales (EDAR) y Estaciones de Regeneración de Agua (ERA) es un paso necesario para observar el funcionamiento de los sistemas involucrados, y controlar contaminantes que están en el punto de mira de las administraciones. Cada vez es más común obtener información de buena calidad a través de sensores y sondas instalados *in situ* y en continuo. La presente investigación ha tenido como objetivo dotar a la sonda spectro::lyser® de más capacidades en la detección y predicción de contaminantes, tanto en agua potable como en agua residual.

A través del espectro Ultravioleta Visible (UV–Vis), utilizando métodos de inferencia estadística avanzada y herramientas quimiométricas, como los motores de selección de variables Forward Selection (FS), Backward Elimination (BE) y Stepwise Selection (SS), y algoritmos de regresión como Multiple Linear Regresion (MLR), Support Vector Machines (SVM) y Artificial Neural Networks (ANN), se desarrollaron diferentes modelos matemáticos para predecir el potencial de formación de trihalometanos (THM FP) durante el proceso de saneamiento del agua potable en ETAP, la concentración de una mezcla de hidrocarburos (tolueno, m-chileno y p-chileno) en aguas residuales urbanas provenientes de influyente de EDAR, y la concentración de coagulante añadido al agua residual durante el proceso de coagulación–floculación en ERA.

Los resultados demuestran que el mejor sistema de selección de variables fue SS, combinado con una selección manual posterior usando el Variance Inflation Factor (VIF). En la modelización llevada a cabo para predecir el THM FP se concluyó que ANN (R^2 = 0,92, RMSE = 0,77) fue el algoritmo con mayores capacidades para observar las relaciones no–lineales entre las longitudes de onda. Por otra parte, en la modelización desarrollada para predecir la presencia de una mezcla controlada de hidrocarburos en agua residual, el algoritmo de MLR obtuvo muy buenos resultados (R^2_{MLR1} = 0,82; $RMSE_{MLR1}$ = 0,22; R^2_{MLR2} = 0 ,87; $RMSE_{MLR2}$ = 0,21; R^2_{MLR3} = 0,79; $RMSE_{MLR3}$ = 0,24). Para predecir la concentración óptima de coagulante a añadir durante el proceso de coagulación-floculación, se concluyó que ANN es el algoritmo que da mejores resultados (R^2 = 0,86, RSE = 0,02).

Los algoritmos que se desarrollaron son específicos de cada punto estudiado, y pueden utilizarse como una herramienta para dar una respuesta rápida y eficaz, en caso de que los operadores de planta tengan necesidad de ello.

Abstract

Water monitoring during the purification and sanitation processes in Drinking Water Treatment Plants (DWTP), Wastewater Treatment Plants (WWTP) and Water Reclamation Plants (ERA) is a necessary step to obtain a wide overview of water treatment processes and to monitor pollutants that are in the spotlight of the public administrations. Obtaining high quality information through sensors and probes installed *in situ* and on line is highly increasing. The aim of this investigation is to provide the spectro::lyser® probe more capabilities for contaminant detection and prediction, both in drinking and wastewater.

By using the Ultraviolet Visible (UV–Vis) spectrum, coupled with advanced statistical inference methods and chemometric tools, such as Forward Selection (FS), Backward Elimination (BE) and Stepwise Selection (SS) variable selection motors, and regression algorithms such as Multiple Linear Regression (MLR), Support Vector Machines (SVM) and Artificial Neural Networks (ANN), different mathematical models were developed. This allows to predict the trihalomethanes formation potential (THM FP) during water sanitation in DWTP, the concentration of a selected hydrocarbons mixture (toluene, m-xilene and p-xilene) in urban wastewaters at the WWTP's influent, and the concentration of coagulant added during the coagulation-flocculation process in ERA.

The results show that the best variable selection system is SS, combined with a subsequent manual selection using Variance Inflation Factor (VIF). In the model developed to predict the THM FP, ANN (R^2 = 0.92, RMSE = 0.77) is the algorithm with the greatest capabilities to observe non-linear relationships between wavelengths. On the other hand, in the model developed to predict the presence of a controlled hydrocarbon mixture in wastewater, the MLR algorithm obtained very good results (R^2_{MLR1} = 0.82; $RMSE_{MLR1}$ = 0.22; R^2_{MLR2} = 0.87; $RMSE_{MLR2}$ = 0.21, R^2_{MLR3} = 0.79, $RMSE_{MLR3}$ = 0.24). To predict the optimal concentration of coagulant to add during the coagulation-flocculation process ANN gave the best results (R^2 = 0.86, RSE = 0.02).

The developed algorithms are specific to each particular Plant studied, and can be used as a tool to provide a quick and efficient response when necessary.

Capítol 1 – Introducció

1.1. Presentació

La present tesi doctoral s'ha realitzat a través de la col·laboració del departament de Tecnologies i Avaluació (TiA) de l'Institut Català de Recerca de l'Aigua (ICRA), l'empresa s::can Iberia Sistemas de Medición, S.L.U. i la Universitat de Girona (UdG), sota la direcció dels doctors Oriol Gutiérrez, Wolfgang Gernjak, i Jordi Raich-Montiu. La tesi doctoral ha estat cofinançada per l'Agència de Gestió d'Ajuts Universitaris (AGAUR) en el marc dels Ajuts a Doctorats Industrials (DI) que s'atorguen semestralment per la Generalitat de Catalunya, i l'empresa s::can Iberia Sistemas de Medición, S.L.U. La recol·lecció de dades s'ha dut a terme a través d'empreses i consorcis privats, com Veolia i el Consorci d'Aigües de Tarragona (CAT), entre d'altres, que han permès la instal·lació dels sensors spectro::lyser® de s::can, per obtenir grans quantitats de dades d'alta qualitat i observar les relacions entre paràmetres durant llargs períodes de temps. També s'ha realitzat una estada predoctoral de tres mesos a Viena (Àustria) a s::can Messtechnik GmbH, la seu de l'empresa s::can, amb la qual s'opta a la menció internacional del doctorat.

La present tesi doctoral està redactada com a tesi clàssica, organitzada amb els capítols generals d'introducció, metodologia, discussió, conclusions i referències, on hi ha la informació integrada dels tres capítols que es presenten a l'apartat de resultats. Aquests capítols estan organitzats com a articles científics (resum, introducció, metodologia, resultats i discussió, conclusions) i no s'han enviat perquè els resultats que se'n deriven tenen una component de secret industrial que fa que no hagi sigut possible.

Seguint un fil argumental basat en el cicle de l'aigua, els capítols s'han organitzat en l'ordre de tractaments que segueix l'aigua des de que es potabilitza fins que es depura i es pot reutilitzar, per a recàrrega d'aqüífers i usos urbans. Els capítols d'introducció, metodologia i discussió integrada (capítols 1, 2 i 6) s'han redactat en català, mentre que el resum de la tesi, els capítols específics de resultats (capítols 3, 4 i 5) i les conclusions generals (capítol 7) s'han redactat en anglès per tal d'optar a la menció internacional del doctorat.

Les bases de dades utilitzades per a desenvolupar els algoritmes dels capítols de resultats i el codi creat amb el programari lliure RStudio, 2009–2019 (versió 1.2.5033) no s'han afegit com a annexes a la tesi doctoral, no només per la gran quantitat d'informació que presenta, si no perquè també està subjecta a secret industrial.

A la llista de Referències (capítol 8) s'hi inclouen els treballs citats als capítols generals d'introducció, metodologia, discussió integrada i conclusions, així com als capítols específics de resultats (capítols 3 a 5).

1.2. L'aigua

L'aigua és el recurs natural més important del nostre planeta, tant que sense ella la vida no podria existir (WHO, 2011; Biswas et al., 2014). Actualment, l'aigua ha esdevingut un engranatge imprescindible en el motor econòmic de la nostra societat, essent una part essencial de moltes empreses del sector primari, secundari i terciari (Nickson et al., 2005). Per tant, tenir a l'abast fonts d'aigua en bon estat que siguin segures i fiables és condició *sine qua non* per a la constitució d'una societat sòlida (Vörösmarty et al., 2000; Custodio, 2022). Tot i això, s'ha de tenir en compte que el creixement exponencial de la societat moderna, la sobreexplotació de les fonts naturals d'aigua i la seva contaminació constant repercuteixen directament en el seu estat i en la seva capacitat de regeneració, cada vegada més lenta (UNESCO, 2006; WHO, 2011; MASE, 2015).

L'abocament descontrolat d'aigües residuals i contaminades industrialment té un impacte negatiu en el medi receptor (aqüífers, rius, etc.), i en provoca el seu deteriorament (WHO, 2011; Biswas, et al. 2014). Depenent del tipus de contaminant, l'impacte pot ser en la qualitat del medi aquàtic en general (per exemple, l'abocament d'aigües amb altes concentracions de sòlids en suspensió sedimentables comporta rebliment ràpid i disminució de la filtració natural del sediment, a més a més de la descomposició de la fracció orgànica, que comporta anòxia), o en els processos que s'hi duen a terme, com ara la fotosíntesi (l'abocament de sòlids en suspensió no sedimentables redueix el pas de la llum), entre d'altres.

És per això que, al llarg de l'últim segle, s'han desenvolupat sistemes i tècniques per a poder tractar l'aigua. Aquests tractaments han permès potabilitzar, tant l'aigua superficial i subterrània, com recuperar l'aigua residual urbana i industrial[1], podent-la

[1] *Les aigües residuals són la combinació de líquids provinents de zones urbanes i industrials a les quals, de forma puntual, se'ls poden afegir aigües naturals pluvials, superficials i subterrànies. La qualitat de l'aigua residual recollida*

retornar, en alguns casos, al medi hídric natural per a la seva reutilització (Poleneni, 2020).

L'aigua recollida per a la potabilització (aigües superficials i subterrànies) o sanejament (aigües residuals urbanes i industrials) és transportada, a través de punts de captació o de recol·lecció, i una infraestructura de canonades, fins a les estacions de tractament o depuració. La captació de l'aigua superficial i subterrània es fa des dels rius, llacs, pous o embassaments, i es transportada a les Estacions de Tractament d'Aigua Potable (ETAP) (Poleneni & Inniss, 2013; 2015). La recol·lecció de l'aigua residual es fa a través de la xarxa de clavegueram instal·lada a pobles i ciutats, on es traslllada cap a les Estacions Depuradores d'Aigües Residuals (EDAR).

1.3. Estacions de tractament d'aigües

1.3.1. Estacions de Tractament d'Aigua Potable (ETAP)

Les ETAP són estacions de tractament on es potabilitza l'aigua crua que s'extreu de rius, canals, embassaments o altres fonts d'aigua natural, superficial i subterrània perquè sigui apta per al consum humà. Estan ubicades entre els punts de captació d'aigua i la xarxa de canalització que distribuirà l'aigua cap als dipòsits que són controlats pels ajuntaments i, després, cap la població (WHO, 1997; WHO, 2018; Abu Shmeis, 2018).

En les ETAP existeixen tractaments convencionals i avançats. Una ETAP convencional estarà formada pels sistemes de coagulació–floculació, sedimentació, filtració de sorra i desinfecció amb clor. Els processos de coagulació–floculació i decantació es basen en l'addició d'una concentració coneguda de productes coagulants a l'aigua (normalment policlorur d'alumini – PAC) perquè, a través de la seva estructura química, atrapin la matèria en suspensió, augmentin la seva densitat i, aquesta acabi precipitant al fons del tanc. Una ETAP avançada, a part de tenir els tractaments enumerats anteriorment, també pot aplicar, entre d'altres, tractaments d'ozonització i filtres de carbó actiu a l'aigua (Raich-Montiu, et al., 2013; Poleneni, S. & Inniss, E, 2013; 2015; Poleneni, S., 2020). L'ozó és molt eficaç desactivant microorganismes i microcontaminants, gràcies a la seva força oxidant. Aplicar un procés de pre-ozonització a l'aigua, just després de

al sistema de clavegueres està interrelacionada amb la seva composició i la concentració de contaminants procedents d'efluents industrials (Tchobanoglous et al., 2013).

captar-la de l'embassament o riu, fa que la resta del procés de potabilització sigui molt més efectiu i sostenible (Vigil, 2003; Cabral, 2010).

Finalment, i una vegada obtinguda l'aigua tractada a través de tots els processos descrits anteriorment, es desinfectarà afegint una concentració controlada de clor (entre 1 i 1,5 mg/L), que n'assegurarà la seva salubritat durant el seu transport i ús final.

1.3.2. Estacions Depuradores d'Aigües Residuals (EDAR)

Les EDAR són plantes de depuració d'aigua residual, la funció bàsica de les quals és recollir les aigües residuals de la població i indústria, i reduir-ne la contaminació a través de diversos tractaments (Tchobanoglous et al., 2013).

Els sistemes convencionals de tractament d'aigües residuals d'una EDAR estan formats pel tractament primari i el tractament secundari. El tractament primari conté els processos de pretractament, on se separen els sòlids voluminosos a través de reixes i tamisos, i el decantador primari, on precipiten i sedimenten els sòlids inerts i la matèria orgànica de major densitat (Baeza et al., 2002; Gernaey et al., 2004). El tractament secundari conté els processos del reactor biològic, on s'elimina la matèria orgànica present a l'aigua a través de l'ús de diferents ecosistemes de bactèries. Aquesta barreja s'oxigena contínuament des del fons. Una vegada ha acabat el procés, es separen els microorganismes al decantador secundari.

El tractament terciari d'aigües residuals és un procés addicional per eliminar contaminants en estat col·loidal o en suspensió romanents després dels tractaments previs. Aquests sistemes es duen a terme a les Estacions de Regeneració d'Aigua (ERA) una vegada ha estat tractada pels processos primari i secundari (England & Krenkel, 2003). El tractament terciari redueix la càrrega de contaminants de les aigües residuals, permetent-ne l'abocament al medi natural, i també la seva reutilització com a recurs hídric alternatiu (Sanz, et al. 2015). Hi ha dos tipus de tractaments terciaris, els convencionals (coagulació–floculació, decantació i desinfecció) i els avançats, que es sumen als tractaments convencionals (oxidació avançada, membranes de filtració i desinfecció). L'ús de membranes de filtració (osmosis inversa, membranes d'ultra–filtració, micro–filtració i nano–filtració) permet retenir molècules petites, sals, substàncies orgàniques, microorganismes, etc. fet que millora inestimablement la qualitat de l'aigua final.

1.4. Control de la qualitat de l'aigua i dels processos de potabilització i sanejament

Tots els processos duts a terme tant a les plantes de potabilització (ETAP) com a les de sanejament d'aigua (EDAR i ERA) estan rigorosament controlats. Es controla des de la qualitat de l'aigua d'entrada a la planta, passant pel temps de residència de l'aigua en cada unitat de procés, fins a la qualitat de sortida de l'aigua de la planta, entre d'altres. Aquest monitoratge es du a terme per arribar als llindars de qualitat d'aigua establerts per les normatives espanyola i Europea. Els RD 140/2003 i RD 1620/2007, emparats per la Directiva 2020/2184/CEE i pel Reglament 2020/741/CEE, estableixen els criteris sanitaris de qualitat d'aigua de consum humà i de reutilització d'aigües depurades, respectivament, i obliguen a mantenir un estàndard de qualitat d'aigua de sortida, tant a ETAP com a EDAR i ERA.

Per controlar la qualitat de l'aigua a les plantes de potabilització i sanejament s'han de realitzar analítiques periòdiques. L'anàlisi d'aigua permet obtenir valors veraços dels diferents paràmetres de qualitat, però té moltes ineficiències. La més rellevant és que hi ha una diferència temporal entre l'extracció de mostra i l'obtenció del resultat. És a dir, la informació de qualitat no es rep a l'instant d'obtenir la mostra i, per tant, si s'observa alguna incidència important en la qualitat d'aquell moment, no es pot actuar de forma immediata. Altres inconvenients són el cost de les analítiques d'alguns contaminants presents a l'aigua, tant pel temps d'operació que se n'esdevé (extreure una mostra, traslladar-la i analitzar-la d'una manera concreta) com per la despesa que suposa l'anàlisi en sí (consumibles, reactius, aparells costosos i de manteniment periòdic) (Vigil, 2003; Roda et al., 2006).

Per minimitzar els inconvenients que comporta el monitoratge de la qualitat de l'aigua a través d'analítiques, cada vegada és més habitual instal·lar sensors de mesura continua en-línia. Aquests sistemes obtenen una mesura de qualitat d'aigua pràcticament a l'instant, i donen valors equivalents als obtinguts a través de les analítiques puntuals, permetent ampliar el temps entre mostrejos i proporcionant als operadors de planta la capacitat de donar una resposta ràpida i fiable, en cas de ser necessària. L'enregistrament de dades per part d'aquests aparells és en continu, mostrant tendències de comportament en les dades i enregistrant esdeveniments puntuals que poden ser interessants de monitoritzar, com per exemple, els canvis en la qualitat de l'aigua deguts a alteracions climatològiques extremes o a l'estacionalitat, entre d'altres (Gruber et al., 2005; Xu et al., 2016; Hernandez–Ramirez et al., 2019).

L'ús de sondes també té alguns desavantatges. En alguns casos, l'import de compra i instal·lació és força alt. Necessiten calibratges, neteges i manteniments periòdics ja que, sinó, la mesura que se n'obté pot no ser fiable. Alguns sensors necessiten consumibles, cosa que fa augmentar el cost del seu manteniment anual.

Per tots els motius exposats, les plantes de potabilització i tractament d'aigües es nodreixen de totes les capacitats que els aporten, tant les analítiques de qualitat (que són les dades que es reporten a l'organisme administratiu de control pertinent) com la instal·lació de sondes de mesura en continu. Per una banda, les anàlisis de laboratori els permeten obtenir valors fefaents de la qualitat d'aigua en un instant concret de temps. Per altra banda, els sensors en-línia són eines molt potents, que ben utilitzades, esdevenen imprescindibles en la monitorització de la qualitat de l'aigua a temps real, tant d'ETAP com d'EDAR (Torres & Bertrand-Krajewski, 2008; Raich-Montiu, et al., 2013; 2014; Hernandez–Ramirez et al., 2019). Els valors en continu que els proporciona la monitorització a través de sondes i sensors és una ajuda indispensable per donar una resposta precoç a una possible incidència, tant en els tractaments que es duen a terme, com en la qualitat de l'aigua en tot el procés. A més, els valors que s'obtenen en les analítiques periòdiques permeten realitzar calibratges constants, i mantenir els aparells analítics en-línia en correcte funcionament.

1.5. Espectre electromagnètic, espectrofotometria i spectro::lyser®

L'espectre electromagnètic és un conjunt d'ones electromagnètiques que es coneixen com radiació electromagnètica. Les ones electromagnètiques són ones transversals, formades per ones elèctriques i magnètiques que formen un angle recte entre elles i es desplacen perpendicularment a la direcció de propagació, i poden viatjar a través del buit, de sòlids, líquids i gasos. Aquest, està dividit en longituds d'ona, i les diferents propietats de cada rang definit en l’espectre electromagnètic (Raigs ϒ, Raigs X, Ultraviolat-Visible, etc.) estarà estretament lligades a la seva energia (longituds d’ona més curtes corresponen a energies més elevades) (Figura C1.1) (Burgess, 2017).

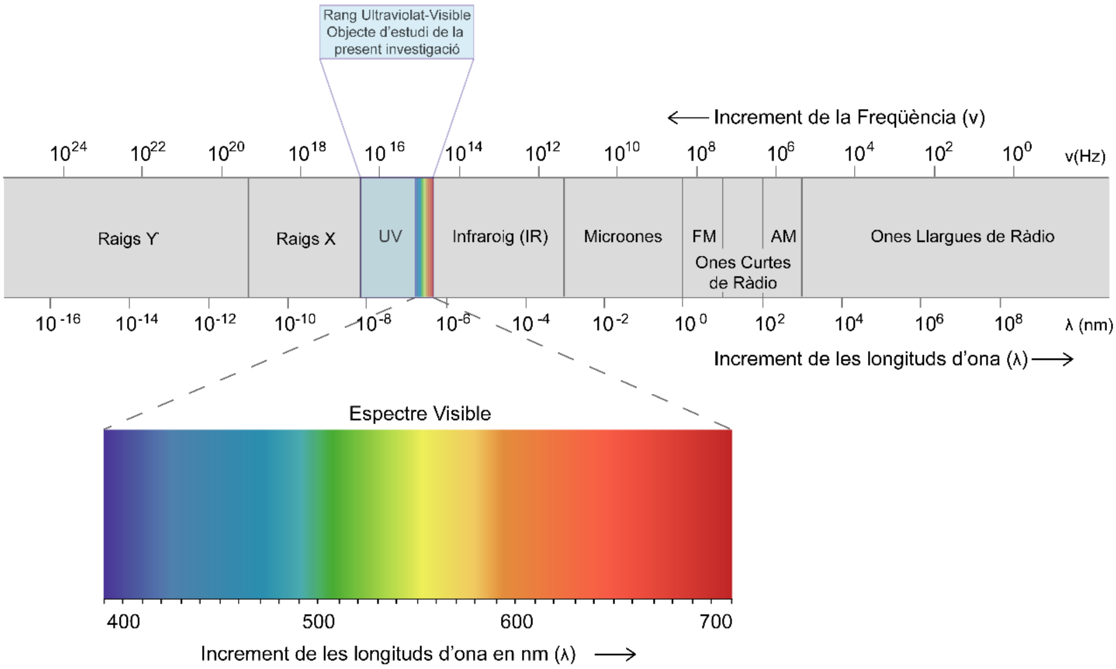


***Figura C1.1** – Espectre electromagnètic on s’observa en detall l’espectre visible. El rang de longituds d’ona objecte d’estudi d’aquesta investigació està en l’Ultraviolat–Visible. Font: Modificat de Burgess, C., 2017*

La idea que la radiació electromagnètica conté una quantitat quantificable d'energia potser es pot entendre millor si parlem de la llum com un corrent de partícules, anomenades fotons, en lloc de parlar d’ona (Burgess, 2017). Si descrivim la llum com un corrent de fotons, l'energia (F1.1) d'una determinada longitud d'ona es pot expressar com:

$$E = \frac{hc}{\lambda} \tag{F1.1}$$

On E és l’energia en KJ/mol, λ és la longitud d’ona en metres, c és la velocitat de la llum en metres per segon ($3{,}00 \cdot 10^{8}$ m/s), i la h és la constant de Planck.

Com que la radiació electromagnètica viatja a una velocitat constant, cada longitud d'ona correspon a una freqüència determinada, que és el nombre de vegades per segon que una cresta passa per un punt determinat (Figura C1.2). Les ones més llargues tenen freqüències més baixes, i les ones més curtes les tenen més altes. La freqüència (s^{-1})

s'indica habitualment en hertzs (Hz), que significa la quantitat d'ones que passen per n mateix punt per segon.

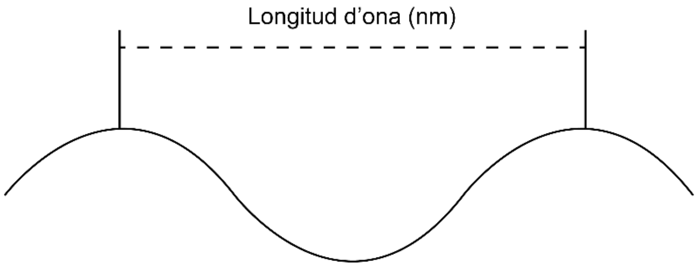


***Figura C1.2 –** Expressió d'una longitud d'ona en nanòmetres.*

Quan parlem d'ones electromagnètiques, podem referir-nos a la longitud d'ona o a la freqüència. Els dos valors es poden transformar mitjançant l'expressió simple:

$$c = \lambda v \qquad \text{(F1.2)}$$

on *v* és la freqüència en segons (s^{-1}). Per exemple, la llum vermella visible amb una longitud d'ona de 700 nm té una freqüència de $4{,}29 \cdot 10^{14}$ Hz.

L'espectrofotometria molecular és una tècnica que permet determinar quanta llum absorbeix una substància química concreta a través de l'espectre Ultraviolat Visible (190 a 400nm i 400 a 780 nm, respectivament) (Thomas & Causse, 2017). La base científica d'aquesta tècnica és que les molècules presents a la mostra líquida absorbeixen les radiacions electromagnètiques d'algunes de les longituds d'ona (Kaur et al., 2021). Com a resultat, salten d'un estat fonamental de baixa energia a un estat excitat, d'energia més alta. No totes les longituds d'ona seran absorbides, sinó que una determinada molècula absorbirà específicament aquelles longituds d'ona que tinguin energies corresponents a la diferència d'energia de transició cap a l'estat excitat que s'està produint. Així, si la transició implica que la molècula salti de l'estat fonamental A, a l'estat excitat B, amb una diferència d'energia de ΔE, la molècula absorbirà específicament la radiació amb una longitud d'ona corresponent a ΔE, i al mateix moment, hi haurà altres longituds d'ona que passin sense absorbir. Això significa que les longituds d'ona on absorbeix una molècula dependran directament de la seva estructura molecular i de les condicions del medi (p. ex., força iònica) (Figura C1.3). La quantitat de llum absorbida per una molècula a cada longitud d'ona és l'*absorbància*, que dependrà linealment de la seva concentració en solució (IUPAC, 2006). Per a calcular l'absorbància d'una molècula s'ha de fer a través de la Llei de Lambert-Beer (F1.3). Mayerhöfer et al., 2020 en fan una molt bona dissertació. Aquesta llei afirma que la llum que incideix sobre una mostra pot disminuir degut a tres fenòmens: 1) la concentració de les diferents substàncies en la mostra, 2) la distància del camí òptic i 3) l'absorbència (o coeficient d'extinció), que és la probabilitat de que un fotó de llum amb una longitud d'ona particular pugui absorbir-se pel material en qüestió. Aquesta relació es pot resumir com:

$$A = \varepsilon c d \qquad \text{(F1.3)}$$

On *A* és l'absorbància (valor adimensional), *ε* equival al coeficient molar d'extinció (L·mol^{-1}·cm^{-1}), *d* és el camí òptic (en cm o m) i *c*, en aquest cas, és la concentració molar de la mostra. (mol·L^{-1})

Quan un feix de llum travessa un medi que l'absorbeix parcialment, aquest feix es descriu com feix incident. El feix que passa a través del medi, pot veure disminuïda la seva intensitat, si s'acaba produint alguna absorció durant el trajecte. La relació entre les dues intensitats (F1.4) es pot expressar com:

$$T = \frac{I_1}{I_0} = i^{-\varepsilon c d} = e^{A} \qquad \text{(F1.4)}$$

On T és la transmitància òptica, I_o és la intensitat entrant a la mostra que es vol mesurar, I_1 és la intensitat que en surt, i A és la absorbància que s'ha indicat a l'equació F1.3. Seguint la relació anterior, el càlcul de l'absorbància (F1.5) també es pot expressar com:

$$A = -log_{10}\,\frac{I_1}{I_0} = -log_{10}(T) \qquad \text{(F1.5)}$$

La Llei de Lambert-Beer postula la relació exponencial entre la transmitància d'una substància i la seva concentració, així com també amb la longitud del camí òptic utilitzat.

En la present recerca, s'ha utilitzat la sonda spectro::lyser®, desenvolupada i comercialitzada per l'empresa s::can Messtechnik GmbH (s'aprofundeix el seu funcionament al **capítol 2**). A través de l'espectrofotometria molecular UV – Vis (metodologia descrita en els paràgrafs anteriors) aplicada a l'aigua, pot mesurar una gran quantitat de substàncies orgàniques i inorgàniques necessàries de monitoritzar en ETAP, EDAR i ERA.

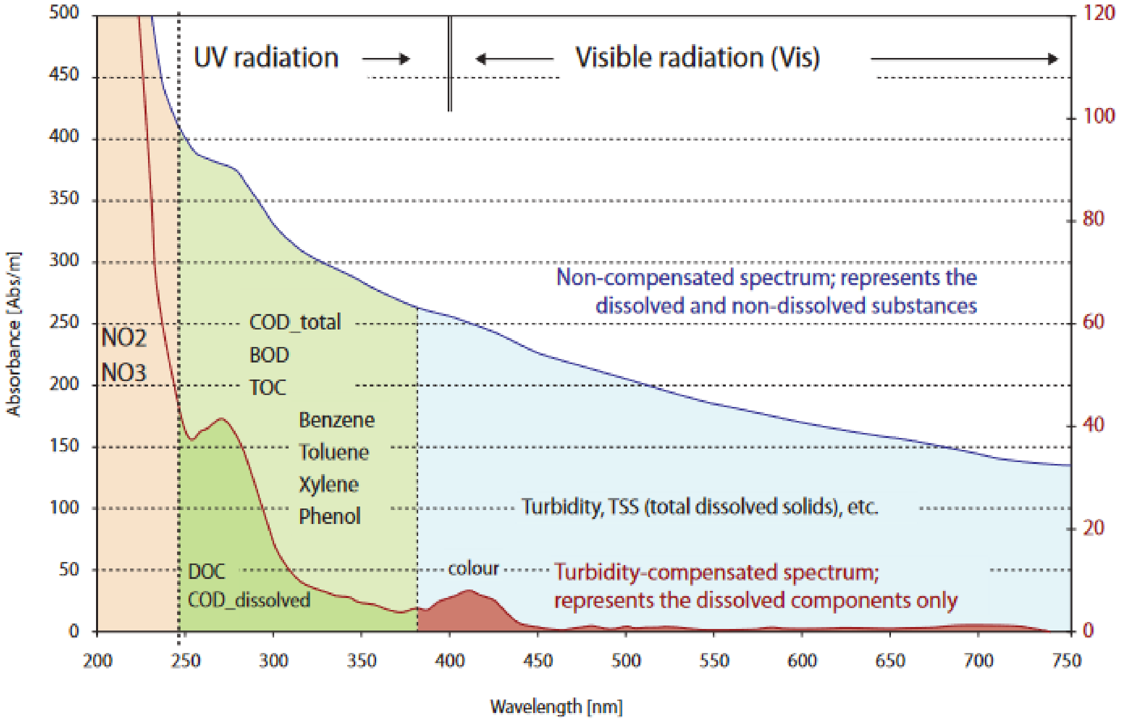


***Figura C1.3** – Espectre òptic Ultraviolat - Visible on es mostren les longituds d'ona de 200 a 750 nm. En aquest gràfic es pot observar els rangs d'absorció d'algunes de les molècules interessants de monitoritzar en qualitat d'aigua. Font: s::can Messtechnik Drinking water catalogue.*

Alguns dels paràmetres que pot monitoritzar són Sòlids Totals en Suspensió (STS), Sòlids Totals (ST), terbolesa, color, Carboni Orgànic Total (COT), Carboni Orgànic Dissolt (COD), la Demanda Química d'Oxigen (DBO) i la temperatura, entre molts altres (Figura C1.3).

Una altra qualitat molt interessant de la sonda espectrofotomètrica spectro::lyser® és que registra els paràmetres cada dos minuts, donant una informació detallada i en continu de la qualitat de l'aigua del punt a monitorar. Això fa que es pugui aprofundir en la millora constant de la seva capacitat de detecció de contaminants pels quals encara no està preparada (p. ex. Potencial de formació de trihalometans).

La gran quantitat de dades de qualitat d'aigua que s'aconsegueixen a través de la sonda espectrofotomètrica spectro::lyser® pot ser molt útil si es combina amb eines d'aprenentatge automàtic (a partir d'ara ML que prové de l'anglès *Machine Learning*) i eines de quimiometria, per a crear models matemàtics de detecció i predicció de contaminants que estiguin en el punt de mira de les normatives vigents i futures.

1.6. Quimiometria, aprenentatge automàtic i modelització matemàtica

La quimiometria és una disciplina de la química analítica amb una importància remarcable en la ciència, ja que interactua amb moltes altres àrees com la física o la bioquímica (Hanrahan et al., 2005; da Costa et al., 2020). És intrínsecament interdisciplinària ja que utilitza l'estadística, les matemàtiques i la programació informàtica per a desenvolupar procediments d'inferència que possibilitin l'obtenció de la informació més rellevant de les dades químiques obtingudes. És a dir, s'utilitza per a crear models descriptius o predictius, que propicien la detecció de característiques rellevants i relacions entre les variables seleccionades d'un conjunt de dades d'un sistema.

L'aprenentatge automàtic (ML) és una branca de la intel·ligència artificial (AI, de l'anglès Artificial Intelligence) que permet que els models matemàtics i estadístics aprenguin i identifiquin patrons sobre grans volums de dades per obtenir-ne prediccions i suggerint-ne comportaments futurs (Zhu et al., 2022). Aquesta tecnologia és present a la vida quotidiana de les persones, per exemple, a les recomanacions dels serveis d'streaming en funció dels gustos del consumidor o en les respostes intel·ligents del correu electrònic en funció dels correus rebuts.

Al llarg de les últimes dècades, els algoritmes de ML s'han anat introduint en quimiometria, i se n'obtenen resultats prometedors. La simbiosi entre models matemàtics avançats, i aparells d'adquisició de dades robustos i en desenvolupament constant, permeten la materialització d'instruments analítics molt eficaços en la detecció i predicció de contaminants. En aplicacions d'aigua, tant potable com residual, s'espera que les tecnologies d'instrumentació intel·ligents modelin sistemes complexos mitjançant la seva generalització, i en relativa facilitat per aconseguir optimitzar-ne els processos.

Les tècniques de ML i AI més emprades en el sector de l'aigua per a la monitorització intel·ligent de tractaments d'aigua són algoritmes de regressió i classificació com Regressió Lineal Múltiple (MLR, de l'anglès Multivariate Linear Regression), Màquines de Vectors de Support (SVM, de l'anglès Support Vector Machines), Xarxes Neuronals Artificials (ANN, de l'anglès Artificial Neural Networks) i Arbres de Decisió Aleatòria, o Boscos Aleatoris (RF, de l'anglès Random Forest), entre d'altres. La hibridació d'aquestes tècniques també ha resultat ser un bon fil d'investigació al llarg de l'última dècada, ja que proporciona un coneixement més profund i essencial de les relacions intrínseques entre els paràmetres de qualitat d’aigua monitoritzats i l’eficiència dels sistemes de tractament.

En aquesta tesi, s’han emprat tècniques de ML i AI, com MLR, SVM i ANN, així com també tècniques d’inferència estadística avançada en dades d’absorbància obtingudes a través de la sonda spectro::lyser® (Torres & Bertrand-Krajewski, 2008). Aquest tipus de sonda, s’utilitza àmpliament en el control de la qualitat d’aigua tant en EDAR, com en ERA i ETAP, ja que se n’obté dades que indiquen la qualitat de l’aigua d’una forma senzilla, i permeten observar, de manera eficaç i a temps real, com estan treballant els sistemes de tractament. La sonda spectro::lyser® registra una gran quantitat de dades de qualitat d’aigua al llarg del temps (cada dos minuts, registra una observació), generant bases de dades amb un volum molt gran d’informació. Això és un factor crucial a l’hora de poder generar models matemàtics de ML de bona qualitat ja que són necessàries grans quantitats de dades perquè aquests models puguin aprendre i predir les diferents interrelacions entre paràmetres.

La base de la present investigació, i per tant, de la creació d’aquests models, és augmentar les capacitats-aplicacions de la sonda spectro::lyser®. Per exemple, actualment, encara no pot detectar o predir alguns pol·luents com els trihalometans, que sí que es poden quantificar a través d’analítiques concretes i força costoses. A través de la modelització matemàtica de les dades d’absorbància obtingudes amb la sonda spectro::lyser® s’han obtingut resultats importants per a la millora d’aquesta eina de monitorització de la qualitat de l’aigua.

1.7. Plantejament general de la investigació duta a terme

Tal i com s’ha esmentat anteriorment, en aquesta tesi s’ha volgut aprofundir en la millora de la sonda spectro::lyser® en termes de monitorització de contaminants, com la predicció del potencial de formació de trihalometans a la sortida de les ETAP.

La combinació de sistemes de sensorització, de tècniques de ML i de modelització matemàtica són mètodes que es poden emprar en la monitorització de la qualitat d’aigua de consum humà. Les normatives nacionals (RD 140/2003 i RD 1620/2007), i internacionals (2020/2184/CEE i 2020/741/CEE), cada vegada més restrictives pel que fa a la qualitat d’aigua de consum humà, obliguen a les ETAP a controlar cada vegada més compostos presents en l'aigua potable, i que podrien ser perjudicials per la salut pública a mitjà i llarg termini, si s'hi està exposat recurrentment. Un exemple clar és el que fa referència a la presència de trihalometans (THM) (Villanueva et al., 2012; 2015). Els THM són compostos orgànics volàtils, fruit de la reacció del hipoclorit de sodi amb matèria orgànica no tractada anteriorment, que es creen durant el procés de potabilització (Redondo-Hasselerharm et al., 2022). Molts d'aquests compostos es consideren perillosos pel medi ambient i per la salut humana, ja que tenen un factor carcinogen important. Tot i així, el fet de clorar l'aigua en concentracions controlades és molt important, ja que n'assegura la salubritat a la sortida de planta i durant tot el recorregut per la xarxa d'aigua, fins al punt final de consum. Per tant, un dels punts crítics de control de la concentració i potencial de formació de THM és en l'aigua de sortida de les ETAP. Per controlar-ne el potencial de formació, s'han de fer analítiques que requereixen uns equips concrets molt costosos i s'han de dur a terme per personal qualificat. Són analítiques altament complicades i laborioses, i els resultats no s'obtenen ràpidament, podent tardar fins a 4 dies.

En la present investigació s'ha estudiat la qualitat d'aigua de l'estació de tractament d'aigua potable del Consorci d'Aigües de Tarragona (ETAP – CAT), que capta, tracta i distribueix l'aigua als municipis i indústries de la província de Tarragona (Catalunya, Espanya). Tal i com s'explica al **capítol 3** *Comparative analysis of Multivariate Linear Regression and Artificial Neural Networks for predicting trihalomethanes formation potential in a Drinking Water Treatment Plant*, es va monitoritzar el tren de tractament de la potabilització de l'aigua a través de la instal·lació de tres sondes spectro::lyser®. El primer es va instal·lar a l'entrada de la pre-ozonització, per registrar dades d'absorbància de l'aigua d'entrada a la planta, el segon es va instal·lar a la sortida de la post-ozonització, enregistrant dades dels canvis en la qualitat de l'aigua al llarg de tot el tren de tractament i l'últim es va instal·lar al bombament de sortida, analitzant els canvis d'absorbància de l'última fase del tractament de l'aigua, la cloració. La recol·lecció de dades es va realitzar durant un any i de forma contínua (cada dos minuts), i això va permetre la creació d'un model matemàtic per a la predicció del potencial de formació de THM a la sortida de l'ETAP. Aquest model es va desenvolupar pensant en la necessitat dels operadors de planta de tenir informació fefaent i immediata de la formació d'aquests compostos, permetent-los donar una resposta àgil en cas de necessitat.

La combinació de sistemes de sensorització, de tècniques de ML i de modelització matemàtica també es poden combinar per a la detecció i predicció d'hidrocarburs en aigües residuals, ja que és una problemàtica en augment a l'entrada de les EDAR.

Els hidrocarburs són compostos orgànics formats per hidrogen i carboni, incolors i hidròfobs, d'olor agradable. Es normal detectar-los a baixa concentració a les aigües residuals urbanes, ja que, quan plou, i a través de l'escorrentia superficial, es poden arrossegar de benzineres, carrers o carreteres. Aquests contaminants, però, també poden provenir de polígons industrials, on s'elaboren productes com dissolvents o pintures, i poden acabar al clavegueram, barrejats amb les aigües residuals urbanes. Les EDAR urbanes estan dissenyades per poder tractar petites concentracions d'hidrocarburs, que no afectarien al seu correcte desenvolupament. No obstant, en altes concentracions, poden afectar directament als sistemes de sanejament fent-ne disminuir l'efectivitat i podent esdevenir un problema de salut pública. La sonda spectro::lyser® monitoritza paràmetres importants que donen informació sobre el correcte desenvolupament del tractament d'aigua residual. Tot i així, no pot detectar hidrocarburs en l'aigua residual, excepte quan es presenten en concentracions molt elevades.

En aquesta tesi s'han agrupat dades d'absorbància d'aigües residuals registrades durant llargs períodes de temps, provinents de diferents punts de l'estat espanyol. Tal i com s'explica al **capítol 4** *Multivariate Linear Regression approach to predict hydrocarbon mixtures in urban wastewaters matrices using spectrophotometric sensors*, a través d'aquestes dades s'han generat diferents models matemàtics que permetrien detectar i predir concentracions baixes d'hidrocarburs coneguts (toluè, m-xilè i p-xilè) en aigües residuals d'influent i per tant, ajudar en una possible resposta per part dels operadors de plantes EDAR, en cas de ser necessari.

La combinació d'instrumentació analítica com la sonda spectro::lyser®, les tècniques de ML i la modelització matemàtica també serien molt útils per monitoritzar i controlar processos de sanejament. Un exemple clar de la necessitat de combinar aquestes tecnologies ocorre en el sistema de coagulació–floculació, que es duu a terme a les EDAR i ERA. Per a obtenir un bon control d'aquest procés es fan anàlisis de gerra[2] o *jar–test*. Aquest tipus d'anàlisi permet saber la concentració de coagulant que s'ha d'afegir a l'aigua residual però no s'executa de manera continuada, ni en línia. Tot això comporta que, en molts casos, els operadors hagin de decidir quina és la concentració òptima de coagulant que s'ha d'afegir als cicles de coagulació–floculació en funció de l'aspecte de l'aigua, convertint aquest pas del tractament en una operació feixuga. Aquesta manca d'informació real suposa una reducció de l'eficiència del procés i, per tant, un augment dels costos, tant en els productes químics emprats, com en l'energia consumida durant el procés.

En la present investigació s'ha estudiat el procés de coagulació–floculació de la planta de regeneració d'aigües del Camp de Tarragona. Aquesta planta la conforma un procés terciari que rep l'aigua prèviament tractada pels sistemes primari i secundari provinents de les EDAR de Vilaseca i Tarragona (província de Tarragona, Catalunya, Espanya) i que, una vegada regenerada, nodreix el Parc Industrial de la petroquímica del Camp de Tarragona.

Tal i com s'explica al **capítol 5** *Application of Advanced algorithms for the prediction and improvement of coagulant dosatge in WRP's considering two scenarios of training*, s'han enregistrat dades d'absorbància de l'aigua residual d'entrada a la planta, així com de paràmetres de qualitat com ara la terbolesa, a través de l'ús de dues sondes spectro::lyser® instal·lades a l'entrada del sistema terciari i després del sistema fisicoquímic de coagulació-floculació anomenat Actiflo®. L'entrenament de models matemàtics a partir de dades d'entrada i de sortida del sistema Actiflo® s'ha utilitzat per predir les concentracions idònies de coagulant a cada moment, i així poder controlar d'una manera més estable la seva addició durant el procés i evitar-ne un sobre consum.

1.8. Objectiu

L'objectiu principal d'aquesta tesi és dotar a la sonda spectro::lyser® de més capacitats en la detecció i predicció de contaminants, tant en aigua potable (THM FP), com en aigua

[2] *L'anàlisi de gerra consisteix en agafar una mostra d'aigua d'entrada a la planta, dividir-la en submostres i afegir una concentració específica de coagulant a cada una, que pot variar de 0,5mg/L entre mostra i mostra. Una vegada s'ha afegit el coagulant, s'executen barreges ràpides (70 rpm durant un minut) i lentes (15 rpm durant 15 minuts) a través d'un conjunt de paletes que conformen l'agitador múltiple amb les que s'aconsegueixen condicions hidràuliques similars a totes les mostres. Després es deixa sedimentar durant 20 minuts (Franceschi, et al., 2002).*

residual (mescla de m-xilè, p-xilè i toluè), a través de l'ús d'eines quimiomètriques d'inferència estadística avançada i de la comparativa entre algoritmes matemàtics de ML i AI (MLR, SVM i ANN, entre altres). El desenvolupament d'aquests nous models de detecció i predicció vol servir d'ajuda en la resposta precoç dels operadors de plantes de tractament d'aigua (ETAP, EDAR i ERA) en cas de que sigui necessari.

Capítol 2 – Metodologia

2.1. Sonda spectro::lyser® i con::cube®

La sonda spectro::lyser® es pot equiparar a un espectrofotòmetre portàtil que utilitza l'espectrofotometria com a mètode de mesura (Figura C2.1). Té un pas òptic, anomenat *finestra òptica o de mesura*, que conté una unitat emissora de llum i una de receptora.

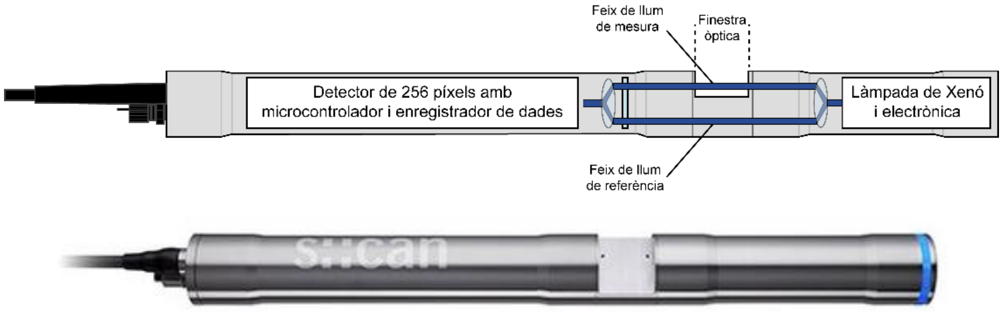

Feix de llum de mesura
Finestra òptica
Detector de 256 píxels amb microcontrolador i enregistrador de dades
Làmpada de Xenó i electrònica
Feix de llum de referència
s::can

***Figura C2.1** – Detall esquemàtic de l'espectrofotòmetre spectro::lyser® desenvolupat per s::can Messtechnik GmbH on es poden observar les diferents parts que el conformen. A la part esquerra de la finestra òptica hi ha la part emissora de llum i a la dreta, la receptora. Font: s::can Messtechnik Drinking water catalogue.*

La llum emesa passa a través de la matriu d'aigua a analitzar, continguda en el pas òptic. La presència de diferents substàncies a l'aigua debilita el pas del feix de llum projectat per part de la font emissora. spectro::lyser® enregistra la diferència entre el feix de llum projectat i rebut, i la transforma en una mesura d'absorbància (Llei de Lambert-Beer descrita en l'apartat 1.5 del capítol 1). Internament, el sensor emet un altre feix de llum que s'utilitza com a referència. Aquesta referència s'enregistra inicialment utilitzant aigua desionitzada per crear un blanc. Això permet compensar, per a cada mesura realitzada, qualsevol defecte instrumental que pogués influenciar la qualitat de la lectura. De cada observació feta per la sonda se n'obté un espectre d'absorció complet (200 a 400 nm per sondes UV, i 200 a 750 nm per sondes UV–Vis), permetent observar, en major o menor mesura, la qualitat de l'aigua en aquell instant de temps. Això la fa una sonda molt versàtil, ja que pot monitoritzar paràmetres de qualitat com ara Sòlids Totals en Suspensió (STS), terbolesa, color, Demanda Biològica d'Oxigen (DBO), Demanda Química d'Oxigen (DQO), temperatura i nitrats (NO_3^-), entre molts altres, en funció de l'aplicació (aigua crua, potable, residual i recuperada).

L'adquisició de dades d'aquesta tesi s'ha fet mitjançant la instal·lació de sondes espectrofotomètriques spectro::lyser®. Depenent de l'aplicació que es volia monitoritzar s'han utilitzat sondes amb una finestra òptica que anava des de 5 a 100 mm d’amplada. Aquestes sondes portaven integrat un sistema de neteja per raspalls o aire a pressió, que s’activaven just abans de fer la mesura per evitar qualsevol tipus d’obstacle en l’observació o adhesió d’embrutiment que, amb el temps, comportés una deriva de la lectura. L’enregistrament de les dades d’absorbància es va dur a terme amb un controlador lògic programable d’última generació anomenat con::cube® (Figura C2.2). Els paràmetres enumerats anteriorment es podien visualitzar en aquest aparell, podent monitorar-ne la qualitat. També es va utilitzar per programar la recurrència de les mesures de les sondes i de les neteges dels raspalls integrats.

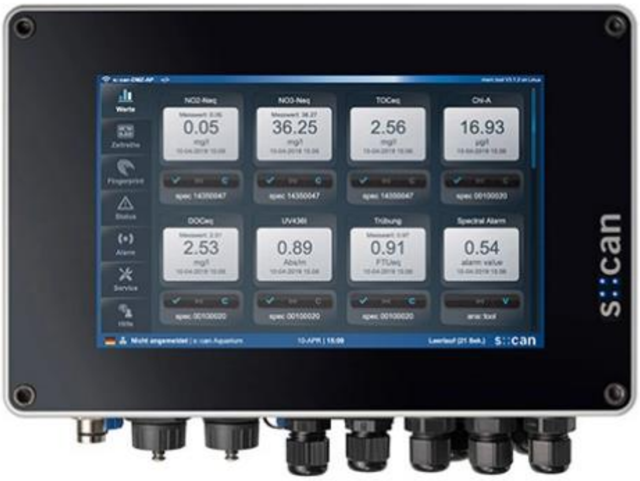


Figura C2.2 – *Imatge de detall del PLC con::cube® desenvolupat per s::can Messtechnik GmbH on s’observen alguns dels paràmetres de qualitat d’aigua que mostra a la pantalla principal. En les pestanyes que es mostren a l’esquerra de la pantalla n’hi ha una que permet mostrar l’espectre òptic enregistrat per l’spectro::lyzer®. Font: s::can Messtechnik Drinking water catalogue.*

Una vegada obtingudes totes les dades dels diferents estudis duts a terme en aquesta tesi, es va pensar en una estructura de bases de dades i de gestió matemàtica senzilla de reproduir i d’entendre.

2.2. Bases de dades i estructura

Tal i com s’ha especificat anteriorment, les bases de dades utilitzades durant aquesta tesi s’han adquirit a través de la instal·lació de sondes espectrofotomètriques spectro::lyser® i de projectes realitzats per part de l’empresa s::can Iberia Sistemas de Medición S.L.U.

Les sondes van enregistrar l’espectre UV i UV–Vis, i també paràmetres de qualitat d’aigua derivats d’aquest espectre, de diferents punts de la planta de tractament d’aigua potable del Consorci d’Aigües de Tarragona (CAT) durant un any, i del tractament terciari de la Planta del Camp de Tarragona durant sis mesos. En tots dos casos, es van instal·lar cobrint els tractaments rellevants a modelitzar, tenint en compte les necessitats d'informació per a la posterior creació de models matemàtics de predicció.

Per altra banda, es van recol·lectar bases de dades provinents de projectes on s'havien instal·lat sondes espectrofotomètriques a diferents influents de depuradora d'aigües residuals urbanes per adquirir-ne l'espectre.

Tal i com s'ha explicat en l'apartat anterior, la sonda spectro::lyser® és una sonda espectrofotomètrica que s'ha desenvolupat per a poder controlar la qualitat de l'aigua. Les longituds d'ona i els paràmetres de qualitat d'aigua enregistrats es van utilitzar com a variables independents per a la creació de models matemàtics. En l'espectre UV s'enregistraven 200 variables (longituds d'ona de 200 a 400 nm), i en l'espectre UV–Vis, 550 (longituds d'ona de 200 a 750 nm). L'absorbància de totes les longituds d'ona i la concentració dels paràmetres es mesuraven cada dos minuts. Al llarg de tot un any, es podien arribar a enregistrar més d'un milió d'observacions per a cada variable.

Totes les dades recol·lectades al llarg de la tesi s'han tractat seguint el mateix patró. Inicialment s'ha fet una neteja i un cribratge d'observacions nul·les i zeros. Una vegada filtrades, les variables s'han estandarditzat, és a dir, s'han re-escalat per obtenir una mitjana de 0 i una desviació estàndard d'1. El procés d'estandardització s'ha aplicat tenint en compte que les longituds d'ona podien tenir magnituds d'absorbància molt diferents entre elles. Quan s'apliquen mètodes automàtics de selecció de variables s'ha de tenir en compte el pes de cada variable en la base de dades. Una diferencia molt gran dels valors d'absorbància entre longituds d'ona, pot generar biaixos en la seva selecció. Estandarditzar les variables fa que se'n reparteixi el pes, i per tant, aquest biaix deixaria d'existir. Per a l'estandardització s'ha aplicat la funció *preProcess* de la llibreria *CARET* (Kuhn, 2008; 2013). Aquesta funció permet seleccionar el mètode d'estandardització que es vol fer servir. En el nostre cas, s'ha seleccionat el centrat i l'escalat de cada variable, per obtenir una mitjana ponderada de 0 i una desviació estàndard d'1.

Una vegada estandarditzades les variables, es seleccionaven les més representatives utilitzant mètodes automàtics de selecció. Aquests mètodes inclouen *Forward Selection* (FS), *Backward Elimination* (BE) i *Stepwise Selection* (SS). Els noms van lligats a la direcció que pren el mètode de selecció. Per exemple, FS comença sense cap variable seleccionada. Durant els passos posteriors, s'avalua si cada variable escollida millora algun criteri estadístic prèviament seleccionat, i repeteix aquests passos fins que cap de les variables restants millora el criteri. Una vegada revisada una variable, es descarta i no es té més en compte. BE comença amb un model complet, és a dir, amb totes les variables. A cada pas, elimina la variable considerada menys important i que no compleix el criteri seleccionat. Una vegada s'ha eliminat una variable, no es torna a revisar. El mètode SS combina els mètodes FS i BE. Cada iteració realitzada per a comprovar la validesa o no d'una variable es considera un model matemàtic diferent. En aquest s'avaluen, a més, el coeficient de determinació (R^2) i l'error quadràtic mitja (RMSE, de l'anglès *Root Mean Squared Error*). L'aplicació de mètodes de selecció automàtics de variables es va fer amb les funcions *forward*, *backward* i *both* de la llibreria Stats. Aquesta llibreria permetia escollir la base de dades de la qual se'n volia seleccionar les variables, i també escollir el criteri llindar de selecció de variables i models. En el nostre cas, el criteri llindar escollit perquè els motors de selecció consideressin o no una variable com a significativa va ser el *valor de significança* o el *p-valor*, que és el valor que s'obté a cada iteració per a cada variable comprovada com a vàlida per a cada model generat. Les variables escollides havien de tenir un p-valor igual o inferior a 0,05.

Per escollir el millor model generat a partir dels diferents motors de selecció de variables es va utilitzar el Criteri d'Informació d'Akaike (AIC, de l'anglès *Akaike Information Criterion*) (Akaike, 1978; 1979). L'AIC és un estimador de l'error de predicció i, per tant, de la qualitat relativa dels models estadístics per a un conjunt de dades determinat. Donada una col·lecció de models per a una mateixa base de dades, s'estima la qualitat de cada model a través del valor de R^2, RMSE i el número de variables, i selecciona el que conté el valor d'AIC més baix.

Una vegada es seleccionaven les variables necessàries de forma automàtica per a cada cas, s'estudiava la correlació entre elles. Aquest pas es feia a través de la funció *ggcorr* de la llibreria *ggplot2*. L'estudi de correlació entre longituds d'ona és un pas previ important abans de començar una modelització que impliqui, per exemple, un model de MLR. Les longituds d'ona contenen informació solapada entre les seves anteriors i les posteriors, i són considerades un paràmetre numèric continu. Aquesta característica està estretament lligada a la multicol·linealitat, que és resultat de la compartició d'informació entre paràmetres o variables que s'han suposat independents. Dit amb altres paraules, el problema de la multicol·linealitat consisteix en l'existència de relacions lineals entre dues o més variables considerades independents, on el coeficient de determinació (R^2) entre aquestes variables és 1 o pròxim a 1. Aplicar un estudi de correlació a les variables seleccionades automàticament evita observar resultats incongruents en les modelitzacions.

En tots els casos que es van utilitzar sistemes de selecció automàtica es van detectar variables altament col·lineals (multicol·linealitat perfecta o R^2 igual, o pròxima a 1). Per observar la importància de cada variable i l'impacte de la seva col·linealitat en el grup de variables de forma numèrica, es va utilitzar un criteri de viabilitat anomenat Factor d'Inflació de Variància (VIF, de l'anglès *Variance Inflation Factor*). El VIF quantifica la intensitat de la multicol·linealitat en un anàlisis de regressió de mínims quadrats convencional. És una de les eines més utilitzades per a detectar l'absència d'ortogonalitat de les variables d'un model, el grau d'aquesta absència i el seu impacte en el model matemàtic final. El valor de VIF és un escalar, que comença per 1 i no té límit final. Com més alt és el valor, més col·lineals són les variables. En general, les variables es consideren independents amb un $5 \geq \text{VIF} \leq 10$. Variables amb VIF més gran de 10 s'han de reconsiderar, ja que mantindrien una correlació molt elevada amb una o més variables del grup, i podrien fer trontollar els resultats predictius finals. Per a calcular el VIF de cada variable en un grup de variables concret es va utilitzar la funció *vif* de la base de dades *Stats*.

La selecció de variables automàtica és una molt bona eina de selecció inicial sobretot en bases de dades on és molt complicat fer la tria manualment (elevat nombre de variables). Aquests motors permeten fer un cribratge ràpid i seleccionar les variables més significatives del grup general. Tot i això, és imperatiu fer una segona selecció manual, per evitar incloure variables que podrien comprometre els resultats finals de predicció.

Una vegada escollides les variables a modelitzar es van separar les bases de dades generals en tres: entrenament, proba i validació. Les dades de validació es van separar primer, i conformaven les últimes mil files de cada base de dades general. Els grups d'entrenament i proba es van crear de forma aleatòria un cop separat el grup validació, de manera que les observacions es barregessin. L'homogeneïtzació de les dades millora la capacitat d'entrenament dels models, permetent que aquests es topin amb un ventall de variabilitat més ampli dins de les dades d'estudi. Per a crear les bases de dades d'entrenament i proba es va utilitzar la funció *createDataPartition* de la llibreria *CARET*. L'ús d'aquesta funció permet decidir quin tant per cent de dades s'inclouen a la base de dades d'entrenament. Les dades restants, s'assignen a la de proba. En tots els capítols de tesi, i una vegada sostreta la base de dades de validació de la base de dades general, s'han utilitzat el 75% de les dades restants en l'entrenament, i la resta en la proba. Una vegada creades les diferents bases de dades, s'apliquen els algoritmes corresponents per a crear els models matemàtics.

2.3. Desenvolupament de models

En aquesta tesi s'han avaluat tres algoritmes matemàtics per a la predicció de contaminants en diferents matrius d'aigua. Concretament, aquests són la regressió lineal múltiple (MLR), les màquines de suport de vectors (SVM), i les xarxes neuronals artificials (ANN). Per a cada cas d'estudi s'han aplicat els models utilitzant les mateixes funcions, i seguint una mateixa metodologia, però tant la variable de resposta que es volia predir, com les variables independents, com les dades registrades, eren molt diferents.

Taula C2.1 *– Resum executiu dels sistemes de selecció de variables, les variables de resposta i independents i els models matemàtics utilitzats per a cada capítol de tesi.*

Capítols	Selecció variables	Variable Resposta	Variables independents	Models matemàtics	Anàlisi bondat
3	FS, BE, SS	Potencial de Formació de THM	Longituds d'ona seleccionades	MLR ANN	R^2 RMSE
4	SS	Concentració de toluè, m-xilè, p-xilè en mescla	Longituds d'ona seleccionades	MLR	R^2 RMSE MAE
5	RF	Concentració de coagulant	Paràmetres de qualitat d'aigua	MLR SVM ANN	R^2 RSE

A la Taula C2.1 s'hi pot observar un breu resum de les eines de selecció de variables, així com la variable de resposta i les variables independents utilitzades per a crear els models matemàtics per a cada cas.

En l'estudi de modelització del **capítol 3** es van instal·lar tres sondes spectro::lyser®. La primera a l'entrada de la planta, abans del tractament de pre-ozonització. La segona a la sortida del tractament de post-ozonització i després del tren de tractaments de l'ETAP. L'última sonda es va instal·lar a la sortida de la planta, una vegada s'havia afegit el clor.

La variable dependent que es va utilitzar per aquest estudi va ser el Potencial de Formació de THM (THM FP, de l'anglès *Trihalomethane Formation Potential*) obtingut a través de la creació d'un paràmetre virtual utilitzant la formula d'Amy, 1998. Aquesta formula es va introduir al con::cube® instal·lat al laboratori de bombament d'aigua tractada del Consorci d'Aigües de Tarragona i registrava la capacitat de formació de THM a les 48h en funció de la qualitat d'aigua que es registrava contínuament. Les variables independents van ser les longituds d'ona escollides després d'aplicar la selecció automàtica dels motors de selecció i l'estudi manual de VIF descrit en l'apartat anterior.

En l'estudi de modelització del **capítol 4** es va crear un algoritme d'addició matemàtica d'espectres òptics dels hidrocarburs d'estudi. Aquesta addició espectral es feia a través d'un disseny factorial de tres nivells, on es podia controlar diferents intervals de concentració d'una mescla coneguda d'hidrocarburs (toluè, m-xilè i p-xilè). El dopatge es realitzava de manera automàtica: l'espectre òptic resultant de la mescla d'hidrocarburs es sumava a l'espectre òptic de cada observació registrada a la base de dades d'aigües residuals urbanes provinent de diferents projectes d's::can Iberia Sistemas de Medición, S.L.U. Aquest algoritme d'addició es va crear específicament per aquest cas d'estudi. La variable de resposta era la concentració d'hidrocarburs afegida a cada observació de la base de dades d'espectres d'aigua residual urbana. Les variables independents eren les longituds d'ona dels espectres d'aigua residual prèviament seleccionades pel motor de selecció SS.

En la modelització del **capítol 5** es van instal·lar dues sondes spectro::lyser®, una a l'entrada del sistema terciari i l'altra a la sortida del sistema Actiflo® de Veolia. Com a variable de resposta es va utilitzar la concentració de coagulant registrada durant sis mesos pel sistema Actiflo®. Les variables independents van ser les seleccionades com a més rellevants per l'algoritme RF, i van ser diferents longituds d'ona i paràmetres de qualitat d'aigua.

2.3.1. Regressió Lineal Múltiple

Els diferents models de MLR es van crear utilitzant la funció *lm* de la llibreria *Stats,* seleccionant la variable de resposta i les variables independents, en cada cas. Una vegada creat el model, es cridava la funció *Summary* de la llibreria Stats per observar el seu coeficient de determinació (R^2), l'Error Residual Standard (RSE, de l'anglès *Residual Standard Error*) i els graus de llibertat. A més, permet observar els coeficients obtinguts per a cada variable independent i la seva significança en el model a través del *p*-valor.

Una vegada realitzat el model matemàtic de MLR, es van comprovar les hipòtesis d'homoscedasticitat, multicol·linealitat, linealitat i normalitat a través de l'estudi de la gràfica de correlació entre els valors observats i els predits, la gràfica Quantil-Quantil (gràfica Q-Q) i la gràfica de punts de palanca o gràfica de Cook. Per a la creació d'aquests tipus de gràfiques es va utilitzar la funció *plot* en el model de MLR creat prèviament. Per a estudiar l'homoscedasticitat i la importància de les variables seleccionades es va utilitzar la mètrica de Breusch-Pagan (BP) (Breusch & Pagan, 1979).

Finalment, i per estudiar la bondat del model es van estudiar l'error quadràtic mitjà (RMSE, de l'anglès, *Root Mean Squared Error*) i l'error absolut mitjà (MAE, de l'anglès *Mean Absolute Error*), a través de les funcions *RMSE* i *MAE* de la llibreria *Metrics*, respectivament.

Per a validar els resultats obtinguts amb el model de regressió lineal, es va utilitzar la funció *predict* de la llibreria *stats*, seleccionant la base de dades de validació separada inicialment, abans de crear les d'entrenament i proba.

2.3.2. Màquines de Suport de Vectors

La tipologia de modelització de SVM seleccionada pel cas d'estudi va ser la ε – regressió SVM, amb la funció lineal Kernel com a algoritme de supervisió (per a més informació revisar el **capítol 5**) (Cortes & Vapnik, 1995). Per a crear el model de regressió a través de SVM es va utilitzar la funció *svm* de la llibreria e1071. En aquesta funció del programa RStudio s'havia d'especificar la base de dades a utilitzar, els paràmetres de la funció

Kernel, que són el cost (C) i gamma (ϒ), la variable de resposta i les independents. Per a la selecció de la millor combinació dels paràmetres C i ϒ es va utilitzar la funció *tune.grid* de la llibreria *e1071*. La funció *plot, point* i *abline* de la llibreria *stats* es va utilitzar per a la visualització dels resultats. La funció *predict* de la mateixa llibreria es va utilitzar per validar el model a través dels resultats de la base de dades de validació, separada inicialment, abans de crear les bases de dades d'entrenament i prova.

2.3.3. Xarxes Neuronals Artificials

Els models de ANN es van crear utilitzant la funció *neuralnet* de la llibreria *neuralnet*. Per a la seva creació s'havia d'especificar la base de dades a modelar, la funció de càlcul que s'utilitzaria, la o les variables dependents i independents, així com el número de capes ocultes i el número de neurones per capa que conformava la xarxa neuronal (per a més informació revisar el **capítol 3 i 5** d'aquesta tesi). La funció escollida per al càlcul dels models ANN va ser la Funció Sigmoide. Es van seleccionar dues capes ocultes seguint la metodologia descrita de diferents autors al llarg de les últimes dècades (Kröse et al., 1993; Ke et al., 2008; Vujicic et al., 2016). Les neurones de cada capa es van seleccionar a través d'un algoritme desenvolupat específicament per a aquest cas d'estudi, en el qual es provava diferents combinacions de capes (dues) i neurones (de una a deu per a cada capa). La visualització dels resultats es va fer mitjançant la funció plot, *points* i *abline* de la llibreria *Stats*. En alguns casos específics de visualització, també es va utilitzar la funció *ggplot* de la llibreria *ggplot2*.

Per a validar els resultats obtinguts amb els models desenvolupats d'ANN es va utilitzar la funció *predict* de la llibreria *stats*, seleccionant la base de dades de validació separada inicialment, abans de crear les d'entrenament i proba.

Capítol 3 – *Comparative analysis of Multivariate Linear Regression and Artificial Neural Networks for predicting trihalomethanes formation potential in a Drinking Water Treatment Plant*

Abstract

It is becoming increasingly important to have comprehensive control of drinking water. Trihalomethanes (THMs), which are harmful to human health if consumed or inhaled, are produced when organic matter reacts with chlorine. In the present study, the predictive capacity of the THM formation potential (THM FP) of a Multivariate Linear Regression (MLR) and an Artificial Neural Networks (ANN) models have been compared with real-time field-scale data from the Drinking Water Treatment Plant (DWTP) of the Consorci d'Aigües de Tarragona (CAT), Spain, using spectro::lyser® probes, installed in several treatment steps of the plant. The models were created utilizing direct relationships between wavelengths selected with Stepwise Selection (SS) method. Following the fitting of the investigated models, ANN demonstrated a precise goodness of fit (R2 = 0.92; RMSE = 0.77), clearly outperforming the MLR model (R2 = 0.30; RMSE = 1.65) in this regard. The effects of severe multicollinearity among wavelengths are responsible for the difference in the model's accuracy. Even though it was reduced by a prior study on the variance inflation factor (VIF), it was still very high for some of the remaining wavelengths That effect resulted in large fictitious correlations that directly affected the MLR model's prediction capability (R^2 = 0.35 in the validation set). e(The ANN model, however, was unaffected by this element and made accurate predictions even for data that it had never seen before (R2 = 0.72 in the validation set). The results showed that the application of Machine Learning (ML) models in ANN can enhance a critical response, becoming an essential element for operators in the daily management of DWTP when required.

3.1. Introduction

One of the most important developments in human health during the past century is water disinfection. Good drinking water quality is assured for millions of people daily from their public water systems. Every day, millions of people benefit from municipal water systems that provide clean drinking water. Chlorine disinfection is a crucial stage in drinking water treatment plants (DWTPs), ensuring the microbiological safety of water not only immediately following treatment but also throughout all its transport, ensuring good quality for all customers. (Richardson et al., 2007; Alver et al., 2018).

Toxic by-products (also known as disinfection by-products, or DBPs) can be produced by chlorinated disinfectants in addition to effectively removing microorganisms, Natural Organic Matter (NOM), and some pollutants (Awad et al., 2015; Alver et al., 2018).

According to some studies, DBPs may cause public health problems such as bladder cancer and effects on reproduction and development (Waller et al., 1998; Nieuwenhuijsen et al., 2000; Bove et al., 2002). DBPs can be ingested, inhaled, or exposed through the skin while washing or taking a bath (volatilization from a rise in temperature), and they can also be ingested through food, although drinking water is the main way that people are exposed to them (Waller et al., 1998; Villanueva et al., 2004; Savitz et al., 2005; Cantor et al., 2010).

In general, some factors that are intrinsically associated with the quality of raw water are related to the presence of DBPs in drinking waters (Mayer et al., 2017; De Castro et al., 2019). Several relevant factors, such as the physicochemical properties of NOM, the source and quality of water (carbon to nitrogen ratio), seasonality (pluviometry, droughts, etc.), the residence and transportation time, and the disinfection treatment used, or the dose added all contribute to enhancing the formation of DBP. (Mayer et al., 2017; Qi et al., 2018).

Chlorine, chlorine dioxide, chloramines and ozone are the most common disinfectants nowadays, and each produces its particular set of DBPs. More than 600 DBPs have been reported in the literature, and only a small number have been evaluated. The most common DBPs found are Trihalomethanes (THMs), Haloacetic acids (HAAs), Haloacetonitriles (HANs), Halonitromethane (HNM), Halopropanones, Haloaldehydes, Haloacetamides, Cyanogen halide (CNX), Nitrosamines and Halogenated furanones (MX).

This study will focus on Trihalomethanes (THMs), which are the best-understood indicator of DBP formation. Among the various organic DBPs, only THMs are now regulated by present directives (Golfinopoulos et al., 2002; Toroz et al., 2005). These authors explored correlations between parameters, always using bench-scale data. Batch-prepared models can be accurate for limited situations, but only if the data used in training is pretty similar to that used in the batch (Rodríguez et al., 2001). Other models have been based on chemometric approaches (Platikanov et al., 2007; 2012).

The most recent European Directive, EC 2020/2184, however, mandates that member states of the EC also control HAAs in drinking water. Organohalogen compounds are known as THMs, such as chloroform (CFM; CHCl3), bromodichloromethane (BDCM; CHCl2Br), dibromochloromethane (DBCM; CHClBr2), and bromoform, are included in this group (BFM; CHBr3). This group of THMs is also sometimes referred to as THM4. It contains the four possible combinations of bromine and chlorine substituents as opposed to the THM9 group, which includes all possible combinations containing bromine, chlorine, and iodine substituents (Alver et al., 2018). They are created when chlorine oxidizes NOM, primarily the humic and fulvic acids that are naturally present in water, during chlorination (Richardson et al., 2015). Due to the potential health risks associated with THMs in drinking water, most countries have established regulations for controlling their concentration, minimizing population exposure to potentially harmful chemicals, and maintaining adequate water disinfection (i.e., Spain has the RD 140/2003) following the minimum standards of European Directive, which limits the maximum THM concentration in 100 µg/L (EU 2020/2184).

In accordance with the restrictions on THMs' presence in drinking water, the DWTPs must perform periodic analyses to control its concentration and supply drinking water in safety conditions (RD 140/2003).

Gas chromatography and ECD detection are required to analyse THMs. These analyses are costly and time-consuming. Moreover, they are at-line experimental processes, making the plant operators unable to react to a critical situation quickly (Elshorbagy et al., 2000). Consequently, it could be an answer to developing predictive models based on raw water quality parameters (Rivadeneyra, 2014). Many authors used important water quality parameters that influence the creation of THMs to describe their kinetics and formation using empirical models (Amy et al., 1998; Abdullah et al., 2003; Sadiq et al., 2004; Chowdhury et al. 2009; Lin et al., 2018). Using only bench-scale data, these authors investigated correlations between parameters. Batch-prepared models can be accurate for limited situations, but only if the data used in training is pretty similar to that used in the batch (Platikanov et al., 2007; 2012). Other models have been based on chemometric approaches (Rodríguez et al., 2001; Golfinopoulos et al., 2002; Toroz et al., 2005; Platikanov et al., 2007; 2012). Several authors formulate their predictive models using Partial Least Squares Regression (Serrano, 2007). One of the most reliable predictors of THMs FP, according to Korshin and Fabbricino, is the use of spectrophotometer probes with differential wavelengths (Korshin, 2007; Fabbricino et al., 2009). Non-linear methods and black-box algorithms, like Neural Networks, are becoming increasingly popular (Milot et al., 2002; Rodríguez et al., 2004; Godó-Pla et al., 2021). Although these methods can prove to be more predictive, they can also have problematic blind spots since they do not provide information about the structure of the function being approximated. Chowdhury and Sadiq provided a good review of non-linear models that have been created (Chowdhury et al., 2009; Sadiq et al., 2019). Robust and reliable mathematical models are still needed for the online prediction of THM FP. These models could give accurate information, provide online control to plant operators, and shed light on the current needs (Rodríguez et al., 2001; Platikanov et al., 2012).

Large amounts of data are needed for each of the models mentioned above. Each water quality parameter that a mathematical model specifies as a predictor must have a sensor or probe installed in by-pass (Torres & Bertrand-Krajewski, 2008). This analytical machinery requires routine calibration, periodic maintenance, and replacement of consumables on a recurrent basis; these tasks are time-consuming and costly.

The spectrophotometric probes, such as the i::scan® and the spectro::lyser® from s::can Messtechnik GmbH, are becoming increasingly important for controlling water quality in DWTPs and eliminating the aforementioned tedious tasks. These devices also feature plug-and-play installation, no periodic calibration, and no consumables. spectro::lyser® and i::scan® measure how much light is absorbed by a chemical substance dissolved in a solution. Since it can provide a lot of knowledge about dissolved matter throughout the spectrum, the absorbance of water is a very valuable indicator of its quality.

Additionally, both sensors can also be used to create mathematical formulas that predict nitrates, dissolved organic carbon (DOC), or total organic carbon (TOC), among other water quality parameters. There are numerous of THMs FP prediction models based on the wavelengths 254 and 272 nm (characteristic wavelengths for organic matter absorption). Better models based should be possible to develop, for example, on wavelengths carefully chosen from the entire spectrum of a spectro::lyser®.

For all the reasons mentioned above, the major objective of this study was to create a mathematical model to predict the THMs FP using only the absorbance spectrum in the UV-Vis range from multiple spectro::lyser® probes installed in a DWTP and the virtual formulas using an i::scan®.

3.2. Methodology

3.2.1. Consorci d'Aigües de Tarragona Drinking Water Treatment Plant

Consorci d'Aigües de Tarragona (CAT) is a non–profit entity founded in 1985 that sources Ebro River water to treat and distribute it in the Tarragona province (Catalonia, Spain).

CAT–DWTP is divided into different tanks and treatments (Figure C3.1). The water capture is done at Camp-Redó, fifteen kilometres from Amposta, Tarragona province (Figure C3.2). The water circulates to a raw water tank with a storage capacity of about 175.000 m^3. This tank's primary function is to regulate the entire potabilization process and guarantee the populational water supply between 12 and 24 hours.

The first treatment the water receives is pH regulation with CO_2 injection. After that, water undergoes a purification process with ozone. Ozone has high effectiveness in deactivating microorganisms and micropollutants and can be substitutive for chlorination. Once pre-ozonation has been applied, water is directed to a distribution chamber. After that, water continues into the flocculation and decantation system. There, poly-diallyl dimethylammonium (shortened by polyDADMAC) is added as a flocculant. Turbines mix the water to enhance floc formation, capturing colloidal particles suspended in the water. Flocs are then separated by gravity in a lamella clarifier.

All sludge separated in the settling process is treated again to extract the maximum amount of water. After all, sludge is collected, it is first sent to the mud thickener and pumped into the dehydration chamber. The sludge is centrifugated, separating the portion of interstitial water. This non-potable water is first sent to the recovery chamber, which is a tank where water utilized in various DWTP processes is collected, and then to the raw water tank to begin the potabilization processes all over again.

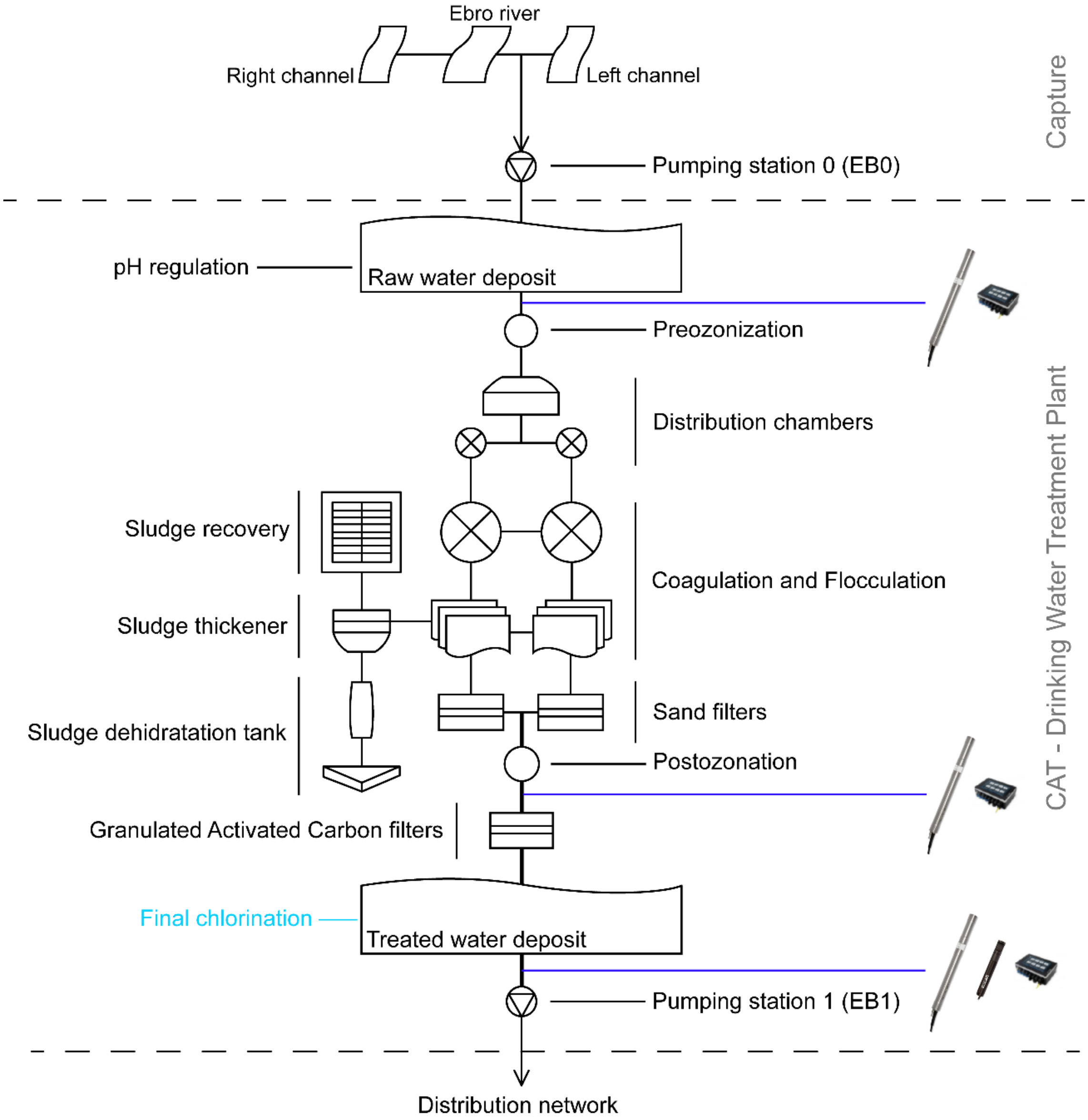


Figure C3. 1 *– Water treatment schema of CAT-DWTP. Water undergoes treatment in several unit processes to achieve a good quality for consumption. Raw water is treated by pre-ozonisation, followed by coagulation, flocculation, and sand filtration. Then a second ozonation process oxidises organic compounds remanent in water before GAC completes treatment before final disinfection with chlorine.*

The next step is sand filtration to trap particles not eliminated by sedimentation. Then, the water is subjected to the second ozone treatment. The main objective of this second ozone application is to deactivate microorganisms and oxidize and fragment organic compounds that will be eliminated in the following Granulated Activated Carbon (GAC) filtration. Carbon is an active agent against organic micropollutants, heavy metals, and other compounds. After all processes have been completed, sanitised water is placed in treated water tanks. Prior to the water distribution, it is chlorinated for final disinfection.

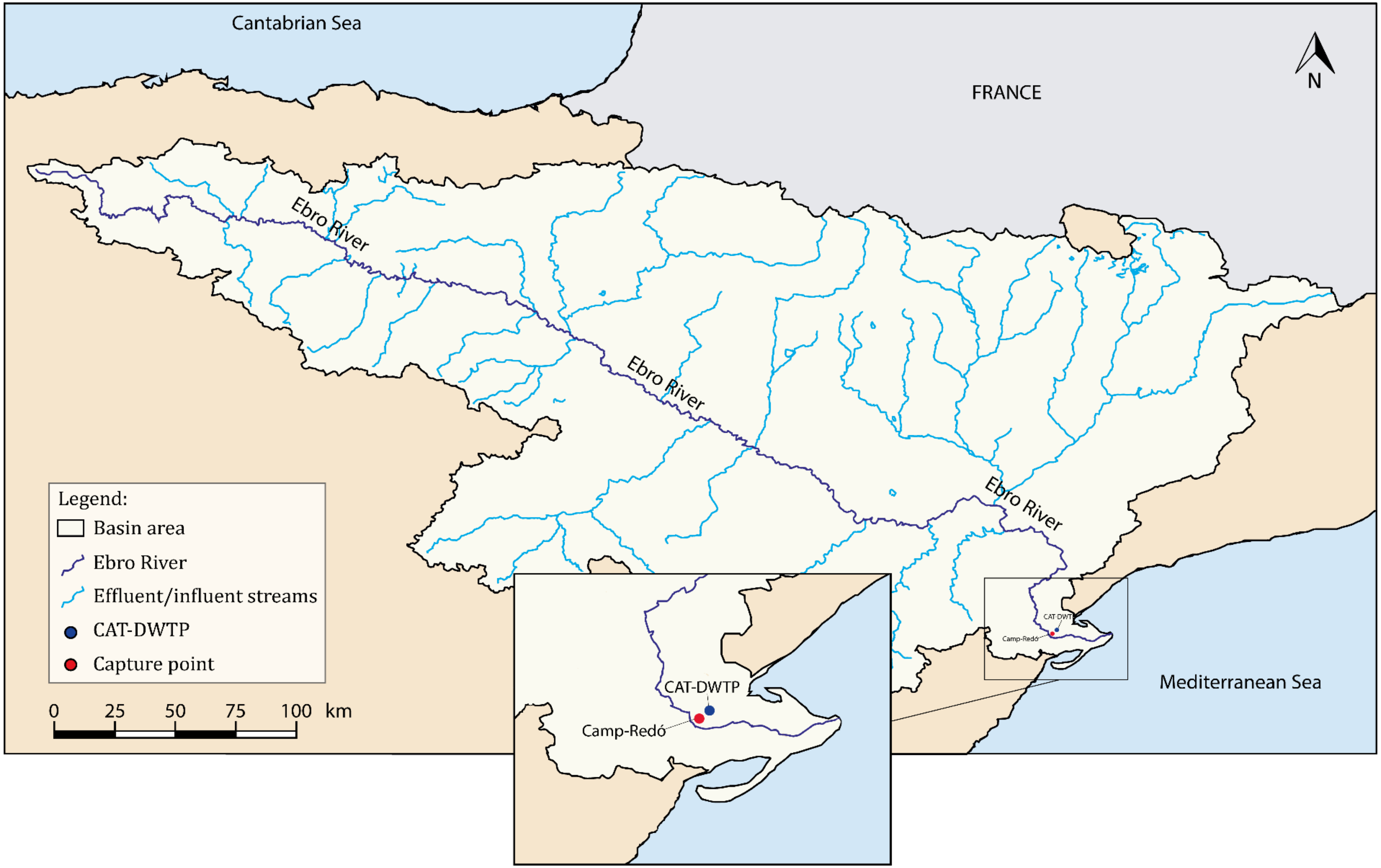


***Figure C3.2 –** CAT-DWTP’s capture point is close to the Ebro’s River delta*

3.2.2. Ebro River water quality

The Ebro River, which has a length of 930 km and a catchment area of 86.100 km^2, is the longest river in Spain and the second longest river in the Iberian Peninsula. It crosses 8 autonomous communities of Spain: Cantabria, Castilla y León, Euskadi, La Rioja, Navarra, Aragón, País Valencià and Catalunya.

The Ebro River mouth is located in Catalonia, forming a delta due to the high alluvial sediments transported along its basin. The Ebro delta extension is about 328 km^2, covered by rice paddies. Figure C3.2 shows the Ebro River's basin area extension and its affluent streams. The magnified detail shows the water capture point in Camp-Redó near the CAT-DWTP.

The parameters shown in Table C3.1 are just a few of the many parameters that are evaluated using an average of two control samples every week in the CAT-DWTP (cations, anions, nitrites, etc.).

During extreme rain events, it can be challenging to ensure meeting the water quality criteria of the RD 140/2003. Heavy rainstorm episodes, closely related to seasonality, can provoke a rapidly rising river level, dragging large amounts of sediment and plants. Turbidity, salinity, and total suspended solids values are up to 50% higher in the wet season (spring and autumn) than during dry intervals However, even with the occasional episodes aforementioned, the high-water quality obtained after an entire treatment is remarkable.

3.2.3. Sensors and data collection

Water quality has been monitored through three spectro::lyser®. The sensors were used to gather information on different CAT-DWTP treatments.

One spectro::lyser® (PreO$_3$) monitored the raw water tank before the pre-ozonation treatment. This probe registered the UV spectrum (wavelengths from 200 to 400 nm). The second spectro::lyser® (PostO$_3$) monitored the water after the post-ozonation process (wavelengths from 200 to 750 nm). And the last spectro::lyser® (EB1) monitored the final chlorination. PostO$_3$ and EB1 registered the UV-Vis spectrum (wavelengths from 200 to 750 nm) (Figure C3.1).

Two of the three spectro::lyser® units ($PreO_3$ and $PostO_3$) were installed in the chemical analysis lab, which has a variety of taps that are used for routine water quality control. After the final chlorination disinfection, the other probe (EB1) was installed in the treated-water laboratory near the pumping station. Figures C3.3A and B exhibit the installation of the probes in both laboratories.

The three probes had integrated an automatic cleaning. The brushes attached were programmed to clean up the window path before every measurement. Water flow passes through the window path, dragging some organic and inorganic particles that can be stuck in it. The more particles adhered to the window, the worse the measurement can become, generating fouling. Cleaning brushes were installed to avoid it.

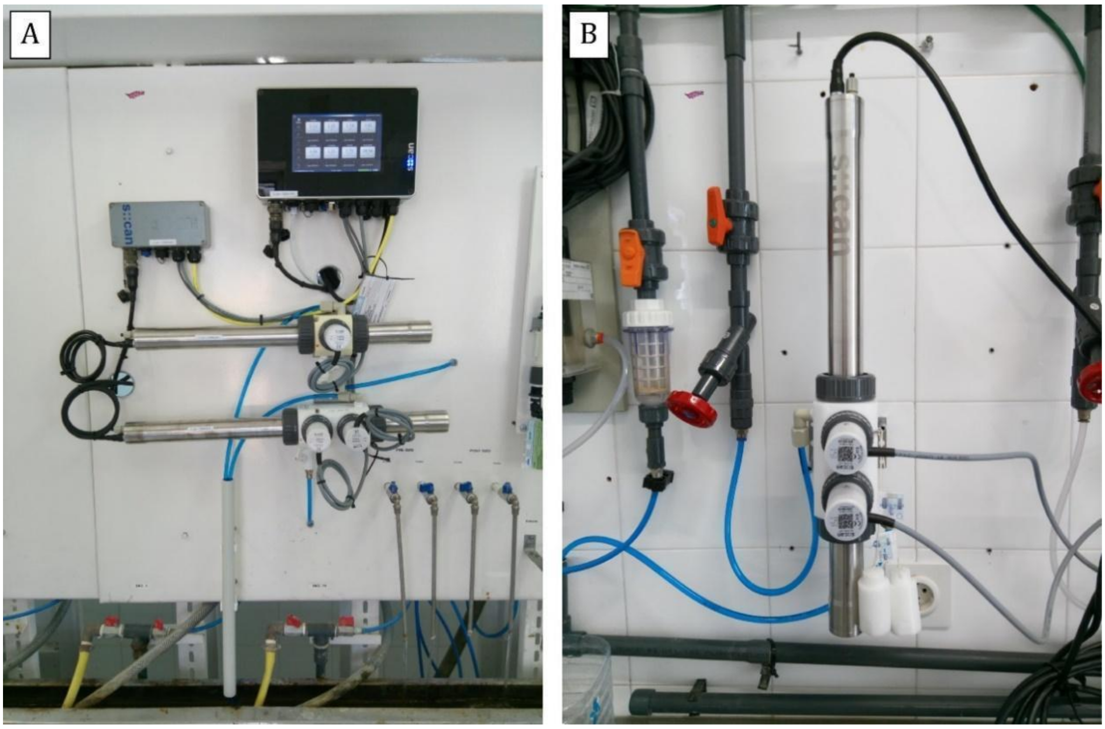


Figure C3.3 – *spectro::lyser® probes installed at CAT–DWTP laboratories. As seen in A, the sensors are connected to a con::cube®, the PLC that records the data, and from where the maintenance of the sensors is managed. All sensors have built-in cleaning brushes, modulated according to the measurements, and clean the measuring paths to ensure high-quality data.*

All probes were connected to a con::cube®, a compact and versatile terminal for data acquisition and water station monitoring. It offers flexible options for remote water quality monitoring and can be connected to SCADA or any central database system. The con::cubes registered and sent high amounts of data gathered by spectro::lyser® for one year, controlling the cleaning system, registering and notifying probes maintenance requirements or critical water quality events. Another interesting con::cube® feature is the capacity to create virtual formulas. A virtual formula known as parameter combination provides essential information that would otherwise be impossible to acquire.

In the treated–water laboratory, CAT–DWTP operators installed an i::scan® probe. This miniature multi-parameter spectrophotometer can estimate water quality parameters such as TOC, DOC, UV_{254} or temperature measuring absorbance at several wavelengths. Using a specific parameter combination recorded with the i::scan®, a trihalomethanes formation potential (THMs FP) virtual parameter was created. The formula used is a nonlinear regression developed by Amy et al., 1998 (F3.1) to calculate THM formation (µg/L) after 48 hours. For lack of a better-suited expression, the term THM formation potential (THM FP) will be employed in this PhD thesis henceforth in this context. The formula is:

$$THM\,FP = 0.00253(DOC)^{1.22}(Cl_2)^{0.442}(T)^{1.34}(pH)^{1.75}(t)^{0.34} \qquad \text{(F3.1)}$$

Where THM FP is the formation of trihalomethanes expressed as µg/L, DOC is the Dissolved Organic Carbon present in water expressed as mg/L, Cl_2 is the concentration of free chlorine expressed as mg/L, applied at the beginning of the treatment, T is the water temperature expressed as Celsius degrees, pH that is a measure of how acidic/basic water is, and t is the time, expressed in hours.

Every week, data from the certified laboratory of CAT-DWTP were compared to those from the virtual parameter to determine whether the THM FP concentration was suitable. In addition to this analysis, Institut Català de Recerca de l'Aigua conducted six sample campaigns (ICRA). The obtained THM concentrations in these formation potential tests were analysed following the standard operating procedure of EPA Method 501.3.

Over a year, the three spectro::lyser® took measurements every two minutes. The UV ($PreO_3$ spectro::lyser® ranged from 200 to 400 nm) and UV-Vis spectrum ($PostO_3$ and EB1 spectro::lyser® ranged from 200 to 750 nm) were recorded and, using s::can commercial virtual formulas, parameters such as turbidity, nitrate (NO_3^-) and Total Organic Carbon (TOC) have been measured. Table C3.1 gives a general overview of the data recorded.

Table C3.1 – *Water quality recorded during the project in three different locations of CAT–DWTP. Raw water quality is recorded before pre-ozonization treatment. Water quality after the intermediate treatments was recorded at the outlet of postozonizaation. Treated water was registered just before the chlorination, after the BAC filter.*

Monitored water	Sensor installed	Params.	Units	Min.	Max.	Avg.	σ
Raw water	spectro::lyser®–v2 UV	Turbidity	NTUs	1.9	42.1	5.6	2.2
		Nitrate	mg/L	1.8	16.7	13.8	0.3
		TOC	mg C/L	1.2	15.0	2.8	0.7
		UV_{254}	Abs/m	1.3	35.9	5.4	1.8
		Temperature	°C	9.4	28.4	20.9	5.0
		Fingerprint	Abs/m	-	-	-	-
Post-ozonated water	spectro::lyser® –v2 UV–Vis	Turbidity	NTUs	0.04	1.2	0.1	0.1
		Nitrate	mg/L	10.4	16.1	14.8	0.3
		TOC	mg C/L	1.9	2.9	2.0	0.1
		UV_{254}	Abs/m	1.7	5.9	2.2	0.6
		Temperature	°C	14.5	31.9	25.2	5.0
		Fingerprint	Abs/m	-	-	-	-
Water before chlorination	spectro::lyser® –v2 UV–Vis	Turbidity	NTUs	0.1	0.8	0.1	0.1
		Nitrate	mg/L	8.9	14.1	11.1	1.2
		TOC	mg C/L	1.7	2.4	1.9	0.2
		UV_{254}	Abs/m	1.9	4.1	2.6	0.6
		Temperature	°C	10.9	27.6	21.7	5.1
		Fingerprint	Abs/m	-	-	-	-
	i::scan	THM FP (48h) Virtual Param.	µg/L	34.6	55.0	47.8	3.2

Figure C3.4 exhibits the differences of absorbance between waters monitored in the CAT-DWTP. As it can be seen, PreO3 is related to raw water extracted directly from the river, as the absorbances are higher than others observed in PostO3 and EB1. Absorbances registered in EB1 (outlet of the plant) are lower than 7 Abs/m in the region of organic matter (254 nm), which it is an indication of a proper water sanitation.

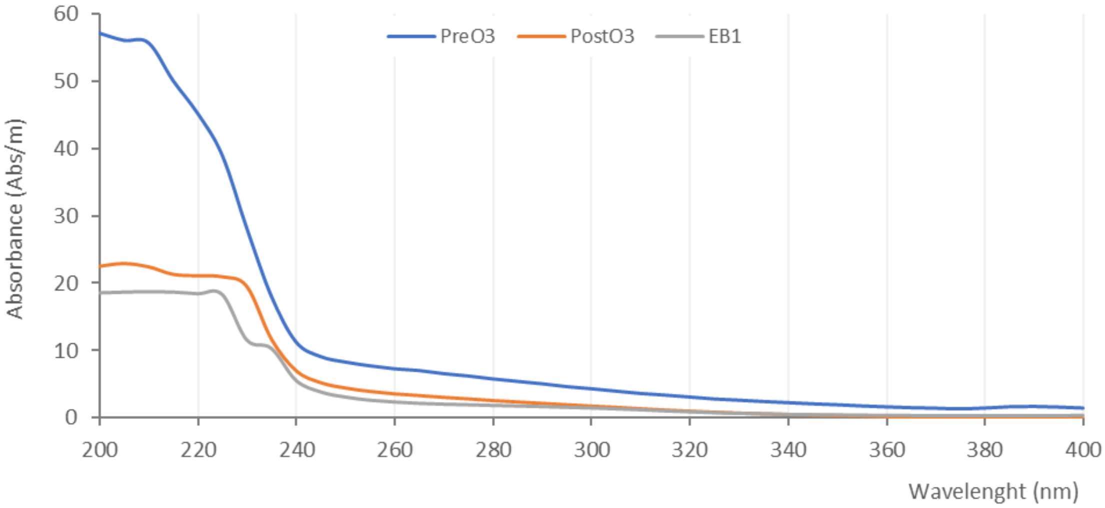


Figure C3. 4 – *Representative UV-Vis spectrum of all three sampling points.*

3.2.4. Data structure

All statistical studies were performed using different libraries from RStudio, v1.2.5033. Figure C3.5 shows the sequence of the procedures performed in this chapter. As it shows, the first step was gathering all data from the spectro::lyser® probes installed. The wavelengths recorded were used as predictors (columns). $PreO_3$ registered 200 wavelengths, and $PostO_3$ and EB1 registered 550 wavelengths. The response variable was the THM FP virtual formula produced by the i::scan®. Every two minutes, each of the three probes recorded one observation of the water quality as a row. A preliminary database with almost 400,000 rows and 1,302 columns was created.

1,301 variables recorded by the three probes were available for modelling. The first step was a deep data cleaning. The general data was explored seeking erroneous information and zeros. In addition, a manual wavelength pruning was made initially, eliminating long wavelengths (from 400 nm), and keeping every 5 wavelengths to enable a quicker automatic variable selection. Once the data cleaning was applied, the general dataset had a remanent of 84,140 rows and 123 columns. Before the automatic variable selection, a centring (mean of independent variables was subtracted from all the values – all independent variables had zero mean) and scaling (predictors are divided by their standard deviation – all independent variables had a standard deviation of one) transformation was applied.

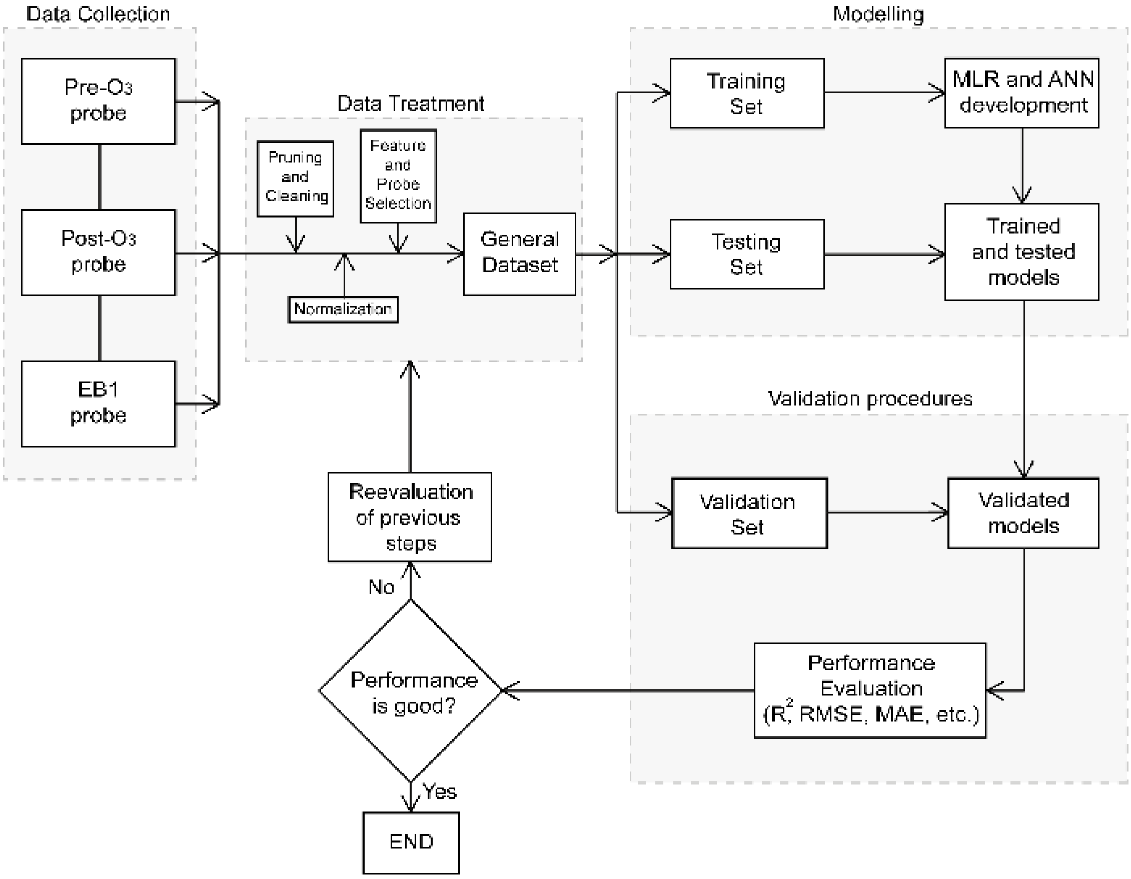


***Figure C3. 5** – Flowsheet outlining the whole model development. Data collected was normalised and divided into three datasets: training, testing, and validation sets. Several indicators of performance determined the goodness of the models.*

The automatic feature selection was applied using Forward Selection (FS), Backward Elimination (BE) and Stepwise Selection (SS) motors (for more detailed information, see next section). The most suitable feature selection method was examined considering all possible sensor combinations to obtain the best variable subset for predicting THM FP. Once the best subset of variables was automatically selected for each sensor combination, Variance Inflation Factor (VIF) was applied as a selected variable validation method.

Once the general databases were created, the correlation between variables was investigated for each combination of sensors by applying the *ggplot2* library. The *ggcorrplot* function calculates the correlation between variables and graphs it automatically. The general and temporal tendencies (i.e., drift) and outliers were explored throughout the *plot* function from the *Stats* library.

The last step of data management was the splitting of the general databases for each sensor combination into training, testing, and validation. First, a validation set was built by separating the final 1,000 rows. The purpose of this separation is for setting a different time period between training-testing and validation sets, considering a great number of observations to properly validate the model prediction capacity with non-trained data. The training and testing sets were created from the remaining data. These were used to create the models and verify their performance, respectively. The partitions were built using the function *createDataPartition* from the *CARET* library (Kuhn, 2008; 2013). In this case, 75% of the remaining data went to the training set, and the last 25% to the testing.

3.2.5. Variable selection methods

Feature selection motors were applied to counteract the dimensionality problem typically met in data science. The approaches suggested in this thesis seek the simplicity of calculations and application. They are Forward Selection (FS), Backward Elimination (BE) and Stepwise Selection (SS). FS, BE and SS motors identify the best variable subset in a high-dimensional data frame, achieving the simplest model with the best outcome (Mehmood et al., 2012).

FS begins with an empty model (without variables) called *Null Model*. The variable that results in the model with the highest correlation is then added one by one by the algorithm. This model is saved and in the next iteration via the same procedure, another variable is added. The algorithm stops when a specified threshold is reached.

BE begins with a model that contains all variables (in short, the *Full Model*). The variable that least reduces correlation in this stage is then removed one by one in an iterative process by the algorithm. Once the specified threshold is reached, the algorithm stops.

SS is a mixed method between FS and BE. It begins with a *Null Model*. The algorithm starts adding the most significant variables. However, it acknowledges that adding a variable can make the contribution of a previously added variable less relevant. The algorithm, therefore, does a new analysis of the chosen variables at each iteration. The approach removes any non-significant aggregated variables from the subset while keeping the significant ones. Once the specified threshold is reached, the algorithm finally stops.

The threshold for selecting the best variable at each iteration for every subset automatically selected was the *p*-value. The *p*-value (probability – value) exposes the likelihood of a new selected variable regarding the selected and the unselected ones. The smaller the *p*-value, the stronger the variable fits in the subset of selected variables.

The threshold for selecting the best subset variable group was selecting the minimum Akaike Information Criterion (AIC) (Akaike, 1978; 1979) value. AIC can be used to more generalised models and applied in a multicollinear context (Yamashita et al., 2003).

AIC is a goodness-of-fit because it estimates the quality of each subset of variables selected. Given a dataset, several candidate models can be classified according to their AIC. The smaller this value, the better the model is. In general, the AIC (F3.2) is defined as:

$$AIC = 2k - 2\,ln\,(L) \qquad \text{(F3.2)}$$

Where *k* is the number of parameters in the model and *ln(L)* is the log-likelihood function for the statistical model. A good mathematical description of this metric and its uses in different fields of science can be found in Box et al., 2008; Aho et al., 2014 and Giraud et al., 2015.

The variable selection methods were implemented using RStudio. The *Stats* library has the *stepAIC* tool, where FS, BE, and SS can be calculated by specifying *forward*, *backward* or *both* functions.

After applying the automatic selection motors, another manual pruning was done considering the Variance Inflation Factor (VIF) for each variable. VIF measures the multicollinearity in a set of multiple features (Neter et al., 2004). A high VIF indicates that the associated group of independent variables is highly collinear and should be reviewed. The formula of VIF (F3.3) is:

$$VIF = \frac{1}{1 - R_j^2} \tag{F3.3}$$

Where R^2_j is the determination coefficient of the variable X_j over the rest of the selected explanatory variables. Specific suggestions for a cut-off VIF are between 5 and 10 ($5 \leq VIF_j \leq 10$). However, it must be considered a $VIF_j = 10$ implies that $R^2_j = 0.9$. A high VIF affects the confidence interval's width for its associated parameter estimate. The higher the collinearity, the closer it will bring R^2_j to 1, and the higher the value of VIF_j. The variables selected with a VIF>10 were reviewed; thus, they had high collinearity, affecting the model and its final prediction capacity.

3.2.6. Model development

The THM FP has been predicted using two types of models: Multivariate Linear Regression (MLR) and Artificial Neural Networks (ANN).

- Multivariate Linear Regression (MLR):

An MLR model is a statistical technique that uses two or more independent variables (or features, since not always are entirely independent) to predict a response variable.

The general linear regression model (F3.4) takes the form of:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ik} + \varepsilon_i \quad i = 1, \dots, \tag{F3.4}$$

Where y is the response variable, β_0 is the y-intercept (the value when X_{ik} are 0), $\beta_1, \dots, \beta_p$ are the regression coefficients. $X_{i1}, X_{i2}, \dots X_{ik}$ are the features of the model, and ε is the residual term (model's random error) of the model. A profound explanation of MLR mathematics can be found, for instance, in Geladi et al., 2003 and Kumar et al., 2014.

To create an MLR model in RStudio, the *lm* function of the *Stats* library was used, which allows obtaining regression results in a simple and agile way. The *summary* function was used to extract the regression results obtained. Based on residual values, standard error, and t-value, this function indicates whether the variables are conforming or not. It also gives information about the final correlation (R^2). The functions *plot*, *points*, and *abline* from the *Stats* library were used for data visualisation.

- Artificial Neural Networks (ANN):

An artificial neural network (ANN) is a set of algorithms inspired by the human brain and how neurons work together to understand daily inputs. In other words, an ANN algorithm tries to recognise, analyse, and process information from a dataset (i.e., patterns or relationships) as it would a human brain. An ANN model has three minimum parts: i) an inlet layer with all the input parameters selected, ii) a hidden layer with an activation function, and iii) the output layer with the response variable.

The *neuralnet* library was used to generate the ANN models, and it requires the specification of a database, the response variable, and all the independent variables. The number of hidden layers of the model must be selected as well as the neurons in each layer. As their importance in the model's overfitting (or underfitting), hidden layers play a significant role in the architecture of an ANN model. Many authors (Kröse et al., 1993; Ke et al., 2008; Vujicic et al., 2016) concluded that choosing the right number of hidden layers and neurons can be done by trial and error, but it is also considered that two layers are the ideal amount to create a simple and adaptable ANN model. In this chapter, there were two layers, and a self-designed algorithm that allows for trial and error was used to choose how many neurons should be in each layer. For each layer, up to 10 different possible combinations of neurons were tested.

To visualise the results, *plot*, *points* and *abline* functions were used from the library *Stats*. The ANN Activation Function used was the Sigmoid Function (F3.5), and it is as follows:

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{e^{x} + 1} \tag{F3.5}$$

To validate the models, the *predict* function of the *stats* library has been used.

- Goodness of models

The goodness of models developed has been observed by correlation (actual against predicted values) and RMSE (Root Mean Squared Error). Based on these criteria, predictability and residual error can be calculated. We applied the function *cor* from the *Stats* library to obtain correlation results between the tested and validation values. A function to find the RMSE (F2.6) was created using the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}} \quad \text{(F3.6)}$$

The graphical design of the results obtained was made using two different libraries. The graphics have been made with the *plot* function of the *Stats* library and the *ggplot2* library; the function used for these cases is the *ggplot*.

A final edition for all graphs of this chapter was made using vector-based software.

3.3. Results and discussion

3.3.1. Variable and sensor selection

Variables were selected using Forward (FS), Backward (BE) and Stepwise (SS) Selection motors prior to the construction of the mathematical models. The relevance of each sensor was also evaluated. These results can be seen in Figures 5 to 10. The correlation (R^2) and the Root Mean Square Error (RMSE) can be seen in all figures. All graphics use dotted lines for FS, dashed lines for BE, and straight lines for SS.

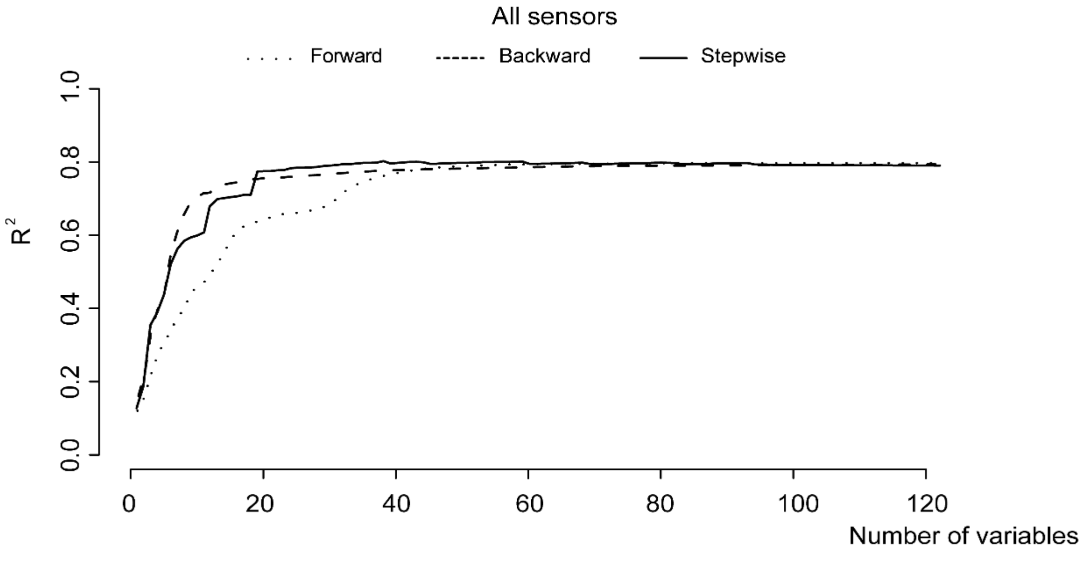


Figure C3.6 *– Determination coefficient results (R2) for all sensors. The highest R^2 obtained was 0.8.*

Figure C3.6 illustrates the behaviour of the determination coefficient (R^2) of the three variable selection motors using data from all spectro::lysers®. The abscissa axis refers to the number of variables selected by each algorithm. All three motors display an increasing correlation trend (from 0.1 to 0.8). SS can hardly improve the coefficient of determination from the twentieth variable selected. By contrast, FS will need more than 40 variables to achieve the same result as SS. On the other hand, BE obtained a high determination coefficient (0.78) in less than fifteen variables. However, it would need ten to fifteen more variables to achieve the maximum R^2 value.

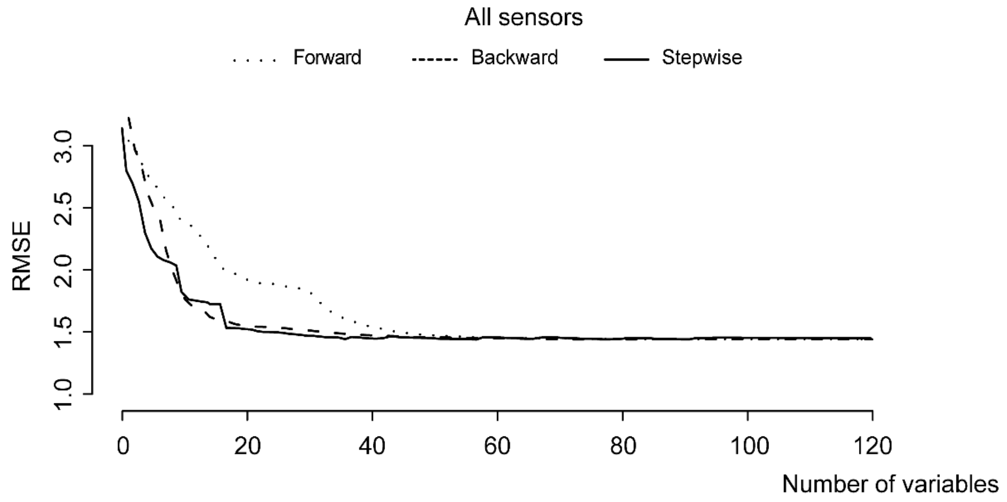


***Figure C3.7 –** Root Mean Square Error (RMSE) for all sensors. The lowest RMSE value obtained was 1.5*

Figure C3.7 exhibits the Root Mean Squared Error (RMSE) obtained from FS, BE and SS motors. The error decreased as many variables were added to the final model (from 3.0 to 1.5). As can be seen, SS reached the minimum error with fewer variables than FS and BE.

The differences between BE and SS are subtle. In other words, between 18 and 40 variables, SS is slightly higher. With up to 18 variables, BE can, however, be more reliable and predictable. In this case, the best selection motors would be both, SS and BE. However, SS achieved the best results while requiring fewer variables. In contrast, FS scoped R^2 and RMSE require twice the number of variables.

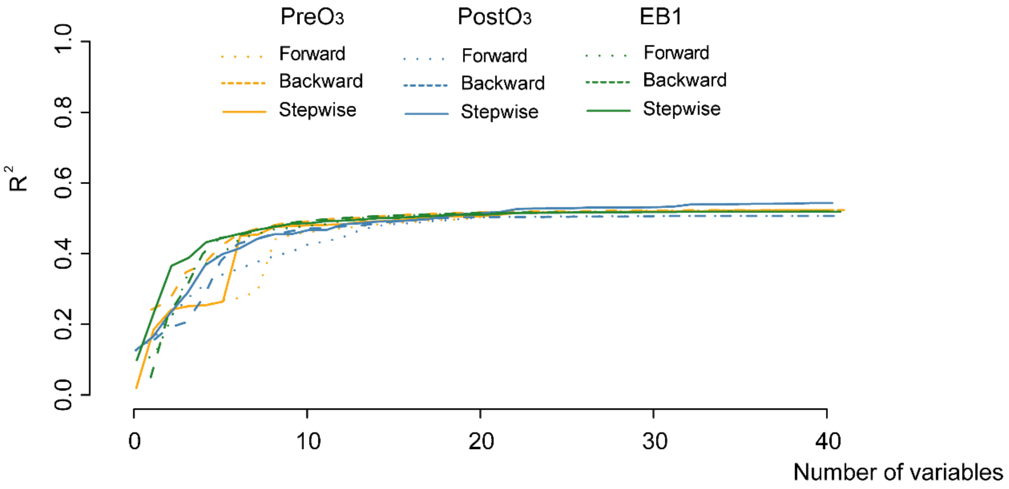


Figure C3.8 – Coefficient of determination results (R^2) for all sensors, modelling them separately. The highest R^2 obtained was 0.55

Figures C3.8 and C3.9 show the Determination Coefficient (R^2) and the Root Mean Squared Error (RMSE) obtained when the sensors are used independently. Even when all the variables are selected in this scenario and for all selection techniques, the R^2 does not reach 0.55. (Figure C3.8).

All sensors give a similar trend. However, for all three selection procedures, PostO3 (blue lines) has the worst determination coefficients. PreO3 and EB1, on the other hand, reach the maximum R^2 with less than 15 variables with the SS method. In this case, PreO3 changes the slope abruptly at the sixth step; whereas, EB1 maintained a steady trend.

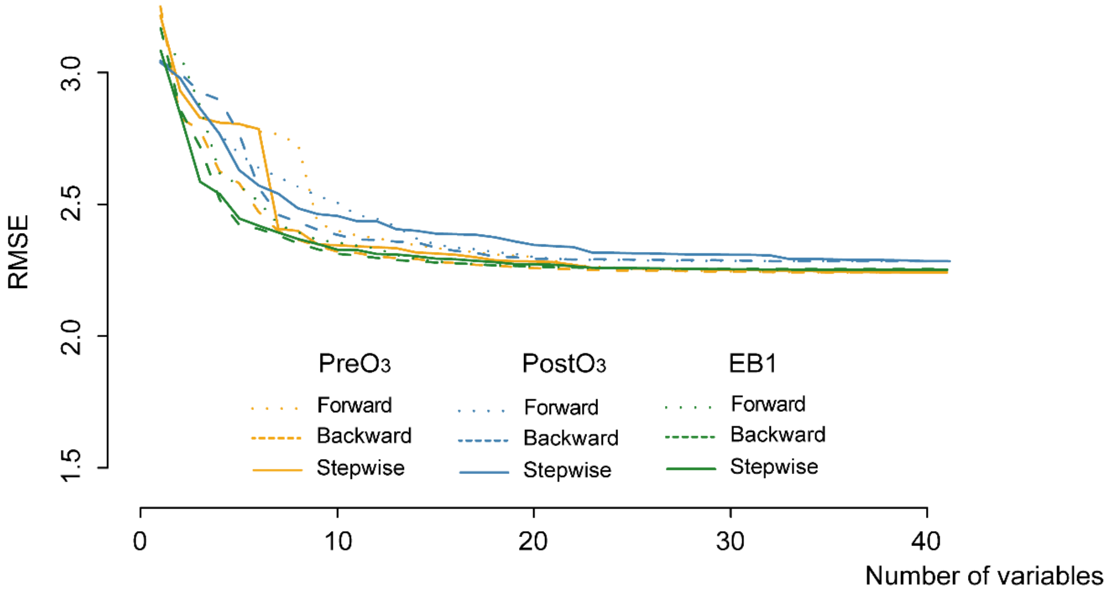


Figure C3. 9 – *Root Mean Square Error (RMSE) for all sensors, modelling them separately. The lowest RMSE value obtained was 2.5*

The lowest RMSE obtained is 2.3 for all three methods in each case (Figure C3.9). However, as discussed in Figure C3.8, PreO3 and EB1 need fewer variables than PostO3 to reach less RMSE. Additionally, EB1 achieve the same RMSE value (2.4) as PreO3 in the sixth step with an unwavering behaviour. In this case, the variables selected by SS were slightly different for EB1 and PreO3.

Figures C3.10 and C3.11 show the R^2 and RMSE combining two sensors. Blue lines correspond to PreO3 and EB1, green lines are related to PostO3 and EB1, and the orange ones are associated with PreO3 and PostO3.

As can be seen in Figure C3.10, the best combination is PreO3 and EB1 (blue lines) since it registers the highest correlation values (0.80), with the smaller number of variables selected (less than 20). The combination of $PreO_3$ and $PostO_3$ has the worst results, obtaining an R^2 of around 0.60, setting more than 30 variables. $PostO_3$ and EB1 combination would achieve a slightly larger R^2 (0.63) than $PreO_3$ and $PostO_3$.

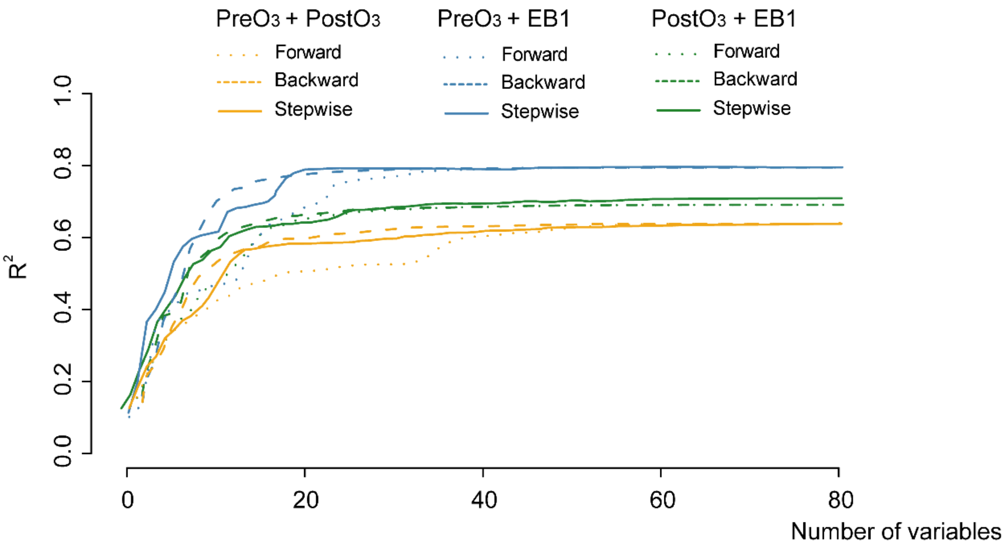


***Figure C3.10 –** Coefficient of determination results (R^2) using combinations of two sensors. The highest R^2 value obtained was 0.80*

Figure C3.10 exhibits the same selection efficiency of FS as Figures C3.6 and C3.8. It is the selection motor with the worst results for all three combinations since it would need between 15 and 20 more variables than the other two selection motors to reach the same results. BE and SS give similar results, but SS would reach the same result with fewer variables than BE.

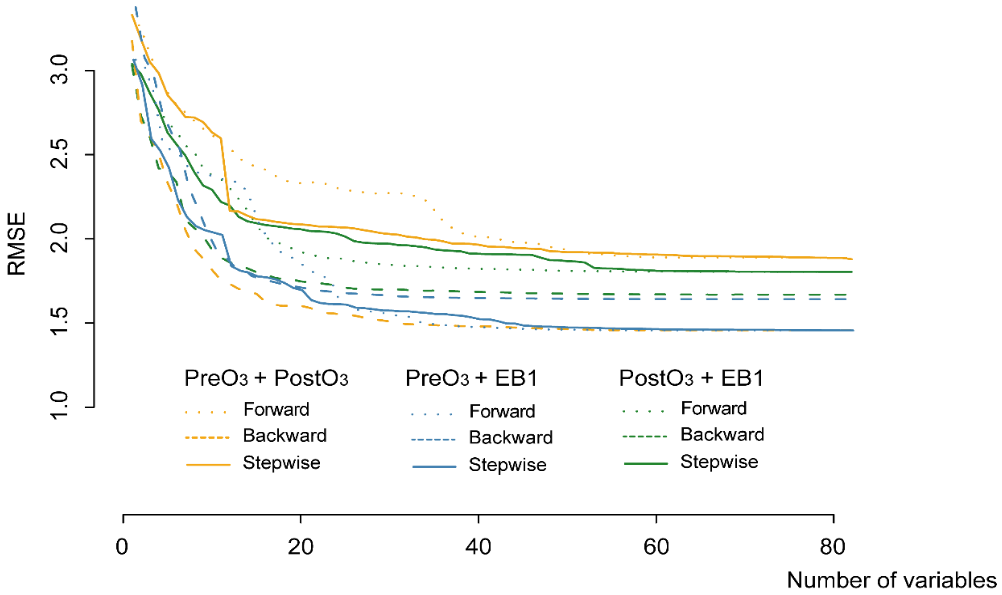


Figure C3.11 – Root Mean Square Error (RMSE) using combinations of two sensors. The lowest RMSE value obtained was 1.5

Suboptimal results obtained from FS shown in Figures C3.6 to C3.11 for all cases would be related to the impossibility of eliminating variables already selected. FS cannot observe the significance of all variables at once and therefore lacks to identify less predictive individuals.

Something similar happens with the SS motor. When the variable selection begins, it is done from an empty model, and the variables are added or subtracted without identifying their individual consistency in a global model. The final improvement of SS results is because of the capability of eliminating variables that worsen the overall model significance at any given time. However, it can avoid choosing other features that would have improved the final result if they were selected first. While one particular subset of a set of variables might have significant predictive power, another subset might not. This behaviour is called *the suppressor effect* and occurs when the predictors are only significant when another predictor or a subset of predictors is held constant (Heinze et al., 2020).

On the other hand, BE performs an initial study of the predictive capacity of each variable, understanding the significance of the global set (Heinze et al., 2020). This method can identify the most and least predictive individuals and select them or not, considering their individual and collective significance. In other words, this kind of selection motor keeps the variables that are most important in the model and eliminates the ones that would have less of an impact. This behaviour is related BE curves' smooth shape observed in all graphics. The logical explanation regarding the differences in BE's and SS's curve shape is the number of iterations of each motor. SS took 18–20 steps to achieve the maximum R^2 and the minimum RMSE but BE performed approximately 100 steps—a time-and energy consuming process—to achieve nearly the same results.

Due to the first motor's ability to re-evaluate those features that have been deleted, the final findings show that SS performs better than BE. Once a variable has been removed by BE it can no longer be chosen. This factor can also generate errors in the selection behaviour previously identified for the FS method. Although there are some drawbacks to FS, BE, and SS methods, the evidence in the results concludes that SS can achieve better results by selecting fewer variables in fewer iterations for all the cases studied.

As shown in Figures C3.8, C3.9, C3.10, and C3.11, the combined information from the EB1 and PreO$_3$ sensors gives the best results. That would seem logical since these two sensors are placed further away in the treatment train. The other combinations of sensors (PreO$_3$ and Post O$_3$; PostO$_3$ and EB1) do not reach an equivalent level of goodness in the values obtained because sensors placed closely together in the treatment train produce data with higher collinearity. Although the EB1 sensor individually gives the best results (Figures C3.8 and C3.9), its maximum prediction value achieved is relatively low (0.55). All the sensors combined (Figures C3.6 and C3.7) yield very similar results concerning the R^2 (0.80) and RMSE (1.50) as if both PreO$_3$ and EB1 sensors were combined.

Briefly, the information gathered at the inlet and the outlet of the CAT-DWTP is essential. This fact seems logical because the THMs FP is directly related to the amount of total organic matter present at the plant's inlet, which is oxidated by chlorination and ozonation processes at the outlet. Finally, observing all the results obtained, it can also be concluded that the information gathered from the PostO$_3$ sensor is less relevant to predicting THMs FP in CAT-DWTP.

A correlation analysis was performed for the best sensor combination (PreO$_3$ and EB1) for determining the THMs FP. Figure C3.12 shows the correlation matrix of all wavelengths for the two sensors. First, there are the wavelengths of the PreO$_3$ sensor and then those of the EB1. There is no direct correlation between PreO$_3$ wavelength and EB1, with ratios ranging from –0.2 to 0.2. Negative correlations occur with some variables (light-blue-coloured areas), for instance, between variables from the EB1 sensor (correlations between λ240 to λ400 and λ200 to λ240). This negative correlation would be related to the disinfectant presence (sodium hypochlorite), which absorbs at low wavelengths. The more sodium hypochlorite, the higher the absorbance on low wavelengths, and the lower absorbance on higher wavelengths (bleaching effect).

The PreO$_3$ sensor registers correlations ranging from 0.5 to 1.0. A similar trend is found in EB1, especially in wavelengths near each other. High correlated variables cause *severe multicollinearity*, directly affecting an MLR model's efficiency. It can increase the variance of the coefficient estimates, making them very sensitive to minor changes and weakening the statistical power of the regression model.

A critical drawback of automatically variable selection is no motor can observe the multicollinearity factor. Wavelengths contain a fraction of the same information since they are continuous variables, though they are often treated as discrete. This fact increases the possibility of having severe multicollinearity and difficult automatic selection. Therefore, a final manual pruning was mandatory once the most significant variables were selected automatically to minimise the effect of severe multicollinearity dragged. The Variance Inflation Factor (VIF) was used to quantify the multicollinearity.

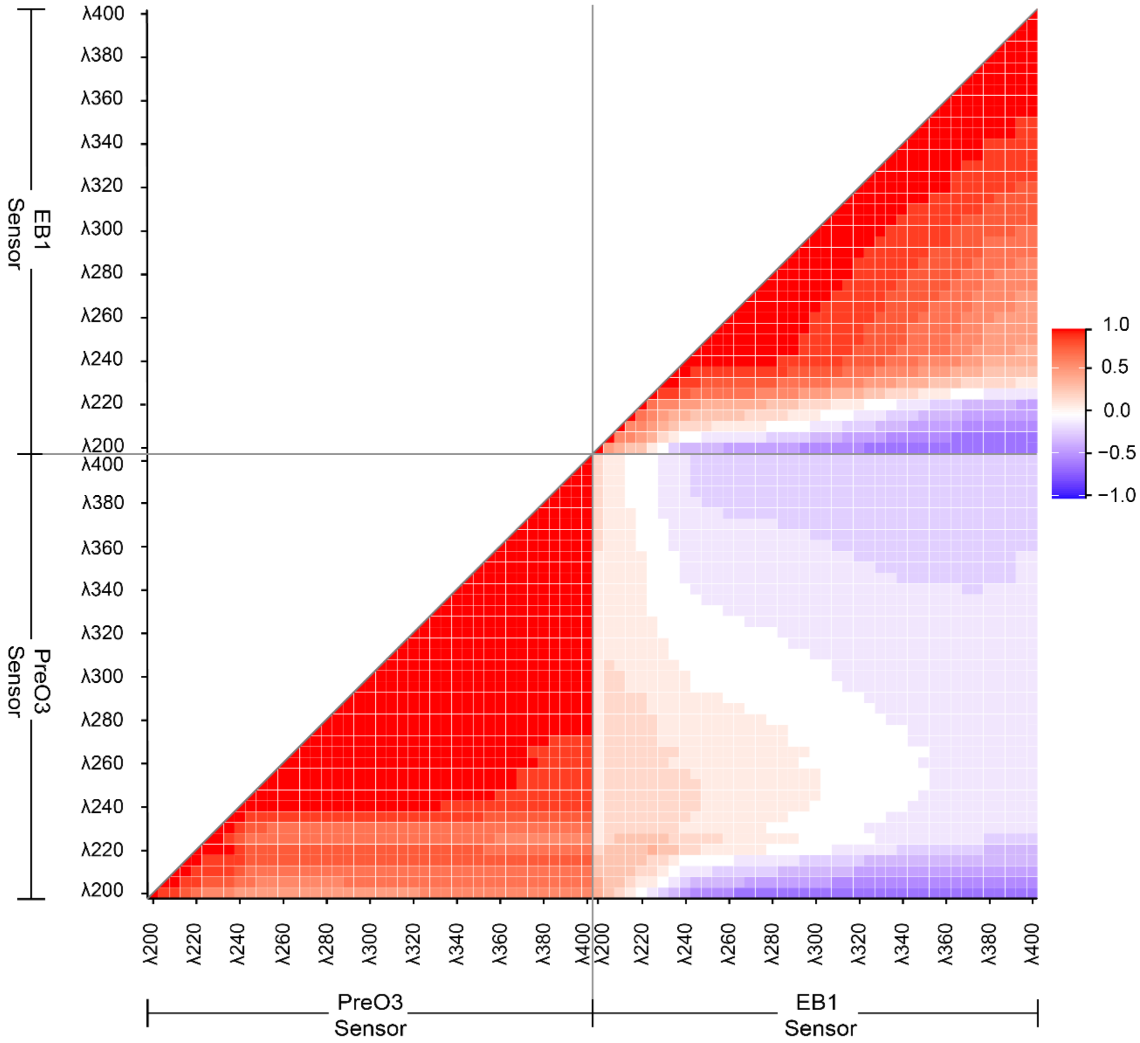


Figure C3.12 – *Correlation matrix between PreO$_3$ and EB1 wavelengths. PreO$_3$ wavelengths have important multicollinearity, and they must be evaluated.*

The VIF of every remaining variable was checked for a correct final selection. The first overall examination exhibits a general high VIF for all variables (VIFs ≈ 2000). The variable eliminated first was that with the higher VIF value. After the variable was eliminated, the VIF of the remaining variables was re-checked for evidence of the overall improvement or recession. As the number of variables in the models decreased, less information was available, worsening the determination coefficient (R^2), and the Root Mean Squared Error (RMSE). All the variables were examined individually, selecting only those that improved the significance of the final group, decreasing the overall VIF.

The table below (Table C3.2) illustrates the VIFs for the final variables selected manually. As previously mentioned, larger VIFs have a direct effect on the precision of the estimated coefficients, which weakens the statistical power of the regression. The final variables selected were those that give the minimal VIF to the group. As can be seen in Table C3.2, there are some variables with critical VIF (from $PreO_3$: λ265, λ395; and from EB1: λ255, λ300). Kutner et al., 2004 explain mathematically and agilely the effect of multicollinearity and VIF value in linear regression.

Table C3.2 – *Variance Inflation Factor (VIF) of the selected variables.*

	Variable	**VIF**
$PreO_3$ Sensor	λ235	5.21
	λ265	40.69
	λ395	39.09
EB1 Sensor	λ300	53.31
	λ400	8.62
	λ235	10.99
	λ255	51.90

The final selected variables from the $PreO_3$ sensor for the modelling are λ235, λ265, and λ395, and for the EB1 sensor were λ235, λ255, λ300 and λ400.

3.3.2. Data organization

The information gathered from the combination of $PreO_3$ and EB1 sensors was used to generate two prediction models of THM FP, an MLR and ANN. The performance of both models was then compared. The general database was divided into training, testing, and validation. Figure C3. 13 shows the organisation of the data.

First, the validation set was separated from the general database. One thousand rows were set apart. Those rows were separated from the end of the general database. The remaining data were randomly divided in a proportion of 75-25 to generate the training and testing sets, respectively. The training set was used to build the models, which were then used to predict how the observations in the testing set would respond. The error rate obtained (RMSE) estimates the error in the model predictions. Additionally, applying the testing set to the trained model can help fine-tune it, evaluating the competence of the selected wavelengths. The testing set contained different data from the training set during the same time frame as the validation set, which was how the two sets were different. The validation dataset, on the other hand, uses observations from a very different time period, that were not used in developing or fine-tuning the model, thus providing an unbiased assessment of the model's performance.

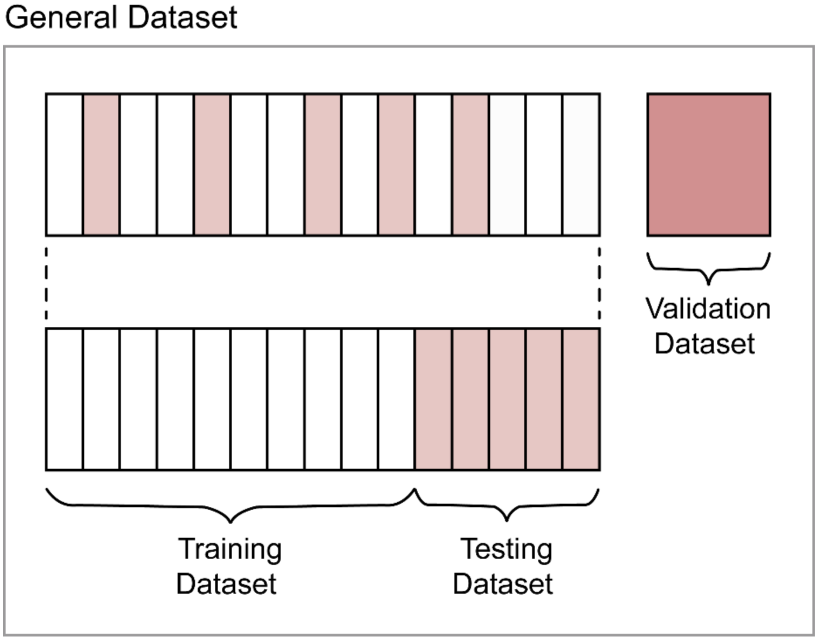


Figure C3.13 – *Schematic data distribution into three datasets: training, testing and validation sets. The first and second datasets were created by generating randomness to give robustness to the model. The third has been separated from the very beginning. This step allows observing the actual predictive performance of the model.*

THM FP variable was obtained by incorporating Amy’s formula (Amy et al., 1998) in the con::cube installed at the chlorination laboratory (EB1). There were different stations with several quality probes (pH::lyser, i::scan, condu::lyser, chlorine analysers, etc.) from which the necessary information was extracted. The THM FP values obtained applying that formula were validated by performing laboratory analyses in the CAT-DWTP (weekly) and Institut Català de Recerca de l'Aigua (monthly).

To develop the formula, Amy et al., 1998 gathered 1,170 samples from 13 natural waters. The experiments were performed considering many cases, from baseline to high-extreme and low-extreme conditions. They analysed the pH and DOC. DOC was reported to define the THM precursor content. Chlorination was made considering a concentration in which a positive chlorine residual was maintained over the time (168 h) from 1.5 to 69 ppm. The temperature, which ranged from 10 to 30 degrees Celsius, was recorded.

More recently, other authors (Sadiq et al., 2019) give a comprehensive review of all empirical models created throughout the decades for predicting THMs FP in drinking waters. Godó-Pla et al., 2021, for instance, created and validated a log-linear model using critical drinking water quality parameters such as UV254, TOC, pH and temperature, among others.

The necessity to adhere to a non-linear formula for obtaining the THM FP concentration is due to a clear limitation in generating large amounts of information on THM FP experimentally. These experiments are time-consuming and expensive, and technical expertise is also required because of an inherent inaccuracy. Additionally, calibrating an empirical model to predict the THM FP requires many different sensors. For instance, the calibration of Amy's or Godó-Pla's models needed at least four sensors to provide THM FP values. This thesis chapter aims to develop a mathematical model capable of real-time THM FP prediction with a minimum number of sensors required.

3.3.3. Model development

MLR and ANN algorithms were used to develop the mathematical models to predict THM FP. The seven variables selected from $PreO_3$ and EB1 sensors were used.

- <u>Multivariate Linear Regression model</u>

The structure of an MLR model is based on the independent variables (wavelengths selected previously) against the response variable (THM FP from Amy's formula). The

MLR results are presented in Table C3.3. The first step in interpreting the multiple regression is observing which predictors are significant. The last column (p-value by predictor) shows a larger significance for all predictors since the values are near 0. As can also be observed by the asterisks, which are the *Significance Codes* related to each estimate, all the predictors chosen are directly associated with the response variable. Three asterisks represent a highly significant p-value).

Table C3.3 – *Information obtained from the MLR model performed.*

		Estimate	Std. Error	t value	Pr (>|t|)
	Intercept	43.19	0.21	208.3	<2e^{-16}
PreO$_3$	**λ235**	0.13	0.01	19.9	<2e^{-16}
	λ265	-1.30	0.03	-45.56	<2e^{-16}
	λ395	1.16	0.03	39.42	<2e^{-16}
EB1	**λ300**	21.11	0.31	67.60	<2e^{-16}
	λ400	-6.49	0.25	-25.75	<2e^{-16}
	λ235	1.44	0.04	34.20	<2e^{-16}
	λ255	-13.02	0.18	-73.15	<2e^{-16}
		RMSE	1.65	**Adj. R^2**	0.30

The column with the estimations informs about the β-coefficients for each predictor, and for our case study, all values are significantly different from 0. Assuming all other predictors remain constant, the B-coefficient for each given predictor can be interpreted as the effect of a one-unit increase in the response variable. In this case study, variables such as λ255 and λ300 from the EB1 sensor have more weight in the THM FP prediction than λ235 or λ395 from the PreO$_3$ sensor; thus, a slight change in the absorbance of those wavelengths has a significant impact on the final THM FP observe. The model, however, is not of a high level of quality, exhibiting a coefficient of determination (R^2) of 0.30. R^2 represents the proportion of variance explained by the independent variables selected (James et al., 2021). An R^2 close to 0 indicates that the model is unable to explain a substantial proportion of the outcome and, as a result, is unable to make accurate predictions.

- Artificial Neural Network model

The architecture of a Neural Network is based on layers. Seven neurons comprise the input layer of this case study, corresponding to the selected wavelengths (Figure C3.14). Optimal hidden layers and neurons were determined using a self-made algorithm (Table C3.4). The algorithm calculates the determination coefficient of ANN models with up to two layers and ten neurons for each layer.

Table C3.4 exhibits the results obtained for two layers and ten neurons. Two layers in an ANN model are the extended criterion to develop a proper network since more complex models are intended to fail into overtraining. It is also essential the number of neurons for each layer. The final selection for the ANN model was two layers with five neurons per layer (light green). An ANN model pursues the best predictive accuracy. However, the maximisation of training by selecting a great number of neurons could finish in a non-useful model, as rather than *learning* the inherent behaviour of the wavelengths, it may *memorise* their unimportant trends (i.e., noise) (Bilbao et al., 2017; de Sá et al., 2017). As shown in Table C3.4, the more neurons in a layer, the better the coefficient of determination (see, for instance, the 10-10 neurons model, with a determination coefficient of 0.96).

Nevertheless, complex sizing not necessarily means better results. The final result was contrasted with those with lesser and higher determination coefficients (light red), comparing their prediction capacity with data the model had never seen before. The results were more liable to overfit, producing predictions with bigger errors in outside data, in models with more than five neurons per layer. In contrast, the capacity prediction of ANN models with fewer neurons did not consider the key patterns to predict correct outcomes in outside data or central data; it is important to avoid undertraining the model as well as overtraining it.

Table C3.4 – *Tests performed to select the best architecture for the Artificial Neural Network model using the self-made algorithm. The more suitable combination was two layers with five neurons per layer.*

		Layer 2									
		N1	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**
Layer 1	**N1**	0.01	0.50	0.64	0.66	0.88	0.78	0.78	0.03	0.88	0.82
	N2	0.32	0.75	0.72	0.63	0.87	0.83	0.52	0.09	0.89	0.83
	N3	0.54	0.84	0.78	0.76	0.88	0.83	0.77	0.81	0.87	0.86
	N4	0.76	0.85	0.88	0.79	0.90	0.84	0.88	0.89	0.89	0.89
	N5	0.80	0.88	0.87	0.81	0.92	0.85	0.89	0.90	0.90	0.89
	N6	0.81	0.89	0.87	0.82	0.90	0.84	0.89	0.93	0.92	0.92
	N7	0.81	0.88	0.82	0.81	0.90	0.85	0.91	0.96	0.93	0.97
	N8	0.79	0.89	0.88	0.85	0.88	0.87	0.92	0.93	0.92	0.94
	N9	0.82	0.89	0.89	0.85	0.89	0.85	0.93	0.96	0.92	0.91
	N10	0.81	0.87	0.85	0.87	0.90	0.85	0.90	0.95	0.95	0.96

Figure C3.14 exhibits the final architecture for the ANN model. As can be seen, the information of each wavelength was distributed to the five neurons of the first layer. The Sigmoid Function, also called *Activation Function* (5), received the information from each input neuron). Its application transformed a linear response into a non-linear one. All this process was repeated for each neuron. The Activation Function was used as a threshold; if the final calculation for each neuron reached the threshold, the neuron was activated, sending the information to the next layer, and starting the calculations again.

The output layer provided a THM FP result once all the calculations were completed in the network. All the computations and mathematics involved in an Artificial Neural Network have an inherent difficulty since it works as a black box (Malekian et al., 2021). The complexity of understanding all the information transmission between layers and neurons restrained the comprehension of where the final results were derived from. However, a detailed and understandable-friendly description of neural networks' operation, calculations, and application can be found in Walczak et al., 2003 and Malekian et al., 2021.

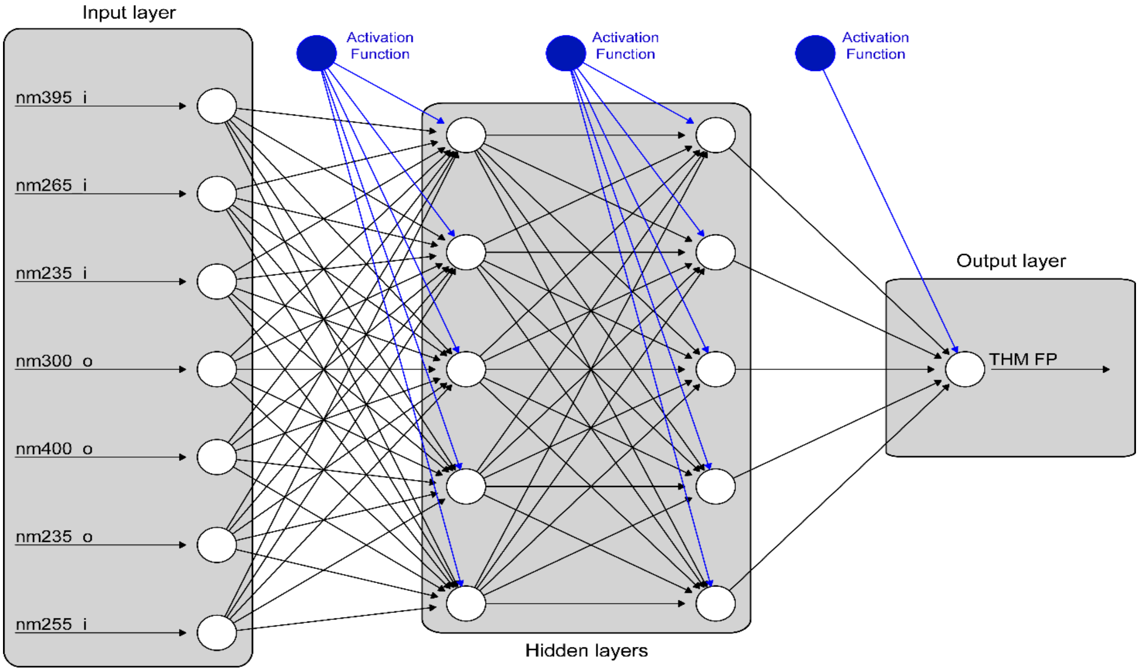


***Figure C3. 14** – The wavelengths selected were at the Input layer, and the THM FP was at the Output layer. This architecture was formed by two layers and five neurons per layer.*

3.3.4. THM FP prediction capacity using MLR and ANN

After developing the MLR and ANN models, the coefficient of determination (R^2), the correlation between predictions and *actual* values (R_{Test} and R_{Val}), and the RMSE were examined.

Table C3.5 exhibits these results. The determination coefficient obtained in the training set (R^2) of the ANN model (0.92) was much higher than that of the MLR (0.30). Additionally, RMSE was lower for ANN (0.77) compared to MLR (1.65). The correlation between testing values and those from Amy's formula (R_{Test}) gave a weak correlation for

MLR (0.32) and a stronger one for the ANN model (0.89). According to the correlation between the validation results and those from Amy's formula (R_{Val}) gave a worsening for the ANN model (0.72), and it followed the same trend as the previous results for the MLR model (0.35).

Table C3.5 – *R^2, RMSE and correlation between tested and validation values versus those obtained using Amy's formula. The ANN model is the one that obtains the best results (0.92, 0.77, 0.89 and 0.72, respectively). On the other hand, MLR does not fit (0.30, 1.65, 0.32 and 0.35, respectively).*

	MLR	**ANN**
R^2	0.30	0.92
RMSE	1.65	0.77
R^2_{Test}	0.32	0.89
R^2_{Val}	0.35	0.72

Figure C3.15 shows the MLR prediction capacity. The training data (red dots) was distributed horizontally, away from the diagonal line. Testing data (blue dots) overlapped with training data, following a similar trend. That occurs when data is randomly selected to create the training and testing datasets. The patterns that the model trains were the same as those it predicted. That means the model failed to *filter* the information that causes a change in THMs FP concentration.

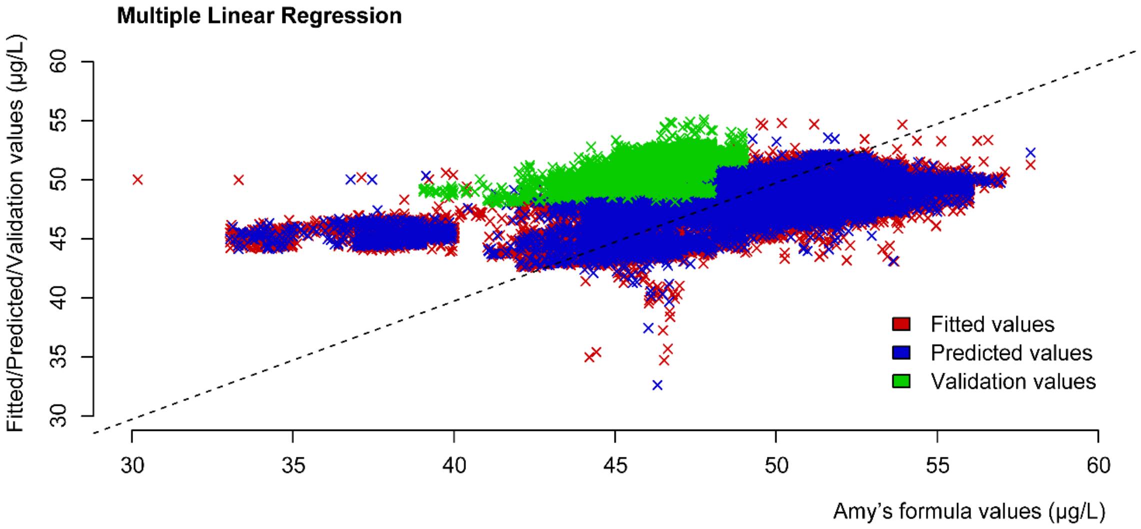


Figure C3.15 – *Correlation results of the MLR model to predict THM FP (µg/L). The validation data give a reliable idea of the low inference capacity of the model.*

It is also noteworthy that the MLR model exhibited inaccuracies when predicting data that it had never trained before (green dots). The validation data had a large vertical scattering tendency and was centred. That would happen because this model trains the core data (THM FP from 45 to 50 µg/L) more effectively and produces inaccuracies predicting the values on the edges. The THM FP values commonly recorded in the plant were those between 45 and 50 µg/L, which can create a bias in predicting values found at the edges. It is also noted that all validation points were located in the upper area of the diagonal. The validation data were therefore overestimated, and the THM FP predictions were higher than they should have been.

Figure C3.16 exhibits the performance of the ANN model. As can be seen, a large part of the training values (red dots) followed the diagonal line, which presupposes a proper linear distribution. The testing data (blue dots) was superimposed on the training data, which means the model has correctly acquired the patterns mimicking them in the predictions.

On the other hand, the ANN model was considerably better at predicting data it had never seen before. Validation data (green dots) were close to the regressed diagonal line with a greater uniform dispersion than the MLRs. This concludes that ANN not only had stronger core values (THM FP from 45 to 50 µg/L) inference but also extreme (values with a THM FP concentration less than 40 µg/L and higher than 50 µg/L) than MLR. The validation cloud in Figure C3.16 was also shown in the upper area of the diagonal.

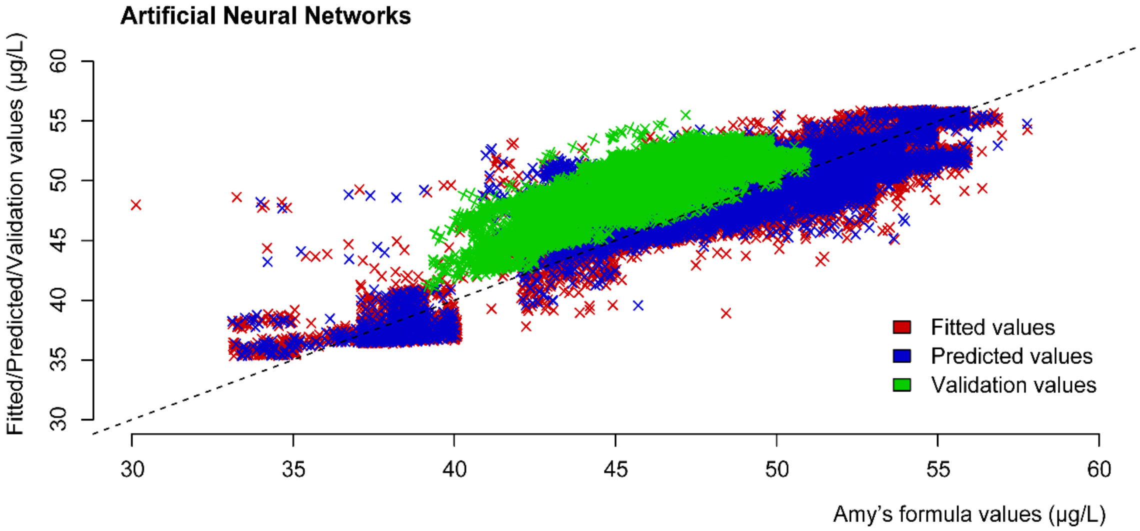


Figure C3.16 – Correlation results of the ANN model to predict THM FP (µg/L). The validation data give a reliable idea of the proper inference capacity of the model.

The factors that drive a change in THM FP seem to be correctly interpreted by the ANN model since its inference capacity for non-trained data had high accuracy. There is an overestimation of validation values (≈ 5 µg/L) which is similar to the results from the MLR model (Figure C3.15). However, the offset of the MLR model was higher due to the coefficient of determination obtained (0.30) and resulting from that, the error dragged was also higher.

The differences in inference capacity between both models could be related to the severe multicollinearity described before. As mentioned previously, the MLR model exhibits strong multicollinearity at some wavelengths (Table C3.2). As more wavelengths were selected, both the R^2 and the RMSE increased rapidly, as did the VIF. Whenever all the variables with elevated VIFs were removed, the R^2 and RMSE of the model drastically worsen, making it impossible to obtain good predictive performance for the linear model.

This severe multicollinearity, which significantly impacted the MLR model, had little to no effect on the ANN model. ANN approached the problem of multicollinearity in a very different way. According to the vast parameterisation, a neural network's coefficients (or weights) are inherently difficult to interpret. However, this same redundancy makes individual weights unimportant. It means that at each level of the network, the inputs are linear combinations of the inputs of the previous level.

The final output is a function of many sigmoid combinations involving high-order interactions between the original variables, transforming the initial linear relationships into non-linear ones (Barron et al., 1992; De Veaux et al., 1994). Thus, the ANN model developed was not so influenced by the effect of inherent multicollinearity of spectrophotometric data. By its physicochemical nature, it always will have a strong degree of collinearity, as one wavelength shares a part of information with its contiguous. Although this was a positive factor for predictive results, it negatively impacts the interpretability of the model (Walczak et al., 2003 Malekian et al., 2021).

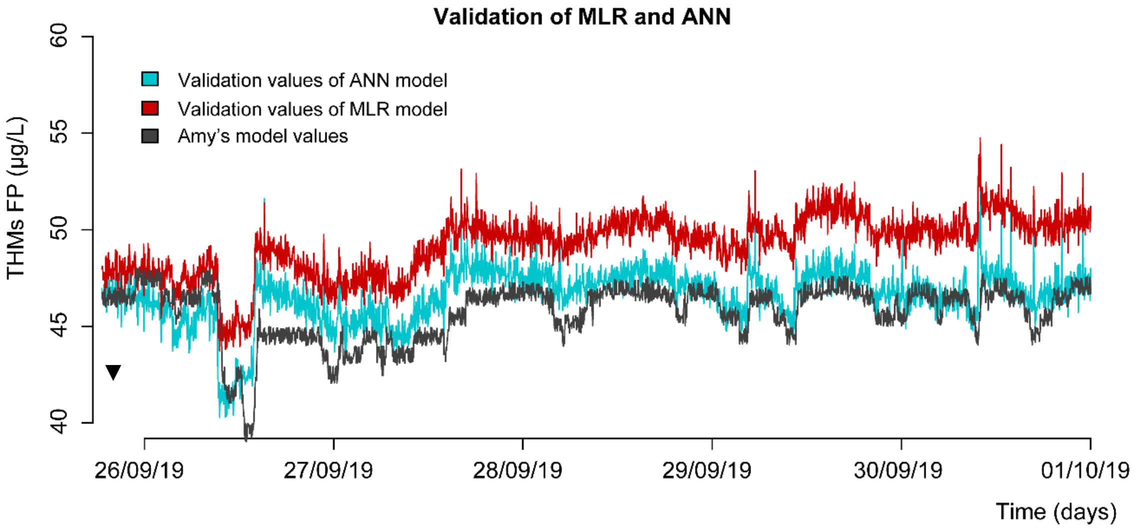


Figure C3.17 – *MLR and ANN validation results over time.*

Figure C3.17 displays the validation versus Amy's formula data over time. We randomly selected six days and compared the results for both models. The two models could predict data patterns correctly. However, the inference capacity of ANN was clearly superior. Although it is less for ANN, there was a tendency for both models to overestimate the THM FP. As observed in Figures C3.14 and C3.15, MLR produced more than 20% inaccuracy and failed to accurately predict extreme data (black arrow). In contrast, ANN inferred better extreme values, such as abrupt changes in the trend.

3.4. Conclusions

The development of a machine learning model could lead to an improvement in the control of THM FP at the outlet of Drinking Water Treatment Plants (DWTP). Particularly, it can become an essential tool for operators in their daily work.

It was proven that it is necessary to install at least two spectro::lyser®, one at the beginning of the treatment train (before the pre-ozonization system) and the other at the end (final chlorination).

spectro::lyser® are an effective instrument for controlling water quality. However, creating a model using the total spectrum obtained from them has some mathematical requirements:

i. It is necessary to perform an initial automatic variable selection to trim the ineffectual variables from the useful ones. The most suitable method to achieve it is the Stepwise Selection motor compared to the Forward Selection and Backward Elimination selection methods.

ii. Wavelengths share a critical part of their own information with the contiguous, generating severe multicollinearity, directly impacting the mathematical model. Once the critical variables are automatically selected, a manual pruning must be performed to decrease the Variance Inflation Factor. Finally, from a total of 751 wavelengths, only 7 remained (3 from the PreO$_3$ spectro::lyser® and 4 from the EB1 spectro::lyser®).

After selecting the variables, MLR and ANN models were developed, and their prediction capacity was compared. MLR has extreme difficulties inferring central (45 to 50 µg/L) and extreme data (30 and 60 µg/L). In other words, the MLR model is, in essence, ineffective at capturing what causes changes in THM FP. The ANN model, on the other hand, can accurately predict the central values of THM FP and the extremes, providing findings that are more reliable and robust than those provided by the MLR.

The incapacity of MLR to predict proper values of THM FP may be directly related to the severe multicollinearity that some wavelengths dragged; therefore, it is concluded that ANN is the best way to predict THM FP because it is unaffected by this problem.

Capítol 4 – *Multivariate Linear Regression approach to predict hydrocarbon mixtures in urban wastewater matrices using spectrophotometric sensors*

Abstract

Correct management of wastewater treatment plants (WWTPs) requires regular monitoring of the quality of the wastewater and the detection of pollutant leaks. Hydrocarbons, such as benzene (B), toluene (T), ethylbenzene (E) and Xylene isomers (X), are increasingly detected in wastewater matrices. Portable probe spectro::lyser® is a reliable probe to monitor wastewater quality parameters through algorithms, such as Total Organic Carbon (TOC) or Dissolved Organic Carbon (DOC). A BTEX real-time quantification in wastewater matrices cannot, however, be provided by any suitable algorithm. In order to demonstrate the capability of cutting-edge statistical inference tools, such as Multivariate Linear Regression (MLR), for detecting and predicting BTEX pollution in wastewater matrices, we gathered spectral data from three urban wastewater influents. The three models developed exhibited a significant Pearson correlation (R^2) coefficient (0.82, 0.87 and 0.79, respectively) and low Mean Squared Error (MSE) values (0.27, 0.25 and 0.35, respectively). They still showed some accuracy differences in their prediction abilities. These differences could be due to variations in wastewater properties, such as the type of wastewater, how it is processed, or its seasonality, among other factors, which would cause instability in the models' capacity for inference. Our results demonstrate how MLR models can accurately predict the presence of low concentration BTEX in urban wastewater matrices using the absorbance of spectro::lyser® probes. The linear models created can be powerful tools for WWTP operators to respond quickly if there is occasional pollution of BTEX.

4.1. Introduction

The presence of benzene (B), toluene (T), ethylbenzene (E) and xylene (X) isomers (BTEX) in urban wastewaters is increasingly common (Zoccolillo et al., 2001; Lima et al., 2011). The considerable risk to human health and the environment has been highlighted by tighter legislation. According to Water Framework Directive 2000/60/EC of the European Comission, the maximum allowed concentrations of BTEX in drinking water (10, 50, 30 and 30 µg/L, respectively) dramatically differ from those targeted by United States Primary Drinking Water Standards (5, 1000, 700, 10,000 µg/L) (US EPA, 2010) and by World Health Organization (10, 70, 300, 500 µg/L), (WHO, 2011). These allowable limits increase the population's exposure to BTEX-contaminated freshwater.

Researchers have been examining the effects of these substances on human health for many years, both in the air and water. They can be detected in rainwater, surface water, soils, sediments, drinking water, wastewater, and aquatic organisms. Long-term exposure to these substances can have negative consequences on the kidneys and the central nervous system, including headaches, fatigue, tremors, incoordination, and dizziness (WHO, 2011; US EPA, 2010).

Wastewater Treatment Plants (WWTP) need to closely monitor the quality of the water entering them since more and more industrial streams, spills, and leaks are introducing toxins like BTEX (Korshin et al., 2017; Douglas et al., 2018; Richardson et al., 2018). The quality control of the influent of a WWTP is generally assessed using physical, chemical, and microbiological analysis, monitoring biochemical oxygen demand (BOD), chemical oxygen demand (COD) and total organic carbon (TOC), among other critical parameters (Langergraber et al.,2003; Gruber et al., 2005). The greatest benefit of these tests is that a detailed blueprint of wastewater quality is obtained. However, there are some disadvantages to laboratory analysis as well. They depend on expensive and time-consuming methods and lack real-time knowledge if a relevant contamination event is detected. This fact restrains an early response from plant operators if needed. (Gruber et al., 2005; Hue et al., 2013; Chong et al., 2013; Yang et al., 2015; Carstea et al., 2016).

For real-time monitoring of influent quality in WWTPs, online water quality probes and sensors, such as the spectro::lyser® s::can spectrophotometer, are becoming more and more popular. Ultraviolet (UV) spectrophotometry is the release of a controlled beam of light passing through a liquid, solid or gas, measuring its absorbance (the negative algorithm of the transmittance, defined as the fraction of light transmitted by the sample when it is irradiated). Plotting the UV spectrum (from 190 nm to 400 nm) against absorbance (Abs/m), the response is typically measured as a function of radiation wavelength (Langergraber et al.,2003; Gruber et al., 2005; Carstea et al., 2016; Korshin et al., 2017; Richardson et al., 2018).

Water quality parameters including DOC, TOC, nitrates, and turbidity, among others, can generally be created by combining certain wavelengths. The capacity to find combinations of wavelengths will enable the creation of new parameters for water quality control, which makes this kind of probe particularly interesting. However, there is no direct combination of wavelengths for all pollutants. Wastewater exhibits strong absorbances (200 to 400 Abs/m) at the first portion of the UV spectrum as a result of its high organic and inorganic matter content (approximately 190 to 400 nm). B, T, X, and E compounds also absorb in the UV range, but with absorbances much lower than those from wastewater (i.e., 0.5 ppm of toluene would be equivalent to 2.10 Abs/m in the wavelength 220 nm), which is entirely covered by the matrix (Norman et al., 1941; Berlman et al., 1971; Quina et al., 1976; Hastie et al., 1992; Du et al., 1998; Karlowatz et al., 2004; Dixon et al., 2005). Therefore, s::can online monitoring spectrophotometer can provide WWTP with effective early detection of several compounds. While low levels of BTEX may be successfully detected in deionized and drinking water, its detection in a wastewater matrix is still challenging.

Multivariate statistical techniques are valuable tools that are increasingly used in understanding environmental pollution issues when combined with Machine Learning techniques (Venables et al., 2002; Rencher et al., 2012; Kuhn & Johnson, 2013). They are used in a variety of contexts, such as forecasting or prediction, mathematical modelling, and statistics. Multivariate Linear Regression (MLR) is a powerful tool that provides detailed information about the hidden relationship between variables, reducing the dimensionality of complex datasets.

The application of MLR models for environmental pollutant monitoring is increasingly common (Cho et al., 2022). Their inherent capabilities make them a flexible and robust tool for identifying possible pollution sources in water. Implementing a well-trained MLR model could facilitate the detection of occasional BTEX discharges, offering a valuable tool for WWTP operators (Putri et al., 2018; Cho et al., 2022). A model that can forecast contaminants in wastewater, such as BTEX, might be created by combining the advanced statistical inference methods stated previously with the data from spectro::lyser®. However, the variability of the wastewater matrices and the low absorbance of those pollutants are significant drawbacks for detecting their presence in wastewater matrices.

For this reason, this research's main objective was to develop a reliable methodology to develop site-specific MLR models for detecting and predicting BTEX in different urban influent wastewater matrices. In order to accomplish our goal, a number of datasets of wastewater sources that were collected using spectro::lyser® probes were taken into account and examined. The *in-silico* spiking of BTEX was made by applying a 3-level factorial design. Once the MLR models were created, an exhaustive examination of their suitability was carried out using several statistical metrics. These models could become an essential tool for providing WWTP operators with an early response to incidental BTEX contamination in the influent waters of WWTPs.

4.2. Materials and Methods

4.2.1. Wastewater databases

Databases for this study were collected from various projects created by s::can Iberia Sistemas de Medición, SLU, between 2014 and 2019. The databases were created using urban influent data from different Iberian Peninsula WWTPs.

Between 7,000 and 50,000 m^3 of urban wastewater are treated daily at these WWTPs. The sanitary systems extend from the sewer networks to the WWTP. The wastewater is treated using several procedures in order to prepare it for reuse in urban settings.

The quality of each WWTP's influent was measured using spectro::lyser®. The quality parameters TOC and DOC, as well as the entire UV-Vis spectrum, were recorded by those probes (from 220 nm to 730 nm, every 2.5 nm). Every two minutes, absorbances were recorded.

4.2.2. Spectral characteristics of BTEX

Organic substances known as BTEX have one benzene ring with either none, one, or two methyl groups attached, as well as ethylbenzene having one ethyl group attached (Figure C4.1). In the wavelength range of 200 – 290 nm, the absorbance of these substances was observed by Jones, et al. 2013 and Khan, 2021, among other authors.

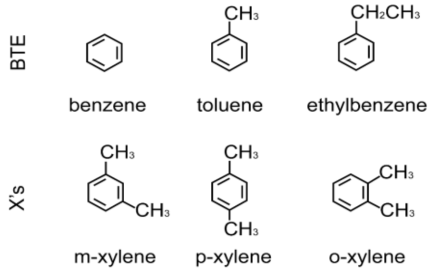


Figure C4.1 *– Structures of Benzene, toluene, ethylbenzene and xylene isomers (BTEX). Source: Adapted from Montero-Montoya et al., 2018.*

Figure C4.2 shows the extinction coefficients of ethylbenzene (blue line), m-xylene (orange line), p-xylene (grey line) and toluene (yellow line). They have the highest extinction coefficient at wavelengths before 220 nm, but there is also an increase in the wavelength range from 250 to 280 nm, with the highest peak between wavelengths 260 and 265 nm (Jones, 1941).

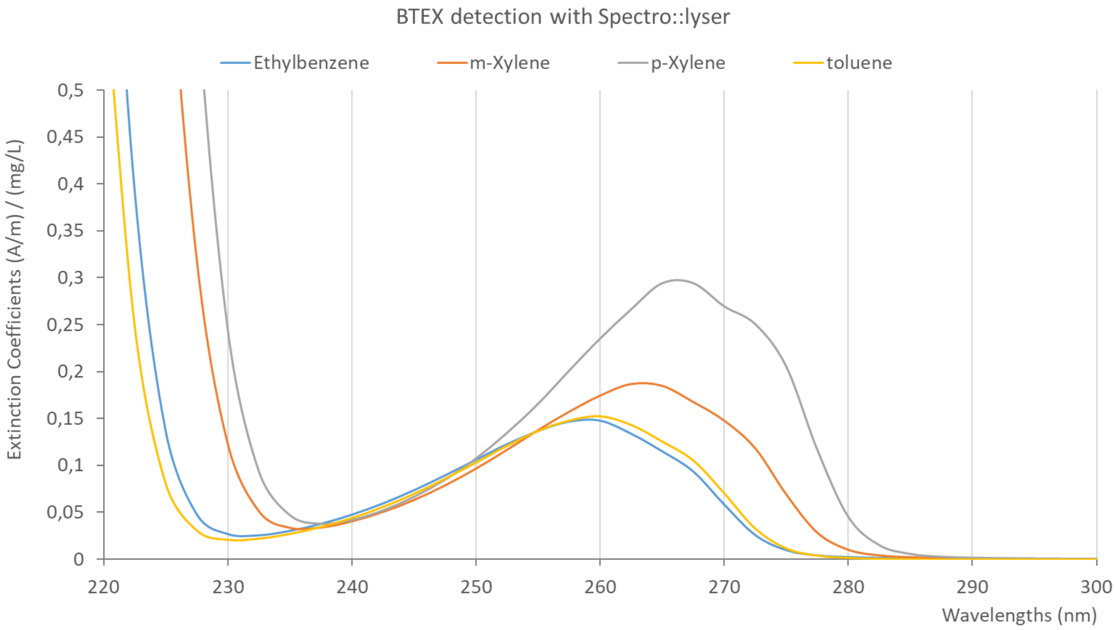


***Figure C4.2 –** Extinction coefficient ((A/m)/(mg/L) of Ethylbenzene (blue), m-xylene (orange), p-xylene (grey) and toluene (yellow) in distilled water recorded with spectro::lyser® UV-Vis.*

4.2.3. Three–level factorial design

Toluene, m-xylene, and p-xylene mixture concentrations in the wastewater matrices must be determined in order to train the mathematical prediction models. No hydrocarbon contamination was in any spectral data collected from s::can projects. As a result, a three-level factorial design was developed as a mathematical spiking method (Table C4.1).

For instance (Table C4.1), the first spike consisted of three spectral concentration levels for each contaminant (0.0, 0.125 and 0.250 ppm). Toluene, m-xylene, and p-xylene discrete absorbance sums were added to a wastewater spectrum for every last k-factor (27, to be exact). This led to 27 different known combinations and concentration ratios of toluene, m-xylene, and p-xylene for 27 different observations of wastewater.

Table C4.1 – *Three-level factorial design to control the spectral addition of toluene, m-xylene, and p-xylene have different combinations and concentration ratios.*

Experiment	Toluene (ppm)	ρ–xylene (ppm)	m–xylene (ppm)	Total (ppm)
1	0.00	0.00	0.00	0.00
2	0.00	0.00	0.125	0.125
3	0.00	0.00	0.250	0.250
4	0.00	0.125	0.00	0.125
5	0.00	0.125	0.125	0.250
6	0.00	0.125	0.250	0.375
7	0.00	0.250	0.00	0.250
8	0.00	0.250	0.125	0.375
9	0.00	0.250	0.250	0.500
10	0.125	0.00	0.00	0.125
11	0.125	0.00	0.125	0.250
12	0.125	0.00	0.250	0.375
13	0.125	0.125	0.00	0.250
14	0.125	0.125	0.125	0.375
15	0.125	0.125	0.250	0.500
16	0.125	0.250	0.00	0.375
17	0.125	0.250	0.125	0.500
18	0.125	0.250	0.250	0.625
19	0.250	0.00	0.00	0.250
20	0.250	0.00	0.125	0.375
21	0.250	0.00	0.250	0.500
22	0.250	0.125	0.00	0.375
23	0.250	0.125	0.125	0.500
24	0.250	0.125	0.250	0.625
25	0.250	0.250	0.00	0.500
26	0.250	0.250	0.125	0.625
27	0.250	0.250	0.250	0.750

In this mathematical spike, the total concentrations achieved for each experiment ranged from 0.0 ppm (any contaminant was added) to 0.750 ppm (all three contaminants were added at a maximum concentration of 0.250 ppm apiece).

This method was used to get a known concentration of toluene, m-xylene, and p-xylene mixture for each observation of the datasets for all urban wastewater spectrums acquired from s::can projects.

4.2.4. Data structure

Figure C4.3 shows a diagram of the sequences performed to create the models. The first step was to collect spectral data on the influents of urban wastewater. More than twenty databases were examined from projects executed by s::can Iberia Sistemas de Medición, SLU. However, only three databases were ultimately chosen for modelling because of their quality (i.e., the quantity of data, connectivity, precision, and correctness).

The columns represented each spectrum wavelength (from 220 to 727.5 nm, at intervals of 2.5 nm). The rows represented each sample that the probes recorded every two minutes. Once the general database for each site was created, a controlled spectrum mixture of toluene, m-xylene and p-xylene was added mathematically to the general database (detailed in the previous section). The concentration of toluene, p-xylene, and m-xylene, obtained from applying the three-level factorial design, was selected as a response variable (0.0 to 2.5 ppm). This entire piece was created using RStudio (Version 1.2.5033).

After the mathematical spikes, the databases were thoroughly examined, and incorrect data, zeros, and extreme outliers were removed. After cleaning and pruning, all features underwent normalisation (centering and scaling). The process of normalisation is essential. The final modelling and the posterior variable selection may be affected as a result of the variability of the wavelength weights. This step was made using the function *preProcess* of the *CARET* library. With the use of this function, the *center* and *scale*, or *nzv*, normalization method could be chosen. Finally, there were 8,052 observations from database 1, 13,883 observations from database 2, and 20,827 observations from database 3.

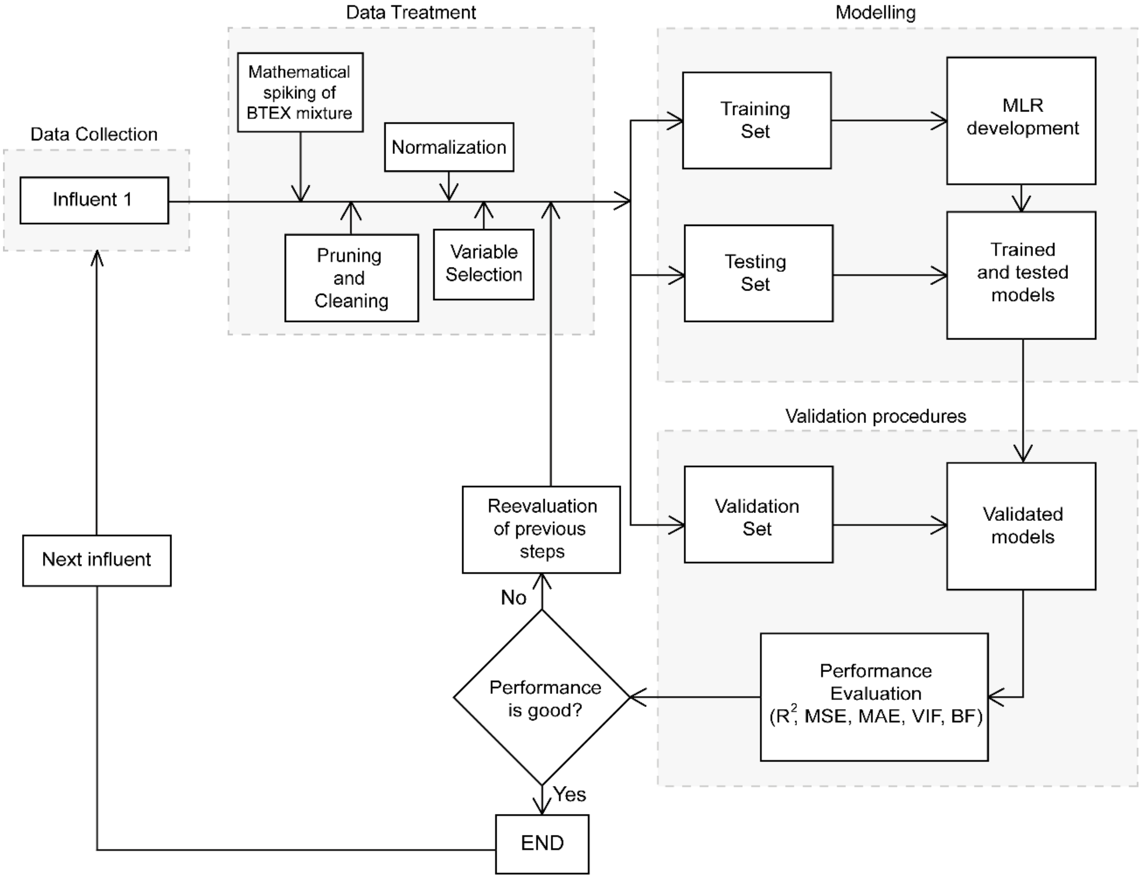


Figure C4.3 *– Flowsheet outlining the whole model development. Data collected was normalised and divided into three datasets: training, testing, and validation sets. Several indicators of performance determined the goodness of the models.*

Variable selection was performed using Stepwise Selection (SS) from the *Stats* library. The function used was *step* with the particular command *both*. After the feature selection, a data partition was applied to create the training, testing, and validation datasets. The validation dataset was first created by separating the last 1,000 rows from each general database. The training and testing sets were created using the *CARET* library (Kuhn, 2013). This library has a function called *createDataPartition*, in which a controlled splitting can be applied. In this case, 75% of the remaining data were for the training sets (5,289; 10,412 and 15,620 rows, respectively), and the remaining portion was for the testing sets (1,763; 3,471 and 5,207 rows, respectively). All rows before the divisions were evenly randomized using this splitting function.

4.2.5. Variable selection methods

A model simplifies reality to promote a clear understanding of the problems we want to address. Therefore, the *art* of model building involves simplifying reality to help us understand the issue at hand. A variable selection method helps identify the best subset of predictors in a high–dimensional dataset. Therefore, this method can choose the most explanatory variables and could be an attempt to find the most basic MLR model we discussed earlier.

In this chapter, the Stepwise Selection (SS) method was tested and applied. Its application was mandatory as running a regression model with many irrelevant variables would lead to a needlessly complex model. Employing a selection method was a way of selecting essential variables to get a simple and easily interpretable model (Thayer et al., 1990; Thomas et al., 1998; Kuhn, 2013).

SS is a method that begins with a model without variables (called the *null model*). It is an iterative, step-by-step model construction method that alternates between adding or removing potential explanatory variables and testing each conducted subset's stability and statistical significance. (For more information, see chapter 1). The *Stats* library's function *both* was used to do SS calculations. A selection motor uses different stopping rules as a concluding threshold. These vary depending on the requirements for each situation. For our case study, the stopping rule selected was a 0.05 *p*-value threshold. Variables with a p-value larger than 0.05 were immediately discarded and only were selected those with a lower p-value improved the general performance of the model.

As said before, the SS algorithm added or removed a wavelength at each iteration. Therefore, adding or removing a variable was considered a new model, with a new R^2, RMSE and MAE. The Akaike's Information Criterion (AIC) (Akaike, 1978; 1979) was applied to select the best model. AIC is a measure of goodness-of-fit because it explains the connection between bias and variance in a model. According to their AIC, several candidate models can be classified given a dataset. The smaller this value, the better the model is. In general, the AIC (F4.1) is defined as:

$$AIC = 2k - 2\ ln\ ln\ (L) \qquad \text{(F4.1)}$$

Where *k* was the number of parameters in the model and *ln(L)* was the log-likelihood function for the statistical model. Box et al., 2008 provided an excellent mathematical explanation of this metric and its applications in several scientific fields.

The SS method was applied for each acquired database corresponding to the three cases studied. SS method selected the essential wavelengths for each database.

In addition to the automated feature selection, which provides us with flexibility and automation, a manual final feature adjustment was performed while taking the Variance Inflation Factor (VIF) and the Breusch-Pagan test into account (BP). These parameters were applied to control the impact of multicollinearity, heightened in models in which continuous, interrelated variables were present.

Neter, Wasserman and Kutner (2004) contemplated a series of indicators to analyse the degree of multicollinearity between the independent variables of a mathematical model. One of them was VIF (F4.2), defined as:

$$VIF_i = \frac{1}{1 - R_i^2} \qquad \text{(F4.2)}$$

Where R^2_j was the determination coefficient of the variable X_j over the rest of the selected explanatory variables. The value of VIF had to be between 1 and infinite ($1 \leq VIF_j \leq \infty$). There was a direct relationship between VIF_j and the variable's estimated determination coefficient variance. The value of VIF_j would increase as the collinearity increased, bringing R^2_j closer to 1. Due to their high collinearity, which would make it difficult to predict future values, it was decided that the variables chosen with a VIF>10 needed to be reviewed.

Breusch and Pagan (1979) developed the BP test (F4.3) to determine the heteroscedasticity of MLR models. It analyses whether the estimated variance of the residuals of a regression, which Ordinary Least Squares estimates at 0, depends linearly on the values of the independent variables. Generally, the BP test is based on the estimation of the equation:

$$\hat{e}_j^2 = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \cdots + \alpha_p X_{ik} + u_i \quad i = 1, \ldots, n \qquad \text{(F4.3)}$$

Where $\hat{e}_j^2$ is the response variable given by residuals from the original model squared, $X_{i1}, X_{i2} \cdots X_{ik}$ are the independent variables of the model and $\alpha_0, \alpha_1 \cdots \alpha_p$ are the residuals of the original model. If it is concluded that $\alpha_0, \alpha_1 \cdots \alpha_p = 0$, it means that the residuals are not a function of the covariates of the model.

Together, these tests provided an estimate of the model's performance, considering the influence of the variables selected and the potential predictions they could make.

4.2.6. Model Development

- Multivariate Linear Regression (MLR)

A Multivariate Linear Regression model (MLR) is a simple correlation between two or more independent variables (also called *dimensions* or *features* because they are not entirely independent) with a dependent variable.

To develop the modelling of this research, three MLR models were performed to quantify the concentration of toluene, m–xylene and ρ–xylene as a mixture in three different wastewater matrices.

The general linear regression model (F4.4) took the form of:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ik} + \varepsilon_i \quad i = 1, \ldots, \tag{F4.4}$$

Where y was the response (or dependent) variable vector. $\beta_0, \beta_1, \ldots, \beta_p$ was a (p+1) dimensional parameter vector of regression coefficients that must be estimated for the sample dataset. β_0 was the intercept term. $X_{i1}, X_{i2}, \ldots X_{ik}$ were the features of the model, and ε was the vector of the error term (or noise) of the model. An insightful explanation of MLR mathematics can be found, for example, in Geladi et al., 2003 and Kumar et al., 2014.

The response variable was the total concentration of toluene, m-xylene, and p-xylene mixture at each observation (row) added mathematically throughout the three-level factorial design. All the wavelengths of the UV-Vis spectrum registered with spectro::lyser® were the predictors of the model (204 variables). The *Stats* library's *lm* function was used to create the MLR model. The functions *plot*, *points*, and *abline* from the *Stats* library were used for data visualisation.

Several assumptions must be considered to develop a good MLR model: linearity, independence, homoscedasticity, and normality. The primary is that the dependent variable should exhibit a *linear relationship* between the features selected. The second assumption is that the data values must not display *multicollinearity*; in other words, the explanatory variables cannot be highly correlated with each other. The selection of variables that positively contribute to the modelling can be challenging due to multicollinearity because they all seem to be necessary, but only a few actually are (Wold et al., 1984). This assumption states that the variance of error terms is similar across the values of the independent variables. *Normality* is an assumption applied only to the residuals. The model's residuals (or error) are the difference between the observed and predicted values of the dependent variable, which should be normally distributed.

The *plot* function applied in an MLR model has the peculiarity that gives three plots for measuring the performance of the least-squares regression model and checking the MLR model assumptions explained above. These plots were fitted VS residuals, the normal quantile-quantile (Q-Q), and the residuals VS leverage. These three plots gave an essential overview of model execution.

Fitted VS residual plot: The fitted VS residual plot investigates the linearity and homoskedasticity assumptions. The residuals must maintain a horizontal pattern with values tending to 0, and their spread must be approximately the same, following an imaginary line across the abscissa axis.

Quantile-Quantile plot (Q-Q plot): The Q-Q plot helps to assess the normality assumption of the actual distribution of a dataset. This plot is a scatterplot of two sets of quantiles. If they came from the same distribution, the scatterplot would show a diagonal roughly straight, fulfilling the linearity assumption.

Residuals VS Leverage plot: The residuals VS leverage plot helps in identifying observations that have a big impact on how well the model fits throughout the Cook's distance. Influential data that exceeds the threshold (Cook's distance is greater than 1) needs to be examined and likely discarded.

4.2.7. Goodness of the models

The error of the MLR modelling is a critical consideration, as it will determine the model's goodness. Fitting a linear model requires estimating the regression coefficients with the lowest error added (F4.5). The error expression took the form of:

$$\varepsilon = y - X\beta \tag{F4.5}$$

Where X was an *n x 2* matrix, y was an *n x 1* column vector, β was a *2 x 1* column vector, and ε was an *n x 1* column vector. This simple statement (2) was the matrix formulation of the MLR model. This notation was used because it was more efficient in a case with many predictors.

The model performance was evaluated throughout the determination coefficient (R^2) and the error term. R^2 is a statistical measure representing the proportion of the variance in the response variable that several predictors can explain in an MLR model. The determination coefficient is a measure given by the *lm* function of the *Stats* library when an MLR model is calculated.

Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) analysis were essential evaluation steps in any machine learning model. The model performs better the lower the error. In the statements below, *p* denotes the predicted values and a is related to actual values. The RMSE (F4.6) is the standard deviation of the residual (prediction errors). It measures the error rate of a regression model; in other words, it measures how spread out the model's residuals are around the line of best fit (diagonal). The evaluation of RMSE is a feature given by the *lm* function of the *Stats* library.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \ (p_i - a_i)^2}{n}} \tag{F4.6}$$

The MAE (F4.7) represents the average of the absolute difference between de actual and the predicted values in the dataset. It measures the accuracy for continuous variables and is a standard method to evaluate the goodness of a regression. The *Metrics* library's *MAE* function was utilized to calculate the MAE for the MLR model.

$$MAE = \frac{\sum_{i=1}^{n} \ |p_i - a_i|}{n} \tag{F4.7}$$

Outliers and leverage points were two additional parameters that were analysed. Outliers are extreme observations in the response variable, whereas leverage points are extreme values in the characteristics that have a high lever (leverage).

4.3. Results and discussion

4.3.1. Spectral characteristics of mathematical spiked wastewater matrices

The shape of the complete UV-Vis spectrum (220 to 720 nm) can exhibit molecules and particles that are present in wastewater. Figure C4.4 shows the UV-Vis spectrum of three different wastewaters. The highest absorbances occur at the lowest wavelengths (220 – 400 nm). The molecules related to TOC, DOC, organic matter, and nitrates generally absorb the strongest in this range. The last part of the spectrum (from 500 to 720 nm) is where the absorbance of solids can be found.

According to the site, the three spectrums displayed differences. The highest absorbance of site 3 reached 157 Abs/m at the lowest wavelengths (220 nm) and gradually decreased to 60 Abs/m at the highest wavelengths (720 nm). It was the same for site 1, but the absorbance did not drop abruptly at the end; it remained around 90 Abs/m. Finally, site 2 had lower absorbances than the other two sites, ranging from 120 Abs/m at the lowest wavelengths (220 nm), and gradually diminishing to 35 Abs/m at the highest (720 nm).

As described in section 1.3.2., the hydrocarbon absorption spectrum has a low signal between 200 – 300 nm. Only a high concentration of toluene, ρ–xylene and m–xylene mixture (more than 10 ppm) can be detected over the wastewater spectrum.

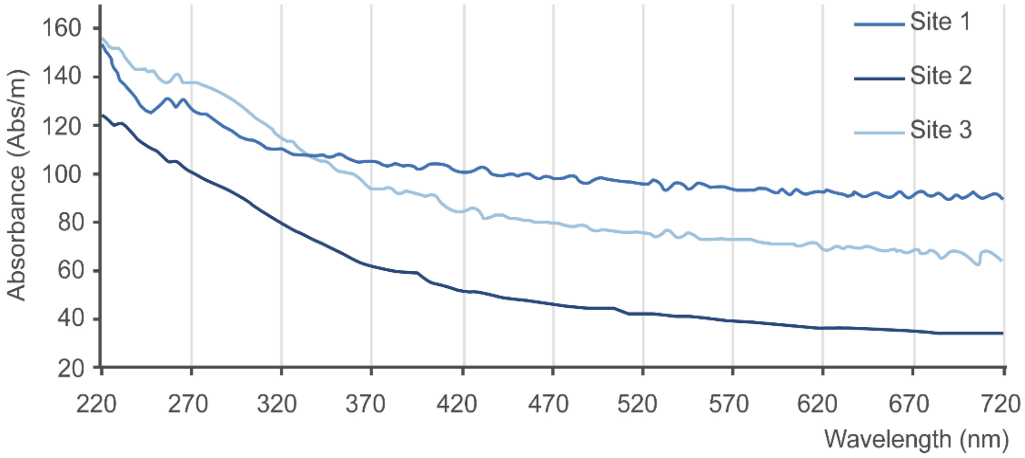


***Figure C4.4 –** Absorbances collected for the three sites studied. As can be seen, the absorbances obtained are directly related to the wastewater source.*

In Figure C4.4, the hydrocarbon mixture was already added mathematically to wastewater spectrums. It shall be noted that the BTEX spectrum (toluene, ρ–xylene and m–xylene) added in this figure had a total concentration of 0.5 ppm. This concentration entailed the addition of 4.2 Abs/m in wavelength 220 nm, which would significate only 2.7% of the total spectrum at that point.

Urban wastewaters have lots of organic and chemical compounds from households and, in some cases, from industrial areas. TOC, DOC, or TSS concentration increases the absorbance registered in the spectrum and completely disguises any slight presence of toluene, ρ–xylene and m–xylene mixture added. Some advanced statistic approaches combined with machine learning methods can facilitate this distinction.

4.3.2. Wavelength selection process

The variable selection process followed in this chapter was Stepwise Selection (SS). As discussed, and concluded in Chapter 1, SS was the most effective motor for choosing highly correlated wavelengths, with p-value and AIC serving as selection criteria.

The variable selection process was done in two steps. The first step was applying the SS method automatically to select the most useful variables (from 204) for this purpose, which met the lowest *p*-values and AIC threshold criteria. This process was applied to the three sites. For each site, around 12 and 15 variables were automatically selected. Adjusted R^2 and RSME values were highly suitable for each model.

The second step consisted of manually pruning the remaining variables and observing the goodness variation of the models at each iteration (adjusted R^2, RMSE, VIF and BP). After analysing the automatically selected variables, an extreme Variance Inflation Factor (VIF) was observed (values ranging from 1,500 to 2,000 VIF). High VIF (VIF > 10) entailed extreme collinearity for one or more predictors. In those circumstances, the regression parameters (i.e., R^2) were excessively inflated, and the coefficients of the independent variables were poorly estimated.

Figure C4.5 shows the correlation between fourteen variables selected automatically by the SS method. The circles' colour and size depicted in this graph were related to the amount of correlation between wavelengths. The greater and darker the circumferences, the higher the correlation between variables.

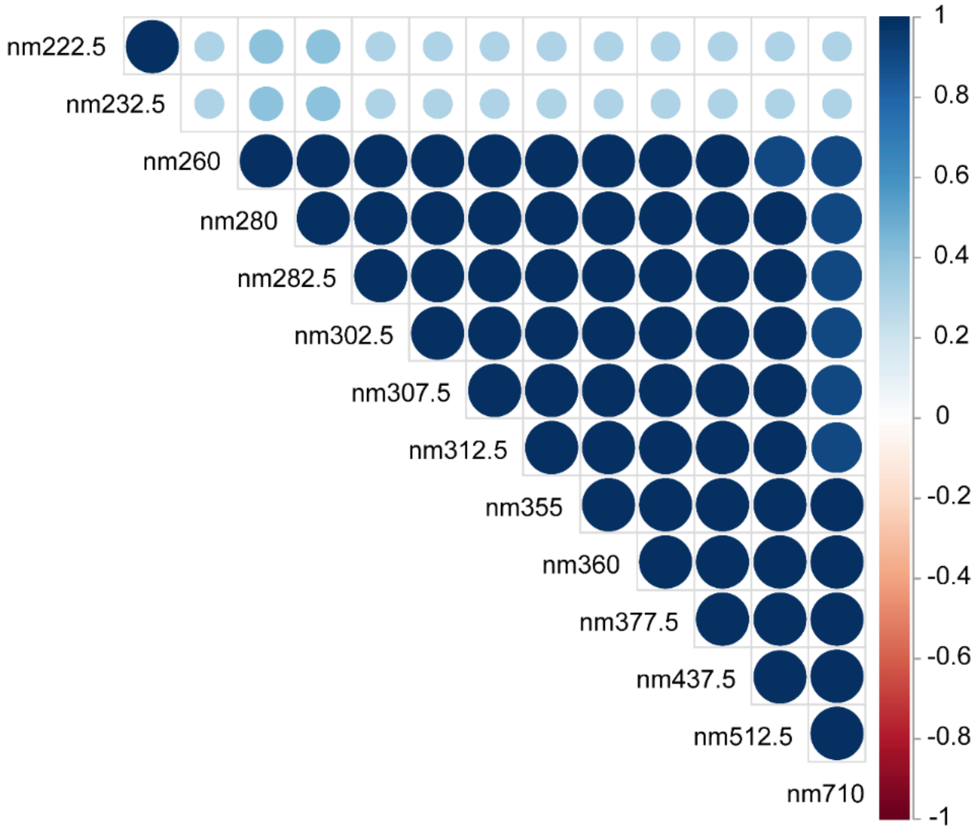


Figure C4.5 – *The Correlation plot shows the regression coefficient for all selected variables in site 1. There is high collinearity among almost all of them in this case.*

As shown in Figure C4.5, there was high collinearity between almost all variables selected automatically. Wavelength 222.5 nm was highly correlated with wavelength 232.5 (regression coefficient was nearly 1). However, those wavelengths exhibited a slight correlation with any other wavelength (regression coefficients from 0.3 to 0.5). Additionally, there was significant multicollinearity among the other included wavelengths. That is related to the continuous numerical nature, which characterises wavelengths (spectral overlap). The leading information of one wavelength is partially shared by the contiguous ones, generating perfect linear combinations (severe multicollinearity). When two or more regressors are expected to be independent but their determination coefficient is close to 1, this is known as *near collinearity* or *severe multicollinearity*. Selecting variables automatically can lead to a group of predictors with poor statistical significance due to their near collinearity. Creating a model with severe multicollinear predictors affects standard errors because it is difficult to separate the effects of one variable from another, worsening the model's prediction capacity. Harrell et al., 2015 and Keith et al., 2015 rejected making an automatic variable selection without further monitoring, as it could cause several issues, especially if the variables are collinear. As an illustration, it results in regression coefficients (R^2) that are very biased and need to be reviewed (Figure C4.5).

A comprehensive manual study of variable impact was performed once the variables were automatically obtained for each site. That study included VIF and BP as multicollinearity revision metrics. Those variables with high VIF and BP, increasing general collinearity, were eliminated. This manual iteration was applied to select the most relevant variables from those automatically chosen to avoid that issue. The final chosen variables were those that gave higher significance from all features chosen automatically and ended up being similar for the three sites (λ222.5, λ237.5, λ280, and λ710).

Due to the significance of several of those wavelengths in the BTEX spectrum, they are used as markers (Figure C4.2). The wavelengths such as λ237.5 are related to the absorbance of all hydrocarbons interchangeably, but others such as λ222.5 and λ280 would not give meaningful information for all hydrocarbons. A plausible explanation for the final variable selection results is that the concentration of the three BTEX compounds is mixed, providing possible features that could be different in the case of only one BTEX compound would be predicted.

Table C4.2 shows the selected wavelengths for each site. Although most of the variables exhibited in the table had a VIF lower than 10, few were set with a slightly higher VIF; the one with the highest value was λ237.5 of the MLR 3 model, with a VIF = 15.53.

BP determined the homoscedasticity of the regression models. If the *p*-values for the BP test were less than the significance level (0.05), then heteroscedasticity would be present in the regression models. For each model developed, the p-values of the BP test were above the significance level (0.27, 0.25 and 0.31, respectively), satisfying the rule of homoscedasticity and, for this reason, the variables with VIF nearly or greater than 10 were maintained.

It has been observed that the previous automatic variable selection becomes less random as more data is fed into the training dataset, frequently selecting a variety of features among the first 50 variables.

***Table C4.2** – Results obtained for each model developed. Each variable's Variance Inflation Factor (VIF) indicates its impact on the model. BP is the Breusch-Pagan test, which shows the model's homoscedasticity. The MSE indicates the mean squared error, and the Adjusted R^2 indicates the goodness of the linear model.*

	Site	Wavelengths (λ)	VIF	BP	MSE	Adjusted R^2
MLR 1	1	222.5 237.5 280 710	4.31 4.41 6.50 5.74	0.22	0.27	0.82
MLR 2	2	222.5 280 710	3.52 4.72 1.70	0.21	0.25	0.87
MLR 3	3	222.5 237.5 280	8.89 15.53 4.29	0.24	0.31	0.79

Additionally, depending on the water matrix, the bulk selected variables may vary This may be due to changes in the slope of the UV-Vis spectrum since each water had a different absorbance pattern, depending on the diluted compounds. However, similar wavelengths remained when the VIF and BP tests were done, all more related to the BTEX absorbance than the wastewater type.

4.3.3. MLR Modelling

Three regression models were created for each site (MLR 1, MLR 2 and MLR 3). Table C4.2 shows their performance parameters (Adj R^2 and MSE). The adjusted coefficients of determination (Adj R^2) showed accurate goodness of fit with 82, 87, and 79% of the dependent variable's variance explained by the independent predictors. Adj R^2 is a good performance evaluator because it is an unbiased estimator that corrects the sample size and the values of estimated coefficients. It captures the fraction of variance of actual values defined by the regression model and gives the model's best picture. In our case study, all three models were created using only particular wavelengths (previously selected and pruned), giving them the capacity to explain all possible variance (< 80% for the three MLR models) with fewer interferences or biases.

The MSE gives relevant information about the residuals of the model. All three models showed a small MSE (0.27, 0.25 and 0.31, respectively). MSE is the average squared distance between the observed and predicted values. It represents the error of a predictive model based on the observations collected. The square of the differences eliminates negative values for the differences and ensures the result is always greater than or equal to zero. It is helpful to obtain the MSE because the squaring penalises more significant errors than smaller ones. If larger errors appeared in the model, larger values of MSE would result from the modelling. In addition, higher error values could result in poorer predictions. Our case study's MSE values provided an overview of the future prediction behaviour for each model. For example, MLR 3 had the highest value of MSE, which means the predicted values resulting from this model could have higher uncertainty than MLR 1 and 2.

Figure C4.6 shows the predicted VS residual plots of the MLR 1, MLR 2, and MLR 3 models. On the abscissa axis are the fitted values, while on the ordinate are the residuals. The residuals are the difference between an observed value of the response variable and the value of the response variable predicted from the MLR model. As can be observed, the residuals showed a pattern of almost equidistant bands of comparable length, ranging from 0.2 to -0.2 for all three examples. These bands were the differences between the *actual* concentration of BTEX computationally added to the wastewater database, done discretely, and the model's predicted values of those concentrations.

Redundant trends and patterns in residuals can result from heteroscedasticity or non-linearity and, in some cases, suggest that the error variances are unequal. Ideally, the points should fall randomly on both sides of the 0 line (red dotted line), with no recognisable patterns (points cloud). The residuals in our plot appeared to exhibit a repetitive pattern and a vertical increment of the variance that remained practically constant all over the abscissa axis. This means that, as the value of the fits increases, the scatter among the residuals is remained uniform. If residuals *fanned out* from left to right rather than exhibiting a consistent spread around the 0, the plot would suggest that the error variances are unequal. As aforementioned, this pattern was motivated by the initial BTEX spiking, but it was not related to any violation of the linearity assumption for any model. It did not primarily affect the model's performance. The coefficients were not biased as a result of heteroskedasticity.

In summary, the three model’s residuals followed a remarkable horizontal trend around 0 (red dotted line). Even though there was a pattern all over the plot, the disposition of the residuals suggested that the error terms' variances were almost equal. The linearity of the residuals was a symbol of homoscedasticity.

Figure C4.6 can also give information about existing outliers and leverage points. Outliers and leverage points are unusual random sample observations that can affect their y-value or x-value, respectively. Outliers and leverage points may seriously jeopardise the regression outcome and must be constrained. Our plot does not exhibit extreme data but does display unusual points (blue circles). Nevertheless, observed separately, those data points cannot be considered an influence because they must be more than three standard deviations from the mean. Therefore, it would not be required to remove them.

Leverage values have also been studied in Figure C4.7 for all models. It shows the tendencies of standardised residuals (ordinate axis) against leverage (abscissa axis) of the explanatory variables for MLR 1, MLR 2, and MLR 3. This plot identifies influential observations (or with high leverage), as they strongly influence the regression model coefficients. Their elimination can noticeably improve the model’s performance since its coefficients would change.

Data exhibited in Figure C4.7 were randomly distributed as a point cloud for all cases, clustered around the 0 over the abscissa axis (the bulk data ranged from 0.001 to 0.002 for MLR 1, from 0.00025 to 0.0005 for MLR 2, and 0.001 to 0.0010 for MLR 3). Some remote points on the far right of the plots (blue squares) could have significant leverage. Standardized residuals must be three standard deviations apart from the mean in order to have leverage. Typically, a standardised residual with leverage greater than two must be removed.

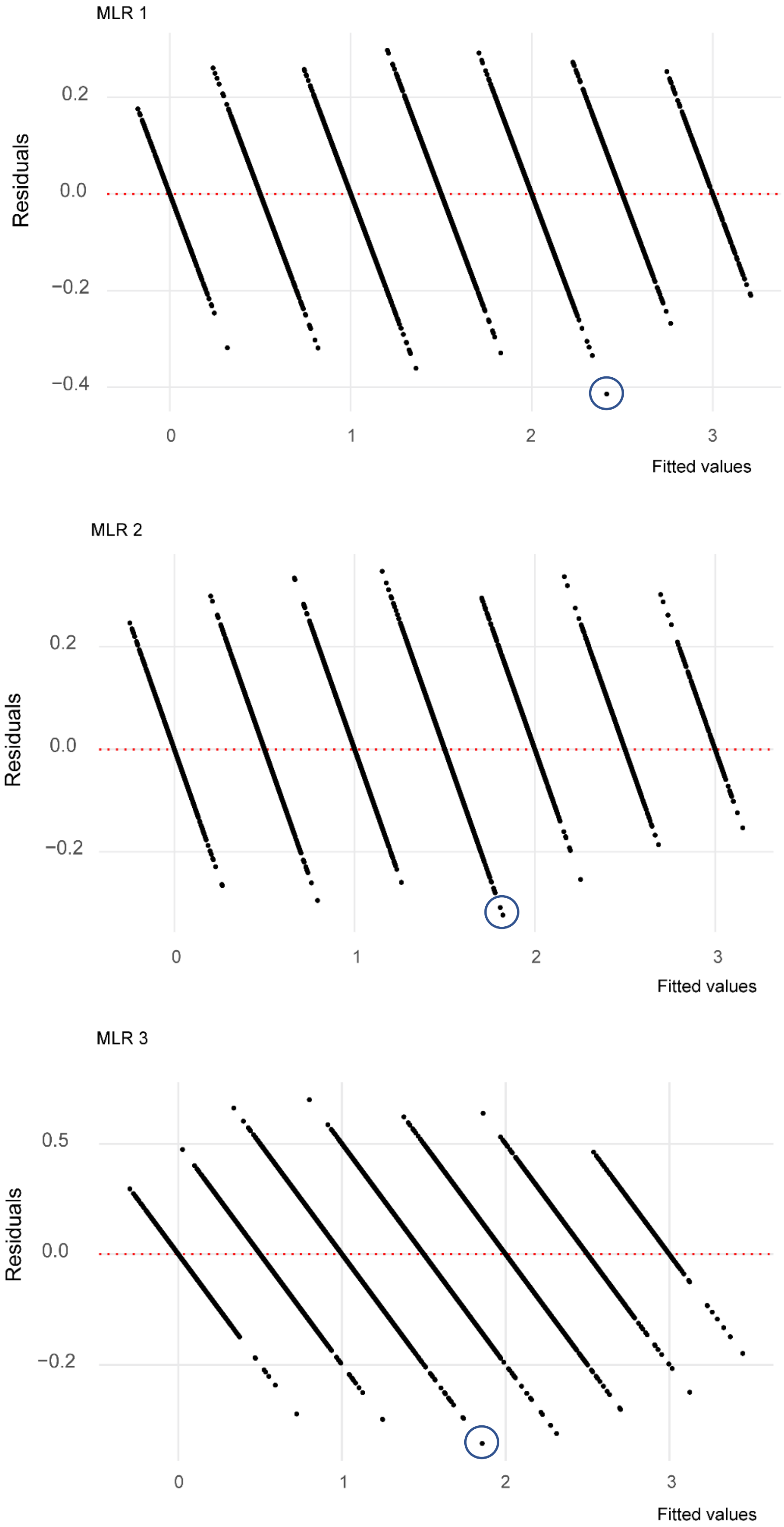


Figure C4.6 – *Fitted VS Residual plots of MLR 1, MLR 2 and MLR 3 models.*

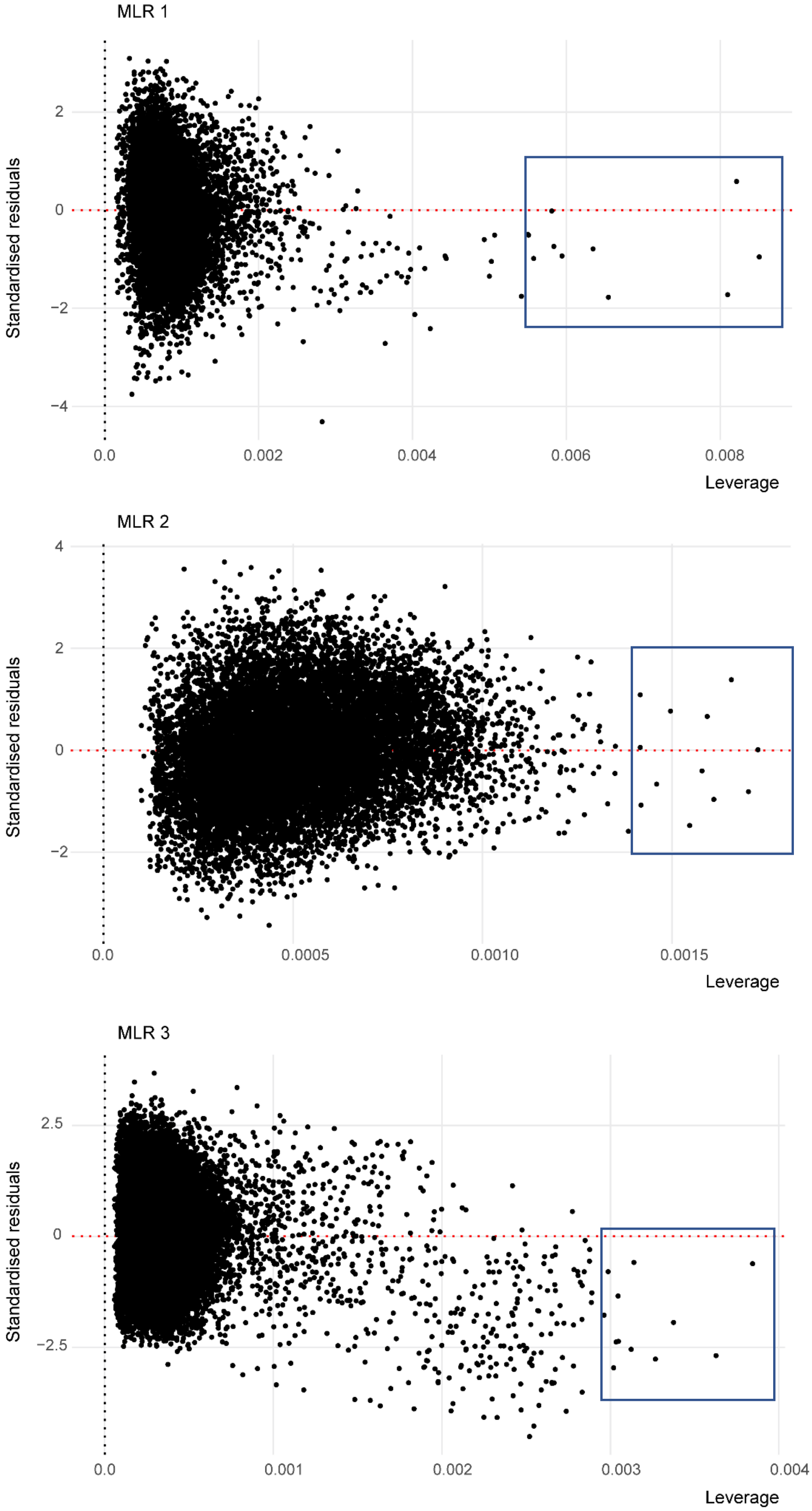


Figure C4.7 – *Residual VS Leverage Plot of MLR 1, MLR 2 and MLR 3 model.*

The highest leverage values were 0.00851, 0.00173, and 0.00384, corresponding to data points 185, 723 and 5271 for models MLR 1, 2 and 3, respectively (Table C4.3). Since they were not greater than three standard deviations, none of these dataset's observations had high leverage.

Figure C4.7 also gives information about the models' heteroscedasticity. The distribution of the points along the x-axis for the three models followed a point cloud trend without critical distribution patterns (i.e., funnel-shaped distribution), concluding that our regression models met the homoscedasticity's assumption. Figure C4.7 can also give information on how many different types of water exists in each dataset. MLR 2 showed a predominant distribution, which fell in leverage between 0.00025 and 0.0005, related to an outlined kind of wastewater. The remainder of the scattered dots might be random disturbances in water or the lenses of the spectro::lyser® since few observations were falling out of the bulk data. MLR 1 and MLR 3 exhibited a different distribution than MLR 2. All scattered observations were slightly altering the tendency of the bulk data, revealing a change in the optical characteristics of the wastewater. This tendency seemed more evident in MLR 3 because more data followed this trend.

Table C4.3 shows the ten higher leverage values of the data. The observations selected are also indicated in Figure C4.7 (blue squares). As shown in Table C4.3, the ten highest leverage observations of the MLR 2 model were recorded in different time frames. Therefore, they must be occasional disturbances while measuring. However, the ten highest leverage observations for models MLR 1 and MLR 3 belonged to the same time frame, exhibiting an area of points with slightly different wastewater quality characteristics in both cases. Even so, the sporadic changes in wastewater quality recorded in MLR 1 and MLR 3 could not affect the models' variance nor the quality of the predictions. Below is a review of the models' predictability.

Figure C4.8 shows the three models' quantile-quantile plot (or Q–Q plot for short). A Q–Q plot is indicative of the theoretical distribution plausibility of a dataset (i.e., normal, exponential). Q–Q scatterplot exhibits two sets of quantiles against one another. In the abscissa axis, theoretical quantiles are plotted. Sample quantiles are displayed on the ordinate axis. Essentially, the Q–Q plot confronts the quantiles from a selection of our actual data with theoretical quantiles, which are normally distributed, to recognise if the firsts follow a normal distribution.

***Table C4.3** – Ten observations with the highest calculated leverage reveal if they impact the models MLR 1, MLR 2 and MLR 3.*

	MLR 1		MLR 2		MLR 3	
	Obs.	**Calculated leverage**	**Obs.**	**Calculated leverage**	**Obs.**	**Calculated leverage**
1	185	0.00851	723	0.00173	5271	0.00384
2	186	0.00821	3367	0.00170	5961	0.00363
3	183	0.00809	13282	0.00166	5957	0.00337
4	180	0.00655	3292	0.00161	5288	0.00327
5	171	0.00635	13415	0.00159	5280	0.00313
6	172	0.00594	11698	0.00158	5971	0.00312
7	184	0.00584	3238	0.00155	5975	0.00305
8	179	0.00581	12053	0.00145	5272	0.00305
9	178	0.00551	3535	0.00146	5268	0.00303
10	173	0.00549	3213	0.00142	5963	0.00302

The data distribution from all the models shown in Figure C4.8 fell about a straight diagonal line, so a normal distribution can be assumed in those populations. MLR 1 and MLR 3 exhibited a marginal curve on the right side of the line, which could mean there were several types of wastewaters registered throughout time. However, the normal distribution of the bulk data was unaffected by the light quality variation detected, even though the standardised residuals could still demonstrate it.

The three models contained several unusual observations at each tail of the lines. Since they followed the diagonal without changing its length or slope, they would not be regarded as outliers or leverage points.

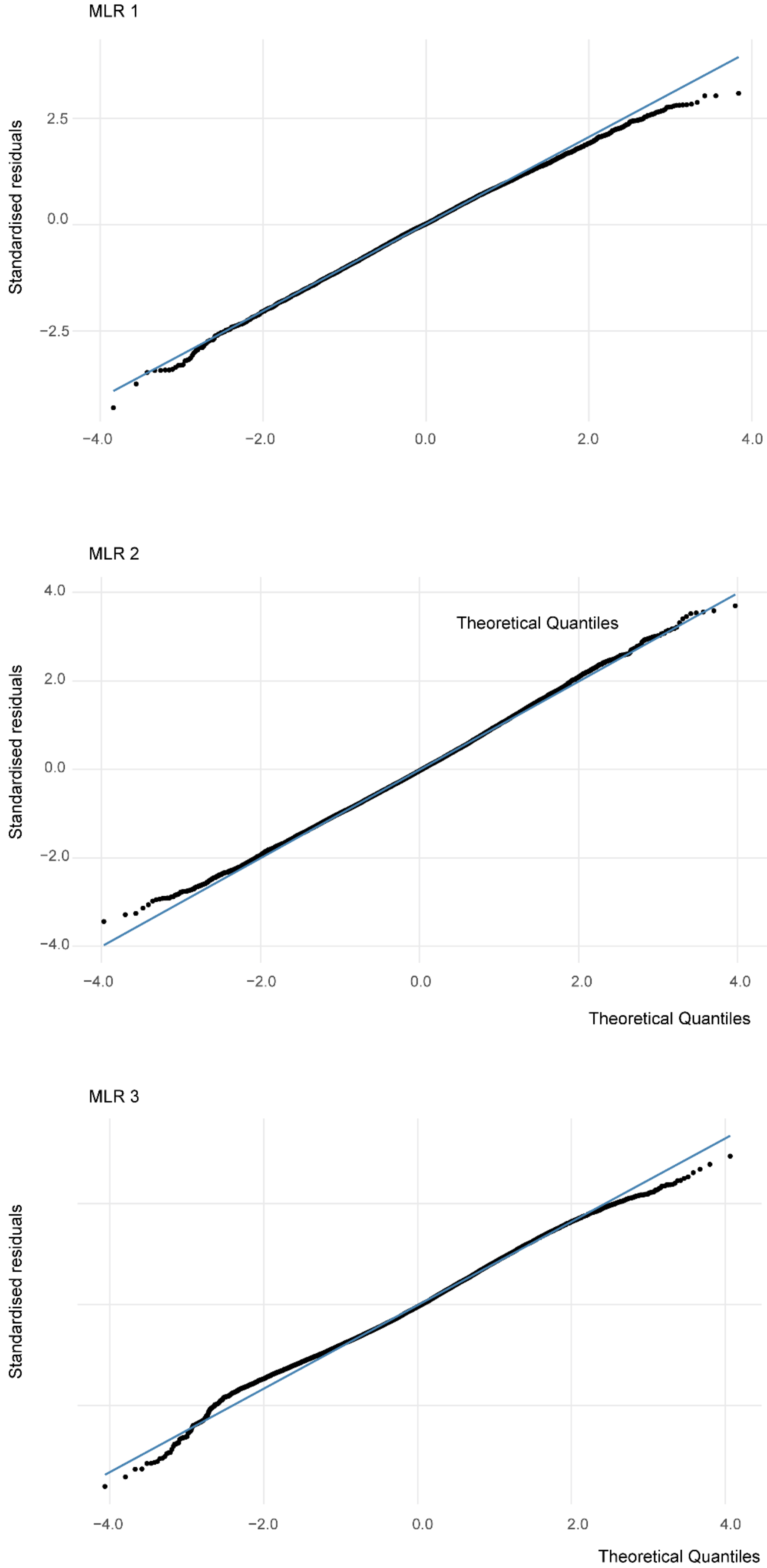


Figure C4.8 – *Q–Q Plot of all three models developed.*

Figures C4.6, C4.7 and C4.8 support the assumptions of linearity, homoscedasticity, normality, and unusual observations (outliers and leverage points) for the linear models developed in this chapter. If any of these assumptions were violated, a review of the models would be required. The three developed models might be used to identify and predict BTEX in wastewater matrices because they were able to meet all the hypotheses.

4.3.4. Prediction capacity

The prediction capacity was evaluated for all three models. Figure C4.9, C4.11 and C4.13 show the plots of actual data (ordinate axis) against predicted data (abscissa axis) for MLR 1, 2, and 3, respectively. Blue and red dots were the actual and predicted values, respectively. In the representative, the assessed concentrations for each model ranged from 0.0 to 2.0 ppm in intervals of 0.5 ppm of a controlled mixture of BTEX.

All models exhibited a notable horizontal dispersion in all prediction intervals. The horizontal distribution in Figures C4.9, C4.11 and C4.13 was already discussed in Figure C4.6.

The residuals are the values obtained by subtracting the observed from predicted values; thus, the residual's horizontal dispersion in Figure C4.6 is equivalent to Figures C4.9, C4.11 and C4.13 for each model.

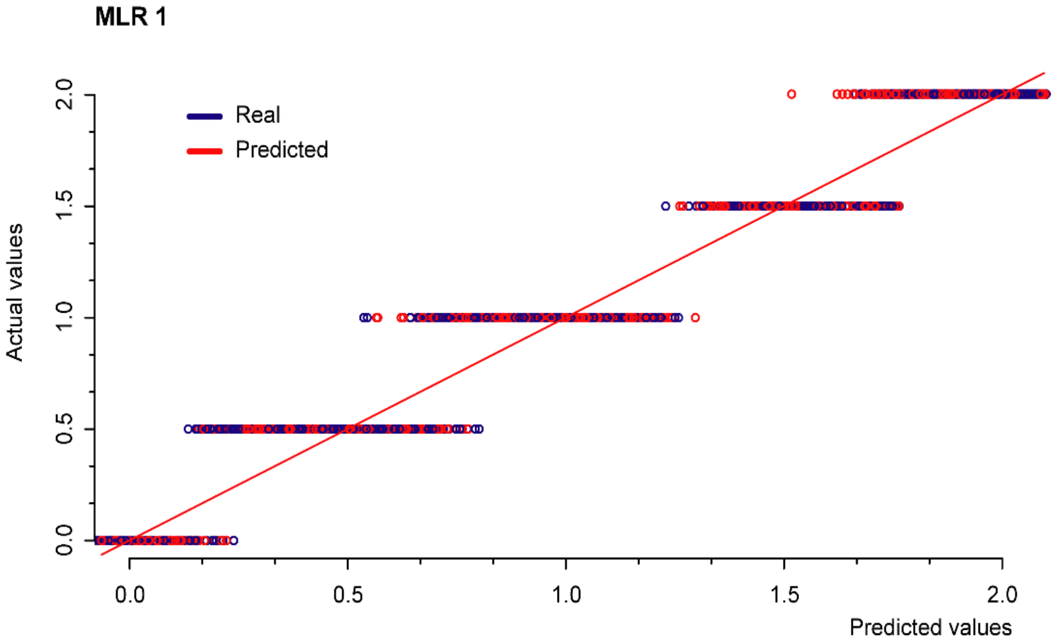


Figure C4.9 – *Scatterplot exhibiting the actual values against predicted values of MLR 1.*

Figure C4.9 shows the predicted against actual values for MLR 1. They showed an acceptable diagonalisation. For MLR 1, the intervals exhibited a regular dispersion. For example, 0.5 and 1.0 ppm predictions ranged from 0.15 to 0.85 ppm and 0.65 ppm to 1.85 ppm, respectively. The interval with accurate predictions was 1.5 ppm because the dispersion was less accentuated (approximately 1.25 to 1.75 ppm).

After studying the error embedded in the predicted values, it can be said that the vast majority of the predicted values had a predictability error between 7 to 8 %. However, there were some points with more significant error (50 %), especially in the 0-ppm concentration (negative values). However, for higher concentrations (1.5 and 2 ppm), the error was comprised from 5 to 10 %.

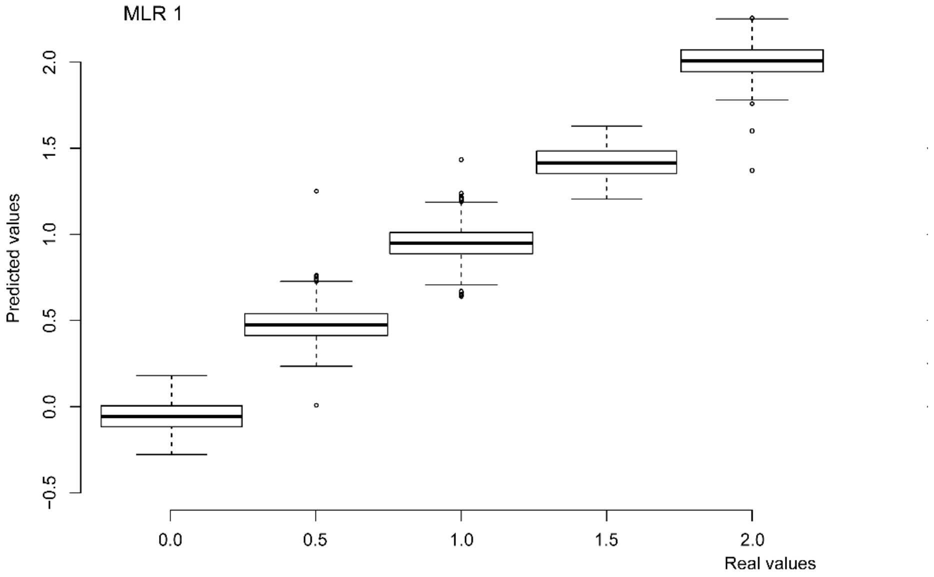


Figure C4.10 – *Boxplot of MLR 1. This plot exhibits the dispersion of predicted values of MLR 1.*

In Figure C4.10, a box plot can be seen. The boxplot box starts at the first quartile (25 %) and ends at the third (75 %). Therefore, the boxes for each concentration in Figure C4.10 represent the middle 50 % of the data. The black line inside the box is the median (the value that occupies the central place of our data when it is ordered from most minor to largest value). The medians of the 0, 0.5 and 1 ppm were slightly offset, which means most predicted values for those concentrations were underestimated. However, the medians of 1.5 and 2.0 ppm of the BTEX mixture were precise. The whiskers (lines outside the boxes) indicate the variability of predicted values of MLR 1 for each concentration of toluene, p-xylene and m-xylene mixture. Dots separated from the whiskers (seen at 0.5, 1.0, and 2.0 ppm) were associated with outliers, but, as stated above, they would not be inconvenient. Briefly, this plot confirms the suitability of almost all predicted values for the MLR 1 model.

The actual against the predicted plot of the MLR 2 model showed an acceptable diagonalisation. Although MLR 2 followed a similar trend as MLR 1, the dispersion for each concentration interval was less noticeable, as shown in Figures C4.11 and C4.12. For example, the predictions of 0.5 ppm could be acceptable, ranging from 0.29 to 0.65 ppm. The intervals with less accurate predictions were 1.0 and 1.5 ppm with an accentuated dispersion (from 0.75 to 1.35 ppm and 1.20 to 1.80 ppm, respectively).

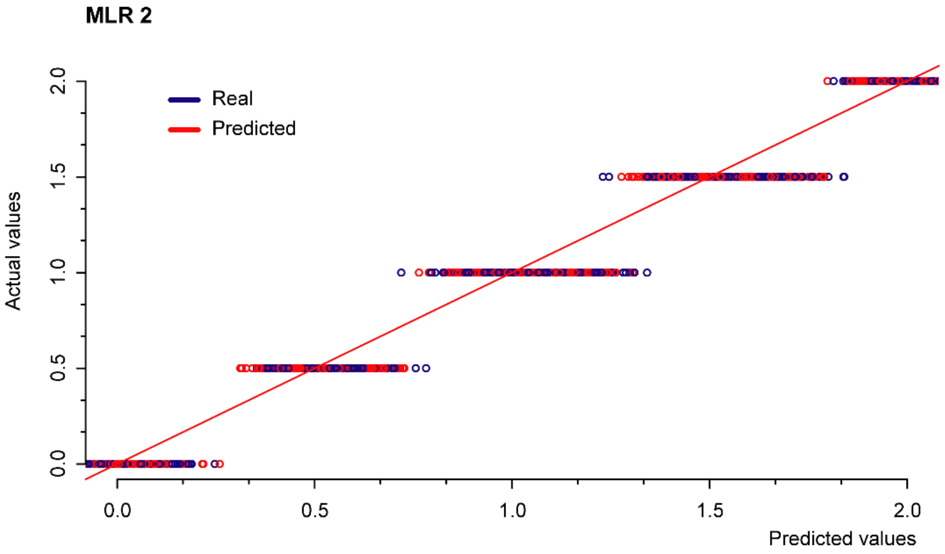


Figure C4.11 **–** *Scatterplot exhibiting the actual values against predicted values of MLR 2.*

In the same way as MLR 1, after studying the error of the predicted values, MLR 2 had at least a 5 % error in the predictions, reaching up to 40 % error. Much of the accumulated error in the predictions was in the 0-ppm range because the most significant number of predicted values fell below 0 (negative values). However, for higher concentrations (1, 1.5 and 2 ppm), the error was comprised from 5 to 15 %.

As discussed above, the dispersion of predicted values of MLR 2 was lesser important than MLR 1. As shown in Figure C4.12, the median in the box plot for each interval exhibited correct behaviour because they laid close to the actual values. The unique median showing a light offset would be for 0 ppm, placed slightly above the actual value. Whiskers displayed in Figure C4.12 for MLR 2 were shorter than MLR 1. There were some outliers at the cue of all whiskers, but they would not compromise the correct performance of the modelling.

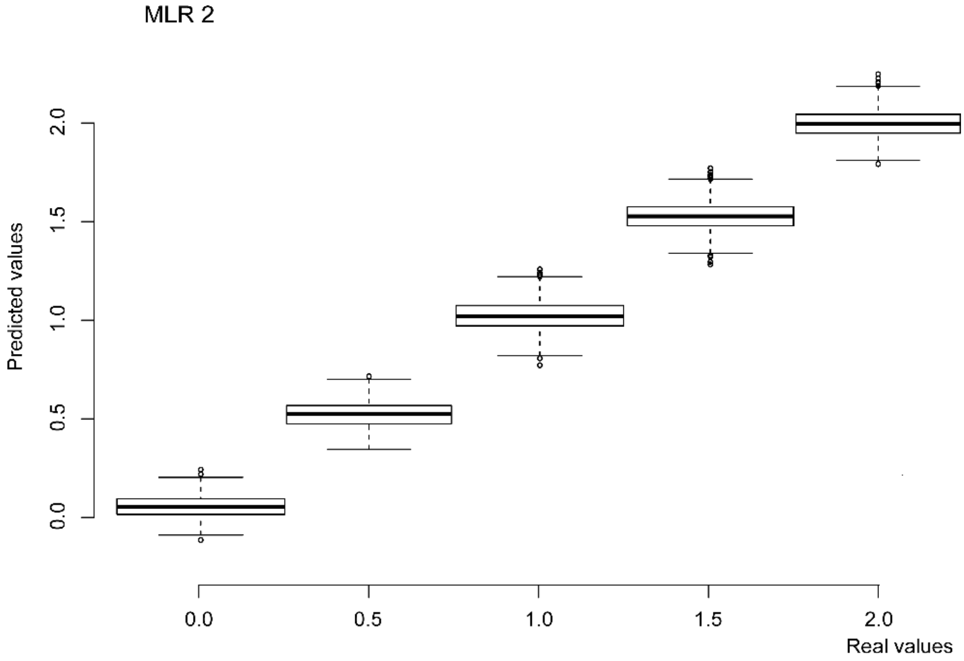


***Figure C4.12 –** Boxplot of MLR 2. This plot exhibits the dispersion of predicted values of MLR 2.*

The predicted against the actual plot for MLR 3 model exhibits the most significant dispersion compared to the other two models. For instance, 0 ppm showed a substantial offset to the left (negative values), meaning many predictions for that concentration were made inaccurately. Something similar holds with the other predicted concentrations. For instance, 1 and 1.5 ppm had a significant variance, ranging from nearly 0 to 1.40 and from 0.60 to almost 2.0 ppm, respectively. The diagonalisation of the predictions could be perceived but was clearly worse than the others exhibited for MLR 1 and MLR 2.

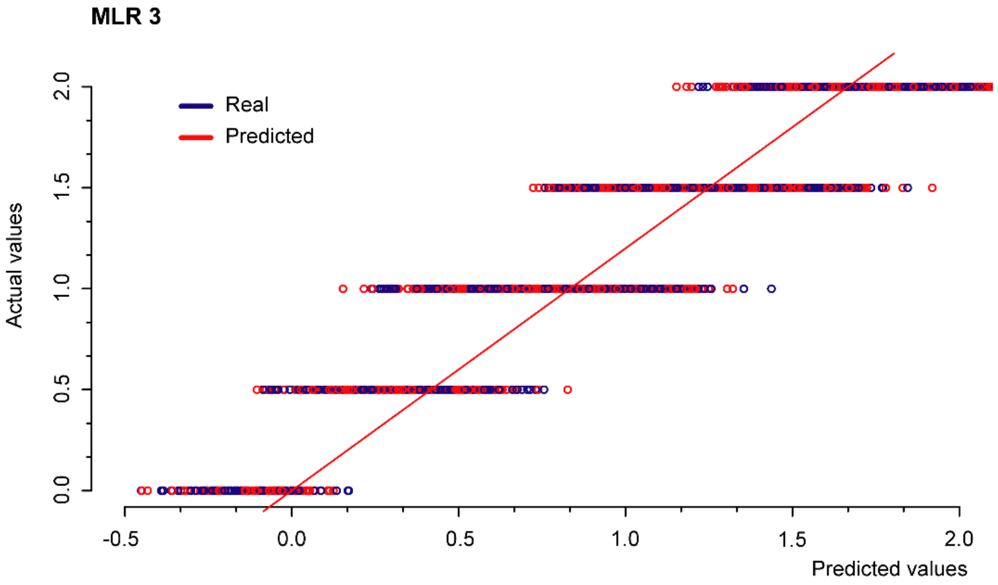


***Figure C4.13 –** Scatterplot exhibiting the actual values against predicted values of MLR 3.*

In the same way as MLR 1 and MLR 2, after reviewing the predicted values' error, we concluded that MLR 3 had the highest prediction error of all three models. An acceptable error for prediction would not be more than 15 %. The lower error registered in the predictions was 15 %, reaching 65 %, distributed among all ranges, being 1 and 1.5 ppm the concentrations with higher predictability error.

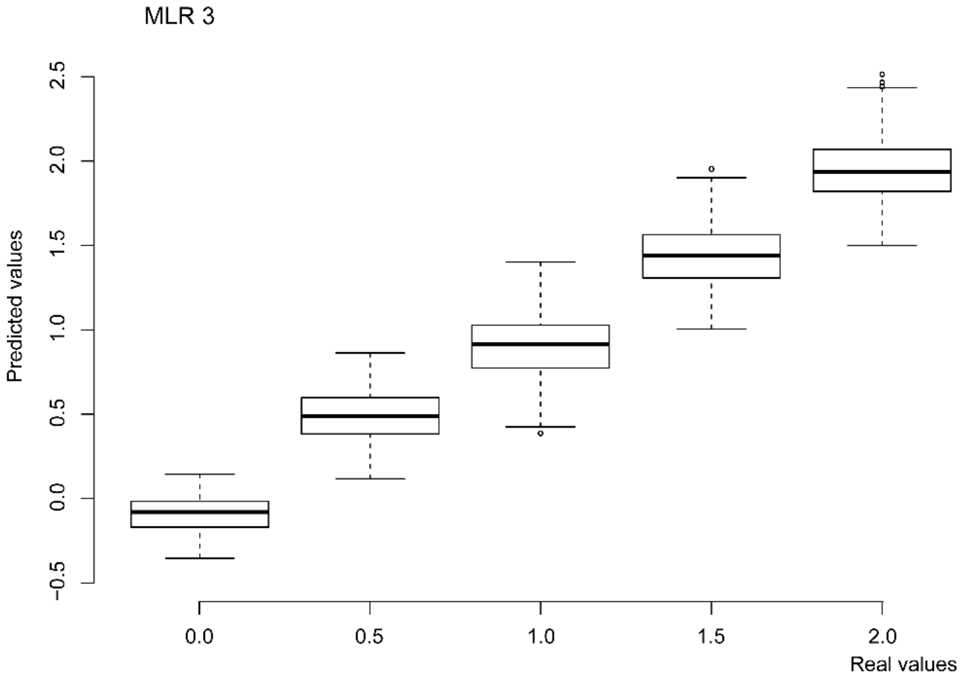


Figure C4.14 – *Boxplot of MLR 3. This plot exhibits the dispersion of predicted values of MLR 3.*

Figure C4.14 exhibits the box plots for each concentration of the MLR 3 model. As aforementioned, the dispersion was notorious. The size of the boxes was greater than the others examined before in Figures C4.10 and C4.12. The same happened with the length of the whiskers. In addition, the medians of each concentration were off-balance, being lower than their actual value. This fact was heavily appreciated in 1, 1.5 and 2.0 ppm, which the medians were far below the actual value of BTEX concentration.

Modelling and predicting contaminants such as toluene, m-xylene, and p-xylene in a wastewater matrix through its absorbance was not easy. Mathematical algorithms such as MLR can properly predict clear relationships between wastewater and BTEX absorbances. However, there are lots of other interrelated characteristics that can impede the modelling (i.e., data variability, availability, etc.). For this project, more than twenty wastewaters were examined and modelled from both urban and industrial. The initial idea was to create a global calibration where all wastewaters with a specific concentration of BTEX mixture could be integrated and modelled. The truth was the difficulty in pooling together in single modelling wastewaters, with huge different absorbances and variability, did not allow it.

It must be considered that all the data used to examine the MLR capacity for toluene, m-xylene, and p-xylene mixture prediction were not registered for that purpose, and hence the data quality, amount, and availability were, in some cases, sparse. That fact made the project balance from the necessity of excellent wastewater absorbance data to the reality of how difficult it was to group and extract a unique calibration from such different wastewaters. The experience concluded with the necessity of constructing site-specific models to achieve the expected modelling results.

As can be seen in the results presented, the predictability of the modelling varied significantly from one model to another. MLR 1 and 2 gave more accurate predictions in all intervals than MLR 3. This particularity can be related to the wastewater quality of each place. Each model was created and performed with singular wastewater, corresponding to distinct quality and sewage management.

The wastewater background absorbances played a particular role in the modelling. The toluene, m-xylene and p-xylene mixture was entirely covered by the high absorbance of the wastewater matrices, impeding their direct spectro::lyser®'s detection. The higher the absorbance of the wastewater matrices, the more complex the modelling to detect and predict the contaminants. That is a logical consequence of a changing signal-to-noise ratio as wastewater background absorbance varies. Only a high concentration of contaminants mixture could stand out from wastewater matrix absorbance. However, this fact would represent a critical environmental issue and would not be realistic.

As aforementioned, the crucial wavelengths for toluene, m-xylene and p-xylene mixture detection were in the range of 200 to 280 nm, and the final wavelengths selected to perform the predictions were λ222.5, λ237.5, λ280, and λ710 nm for MLR 1 and MLR 2, and λ222.5, λ237.5, and λ280 nm for MLR 3. As shown in Figure C4.4, wastewater from site 3 (corresponding to MLR 3) had the highest absorbance from 220 to 350 nm, compared with the other two sites. That fact could be the cause of the poor prediction capacity of MLR 3. By contrast, site 2 showed the lowest absorption across the entire UV spectrum, as well as in the previously mentioned interval (from 220 to 350 nm). The MLR

2 model seemed to exhibit the best prediction capacity, and the wastewater quality of site 2 might help. Briefly, the wavelengths automatically selected by SS seemed a correct combination, at least for MLR 1 and MLR 2, which predicted the known toluene, m-xylene, and p-xylene mixture satisfactorily.

The wastewater absorbance is not the only crucial characteristic to obtaining good modelling predictions. The variability in sewage management is also crucial. Notorious differences related to changes in the source of the wastewater quality (i.e., from urban to industrial in the wastewater collectors) can lead to failure in modelling. To avoid that, the models must be retrained to allow them to assimilate the changeability of the wastewater nature.

Another essential factor impacting the predictability of the models is the amount of solved toluene, m-xylene, and p-xylene mixture: the higher the mixture concentration, the better predictions of the models. The absorbance of toluene, m-xylene, and p-xylene increase as their concentration; therefore, it seems reasonable that the modelling prediction improvement (particularly for MLR 1) might be noticed in concentrations higher than 0.5 ppm. However, that trend was not so notoriously perceived in MLR 2 and not at all in MLR 3.

Table C4.4 shows the correlation coefficient, the MSE and the MAE of the predicted against actual data. The results in this table exhibit how much error the modelling prediction capacity had: the higher the correlation coefficient and the lower MSE and MAE, the better the modelling accuracy.

Table C4.4 – *Correlation coefficient and error metrics (MSE and MAE) of actual against predicted values.*

	Error metrics between actual VS predicted		
	Correlation coefficient	**MSE**	**MAE**
MLR 1	0.88	0.49	0.20
MLR 2	0.90	0.56	0.27
MLR 3	0.79	0.59	0.29

The correlation coefficient between the predicted values against the actual measured degree of intensity between them. Table C4.4 shows the correlation coefficient for MLR 1, 2, and 3. Following the discussed results, MLR 3 was the model with a poorer correlation coefficient between predicted and actual values. MLR 2 was the best one. These results were related to the data variability discussed in Figures C4.9 to C4.14. A higher variance between actual and predicted values originated poor correlation results between actual and predicted—the less the variability, the better the correlation coefficient.

Table C4.4 also exhibits each model's Mean Squared Error (MSE) and the Mean Absolute Error (MAE). These metrics were also used to evaluate the relationship between the predicted against actual values. MSE represents the average squared difference between the actual and predicted values. In summary, it measures the variance of the residuals. The MSE obtained for MLR 1, 2, and 3 are 0.49, 0.56, and 0.59, respectively. The MAE represents the average of the absolute difference between the actual and predicted values, measuring the average of the residuals. The MAE obtained for MLR 1, 2, and 3 are 0.20, 0.27, and 0.29, respectively. As shown in Table C4.4, and following the above results, MLR 1 and MLR 2 had less MSE and MAE than MLR 3. As aforementioned, MSE penalises large prediction errors. MLR 3 had a higher % of error in some predicted values (over 60% in some cases). MLR 1 had an MSE smaller than MLR 2, meaning there were shorter prediction errors. As we discussed above, some predicted values of MLR 1 had at least a 50 % of error. However, instead of MLR 2 having fewer erroneous predicted values (40 %), there were more erroneous values in larger concentrations, which could cause the MSE of MLR 2 to be a bit larger than MLR 1. MLR 2 has a higher MAE than MLR 1 because it had more and larger erroneous predicted values. The values obtained studying this metric validate the last hypothesis.

Concluding, and as it can be seen, MLR 1 and MLR 2 were better models than MLR 3. However, the three models could be helpful as an alarm tool for noticing a possible hydrocarbon spill from an industrial park.

4.4. Conclusions

Real-time absorbance from spectro::lyser® could positively change the control of undesirable spills such as BTEX compounds. The creation of mathematical models using absorbances of wastewater to detect and predict an occasional presence of BTEX could become a helpful tool for operators that would be able to take action in case necessary.

Multivariate Linear Regression (MLR) models created from the absorbance of wastewater using a spectro::lyser® are an effective tool for directly predicting the presence of toluene, m-xylene and p-xylene mixture. However, some particularities must be considered:

i) As a result of the quality of the data recorded, it is often a challenge to collect and select the most relevant data for modelling. In this case, more than twenty databases were examined but only three were ultimately chosen for modelling.

ii) An UV-Vis spectrum has a high number of wavelengths. Thus, an initial automatic selection tool would be necessary to choose the most suitable ones for each case study. Stepwise Selection (SS) is an excellent selector to seek the most relevant variables for the MLR models.

iii) The wavelengths have a continuous nature, leading to collinearity. This fact entails the necessity to apply a final wavelength by-hand selection, using some metrics to decide which one is better for the models' performance. Metrics such as the Variance Inflation Factor (VIF) and the Breusch-Pagan test (BP) can help assess the final selection.

iv) Once the models are developed, the five assumptions of regression must be reviewed. They are linearity, multicollinearity, independence, homoscedasticity and normality. There are several manners to examine these assumptions. Still, the bests ones are the evaluation given by the Residual against fitted plot, Residual against Leverage plot, Quantile – Quantile plot and VIF metric. Neither of these assumptions should be violated. The MLR models can be considered accurate and ready to start predicting data if all are met.

The data variability plays a vital role in the prediction capacity of an MLR model. The three models developed in this chapter were site-specific, yet variability obstacles were detected. However, the prediction capacity of MLR 1 and MLR 2 was accurate, giving a proper predicted value in almost all cases. The vast majority of predicted values for those models had around 5 – 8 % of error (reaching up to 40 and 50 % in some prediction values, respectively). MLR 3 showed a higher lack of predictive capacity, with at least a 15 % error, reaching 65 % in some prediction values.

Machine Learning (ML) models developed in this chapter can learn over time. This can improve the model's adaptability to wastewater matrix changes and toluene, m-xylene and p-xylene mixture prediction. Thus, with constant training, these models could be an excellent tool for an early alert in any Wastewater Treatment Plant (WWTP).

Capítol 5 – *Application of Advanced algorithms for the prediction and improvement of coagulant dosage in WRPs considering two scenarios of training*

Abstract

Coagulation-flocculation process is a chemical water treatment technique. It usually carried out by sedimentation/filtration to improve the removal of solids and turbidity in wastewater reclamation plants (WRP). One of the most critical parameters to control during the process is how much coagulant is added. An optimal coagulant dosage would ensure proper solids removal to comply with legislation while simultaneously reducing the chemical consumption costs. The dosage of coagulants is still being calculated from offline Jar-tests. In this chapter, we evaluated three Machine Learning (ML) methods to develop a mathematical model that can automatically optimize the required dose of coagulant, based on the process' most important variables while maintaining effluent water quality. Multivariate Linear Regression (MLR), Support Vector Machines (SVM), and Artificial Neural Networks were the ML algorithms chosen (ANN). Two spectrophotometric probes spectro::lyser® installed at two crucial locations of the Camp de Tarragona WRP in Catalonia provided six months of online data for training these algorithms. In comparison to the additions being used, the results obtained indicated an optimization in the coagulant dosage. The ANN model produced better predictability and identified the main limitations for sound correspondence. In the best-case scenario, for instance, MLR provided a coefficient of determination of 0.57, SVM provided 0.89, and ANN provided 0.86. The prediction capacity of the three models were 0.71, 0.81 and 0.78, respectively. Although MLR and SVM had good prediction results, they typically provided an unreliable estimate of the coagulant dosage prediction, with the ANN being the most reliable model. In the best scenario, MLR and SVM increase the coagulant dosage by 12 and 7 percent, respectively. However, if an ANN model were used and properly trained to control the dosage of coagulation, 4% of the coagulant could be saved.

5.1. Introduction

One of the most fundamental human needs is to have access to safe and clean drinking water (Clean water and sanitation, 6th Goal of UN for 2030). The water and wastewater treatment industry have the critical challenge of providing access to safe water to a growing population while simultaneously complying with increasingly stringent environmental regulations. Due to a series of factors such as water scarcity and stress, environmental pollution from improper wastewater disposal and recognition of the resource value of wastewater, the water-reclamation approach has gained a lot of importance in the last 20 years. Water reclamation (also known as wastewater reuse, water reuse or water recycling) is the process of reclaiming municipal wastewater (sewage) or industrial wastewater into water that can be reused for several purposes such as urban reuse or, agricultural reuse (irrigation), among others (Sanz, et al., 2015).

A Water Reclamation Plant (WRP) is a Wastewater Treatment Plant (WWTP) that provides high water quality that can be reused. It is also considered a tertiary treatment that regenerates water from WWTP. A WRP typically consists of a number of procedures that, when taken together, can meet the stringent treatment standards in effect today and guarantee the hygienic safety of processed water. Coagulation-flocculation, membrane treatment (ultrafiltration, forward osmosis, reverse osmosis), and ozonation are the most frequently used technologies in a WRP. This approach allows replacing drinking water with reclaimed water in several processes at the nearby Tarragona petrochemical park. A WRP's individual processes are interconnected and consequently have an impact on one another) For instance, the initial pre-oxidation treatment affects the coagulation-flocculation process, which in turn may impact subsequent steps like settling, filtration, ozonation, etc. (Jiang et al., 2015).

The coagulation-flocculation process is a chemical water treatment technique, that is typically applied before a physical separation process. To improve the particle removal capability of the WRP, sedimentation or filtration are typically used. (Sillanpää et al., 2018). It is a simple and economical technique widely employed in mass-scale wastewater treatment plants, where the primary water supply is treated. (Huang et al., 2018; Sun et al., 2020). Some factors directly affect the coagulation-flocculation process and intervene with the interaction between the particle and the coagulant These include water source composition (natural organic matter – NOM – origin, seasonal changes), chemical characterization of the coagulant, coagulant dosage, pH control (at the inlet and outlet of the process), control of the temperature, flow velocity, etc.

The amount of coagulant added during the process is one of the most critical parameters to keep under control. An optimal coagulant dosage would ensure an effective coagulation while simultaneously reducing the chemical consumption and, as a result, operational costs. Higher and faster NOM precipitation is achieved, which results in looser aggregates, if coagulant is added in excess and the pH is controlled. Therefore, lower dewatering properties are observed (Swietlik et al., 2004; Matilainen et al., 2010a; 2010b). An off-line analysis known as a jar test involves mixing the samples to emulate the turbulence regimes of the tanks while exposing a volume of water from the treatment plant's inlet to various doses of coagulant (Patel et al., 2013; Saritha et al., 2017). Once mixing conditions are stopped, the flocs are allowed to settle. The turbidity of the samples is then measured, and the dose with the lowest turbidity is considered to be the optimum coagulant dosage. The simplicity of the jar test is one of its main benefits (Jarvis et al., 2006; Yan et al., 2008s). However, the information is localised and related to the tested wastewater characteristics. If those are stable over time, the coagulant dosage will remain valid. But that is not usually the case, and therefore, the jar test does not provide continuous information on the coagulant dosage that must be added if water quality changes. Together with the large number of parameters affecting the coagulation, this factor complicates the control of the process for WRP operators (Franceschi et al., 2002; Fitzpatrick et al., 2004; Gregory et al., 2004).

On-line sensors such as spectro::lyser® monitor the water quality at the inlet and outlet of the coagulation-flocculation process. These sensors continuously monitor and record large amounts of water quality data such as 200 to 400nm spectrum, water temperature, total organic carbon, etc., therefore, they can be very useful for plant operators. However, even though these sensors can satisfy the demand for specific water quality information, they do not report the required dose of coagulant. In addition, more and more researchers are attempting to predict the optimal dose of coagulant based on the various parameters that have been discussed (inlet and outlet pH, flow rate, etc.), not only to optimize the chemicals dosed but also to obtain the ideal result of the process (Cheng et al., 2010; Olanrewaju et al., 2012). Artificial Intelligence (AI) has become more and more popular for performing this task. Mathematical prediction models could be beneficial for obtaining continuous and accurate coagulant dosage information, but they occasionally require a considerable amount of data to produce reliable results (Olanrewaju et al., 2012). AI prediction methods such as Multivariate Linear Regression (MLR), Support Vector Machines (SVM) and Artificial

Neural Networks (ANN) are increasingly used to assist operators in decision-making. When properly trained, these algorithms can provide an early and decisive response to a possible change in a wastewater reclamation plant process, widely improving the information obtained by conducting only a jar test (Maier et al., 2000; Hsu et al., 2016; Baovab et al., 2018).

The objective of this chapter is to develop a mathematical model that can automatically control the required dose of coagulant based on the variables present during the process and two spectro::lyser® probes. If necessary, plant operators could use the model as a tool to respond quickly to WRP. This has been accomplished by comparing the outcomes of the mathematical algorithms MLR, SVM, and ANN.

5.2. Methodology

5.2.1. Water Reclamation Plant

The research elaborated in this chapter has been developed in the Camp de Tarragona WRP in Catalonia (Spain), (Sanz et al., 2015). WRP Camp de Tarragona is operating by Aguas Industriales de Tarragona, S.A. (Industrial Waters of Tarragona, Inc., AITASA) since 2012. Veolia Water Technologies is providing technical support and advice. The capital investment of the first phase of the water reclamation and reuse project was 47 million euros, jointly provided by EU cohesion funds, the Catalonian Government, and the Spanish Ministry of the Environment.

The raw water comes from the secondary treatment of two industrial parks located in neighbouring towns. The raw water received by the WRP accumulates in a regulation tank consisting of 2 chambers, each with an operating volume of 3,060 m^3 and dimensions of 20x30x5.10 m. The raw water goes to the first regeneration treatment process with a maximum unitary flow rate of 625 m^3/h. Lamellar-lasted decantation, based on the Actiflo process, is the first treatment step in this WRP. It is a compact clarification system that uses microsand as a precursor of flocs with greater specific weight. This design allows for high hydraulic speeds and short retention times (Sanz et al., 2015).

The physicochemical treatment is distributed in coagulation, injection, maturation tanks, and lamellar decantation. The WRP contains filtration, reverse osmosis, and a final disinfection process (Figure C5.1). The WRP's operational chain enables obtaining regenerated water with optimum quality for its industrial reuse.

This plant has supplied 2 hm^3/year since 2013 and it has been gradually increasing the water volume supply to reach 6.8 m^3/year of nominal capacity (Sanz et al., 2015).

5.2.2. Sensors and data collection

Two spectro::lyser® were installed in the WRP. One of them was installed in the regulation tank. The other one was near the coagulation process' outlet (Figure C5.1). To prevent fouling that would obstruct the window path and cause signal drift, both sensors were equipped with an automatic cleaning system based on air-mechanic methods that was displayed every 2 minutes.

Ultraviolet (UV), with a wavelength range of 200–380 nm, was the range that both spectrophotometric sensors covered. They were connected to the Con::cube®, a PLC that allows managing the data gathered, the sensor's conditions and their stability, the frequency of cleanings, etc.

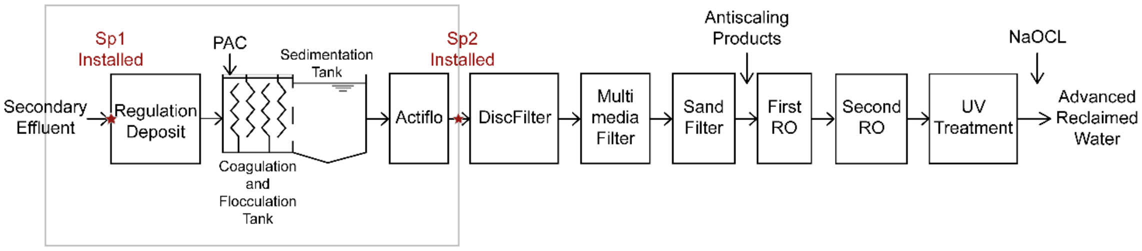


Figure C5.1 – *Scheme to visualise different processes conducted in the WRP. The coagulation, flocculation and Actiflo processes (framed in grey) had been monitored for six months. Red stars mark where the spectro::lyser® were installed. Source: modified from Sanz et al., 2015.*

In addition to UV spectrum, the parameters monitored throughout the two probes mentioned above were turbidity, colour, Total Organic Carbon (TOC), temperature and Dissolved Organic Carbon (DOC). Additionally, the WRP already had several sensors installed in the inlet of the regulation tank and at the outlet of the coagulation system. They monitored pH, flow rate, conductivity, and the dose of coagulant added.

Data was collected between December 2019 and May 2020. Every two minutes, the parameters listed above were all recorded.

5.2.3. Summary of water quality

All the information collected demonstrated several stages of water quality, providing an overview of the changes it underwent. Raw water turbidity varied from 7 NTU under the most favourable quality conditions to over 90 NTU during the episodes of poor water quality. Similarly, raw water conductivity ranged from 1,500 to 2,500 µS/cm.

The variations in raw water quality determined the coagulant dosage amount (aluminium chloride –PAC–). The coagulant applied ranged from 20 to 360 mg/L.

Regarding the parameters registered at the outlet of Actiflo, turbidity ranged from 0.28 to 1.2 NTU. It did not show significant variations, only during isolated intervals of poor water quality, reaching values of 5.7 NTU. The output turbidity is a parameter considered critical for optimal coagulation and flocculation process operation. An appropriate value of turbidity would range between 0.5 and 1.5 NTUs. A higher turbidity output suggests poor coagulant or pH reaction management, and the process should be regulated. Similar trends were observed in the other parameters registered.

Figures C5.2 and C5.3 demonstrate the difference between the UV spectrum recorded at the inlet of the regulation tank and at the outlet of the coagulation treatment at different time intervals of the project, respectively. Figure C5.2 shows different ranges of absorbances (220 to 310 abs/m at wavelength 200 nm and between 30 – 50 Abs/m at wavelengths 240 to 400 nm). The dark red arrow marks an important event of poor water quality (absorbances reaching 120 Abs/m at wavelengths 240 to 400 nm).

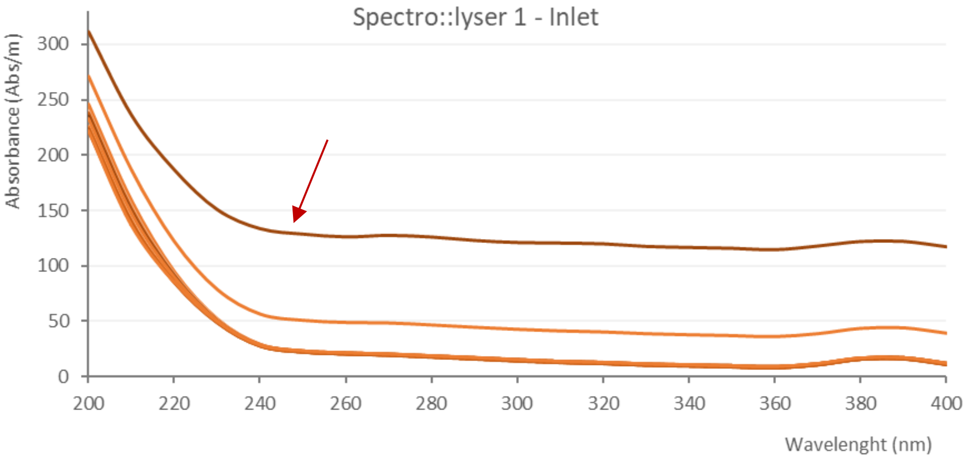


Figure C5.2 – *Absorbances recorded by spectro::lyser® installed at the regulation tank at different project periods. Different water qualities are observed. These events must be controlled to execute the coagulation-flocculation process correctly.*

Figure C5.3 shows wastewater absorbance after coagulation treatment. The red arrow shows the neutralization capacity of the treatment. All UV spectrum was recorded to maintain a similar trace, with absorbances around 200 Abs/m in wavelength 200 nm and decreasing until practically reaching zero from wavelengths 240–260 nm. In this case, the red arrow shows the excellent performance of the coagulation-flocculation process. Thus, that poor water quality episode was correctly controlled and managed.

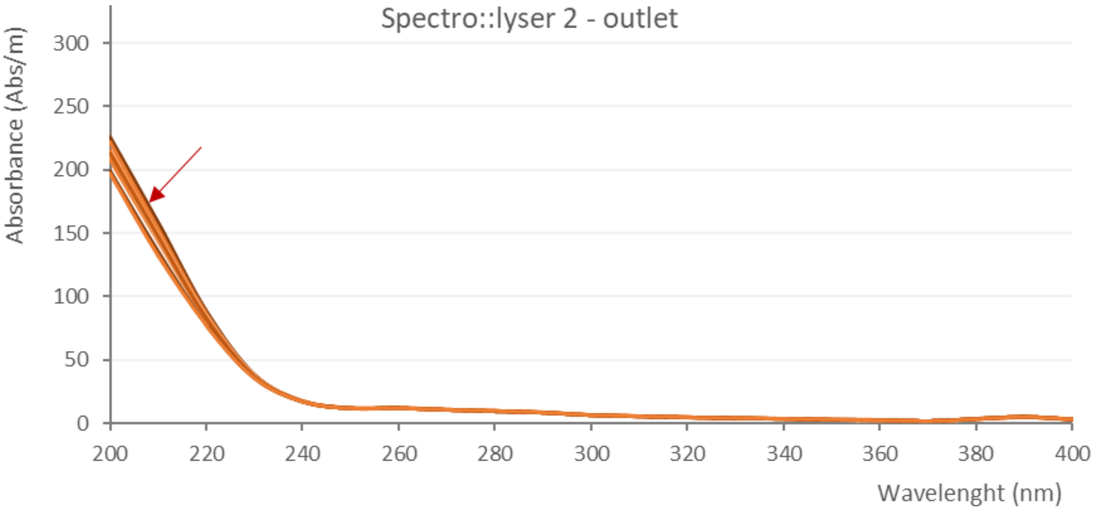


Figure C5.3 – *Absorbances recorded by spectro::lyser® installed at the outlet of the coagulation-flocculation process at different time intervals of the project. The quality has been unified.*

5.2.4. Data structure

Figure C5.4 shows a diagram of the sequence of procedures performed to create the models. As it shows, the first step was data collection and then pruning, and normalisation steps were applied. Once data was prepared, training, testing and validation sets were organised. To do this, RStudio (Version 1.2.5033) was used. In the following paragraphs, a detailed explanation is presented.

Data collected by spectro::lyser® were divided into three data sets: training, testing and validation sets. Each dataset's columns were the registered water quality parameters, and the rows, each sample collected by the sensors every two minutes. Before executing the database divisions, erroneous or zero data were deleted. In total, 4,034 observations remained.

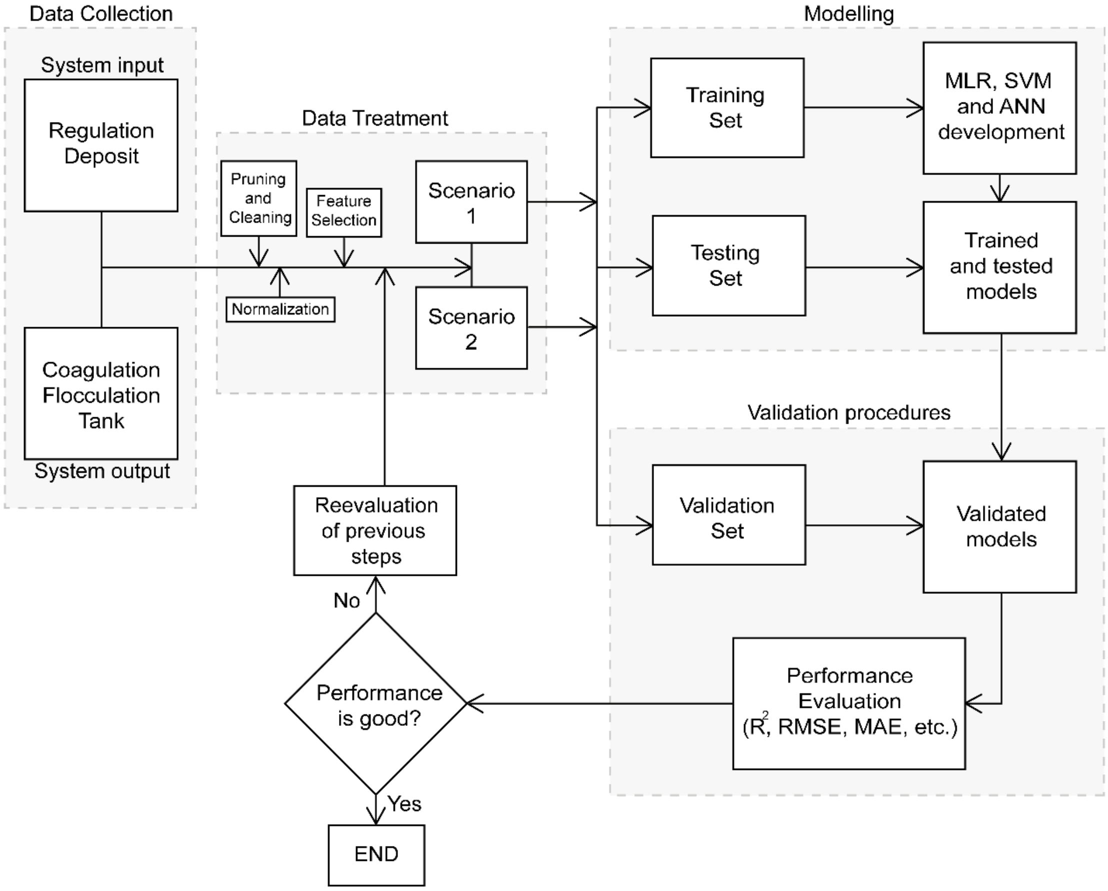


Figure C5.4 – *Flowsheet outlining the whole model development. Data collected was normalised and divided into three datasets: training, testing and validation sets. Several indicators of performance determined the goodness of the models.*

The relationships between the variables were evaluated after the data was organized and pruned. The function *cor* from the *Stats* library was used to create the correlation table (Table C5.1).

Variable selection has been performed using Random Forest (RF) feature importance algorithm. It describes which features relevant to our dataset. This selection was based on *randomForest* and *importance* function of the *randomForest* library. To visualise it, the function *barplot* from the *Stats* library was used.

Coagulant dosage has been selected as a response variable, which varies between 20 – 360 ppm, as a function of the other water quality parameters included in the process. Water quality indicators registered at the flocculation-coagulation process' inlet and outlet were the chosen predictors.

To test and validate the algorithms' ability to predict outcomes based on their initial training, two scenarios were developed. In the first scenario, a training set was created using 50% of the data without randomization, and the remaining 50% was divided equally between testing and validation sets (25 – 25%).

The training, testing, and validation sets, respectively, contained 2017, 1009, and 1009 rows as a result of the application of the data partition described above. All data were scaled linearly between 0.0 and 1.0, and non-critical outliers of all parameters were included in the datasets.

The second scenario was developed using K–means clustering to randomise the training and testing data. First, the validation set (the last 1000 rows of the general dataset) was separated from the bulk data before the randomisation.

The remaining data was subjected to K-mean clustering using the function *kmeans* from the *cluster* library. K-means requires the precise number of clusters to divide the data because it is an unsupervised method. The method Silhouette provided the proper divisions of clusters from the function *fviz nbclust*. The training and testing sets were divided after the clusters had been formed, taking into consideration an equitable distribution of the clustered data, 1975 and 1060, respectively.

The partition in three differentiated datasets avoids overfitting, as the validation set is not used as a part of the model development. The validation set is part of the final confidence analysis.

5.2.5. Model Development

Three algorithms were applied to compare the prediction capacity of coagulation dosage. As stated above, the algorithms are Multivariate Linear Regression (MLR), Support Vector Machine (SVM), and Artificial Neural Networks (ANN).

- Multivariate Linear Regression (MLR)

MLR is a statistical technique able to predict a response using several explanatory variables. The equation of an MLR (F5.1) model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon \qquad \text{(F5.1)}$$

Where y_i is the dependent variable and x_i is the explanatory variable. β_o is a constant term called y–Intercept. All variables must have their slope coefficient (β_p). ϵ is the error term (also known as *residuals*).

The response variable has been the coagulant dosage, while all predictors have been the quality parameters selected at the inlet and outlet of the coagulation-flocculation process. To create MLR models, the function *lm* from the *Stats* library was used. The function *plot*, *points*, and *abline* from the library *Stats* have been used for data visualisation.

- Support Vector Machines (SVM)

SVM is a method capable of analysing data and recognising patterns using supervised learning methods. These methods are mainly used in classification and regression problems.

In our case, we used an ε – regression SVM (Eps–Regression, from now on), with a Linear kernel function as a supervised algorithm (F5.2). A good explanation of the different kernel functions and their impact on an SVM model is widely explained in Cortes & Vapnik, 1995.

$$K(x_i, x_j) = x_i^T x_j \qquad \text{(F5.2)}$$

The *svm* function from the *e1071* library has been used to create Eps-Regression SVM models. The *tune.grid* function has been used to study the cost variable (C) and gamma (ϒ) responsible for obtaining the prediction results. The function *plot*, *points*, and *abline* from the library Stats have been used for data visualisation.

The grid tunning and the general modelling of SVM have been performed following a specific procedure published by Hsu et al., 2016. As this procedure proposes, a simple scaling transformation of the data was performed. The next step was considering one of the four essential kernel functions (linear, polynomial, radial basis function and sigmoid). This choice is directly related to the intrinsic characteristics of the data, and all four functions were examined in order to conduct a thorough modelling study. After choosing the function, cross-validation using a tuning grid was performed. These steps were executed to set apart the best kernel parameters (C, ϒ). Once the best model was created and trained, it was time to test and validate it.

- Artificial Neural Networks (ANN)

An artificial neural network is an algorithm designed to simulate how the human brain analyses and processes information. It is the foundation of Artificial Intelligence (AI). An ANN model has three minimum parts: i) an inlet layer with all the input parameters selected, ii) a hidden layer with an activation function, and iii) the output layer, where the response variable is included.

The *neuralnet* library was used to develop our ANN model. *Plot*, *points*, and *abline* functions from the library *Stats* were used to visualize the results. Resilient backpropagation (rprop+) is the algorithm used to calculate the neural network (Riedmiller and Braun, 1992). The ANN Activation Function equation used is the Sigmoid Function (F5.3), and it is as follows:

$$S(x) = \frac{1}{1+\ e^{-x}} = \frac{e^{x}}{e^{x}+1} \qquad \text{(F5.3)}$$

Input and output layers vary depending on the parameters included in the model. However, hidden layers are always a great topic of discussion. Due to their importance in the overfitting (or underfitting) of the model, hidden layers play a significant role in the architecture of an ANN model. In many cases, the dimension selection of hidden layers is performed by trial–error, enhancing the idea of choosing the correct number of neurons and hidden layers is a hazy topic.

Many authors have described different methods to select the best number of hidden layers and neurons for each case (Kröse et al., 1993; Maier et al., 2000; Ke et al., 2008; Vujicic et al., 2016). For our case, an experimental design was developed to select the appropriate number of hidden neurons for our ANN model. This experimental design comprehended a *for loop* algorithm that added neurons into a hidden layer one by one. The results of all models have been compared to examine the differences between one to two hidden layers with one to ten neurons. This algorithm was applied in both scenarios.

- <u>Goodness of the models</u>

The goodness of the models developed has been observed by correlation (actual against predicted values) and RMSE (Root Mean Squared Error). These criteria allow keeping the predictability and the intrinsic error of the residuals obtained. The function *cor* from the *Stats* library was applied to achieve correlation results. A function to find RMSE (F5.4) was created using the formula:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(Predicted_i - Actual_i)^2}{N}} \qquad \text{(F5.4)}$$

5.3. Results and discussion

5.3.1. Water quality parameter selection

Selecting the right features before the model training can mean the difference between mediocre performance with long training times and outstanding performance with short training times. RStudio provides different tools to automatically organise the features by importance in a dataset.

Figure C5.5 shows the first 20 variables selected by RF. The variables are arranged by this algorithm from those that have the greatest impact on accuracy to those that have the least. The abscissa shows the percentage accuracy.

The blue line shows the threshold to limit the variable importance for the models. The variables that positively impact the model's accuracy are, in descending order, pH_o, flow, $turbidity_o$, $250nm_o$, $220nm_o$, $250nm_i$, $220nm_i$, pH_i, $turbidity_i$ and conductivity.

RF is a supervised classification method. This means that choosing a response variable—in our case, the dosage of the coagulant—is necessary in order to make a proper classification. The variables shown in Figure C5.5 have been chosen mathematically, considering the accuracy of a hypothetical model between each of them as predictors and the coagulant concentration as a response variable.

Besides, RF permutes each feature's values and measures how much the permutation decreases the model's accuracy as if it was cross-validation. Clearly, permuting unimportant variables should have little to no effect on model accuracy, whereas permuting important variables should significantly decrease it. The most significant ones are chosen in this manner. The first three variables (pH_o, flow, and $turbidity_o$) are the only ones that are always chosen first, regardless of the amount of data, after performing various tests to observe how the variables were organized according to the amount of data present in the training set (scenarios 1 and 2). The following water quality variables and wavelengths may vary in position. The selected wavelengths are always between 230 and 260 nm, though they can vary.

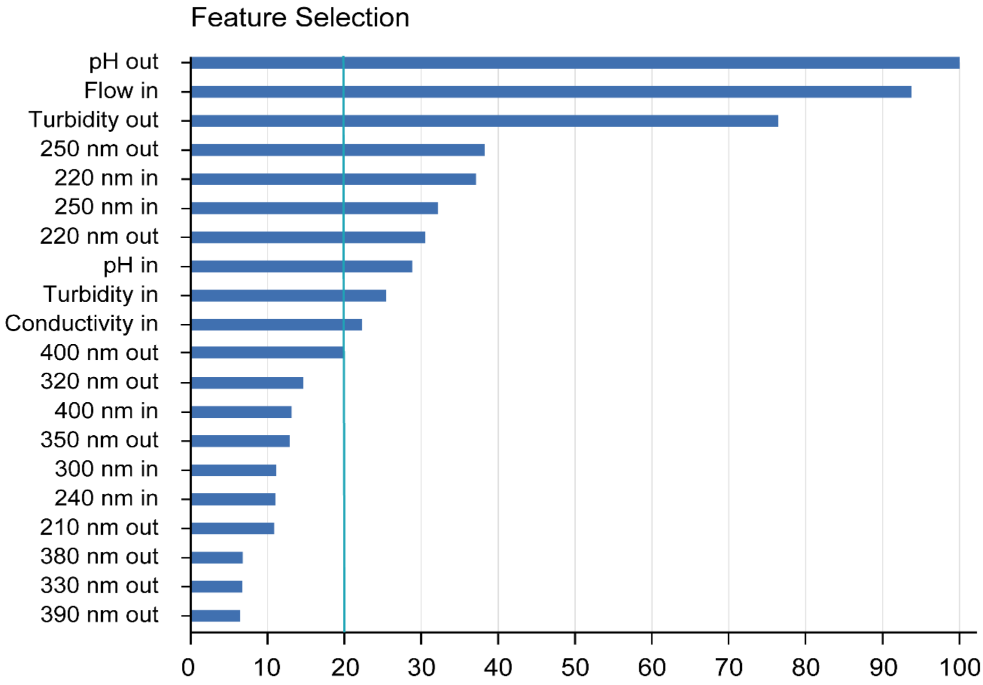


***Figure C5.5 –** Random Forest feature selection. Variables with the best accuracy are at the top of the list. The feature organisation provided by the RF algorithm is made by testing the accuracy of each variable in a hypothetical model, where coagulation dosage would be the response.*

Gauchi et al., 2001 and Menze et al., 2009 widely explain the data dependence in the benefit of a preceding feature selection. A gradual change in the data recorded can be expected in a natural environment. Variations in the response of automatic feature selection algorithms are motivated by the variability between the various data periods. For this reason, it is necessary to apply a feature selection preceded by cross-validation. In this way, the dependencies between variables can then be visualised and correctly selected.

In Chapters 3 and 4, the selection algorithms applied were Forward Selection (FS), Backward Elimination (BE) and Stepwise Selection (SS). In those two particular cases, the number of observations and variables were considerably higher than in this case.

Besides that, the multicollinearity between features was highly relevant. Accordingly, a balance between motor selection capacity, multicollinearity impact, and ease of method application and understanding was sought. However, the main reason for using the RF method rather than FS, BE, and SS was the number of variables against the number of observations. These selection methods cannot be adjusted when the number of predictors is close to the number of observations. Even though the observations recorded for this study exceed the minimum number of features needed, the sum of all inconveniences made the application of RF the most suitable approach for this case study.

Table C5.1– *Pearson's factor between variables used to create the mathematical models. In yellow, the variables highly correlated are observed. Coagulation has not had a direct correlation with any variable.*

	Coag.	**Cond**	**Flow**	**pH$_i$**	**pH$_o$**	**Turb$_i$**	**Turb$_o$**	**220$_i$**	**250$_i$**	**220$_o$**	**250$_o$**
Coag.											
Cond	0.27										
Flow	0.33	0.10									
pH$_i$	0.47	0.44	-0.18								
pH$_o$	-0.54	-0.02	-0.001	-0.15							
Turb$_i$	0.41	-0.03	0.01	0.22	-0.34						
Turb$_o$	0.52	0.28	-0.08	0.44	-0.21	0.51					
220$_i$	0.05	0.38	-0.03	0.15	0.06	0.03	0.18				
250$_i$	0.15	0.16	0.03	0.06	-0.20	0.29	0.22	0.41			
220$_o$	-0.18	0.36	-0.01	0.06	0.22	-0.37	-0.14	0.49	-0.18		
250$_o$	0.60	0.42	-0.002	0.52	-0.26	0.60	0.78	0.19	0.36	-0.30	

The wavelengths selected (220$_i$, 250$_i$, 220$_o$, 250$_o$) provide the suspended organic matter information. The changes registered in UV spectra between inlet and outlet Spectro::lysers® give essential information about the clarification process. RF algorithm detects those changes by selecting a combination of variables that can exhibit those changes.

A correlation table (Table C5.1) was performed to observe the direct relationships between the explanatory and the response variable selected to support this result.

***Table C5.2** – First 10 variables selected using Random Forest classification algorithm. This table also shows what sensors were used to register each parameter.*

	Predictor Selected	**Source**
Inlet – Regulation Tank	Turbidity	spectro::lyser®
	pH	pH::lyser®
	Conductivity	condu::lyser®
	Flow velocity	Vernier
	UV 220 nm	spectro::lyser®
	UV 250 nm	spectro::lyser®
	Vis 400 nm	spectro::lyser®
Outlet – Actiflo system	Turbidity	spectro::lyser®
	pH	pH::lyser®
	UV 220 nm	spectro::lyser®
	UV 250 nm	spectro::lyser®
	Vis 400 nm	spectro::lyser®

Table C5.1 shows low correlations among variables (between 0.2 and 0.6). The highest correlations are between turbidity ($Turb_i$ and $Turb_o$) and wavelengths 250_o. Coagulation does not show a direct relationship with any of the variables selected separately, but that does not mean there is no intrinsic relationship with more suitable sub-groups of variables in the database. In most cases, relationships are evidenced when the model's dimensions exceed the observable human capacity (more than three dimensions), demonstrating AI's effectiveness.

Table C5.2 summarises the parameters selected and the probes that have been used to register all data.

5.3.2. Scenarios evaluated

Two different scenarios were evaluated for the prediction of coagulation dosage.

As explained above, the first scenario was elaborated with a non–randomly split of training, testing and validation datasets. Figure C5.6 shows how the data has been distributed.

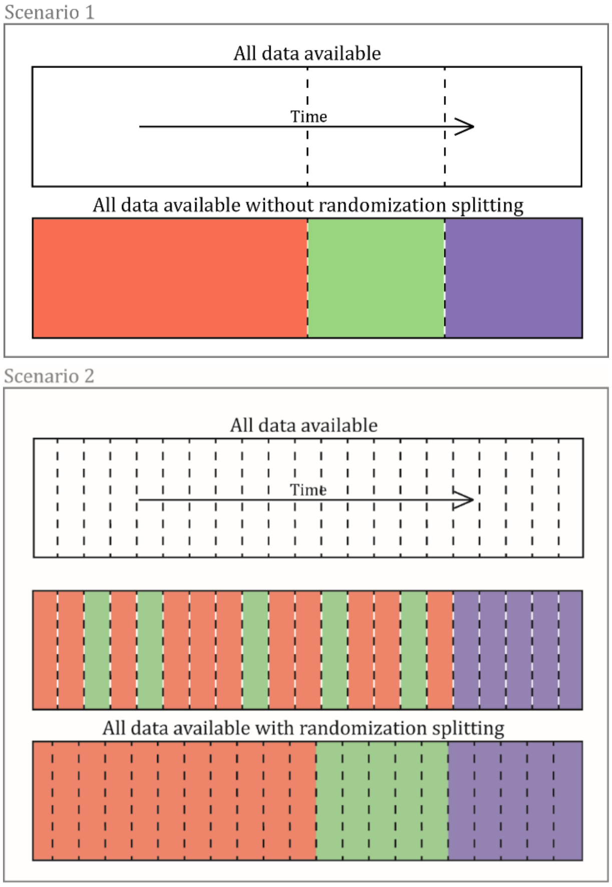

Scenario 1
All data available
Time
All data available without randomization splitting
Scenario 2
All data available
Time
All data available with randomization splitting

Figure C5.6 – *Graphical outline of the data distribution in scenarios 1 and 2. In scenario 1, the data was divided following a temporal distribution. This means there was no previous randomisation between training and testing datasets. In scenario 2, the data were clustered using the K-means algorithm. A training set was created gathering a part of each cluster, and the same for testing. The validation data set was separated before the K-means algorithm application.*

The second scenario was developed using the same data but guarantying the inclusion of all different populations in the training and testing set. Before, the last 1000 rows were separated for the validation process. Figure C5.6 shows an overview of data randomisation. The main idea is clustering the data following its intrinsic features, using the K-means algorithm, and then reorganising it considering a homogeneous aggrupation of information for the training set. A fraction of the information about each cluster must be collected in the training set. Therefore, the training can be as comprehensive as possible.

The K-mean clustering method is helpful for grouping data with similar characteristics and promotes a good distribution of the populations present in a dataset. This method was applied to command a representative sample of all populations recorded in the training set for scenario 2.

As an unsupervised method, the K-means algorithm requires a preceding definition of the number of clusters. Before applying this method, the Silhouette algorithm was executed to ascertain what number of clusters were suitable for the training dataset of scenario 2. Silhouette analysis was also employed to study the separation distance between the resulting clusters. Figure C5.7 is a Silhouette plot. It displays how many clusters will be favourable in the dataset. As shown in Figure C5.7, this algorithm marks 2 as the auspicious number of clusters.

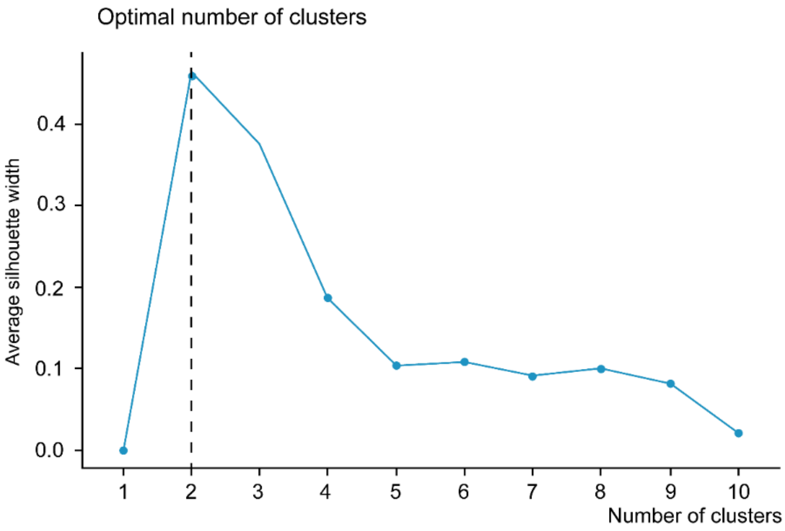


Figure C5.7 – *Silhouette method to select the optimal number of clusters present in the dataset for scenario 2.*

Two divisions were created used K-Mean clustering (Figure C5.8). In our case, the importance of dividing data by clusters lies in having a good information distribution in the training and testing datasets.

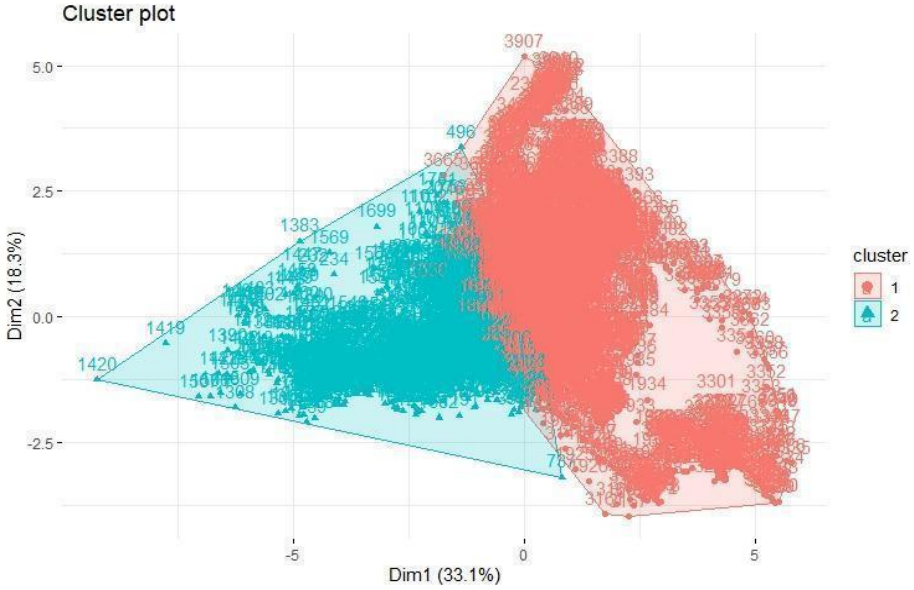


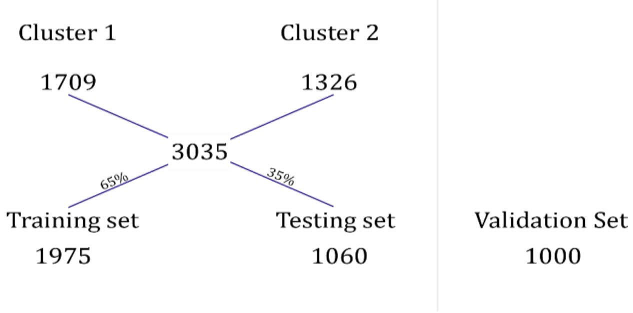
Figure C5.8 – *Data division by clusters. Exist an overlap between clusters. This could be caused because the data divided is continuous, and the water quality varies in an interrupted way in most cases.*

Overlapping clusters could cause problems in classification modelling. In the sight of overlap, the model would be unable to correctly classify the coincidental data, and significant errors would be generated. As the clustering was created to observe various wastewater sources as they were recorded by the sensors for six months, it would not be an issue in our case.

Figure C5.8 shows there have been two contrasting events. Since this data has the greatest distance to each centroid, these two episodes would be the farthest from the overlapping zone. If a very different–quality event had occurred, it would have been observed far from the two centroids under analysis, possibly forming a third data segment.

Training and testing sets have been established with an equitable distribution of both clusters. As beforementioned, the validation set was selected before the clustering process (Figure C5.6). Table C5.3 shows the data distribution for the second scenario. The data used for classification were approximately 76% of the total. Validation data is the remaining 24%. Of this 76% and 65% were selected for training, and the remainder was used for testing (35%).

Table C5.3 – *Data distribution to create the clusters and then, the training and validation databases. 76% of the data was used for clustering; the leftover was separated for validation. Of this 76%, 65% was used for training and the rest for the testing set.*



Clustering is an excellent way to make generalist and well-trained prediction models (Kuhn, 2008; 2013). For this case, it may be an incommodious methodology, as there is limited data. However, in situations where the data has been recorded for years, this system provides reliability to the global training and produces better results.

5.3.3. Model tuning

A previous tuning has been made for all of them to develop the mathematical models.

- <u>Multivariate Linear Regression tuning</u>

MLR models built for both scenarios have been performed in two steps. The first step has been used to observe the variable importance previously selected through the RF algorithm. The second step was used to discern the difference between these variables and decide whether to eliminate them, depending on their final performance.

To observe the weight of each variable in the model, *p-values* were used. The higher this value, the less critical the variable is. Tables C5.4 and C5.5 show the results obtained for each scenario and the two steps performed.

For scenario 1, the wavelengths $\lambda220_i$ and $\lambda250_i$ which exhibit a high p-value, would not be essential variables in this case. For the second scenario, the variables are Turbidity_i and $\lambda220_i$. The process of removing these variables and reperforming modelling is similar to a second verification of the initial variable selection.

As can be seen in both tables, the performance of the models (RSE and R^2) does not vary even if these variables are eliminated. For this reason, it was decided to keep all the variables selected previously in both scenarios.

Table C5.4 – *Results obtained from MLR for scenario 1. The two steps were employed to verify the quality of the variables selected using the RF algorithm.*

	Scenario 1							
	Step 1				**Step 2**			
	Estimate	**Std. Error**	**t–value**	**p-value**	**Estimate**	**Std. Error**	**t–value**	**p-value**
Intercept	0.48	0.02	30.69	$<2\cdot10^{-16}$	0.48	0.02	31.20	$<2\cdot10^{-16}$
Conductivity	-0.06	0.01	-4.81	$1.4\cdot10^{-6}$	-0.06	0.01	-5.15	$2.8\cdot10^{-7}$
Flow	-0.34	0.01	-32.88	$<2\cdot10^{-16}$	-0.34	0.01	-32.89	$<2\cdot10^{-16}$
pH$_i$	0.31	0.02	17.45	$<2\cdot10^{-16}$	0.31	0.02	17.40	$<2\cdot10^{-16}$
pH$_o$	-0.41	0.01	-31.21	$<2\cdot10^{-16}$	-0.41	0.01	-31.23	$<2\cdot10^{-16}$
Turbidity$_i$	0.03	0.01	-3.08	0.002	-0.04	0.01	-3.33	$<9\cdot10^{-4}$
Turbidity$_o$	0.13	0.02	5.73	$1.2\cdot10^{-8}$	0.13	0.02	5.70	$1.4\cdot10^{-8}$
$\lambda220_i$	0.027	0.01	2.03	0.04				
$\lambda250_i$	-0.15	0.02	-0.77	0.44				
$\lambda220_o$	-0.06	0.01	-5.55	$3.2\cdot10^{-8}$	-0.06	0.01	-5.25	$1.7\cdot10^{-7}$
$\lambda250_o$	0.19	0.02	9.36	$<2\cdot10^{-16}$	0.19	0.02	9.84	$<2\cdot10^{-16}$
	RSE: 0.06		**Adj. R^2:** 0.66		**RSE:** 0.06		**Adj. R^2:** 0.66	

Table C5.5 – *Results obtained from MLR for scenario 2. The two steps made were employed to verify the quality of the variables selected using the RF algorithm.*

	Scenario 2							
	Step 1				**Step 2**			
	Estimate	**Std. Error**	**t–value**	**p-value**	**Estimate**	**Std. Error**	**t–value**	**p-value**
Intercept	0.40	0.01	31.36	$<2\cdot10^{-16}$	0.40	0.01	31.53	$<2\cdot10^{-16}$
Conductivity	-0.06	0.01	-5.74	$1.06\cdot10^{-8}$	-0.05	0.01	-5.43	$<6.2\cdot10^{-8}$
Flow	-0.18	0.01	-21.56	$<2\cdot10^{-16}$	-0.18	0.01	-21.71	$<2\cdot10^{-16}$
pH$_i$	0.30	0.02	19.03	$<2\cdot10^{-16}$	0.30	0.02	18.92	$<2\cdot10^{-16}$
pH$_o$	-0.36	0.01	-29.70	$<2\cdot10^{-16}$	-0.36	0.01	-30.05	$<2\cdot10^{-16}$
Turbidity$_i$	-0.02	0.01	-2.08	0.038				
Turbidity$_o$	0.25	0.02	11.51	$<2\cdot10^{-16}$	0.25	0.02	11.37	$<2\cdot10^{-16}$
λ220$_i$	0.01	0.01	0.82	0.41		-		
λ250$_i$	-0.08	0.02	-4.53	$6.12\cdot10^{-6}$	-0.08	0.02	-4.89	$<1.07\cdot10^{-6}$
λ220$_o$	-0.09	0.01	-9.28	$<2\cdot10^{-16}$	-0.09	0.01	-9.80	$<2\cdot10^{-16}$
λ250$_o$	0.07	0.02	3.72	$2.07\cdot10^{-4}$	0.06	0.01	3.41	$<6.71\cdot10^{-4}$
	RSE: 0.06		**Adj. R^2:** 0.57		**RSE:** 0.06		**Adj. R^2:** 0.57	

- Support Vector Machine model tuning

The construction of an SVM algorithm involved the selection of two different values: i) cost value (C) that controls the severity of margin and the hyperplane, and ii) epsilon value (ε) that controls the margin of error tolerance.

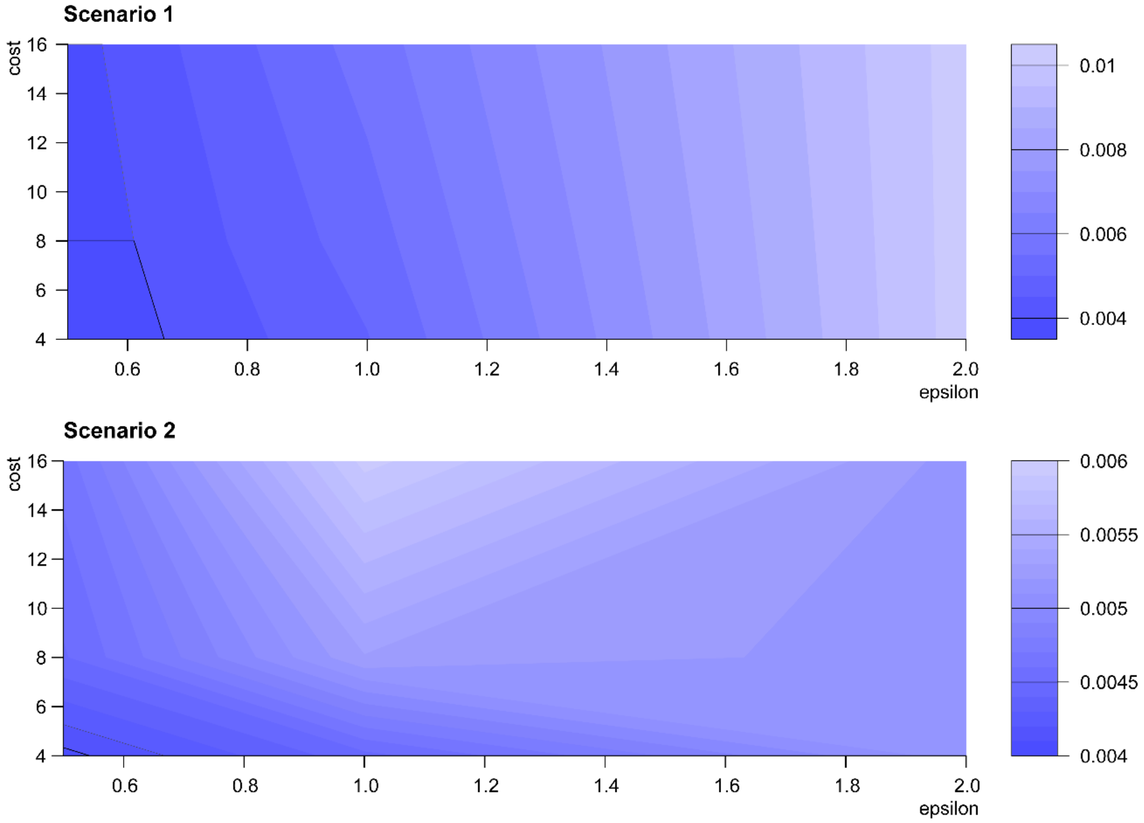


Figure C5.9 – *SVM tunning by cost and epsilon (C and ε) for scenarios 1 and 2, respectively. Tunning is made by applying 10– fold Cross-Validation to select the best modelling. Black squares denote the area where the best hyperplane is located. R selects de best model automatically.*

Figure C5.9 shows the results obtained performing the Grid Search approach on C and ε, using cross-validation. Several values (C, ε) were examined, and the best results were chosen. The colour gradient indicates each model's best performance (C, ε). The darker the area, the better the parameters fit the training set.

For scenario 1, the tunning performed indicated that the best values of C and ε were 4 and 0.2, respectively. For scenario 2, the best values obtained were 2 and 0.5, respectively (Table C5.6).

The results obtained through Linear Kernel allow a more profound and accessible understanding of how the models are built. Other available functions (previously indicated) are more complex and offer more hyperparameters to tune, adding complexity to the results. The complexity of the model directly impacts the possibility of overfitting it (Lameski, 2015).

Table C5.6 – *Parameters obtained tuning the SVM RBF. Costs (C) and Epsilon (ε) are the variables giving the best hyperplane found in the training data.*

Parameters	Scenario 1	Scenario 2
SVM–Type	Eps–Regression	Eps–Regression
SVM–Kernel	Linear	Linear
Cost (C)	4	4
Gamma (ϒ)	0.1	0.1
Epsilon (ε)	0.2	0.5

C and ε parameters are the SVM input and influence the optimisation process. Parameter C defines how hard or soft the classification margin should be. The higher the value of C, the lower the number of points allowed in the error margin.

Following this premise, the tuning function must be used carefully. This function studies the data and creates an adaptable model. Unrealistic and rigid models that are closely related to the data in the training set (overfitting) can be generated by exhaustive tuning or by performing this action repeatedly.

- ANN model tuning

The ANN tuning was performed by developing an experimental design. The selection of an accurate architecture was based on the idea that the model obtained must capture all the relevant information with the simplest topology.

This step was made for scenarios 1 and 2. Table C5.7 shows the analysis done for scenario 1, with all the possible neuron distribution only considering two layers. The correlation obtained for the selected architecture is highlighted in yellow. The final architecture chosen for this scenario was two hidden layers, the first one with eight neurons and the second one with six neurons.

***Table C5.7 –** Pearson's tests were performed to select the best architecture with less computing cost for scenario 1. Results obtained for each architecture are inaccurate. The highest correlation value is 0.38. Most values are rounding 0.1.*

		Layer 2									
		N1	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**
Layer 1	**N1**	0.01	0.06	0.05	0.06	0.05	0.06	0.07	0.08	0.06	0.08
	N2	0.14	0.14	0.1	0.08	0.09	0.15	0.04	0.10	0.10	0.13
	N3	0.12	0.09	0.09	0.18	0.09	0.24	0.10	0.15	0.08	0.18
	N4	0.05	0.05	0.25	0.14	0.09	0.07	0.08	0.11	0.08	0.10
	N5	0.12	0.11	0.16	0.16	0.17	0.10	0.20	0.16	-0.02	0.09
	N6	0.24	0.31	0.12	0.20	0.10	0.11	0.32	0.24	0.10	0.02
	N7	0.24	0.10	0.32	0.12	0.17	0.07	0.18	0.23	0.24	0.37
	N8	0.09	0.08	0.08	0.24	0.22	0.38	0.11	0.24	0.08	-0.01
	N9	0.23	0.08	0.22	0.19	0.23	0.26	0.33	0.31	0.18	0.30
	N10	0.28	0.10	0.19	0.30	0.06	0.26	0.26	0.07	0.29	0.25

Table C5.8 shows the analysis performed for scenario 2. The correlation obtained for the selected architecture is highlighted in green. The final architecture that was selected consisted of two hidden layers, the first of which had three neurons and the second of which had three neurons.

As can be observed in Table C5.7 and Table C5.8, the results obtained for each model are displaying contrasted correlations. Table C5.7 displays weak correlations that do not exceed 0.40. In contrast, Table C5.8 displays high correlations, in some cases, reaching 0.89. The amount of data available and *all-inclusive* training are two crucial points that indicate the future behaviour of a model.

Table C5.8 – *Tests performed to select the best architecture with less computing cost for scenario 1. In general, all results are similar. All values are between 0.85 and 0.87.*

		Layer 2									
		N1	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**
Layer 1	**N1**	0.002	0.03	0.76	0.78	0.78	0.78	0.78	-0.09	0.78	0.78
	N2	0.83	0.83	0.82	0.83	0.82	0.83	0.82	0.83	0.83	0.81
	N3	0.84	0.83	0.88	0.85	0.84	0.83	0.82	0.84	0.83	0.85
	N4	0.86	0.86	0.87	0.85	0.85	0.84	0.82	0.85	0.84	0.84
	N5	0.86	0.85	0.85	0.85	0.87	0.85	0.83	0.84	0.86	0.84
	N6	0.85	0.87	0.86	0.84	0.87	0.84	0.85	0.84	0.85	0.86
	N7	0.87	0.85	0.86	0.88	0.85	0.85	0.87	0.87	0.87	0.87
	N8	0.86	0.87	0.85	0.86	0.87	0.87	0.86	0.84	0.85	0.84
	N9	0.87	0.85	0.87	0.86	0.87	0.85	0.86	0.86	0.87	0.85
	N10	0.87	0.85	0.87	0.88	0.86	0.85	0.88	0.89	0.87	0.86

As shown in the cluster analysis, two different episodes of water quality were observed that have been recorded in different periods. The possibility of obtaining an appropriate prediction is limited if neither is present in the training set, leading to outcomes similar to scenario 1.

Scenario 2 is well trained through clustering. This fact allows for obtaining better results without having lots of data. Scenario 1 would require a topology with more layers. This, in the end, would also be unnecessary since it would be an overtrained model, and the results would serve only for a particular subset of data.

Figure C5.10 shows the distribution of the neural network corresponding to scenario 2. The distribution by layers and neurons can be observed. ANN conform to the basis of deep learning, a subfield of machine learning where the structure of the human brain inspires the algorithms. Neurons from one layer are connected to neurons of the next layer through channels. These channels are assigned to a numerical value known as weight (black values are shown above each arrow in Figure C5.10).

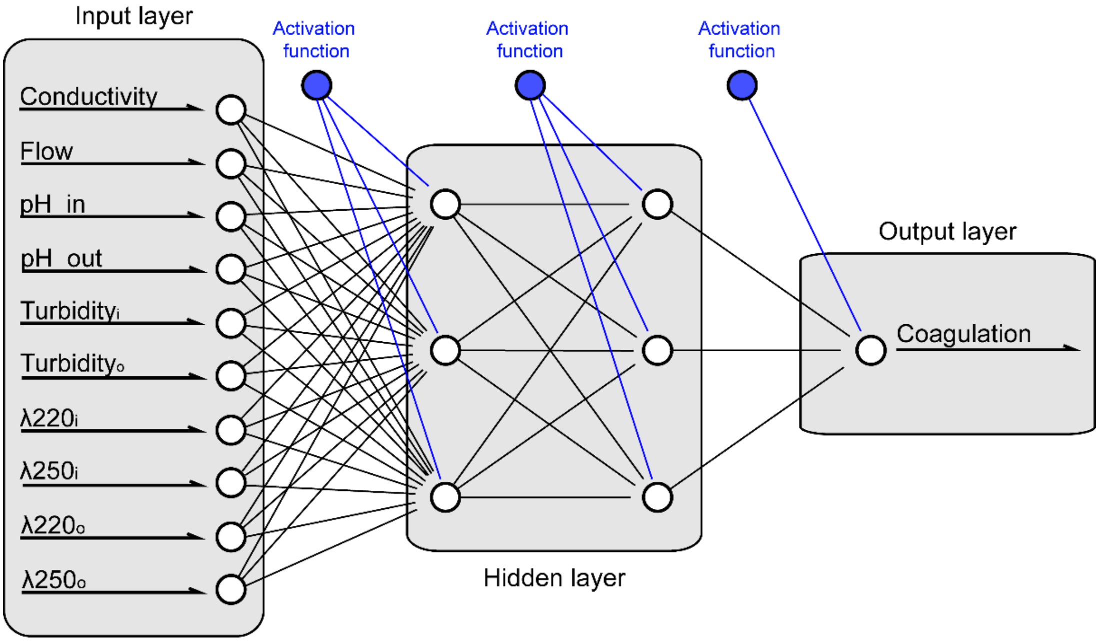


Figure C5.10 – *Artificial Neural Network for scenario 2. All parameters selected are at the inlet layer and the coagulant dosage at the output layer. This architecture is formed by two layers, with three neurons per layer.*

The inputs are multiplied by the corresponding weights, and the sum is then sent as an input to the neurons in the hidden layer. Each of these neurons is associated with a numerical value called bias, which is added to the input sum (values in blue for each hidden layer and outlet layer in Figure C5.10). The result is then passed through a threshold function called Activation Function (F5.3). The results of the Activation Function determine whether or not a particular neuron will get activated. Activated neurons transmit data to the neurons in the next layer, if any, or to the output layer, if none exist.

5.4. Prediction capacity of MLR, SVM and ANN models

The predictability has been studied for all three algorithms in the two established scenarios. Table C5.9 shows the results for both scenarios. The first three columns belong to the MLR, SVM, and ANN algorithms for scenario 1. The following three columns are for the same algorithms for scenario 2.

As can be seen, for scenario 1, the three algorithms tested obtained similar RSE values (between 0.03 and 0.06). On the other hand, slightly higher values were observed for R^2 in SVM and ANN algorithms than the value obtained for MLR (0.72, 0.76 and 0.66, respectively). For all three algorithms, the predictability for the testing set was reduced (0.22, 0.12 and 0.12). Finally, the Validation set showed an improvement (0.75, 0.67 and 0.59).

Scenario 2 exhibited a significant improvement for SVM and ANN algorithms, R^2 (0.89 and 0.86), and prediction values from the testing and validation set. In contrast, the value of R^2 for MLR decreases from 0.66 to 0.57.

***Table C5.9** – Values obtained from RSE and R^2 in testing and validation sets (actual vs predicted). All three algorithms improve their prediction results considerably in scenario 2.*

	Scenario 1			**Scenario 2**		
	MLR	**SVM**	**ANN**	**MLR**	**SVM**	**ANN**
RSE	0.06	0.04	0.03	0.06	0.04	0.03
R^2	0.66	0.72	0.76	0.57	0.89	0.86
R^2 Predicted VS Real	0.22	0.12	0.12	0.73	0.72	0.85
R^2 Validated VS Real	0.75	0.67	0.59	0.71	0.81	0.78

The values obtained seemed logical given the training differences in both scenarios. Scenario 1 was trained with a particular part of the data. However, scenario 2 was trained, including a larger sample of trends present in the data (see Figure C5.6).

An important fact observed in the results obtained is the low predictive capacity for the testing set, and the considerable improvement in the validation set, for the three algorithms in scenario 1. This can be related to the quality of the water recorded since there was an event of poor water quality that lasted a few days. The training made in scenario 1 did not include this event. That would explain the determination indexes in the three algorithms' testing sets (0.37, 0.21 and 0.12, respectively). The resulting models would not be able to predict significantly different ranges of data from those used for training.

The determination indexes from validation data showed an essential improvement for all modelling (0.75, 0.67 and 0.57, respectively). The wastewater quality recorded in the validation set was comparable registered for the training set. This can lead to an improvement in the prediction results for all models.

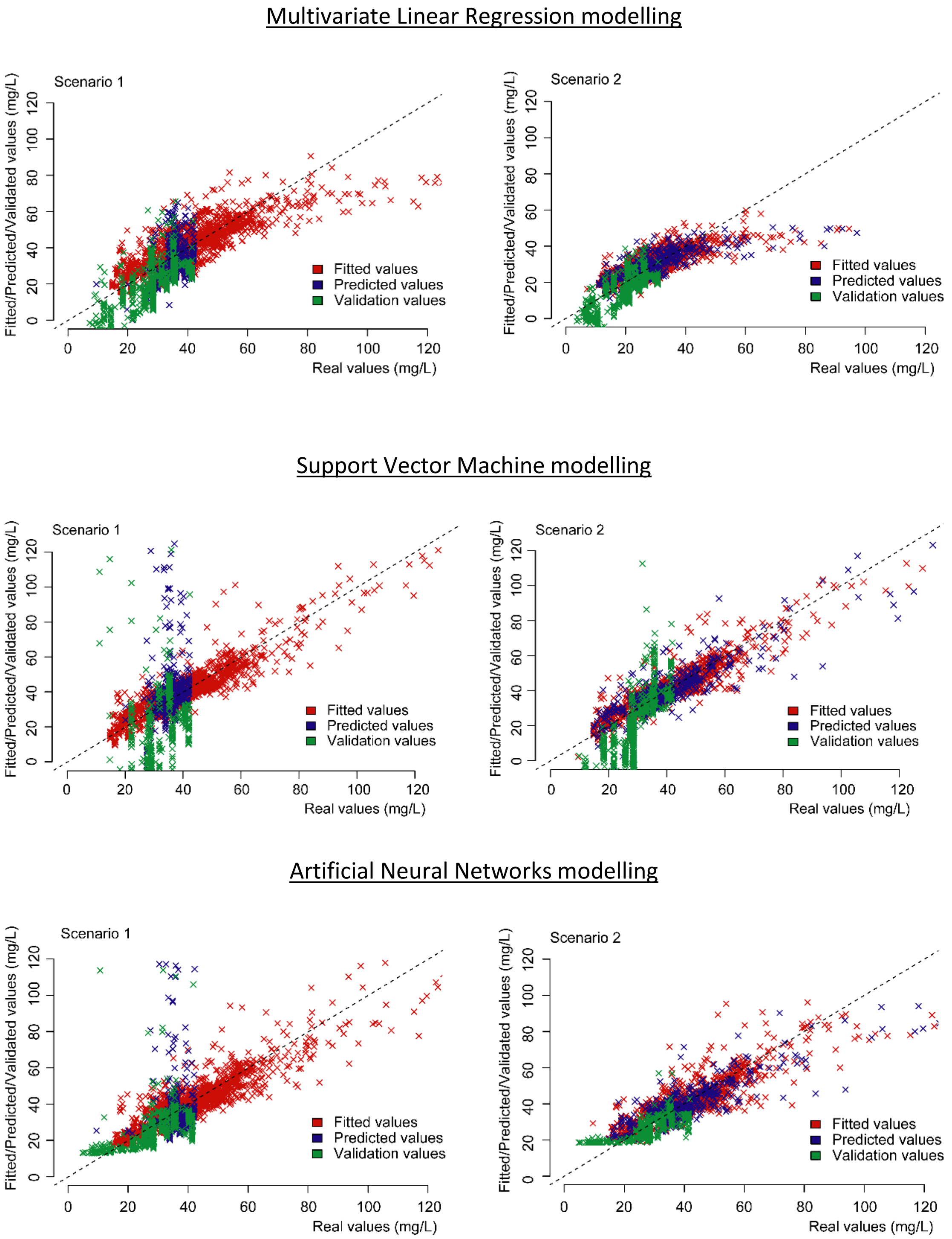
In a general overview, scenario 2 showed a significant improvement in the results obtained for the three models (Table C5.9).

The values obtained for the determination indexes in both the prediction and validation set for the MLR model were high quality (0.73 and 0.71, respectively). The same result was observed for the SVM (0.72 and 0.81, respectively) and ANN (0.85 and 0.78, respectively).

A prediction model has a vital requisite: it must be robust and flexible to distinguish a wide range of relationships between data. This premise is based on making predictions from data and emphasizes the value of developing a training set that contains as many populations present in the database as is practical.

Figure C5.11 shows the correlations for the three models in both scenarios. The red, blue, and green dots represent the training, testing and validation sets, respectively. In a general overview, for scenario 1, data appears a little disorganized, with a slight diagonal tendency but without a strictly straight correlation. SVM would present higher dispersion. Some prediction dots have a vertical trend towards positive values, practically perpendicular to the training data. ANN shows a similar behaviour, with a lesser extent of vertical dispersion for prediction values.

The validation dots have vertical distribution toward negative values for both models. For the ANN model, the dispersion of these validation values is smaller, with nearly all values falling on the diagonal (dotted line). MLR would exhibit the best diagonalisation. Although, a poor prediction values distribution is observed (they are clustered in a small graph area). The MLR model has a reasonable validation values distribution that, despite having a vertical tendency towards negative values, shows some diagonalization of the results.



Multivariate Linear Regression modelling

Support Vector Machine modelling

Artificial Neural Networks modelling

***Figure C5.11** – Comparison of MLR, SVM and ANN prediction models, separated in scenario 1 and scenario 2. The abscissa axis shows the actual values, and the ordinate axis shows the calculated ones. Red, blue and green dots are fitted, predicted, and validation values. As can be seen, the capability of predicting is closely related to the training made, being much better in the figures on the right, graphs belonging to scenario 2.*

Scenario 2 showed better organised and diagonalised trends for the prediction and validation dots in the three graphics depicted. As can be seen compared to the outcomes from scenario 1, each model's correlation of validation values considerably improved.

Overall, the data was better distributed, without significant vertical dispersion nor observed over-grouping.

In general, SVM and ANN provided more accurate predictions when their training efforts were more comprehensive, or the prediction data was similar to the training data. Although, the best inference was obtained by the ANN algorithm because there was not a wide dispersion for its validation values.

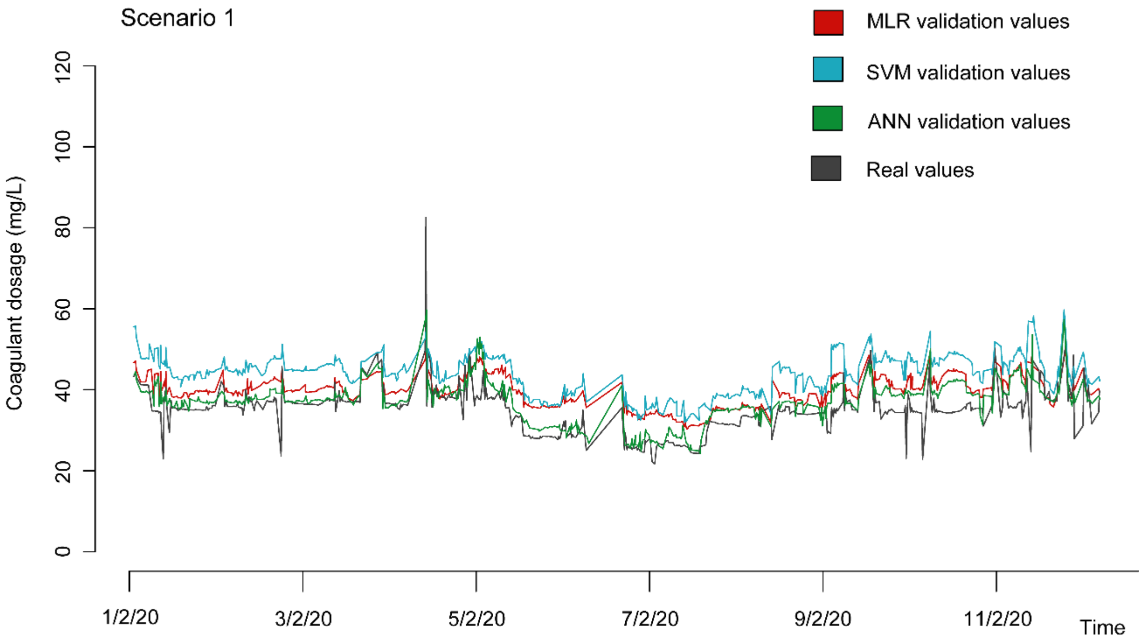


***Figure C5.12** – Coagulant dosage VS time. It shows the inference of the three models in scenario 1. A general overprediction is observed.*

Figures C5.12 and C5.13 correspond to coagulant predictions over time. To compare the MLR, SVM, and ANN models' predictive abilities, the same days were selected for both scenarios. The grey line represents the actual data of coagulant dosage. The red, blue, and green lines represent the outcomes of the MLR, SVM and ANN predictions, respectively.

Figure C5.12 highlights the inference of the three models in scenario 1. As can be seen, MLR (red line) shows a high and non-realistic overestimation. Given the prediction made by MLR, an addition from 30 to 40% of coagulant would have to be applied. Something similar happens with the SVM model (blue line). Most of the predictions made by this model are also overestimated. At some points, the SVM model considers a further addition from 5 to 10% of coagulant, which could be reasonable values, but at some points, it predicts an addition of up to 20% more. On the other hand, the ANN model (green line) predicts plausible coagulant values. Although, in the same way as SVM, the ANN model overestimates the coagulant addition at some points.

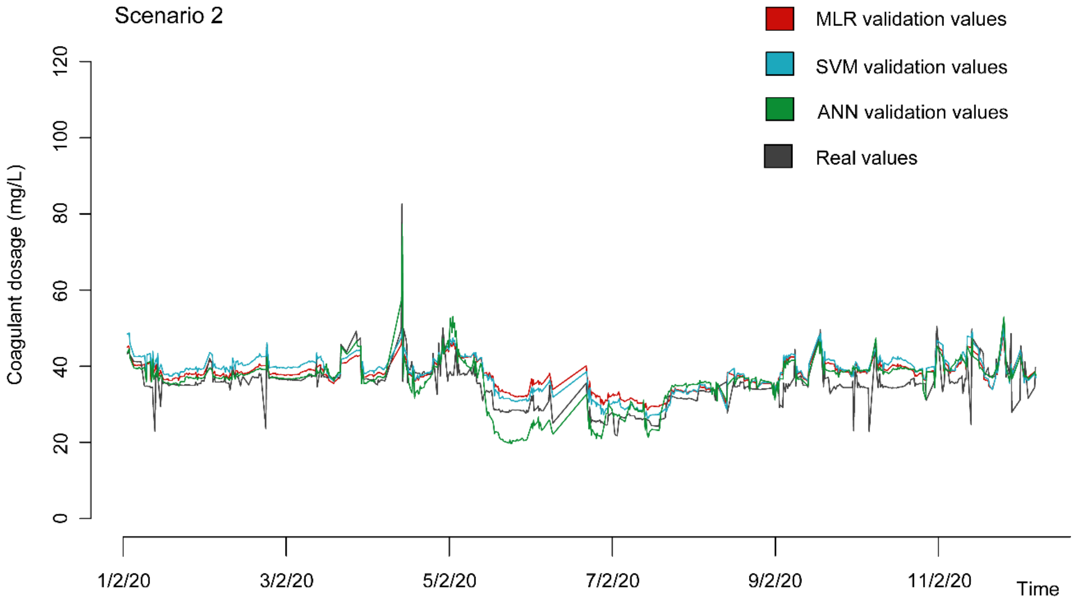


Figure C5.13 – *Coagulant dosage VS time. It shows the inference of the three models in scenario 2. The predictive accuracy has been widely improved. All three models can follow the real trend of the data.*

The effects of a highly contaminated wastewater event reported at the plant at the beginning of November may be one reason why all three models tend to overestimate the actual values in Scenario 1. Over the course of nearly ten days, a high turbidity input was recorded. To control that contamination event, a higher amount of coagulant was added. However, the other quality parameters did not vary correspondingly. Turbidity and colour were both greatly impacted by the contamination event's characteristics, but other parameters like pH and conductivity did not modify their tendency. The models' training learns about the parameter's tendencies and their relationships. An abrupt change in one or two parameters could create an artifact in the model, decreasing the accuracy of the predictions, and producing biased results.

In Scenario 1, predicting such different quality data with little variation in the training data may lead to overestimated predictions. To minimise this situation, it would be necessary to conduct a more comprehensive training that could cover a broader range of situations.

Figure C5.13 shows the predictions of Scenario 2. In general, all three models can remarkably follow the actual data tendency. Trends can be imitated by MLR and SVM (represented by red and blue lines, respectively). However, they typically overestimate the applied values and predict higher coagulant dosages than were added. ANN (green line) is the model that best follows trends from real data. On some days, the actual coagulant concentration is much higher than predicted. The values of coagulant added between February 5 and 7 were about 30 mg/L, whereas the ANN model predicted values of between 20 and 25 mg/L. This meant that the coagulant dosage could be potentially reduced by 25-35% in this period. The explanation for this decrease in the predicted coagulant concentration by the ANN model is related to the input and output turbidity registered at the WRP during those days. Both parameters recorded much lower values than usual. ANN can infer the prediction of coagulation in Scenario 2 better than in Scenario 1, which indicate an apparent reduction in the coagulant concentration that should have been added during those days.

The three prediction models tended to overestimate coagulant addition from February 9 to February 11. The actual values showed an added coagulant concentration of 35 mg/L, whereas the predicted values were 40 mg/L. This would result in an increase of the coagulant addition of almost 12%. All three models indicate this increase in coagulant concentration. That can be related to a trained artifact, a similar situation as explained in Scenario 1.

Figures C5.12 and C5.13 show only twelve days of predictions. Nevertheless, for the entire month of February, the prediction of the added coagulant concentration was performed. On average, in February, MLR, SVM, and ANN predicted coagulant concentrations over 35% higher than actual coagulant doses. For scenario 2, MLR and SVM predicted a rise of 12 and 7% in coagulant addition, respectively. However, ANN exhibited, on average, a 4% coagulant reduction. Presumable, comprehensive training is crucial for obtaining proper predictions. The modelling inference drastically changes if more information is added to the training set.

Consequently, ANN provides a more realistic prediction values for coagulant addition, it is regarded as the best prediction model. On the other hand, SVM and MLR could not predict the actual values in the same way, showing, in most time intervals, overestimated values with a higher error range.

Coagulation dosage is one of the essential parts that must be considered to determine the optimum condition for the performance of the coagulation-flocculation process. The incorrect coagulant dosage would lead to consequences in flocculation performance and high consumption of chemicals and energy, which would produce a negative economic and environmental impact on the WRP. The application of machine learning models such as ANN in the day-to-day WRP work could help reduce chemical overuse. For instance, considering the 4% decrease in coagulant addition predicted by the ANN model, the average amount (485 T) of coagulant added in a regular month of treatment in the WRP and the price of a tone of coagulant (≈ 350 €/T), the application of a well-trained machine learning model in the decision control of coagulant dosing would imply a saving of ≈ 6,800 € per month, which would be equivalent to ≈ 81,400 € per year. The determination of optimum dosage can minimise the dosing costs and the final sludge formation and achieve optimum performance in treatment (Patel and Vashi, 2013; Saritha, 2017).

In conclusion, this type of model can be used to allow operators to respond effectively, diminishing chemical overuse, and maintaining a good quality of water at the treatment outlet.

5.5. Conclusions

Based on the coagulant dosage added and water quality parameters recorded using two spectro::lyser®, this study demonstrated the predictive capacity of ANN over MLR and SVM model in a real application coagulant addition in WRP The ANN model developed could be used by plant operators as a tool for a quick response in WRP if required.

MLR, SVM and ANN were tested in two different scenarios, i) without extended training (scenario 1) and ii) with extended training (scenario 2). The predictability of the three models was found to improve significantly in scenario 2. Comprehensive training of a model was crucial to ensure accurate representation of the models. Water quality can vary over the year in the WRP, and this variation will depend on the weather (heavy rains, droughts, etc.) and occasional discharges. If models have not been trained correctly, it is quite possible that sudden changes in water quality cannot be well captured, leading to prediction that are ineffective.

Variable selection was also an essential step in providing reliable predictions. The predictive capacity of a model was closely linked to this selection because choosing unimportant variables selected would generate noise in future predictions.

The conductivity, water flow, pH_i, pH_o, $Turbidity_i$, $Turbidity_o$, and wavelengths 220_i, 250_i, 220_o, and 250_o were the most critical variables in the coagulation-flocculation process to conduct the prediction models.

As a result, ANN provided more reliable results than SVM and MLR in both scenarios and best fit reality with the least amount of error.

Capítol 6 – Discussió

6.1. Efecte de la variabilitat de les dades en la modelització

La modelització matemàtica està estretament lligada a la qualitat de les dades que la conformen. La dependència de la qualitat d'un model a les dades que el suporten és tan gran que, en molts casos, no es poden arribar a separar dels resultats que se n'obté. Per això és tan important conèixer profundament les fonts de dades, la seva organització i com tractar-les abans de començar una modelització matemàtica.

En aquesta tesi, s'ha seguit una metodologia d'enregistrament, tractament i selecció de dades uniforme, que fos eficient i flexible amb la tipologia d'informació amb que es treballava: l'espectre òptic Ultraviolat-Visible (UV–Vis) d'aigua crua, potable i residual urbana.

L'absorbància de l'espectre òptic UV–Vis conté informació rellevant de la qualitat de l'aigua a investigar, i varia considerablement depenent del tipus d'aigua enregistrat (descripció de la Llei de Lambert-Beer a l'apartat 1.5). En qualsevol espectre òptic tindrem absorbàncies més elevades a les longituds d'ona baixes (on s'observa la matèria orgànica dissolta), i absorbàncies més baixes a longituds d'ona més altes (on s'observa, per exemple, el color), tot i que la magnitud de l'absorbància registrada variarà en funció de la tipologia d'aigua. En aigua crua i potable, l'absorbància registrada serà de magnituds més baixes (p, ex. 20 i 1,2 Abs/m, respectivament), que per l'aigua residual. Aquesta última pot tenir absorbàncies molt elevades (p. ex. de 50 a 2000 Abs/m o més) depenent de la seva procedència (urbana o industrial).

Per a realitzar les modelitzacions matemàtiques d'aquesta tesi, s'han utilitzat aigües de diferents tipologies. Per a la detecció i predicció del potencial de formació de trihalometans (THM FP) s'ha utilitzat la informació de l'espectre òptic d'aigua crua i aigua tractada a diferents punts de l'Estació de Tractament d'Aigua Potable (ETAP) del Consorci d'Aigües de Tarragona (CAT) (L'Ampolla, Catalunya, Espanya) al llarg d'un any. Per a la detecció de toluè, m-xilè i p-xilè en aigua residual, es van utilitzar els espectres òptics d'aigües residuals urbanes de l'influent de depuradora, registrats a diferents punts de l'Estat Espanyol en diferents projectes duts a terme per l'empresa s::can Iberia Sistemas de Medición, S.L.U. Per modelitzar la quantitat òptima de coagulant que s'havia d'afegir al sistema de coagulació i floculació Actiflo® de Veolia es van registrar absorbàncies al tanc d'homogeneïtzació de l'aigua d'entrada a l'Estació Regeneradora d'Aigua (ERA) del Camp de Tarragona (Tarragona, Catalunya, Espanya) i de sortida del sistema Actiflo® durant sis mesos. La Figura C6.1 mostra les diferències d'absorbància entre uns espectres òptics i altres.

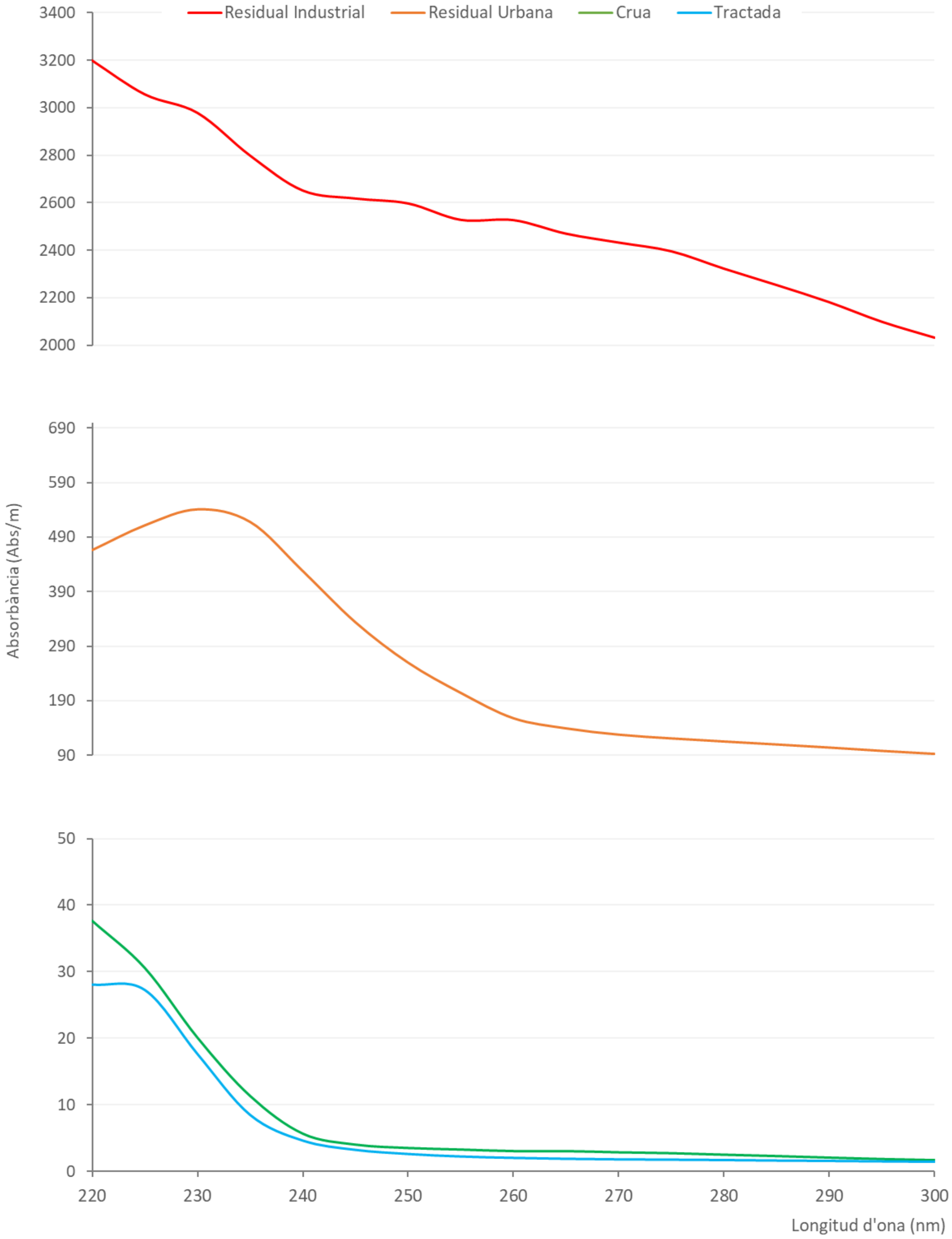


Figura C6.1 – *Exemple d'espectres d'absorbància de diferents tipologies d'aigua estudiades durant la present tesi. Les longituds d'ona estan tallades a 300 nm per motius de visualització.*

En tots els casos d'estudi esmentats, es van enregistrar espectres òptics durant llargs períodes de temps. Això va comportar obtenir molta informació d'un mateix punt o procés, on la qualitat de l'aigua podia variar degut a l'estacionalitat, a possibles canvis en l'origen, o a canvis en els tractaments previs a l'enregistrament, entre d'altres factors. Les diferències de l'espectre UV – Vis enregistrades al llarg del temps en un mateix punt es van considerar part de la variabilitat inherent de les dades dels projectes i per tant, també es van utilitzar en les modelitzacions. Tot i així, en alguns casos concrets, aquesta variabilitat s'ha estudiat com una possible amenaça del bon funcionament predictiu dels models.

Els tres casos d'estudi desenvolupats en aquesta tesi demostren, de manera il·lustrativa, l'efecte de la variabilitat de les dades en models matemàtics de predicció. Pel cas d'estudi en el que es va desenvolupar un algoritme de predicció del THM FP en aigua potable, es va observar que la dinàmica de les absorbàncies registrades no variava abruptament, mantenint-se en uns barems d'entre 20 i 60 Abs/m, pel cas de l'aigua crua, i al voltant de 1,5 Abs/m pel cas de l'aigua tractada. Això va permetre aconseguir resultats de predicció molt estables i fiables en el temps utilitzant un algoritme de xarxes neuronals artificials (ANN).

Per altra banda, i anant al cas d'estudi més extrem, durant el desenvolupament de l'algoritme de predicció d'hidrocarburs en aigua residual, es va tantejar la possibilitat de crear un calibratge global, és a dir, un algoritme capaç de detectar una mínima concentració d'hidrocarburs, fos quina fos l'aigua que s'estigués investigant. Els primers models matemàtics es van realitzar seleccionant més de vint tipologies d'aigües residuals (urbanes i industrials). Les dades d'absorbància presentaven una variabilitat molt gran on, en alguns casos, el diferencial era de 60 Abs/m en les longituds d'ona més altes, i en altres casos podia ser superior a 2000 Abs/m. Després de realitzar les modelitzacions pertinents, i d'estudiar-ne els resultats, es va arribar a la conclusió que un model matemàtic no podia copsar dades tant dispars per predir concentracions tant baixes d'hidrocarburs (més del 75 % d'error en la predicció per a la majoria d'aigües residuals). Aquest fet, va dur el cas d'estudi a la modelització discretitzant per tipologia d'aigua, és a dir, generar un model matemàtic de predicció d'hidrocarburs per aigua residual urbana, i un altre de diferent per aigua residual industrial. Després de diferents proves, van quedar ambdós descartats ja que, tot i que les característiques principals d'absorbància de les aigües que conformaven cada grup estaven més a prop, les diferències de magnitud encara eren massa grans per aconseguir bons resultats de predicció (entre el 30 i el 50 % d'error en tots els casos). Finalment, es va optar per crear un model matemàtic de predicció d'hidrocarburs específic per a cada tipologia d'aigua residual, només en l'àmbit urbà i només utilitzant aquelles aigües amb una quantitat i qualitat de dades que permetés una bona modelització matemàtica.

Tal i com s'ha esmentat en l'apartat previ, la naturalesa canviant dels espectres òptics està lligada a la qualitat de l'aigua que s'enregistra en aquell moment. Donat que, ni les dinàmiques naturals com l'estacionalitat, ni les dinàmiques antropogèniques, són factors controlables, els models matemàtics que sorgeixin de l'estudi d'espectres òptics de l'aigua han d'incloure aquest factor com a rellevant. Les variacions abruptes i sense patró de repetició concret, que poden ser naturalment observables en la qualitat d'aigua, afecten en gran mesura a la capacitat de predicció dels models matemàtics, ja que fomenten una percepció estocàstica de la informació inclosa i n'entorpeixen la qualitat del càlcul en gran mesura.

Per resoldre l'impacte de la variabilitat en la capacitat de predicció del model matemàtic concret s'ha de disposar d'una gran quantitat de dades per entrenar-lo. La quantitat variarà en funció de la necessitat de l'algoritme utilitzat, i també del nombre de variables independents escollides. En tot cas, serà necessari que el mateix model hagi visualitzat i entès la gran majoria de patrons que trobarà al llarg de la seva vida predictiva per obtenir-ne el major profit, i això, en molts casos, serà complicat d'assolir.

6.2. Efecte de la matriu d'aigua i el contaminant en la modelització

La variabilitat de les bases de dades no és l'únic factor que s'ha de tenir en compte a l'hora de crear un model matemàtic robust i a la vegada flexible. Altres factors d'interès a tenir en compte són les característiques òptiques de la matriu d'aigua que es vol treballar, el tipus de contaminant que es vol predir i el rati senyal-soroll (S/S).

En aquesta tesi, i tal i com s'ha explicat anteriorment, s'han utilitzat les absorbàncies de dos punts diferenciats de l'ETAP del CAT per predir el THM FP (**capítol 3**) ja que, els trihalometans no absorbeixen a rangs de longituds d'ona UV–Vis i, per tant, no n’obtenim cap senyal directa observable. Així, per a obtenir el THM FP, va ser necessari estudiar l’espectre de l’aigua captada del riu amb una concentració total de matèria orgànica registrada, i l’espectre d’aigua tractada i clorada, amb una concentració mínima o nul·la de matèria orgànica, pressuposant que tota la matèria orgànica existent reaccionava amb el clor, transformant-se en trihalometans.

En aquest cas, la matriu d’aigua no influiria directament en la modelització, i per això es va utilitzar com a eina per a observar els canvis en l’espectre, i relacionar-los amb els valors de THM FP registrats a través d’analítiques i també a través de la informació de l’algoritme específic d’Amy et al., 1998, avaluat i contrastat per a dur a terme aquest propòsit.

En canvi, en el cas d’estudi on es va voler predir la concentració de toluè, m-xilè i p-xilè en aigua residual urbana provinent d’influent d’EDAR, la matriu jugava un paper molt més rellevant (**capítol 4**). Els hidrocarburs en general, i els seleccionats per l’estudi en particular, tenen una corba d’absorbància concreta, que varia en funció de la seva estructura molecular (senyal d’estudi). Com més concentració de contaminant, més gran és la magnitud de la corba d’absorbàncies que es genera (comportament lligat a la Llei de Lambert-Beer, especificada a l’apartat 1.5). Les longituds d’ona on s’observava més absorbància de toluè, m-xilè i p-xilè era al rang 200 – 220 nm (Figura C4.2), però es va observar que la matriu d’aigua residual (soroll de base) també ho feia. En aquest rang de longituds d’ona, el rati senyal d’estudi – soroll de base (rati S/S) era molt baix, ja que la matriu emmascarava la senyal del contaminant completament, complicant, en gran mesura, la modelització i la capacitat de predicció del model matemàtic. En canvi, en el rang 255 – 270 nm, s’observava un segon pic d’absorbància dels contaminants i, en aquest interval, les matrius d’aigua residual urbana no registraven absorbàncies tan altes perquè, generalment, estaven composades de matèria orgànica, que absorbeix a longituds d’ona més baixes. Aquest fet feia augmentar el rati S/S, millorant una possible detecció d’hidrocarburs en aquest interval.

Tot i observar un augment del rati S/S en algunes parts de l’espectre, és inqüestionable la relació física directa entre contaminant (senyal que es volia estudiar) i matriu (soroll de base), i la limitació en la modelització que això suposa. En casos més complexes, on la matriu d’aigua residual era, per exemple, d’influent industrial, es feia molt complicat obtenir resultats on no s’observés un error de predicció de més del 60%, i pràcticament impossible si el que es volia era crear un model de predicció que englobés més d’una matriu d’aigua residual. En aquests casos esmentats, el rati S/S esdevenia tan baixa, que el model no podia arribar a discernir entre la senyal provinent de l’absorbància dels hidrocarburs d’estudi i el soroll derivat de la matriu. Tot i així, es va observar que l’enfoc multivariat utilitzant diferents longituds d’ona de l’espectre UV–Vis és la clau de volta per a poder obtenir models matemàtics de predicció que puguin ser útils, sempre i quan es desenvolupin tenint en compte el tipus de matriu i els possibles canvis de pendent que aquesta pugui registrar.

En canvi, en la monitorització del sistema de coagulació-floculació Actiflo® de Veolia (**capítol 5**) per a predir la concentració de coagulant que s’ha d’afegir a l’aigua, l’estudi multivariat no només es va dur a terme utilitzant algunes longituds d’ona de l’espectre UV–Vis de la matriu d’aigua, sinó que també es van utilitzar paràmetres fisicoquímics (pH, concentració de coagulant o terbolesa, entre d’altres). En aquest cas d’estudi es va utilitzar la informació de dos punts de l’ERA del Camp de Tarragona per predir la concentració de coagulant i, per tant, el que es pretenia era estudiar el canvi de comportament dels paràmetres fisicoquímics en funció de l’addició de coagulant al llarg del tractament Actiflo®. En aquest cas, la matriu d’aigua també juga un paper important en els resultats de predicció que s’obtenen de la modelització.

Durant els sis mesos d’enregistrament de dades es va detecta un canvi dràstic puntual en l’absorbància registrada a l’entrada del procés de coagulació-floculació. Aquest canvi de tendència va durar uns dies i va ser important. Per a observar l’impacte d’aquest canvi de pendent en l’absorbància de la matriu d’aigua residual es va decidir crear dos tipus d’entrenament: un que fos excloent, tenint en compte un conjunt de dades en les quals no hi havia aquest esdeveniment puntual, i un que fos inclusiu, que les tingués en compte. En aquest cas, els canvis bruscos en l’absorbància de la matriu d’aigua residual podien generar inestabilitats en el model desenvolupat, creant incongruències en l’entrenament i disminuint la seva capacitat de predicció. Els resultats de l’entrenament inclusiu respecte a l’excloent van ser determinants: el model matemàtic tenia molt més poder de predicció si s’entrenava de manera extensiva. Per tant, un dels requisits per generar models matemàtics que siguin flexibles i adaptables als canvis de matriu que pot generar un esdeveniment de mala qualitat en l'aigua residual, és introduir, de manera regular i controlada les noves dades registrades, en el cas que es consideri que aquest esdeveniment podria tornar a succeir.

6.3. Motors automàtics de selecció de variables

En tots els casos d'estudi desenvolupats en aquesta tesi s'han aplicat motors automàtics de selecció de variables com a primer pas per a la reducció de dimensions. Les variables independents utilitzades en els **capítols 3** i **4** són longituds d'ona seleccionades dels espectres òptics ultraviolat-visible sencers, i això que, en ambdós casos, s'han utilitzat entre 220 i 550 longituds d'ona com a variables.

Una de les característiques més importants a l'hora de seleccionar longituds d'ona com a paràmetres independents és que no ho són. L'espectre òptic és fonamentalment continu, és a dir, la informació que conté cada longitud d'ona, la contenen en part les longituds d'ona prèvies i posteriors, i aquest fet es tradueix matemàticament en una alta correlació entre elles, o altrament dit, *multicol·linealitat*. Això s'exemplifica molt bé a la Figura C3.11, on es mostra la matriu de correlació de les diferents longituds d'ona dels sensors PreO$_3$ i EB1, prèviament seleccionades automàticament. Les longituds d'ona enregistrades en un mateix punt (tipologia d'aigua o procés) comparteixen molta informació, i per tant, estan altament correlacionades. Aquest factor és extremadament crític, sobretot en modelitzacions que impliquin l'ús de regressions lineals múltiples, molt sensibles a la multicol·linealitat de les variables seleccionades, que fa que s'hagin d'estudiar en profunditat.

La metodologia emprada en aquesta tesi per a la selecció automàtica de variables van ser els motors de selecció FS, BE i SS (extensament descrits en els **capítols 3 i 4**). Una altra eina també molt utilitzada per a la reducció de dimensions són els Anàlisis de Components Principals, i la Regressió de Components Principals (PCA i PCR, de l'anglès Principal Component Analysis i Principal Component Regression, Respectivament). No obstant, en aquesta tesi, es va apostar per la selecció de variables on no hi hagués un canvi de coordenades i de variables, que dificultés, en gran mesura, saber quines longituds d'ona estaven implicades en la modelització. Una bona descripció del funcionament de PCA i PCR es pot trobar a Kumar et al., 2014.

Una vegada aplicada la selecció automàtica a través dels motors de selecció esmentats, es va proposar una revisió manual com a pas posterior a la selecció automàtica. Això és degut a que, tot i que les seleccions automàtiques de variables van fer la seva funció utilitzant uns valors llindar prèviament seleccionats (p. ex. *p*-valor, AIC) que vetllaven pel bon funcionament del model resultant, no seleccionaven les variables en funció de la seva col·linealitat. Aquesta col·linealitat impactava d'una manera crítica en la flexibilitat final del model, ja que provocava que un petit canvi en les dades d'entrada, afectés a totes les prediccions futures. Això significa que, si es mantenien només les variables automàticament seleccionades, es condicionaven negativament els resultats futurs, si les dades d'entrada no es mantenien permanentment estables, cosa que s'ha demostrat, si no impossible, improbable, en la tipologia de dades que es va treballar.

La selecció manual es va fer emprant l'equació del factor de inflació de variància (VIF). L'aplicació del VIF va permetre estudiar l'impacte de cada variable en el model, eliminant les altament correlacionades una a una, i observant la bondat dels resultats obtinguts a cada pas. Això va permetre establir lligams concrets entre variables, com per exemple, variables molt semblants (o properes) demostraven una multicol·linealitat molt alta, i per tant, mantenint la més significativa era suficient perquè el model observés tota la informació que aportaven les anteriors i subsegüents. En general, es va intentar descartar les longituds d'ona amb un VIF més alt que 10. Es va arribar a un compromís entre el VIF, l´R^2 i l'estudi de l'error quadràtic, en tots els casos. Si una variable mostrava un VIF més alt i a l'eliminar-la feia baixar dràsticament la correlació, però al variar les altres sense modificar la primera, no s'observen canvis, es considerava que la variable en qüestió era explicativa d'aquella tipologia de dades, i per tant, es mantenia al model. Aquest estudi exhaustiu es va aplicar posteriorment de totes les seleccions de variables automàtiques, i tot i ser laboriós, va acabar resultant molt necessari per a obtenir els bons resultats en les prediccions dels models matemàtics generats.

6.4. Selecció de models matemàtics en funció del cas d'estudi

Els tres casos d'estudi d'aquesta tesi es van desenvolupar a través d'una metodologia concreta que va tenir en compte la variabilitat de les dades d'estudi, les característiques òptiques pròpies tant de la matriu d'aigua estudiada com dels compostos a predir, les característiques intrínseques de l'espectre UV - Vis i les possibles dificultats que pot comportar l'ús de longituds d'ona com a variables discretes. Això va fer que, en tots els casos es poguessin comparar els resultats entre els diferents models matemàtics resumits en el **capítol 2**, i entre altres que no s'han inclòs per la irrellevància dels resultats obtinguts.

S'ha observat que l'algoritme de MLR és robust en els casos on, el compost a predir, té absorbància directa en l'espectre, com per exemple, en el **capítol 4**, on l'objectiu era predir la concentració d'hidrocarburs en la matriu d'aigua residual. En canvi, en el cas d'estudi de la predicció del THM FP (**capítol 3**), on era necessari observar les diferents interconnexions entre les longituds d'ona seleccionades prèviament, aquest algoritme no va ser útil. També es va observar que la multicol·linealitat present entre les longituds d'ona afectava visiblement a la capacitat de predicció de l'algoritme de MLR, tot i que, tal i com s'ha observat en ambdós capítols, va impactar d'una manera més directa en el cas d'estudi de predicció del THM FP. Aquest fet podia ser degut a que la magnitud de l'absorbància registrada en les longituds d'ona de l'aigua potable era més baixa que la registrada en l'aigua residual, fet que va fer que hi hagués menys diferència entre les unes i les altres, i que podia amplificar l'afectació de la multicol·linealitat en el model de regressió lineal.

Les xarxes neuronals, en canvi, no es van veure afectades per la multicol·linealitat de l'espectre UV - Vis. Quan es crea una xarxa neuronal artificial, aquesta està basada en diferents càlculs sistemàtics en les neurones de les capes presents en l'algoritme. L'aplicació de la funció sigmoidal, que és la que distingeix una neurona activa, d'una neurona inactiva, limitarà l'afectació de la multicol·linealitat en el model. Aquest fet també fa que les xarxes neuronals tinguin una millor capacitat d'observar les relacions intrínseques entre les variables seleccionades obtenint millors resultats predictius.

L'algoritme de Màquines de Suport Vectorial (SVM), utilitzat en el **capítol 5** per monitoritzar el tren de tractament de coagulació-floculació dut a terme a la ERA del Camp de Tarragona, al igual que les xarxes neuronals artificials, és molt robust. Ambdós (ANN i SVM), tenen la capacitat d'observar i aprendre ràpidament les diferències de comportament entre paràmetres, les relacions intrínseques que s'amaguen darrere de la informació disponible, i se´n pot treure un rèdit predictiu molt gran. Tot i així, són algoritmes que fàcilment cauen en el sobre-entrenament, ja que l'usuari no té control sobre els càlculs entremitjos de l'algoritme i aquests esdevenen una *caixa negra*. Un algoritme *caixa negra* és el que s'aplica a una base de dades i emet un resultat que és molt difícil de rastrejar, és a dir, es perd la capacitat d'observació dels resultats en funció dels càlculs aplicats perquè aquests són massa complexes.

A l'hora d'escollir un model matemàtic com a resultat de les investigacions desenvolupades, no solament es va observar l'error obtingut (MAE, RMSE, etc.) o el coeficient de determinació de Pearson (R^2) com a paràmetres de validació, sinó que també es va tenir en compte el factor *caixa negra* com a limitant. En tots els casos, es va estudiar la diferència de comportament entre els models aplicats, i es va decidir també en funció de la facilitat de càlcul i de comprensió del model. És per això que, en el cas d'estudi de la predicció de la concentració d'hidrocarburs en aigües residuals, es va escollir modelitzar aplicant l'algoritme de MLR, ja que, a l'aplicar inicialment l'algoritme d'ANN no es va observar una millora excepcional de les prediccions.

Els models matemàtics que es van generar en aquesta tesi van donar molt bons resultats, tant en l'adaptació a les dades d'entrenament i de prova, com en l'aplicació de les de validació. No obstant, tot i els bons resultats obtinguts, s'ha de tenir en compte que són específics del lloc estudiat. Per a desenvolupar un model matemàtic predictiu robust que sigui adaptable a qualsevol tipologia d'aigua (dins dels barems pel qual hagi estat creat) es necessita una quantitat/qualitat de dades que no poden ser adquirides *només* durant una tesi doctoral, i en molts casos, ni obtenint aquesta gran quantitat de dades es podria obtenir un model que tingués funcionalitats de model global (p. ex., un model de THM FP per a qualsevol tipus d'aigua potable).

Per crear, per exemple, un model de predicció del THM FP com el que s'ha creat en el **capítol 3** d'aquesta tesi, que sigui tan flexible com per predir un valor de potencial de formació per a qualsevol planta de tractament d'aigua potable, cal una quantitat d'informació molt gran, i en qualsevol cas, una adaptació inicial del model a les dades del lloc on s'instal·li. Pels altres dos casos d'estudi desenvolupats en aquesta tesi, es considera molt difícil que es pugui desenvolupar un model global, ja que existeixen barreres importants en la capacitat d'interpolació i extrapolació dels models matemàtics en relació a la variabilitat de les dades que s'han enregistrat. El que sí que es considera un avenç important, pel que fa aquests dos casos d'estudi en particular, és la creació d'una metodologia analítica prèvia al desenvolupament i implementació dels models matemàtics que fa que sigui molt més lleugera l'obtenció de resultats, sabent ràpidament què se'n pot esperar en un futur (si seran robustos i fiables, si faltaran dades d'entrenament o de validació, etc.).

6.5. Treballs futurs

L'ús de sensorització és cada vegada més important en l'àmbit de la monitorització de la qualitat d'aigua. Aquests dispositius aporten robustesa i flexibilitat en les mesures obtingudes minut a minut. La combinació de sensorització, com la sonda espectrofotomètrica spectro::lyser®, amb eines de ML i AI és una aposta de futur que pot aportar grans avenços en la detecció i predicció de contaminants en diferents matrius d'aigua.

Un altre factor a tenir en compte en aquesta equació és la importància, cada vegada més rellevant, que està prenent el món del Big Data. Les aplicacions *Cloud* amb capacitats federatives, on l'accés a la informació, a models matemàtics i serveis específics estan a l'ordre del dia, i poden donar una empenta important a la capacitat de detecció i predicció de models d'aprenentatge automàtic que, d'altra manera, necessitarien una quantitat de dades prohibitiva per poder-se entrenar.

L'aplicació d'aquest tipus de serveis integrals pot generar un *know-how* circular molt valuós, tan per l'empresa s::can com pels seus usuaris, ja que les dades es podrien utilitzar per crear algoritmes personalitzats de monitoratge de la qualitat d'aigua i del tren de tractament d'ETAPs, EDARs i ERAs, per optimitzar parts del procés de tractament, millorant-ne la circularitat i l'impacte econòmic.

En qualsevol cas, l'adaptació en l'ús d'eines d'inferència estadística avançada, aprenentatge automàtic i intel·ligència artificial és la porta a la creació de models matemàtics més robustos, flexibles i adaptats a les necessitats del client, que pot comportar grans avenços en la millora del monitoratge de la qualitat d'aigua, i que acabarà repercutint, no només en s::can i en el seus usuaris, sinó en tots els usuaris de la xarxa d'aigua potable i reciclada.

Capítol 7 – *Conclusions*

This industrial PhD thesis demonstrates that spectral data (Ultraviolet-Visible, UV–Vis) obtained through the spectro::lyser® probe, in combination with chemometric techniques and advanced statistical inference, can be used to develop mathematical models. Therefore, relevant and instantaneous information on water quality in Drinking Water Treatment Plants (DWTPs), Wastewater Treatment Plants (WWTPs) and Water Reclamation Plants (WRPs) for operational monitoring is provided.

Discrete wavelengths as independent variables for predictive modelling entail a subsequent selection by using automatic variable selection methods. After a comparison of the efficiency of Forward, Backward and Stepwise variable selection (FS, BE and SS, respectively), the best selection motor in all case studies is SS.

Regarding the extreme multicollinearity between wavelengths, an automatic selection must be applied with a subsequent manual review of the remaining variables by applying the Variance Inflation Factor (VIF). This parameter exhibited the most correlated variables, allowing an individual study and a final selection and considering only the positive variables for the final model.

The databases obtained in this study were divided into three: training, testing and validation. This methodology allowed an intensive and monitored model training, also including a subsequent test and validation, in order to observe its final goodness.

After conducting several assessments to analyse the importance of deep training in machine learning algorithms, it is concluded that, in order to obtain proper predictions, algorithms must be trained with as much data as possible. An algorithm fed on a wider range of data provides a better response to potential changes in data trends.

Once the mathematical models are created, the most suitable parameters to observe the model goodness are: determination coefficient (R^2), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

The mathematical algorithms used to create prediction models were Multivariate Linear Regression (MLR), Support Vector Machines (SVM) and Artificial Neural Networks (ANN). In applications where the pollutants observed in the spectrum (hydrocarbon prediction) have a direct absorbance, the MLR algorithm is effective. Otherwise, in the prediction of trihalomethanes formation potential (THM FP), without absorbance in UV

– Vis spectrum and in which the relationship between raw and treated water spectral information was studied, the algorithm with best results is ANN.

The models created in this study are site-specific. The prediction capacity of the models is closely related to the information used in the training. With abrupt changes, this capacity can be reduced exponentially. The divergence between trained and real data may affect their extrapolation ability, producing stochastic predictions.

The development of machine learning models to monitor water quality is highly recommended. This is not only because it provides a wide overview of all water processes in progress, but also because such models can be crucial tools for plant operators when necessary.

Capítol 8 – Bibliografia

Abdullah, M. A., Yew, C. H, Ramli M. S., 2003. Formation, modelling, and validation of trihalomethanes (THM) in Malaysian drinking water: a case study in the districts of Tampin, Negeri Sembilan and Sabak Bernam, Selangor, Malaysia. Water Research, 2003,37: 4637–44. DOI: https://doi.org/10.1016/j.watres.2003.07.005

Abu Shmeis, R. M., 2018. Comprehensive Analytical Chemistry. Fundamentals of Quorum Sensing, Analytical Methods and Applications in Membrane Bioreactors, Vol. 81. Water Chemistry and Microbiology, chapter 1, pp.1–56. DOI: https://doi.org/10.1016/bs.coac.2018.02.001

Aho, K.; Derryberry, D.; Peterson, T. 2014. Model selection for ecologists: the worldviews of AIC and BIC. Ecology, 95 (3): 631–636, DOI: https://doi.org/10.1890/13-1452.1

Akaike, H., 1978. On newer statistical approaches to parameter estimation and structure determination.International Federation of Automatic Control, 3, 1877–1884.

Akaike, H., 1979. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting.Biometrika, 66, 237–242.

Alver, A., Baştürk, E., Kılıç, A., 2018. Disinfection By-Products Formation Potential Along the Melendiz River, Turkey; Associated Water Quality Parameters and Non-Linear Prediction Model. International Journal of Environmental Research, 12 (16). DOI: https://doi.org/10.1007/s41742-018-0145-4

Amy, G., Siddiqui, M., Ozekin, K., Zhu, H.W., Wang, C., 1998. Empirically based models for predicting chlorination and ozonation by-products: Haloacetic acids, chloral hydrate, and bromate. EPA report CX 819579.

Awad, J., van Leeuwen, J., Chow, C., Drikas, M., Smernik, R., 2015. Modelling THM formation potential based on the character of organics-in catchments and drinking water sources. 2015. 21st International Congress on Modelling and Simulation. Gold Coast, Australia.

Baeza, J. A., Gabriel, D., Lafuente, J., 2002. In-line fast OUR (oxygen uptake rate) measurements for monitoring and control of WWTP. Water Sci. Technol. 45 (4-5), 19–28.

Baouab M. H., Cherif S., 2018. Prediction of the optimal dose of coagulant for various potable water treatment processes through artificial neural network. Journal of Hydroinformatics.

Barron, A. R., 1992. Approximation and Estimation Bounds for Artificial Neural Networks. Machine Learning, 14 (1), 115 – 133. DOI: https://doi.org/10.1007/BF00993164

Berlman, I.B., 1971. Handbook of Fluorescence Spectra of Aromatic Molecules. 2nd Ed. Academic Press, NY.

Bilbao, I., Bilbao, J., 2017. Overfitting problem and the over-training in the era of data. Particularly for Artificial Neural Networks. The 8th IEEE International Conference on Intelligent Computing and Information Systems (ICICIS 2017). 173-177. DOI: https://doi.org/10.1109/INTELCIS.2017.8260032

Biswas, R.K., et al., 2014. Assessment of drinking water related to arsenic and salinity hazard in Patuakhali district. International journal of advanced geosciences, 2 (2), 82–85.

Bove, F., Shim, Y., Zeitz, P., 2002. Drinking water contaminants and adverse pregnancy outcomes: A review. Environ. Health Perspective, 2002, 110, 61–74.

Box, Jenkins and Reisel, 2008. Time Series Analysis: Forecasting and Control; John Wiley & SONS.; 4th edition (Jun 30, 2008), ISBN: 470272848

Brereton, R. G., Lloyd, G. R., 2010. Support vector machines for classification and regression. Analyst, 2010, 135, 230–267. DOI: https://doi.org/10.1201/b10911-3

Breusch, T. S., Pagan, A. R., 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica, 47(5), 1287-1294.

Burgess, C., 2017. The basics of Spectrophotometric measurement. UV-Visible Spectrophotometry of Water and Wastewater. 2nd Edition. Elsevier. ISBN: 9780444639004

Cabral, J. P. S., 2010. Water microbiology. Bacterial pathogens and water. International Journal of Environmental Research and Public Health, 7(10), pp. 3657-3703. DOI: https://doi.org/10.3390/ijerph7103657

Cantor, K. P., Villanueva, C. M., Silverman, D. T., Figueroa, J. D., Real, F. X., Garcia–Closas, M., Malats, N., Chanock, S., Yeager, M., Tardon, A., Garcia–Closas, R., Serra, C., Carrato, A., Castaño–Vinyals, G., Samanic, C., Rothman, N., Kogevinas, M., 2010. Polymorphisms in GSTT1, GSTZ1, AND CYP2E1, disinfection by-products, and risk of bladder cancer in Spain. Environ. Health Perspective, 2010, 118, 1545–1550. DOI: https://doi.org/10.2307/40963838

Carstea, E. M., Bridgeman, J., Baker, A., Reynolds, D. M., 2016. Fluorescence spectroscopy for wastewater monitoring: A review. Water Research, 95, 205-219. DOI: https://doi.org/10.1016/j.watres.2016.03.021

Cheng, W.P., Yu R. F., Hsieh, Y.J., 2010. Optimizing coagulant demand by Nephelometric Turbidimeter Monitoring System (NTMS). Desalination and water treatment, 16(1-3), 95-100. DOI: https://doi.org/10.5004/dwt.2010.1048

Cho, Y.-C., Choi, H., Lee, M.-G., Kim, S.-H., Im, J.-K., 2022. Identification and Apportionment of Potential Pollution Sources Using Multivariate Statistical Techniques and APCS-MLR Model to Assess Surface Water Quality in Imjin River Watershed, South Korea. Water 2022, 14, 793. DOI: https://doi.org/10.3390/w14050793

Chowdhury, S., Champagne, P., Mclellan, P.J., 2009. Models for predicting disinfection byproduct (DBP) formation in drinking waters: a chronological review. Sci. Total Environ. 407, 4189–4206. DOI: https://doi.org/10.1016/j.scitotenv.2009.04.006.

Cortes, C., Vapnik., V., 1995. Support-vector network. Machine Learning, 20, 273–297.

Custodio, E., 2022. Considerations on the past, present and future of groundwater in Spain. Ingeniería del agua, 26(1), 1-17. DOI: https://doi.org/10.4995/Ia.2022.16245

da Costa, H., Pinheiro, S., Pinheiro, L., Souza, A. F., de Melo, T., Silva, C., Magno, R., Garcia, K., Castros, D., Sousa, E., 2020. Chemometrics applied in the development of a water quality indicator System for the Brazilian Amazon. ACS Omega 2020, 5, 51, 32899-32906. DOI: https://doi.org/10.1021/acsomega.0c03430

de Castro, L., de Alencar, F. L., S., Navoni, J., A., de Araujo, A. L., do Amaral, V. S, 2019. Toxicological aspects of trihalomethanes: a systematic review. Environmental Science and Pollution Research, 26, 5316–5332 (2019). DOI: https://doi.org/10.1007/s11356-018-3949-z de Sà, A.G.C., Pappa, G.L., Freitas, A.A., 2017. Towards a Method for automatically selecting and configuring multi-label classification algorithms. Proceedings of the 19th annual conference companion on Genetic and Evolutionary computation, Berlin, Germany (GECCO'17 Companion). DOI: 10.1145/3067695.3082053

de Veaux, R. D., and Ungar, L. H. (1994). Multicollinearity: A tale of two nonparametric regressions. Lecture Notes in Statistics, 393–402. DOI: https://doi.org/10.1007/978-1-4612-2660-4_40

Dixon, J. M., Taniguchi, M., Lindsey, J.S., 2005. PhotochemCAD 2. A Refined Program with Accompanying Spectral Databases for Photochemical Calculations. Photochem. Photobiol., 81, 212-213.

Douglas, R. K., Nawar, S., Alamar, M. C., Mouazen, A. M., Coulon, F., 2018. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. Science of the Total Environment, 616-617, 147-155. DOI: https://doi.org/10.1016/j.scitotenv.2017.10.323

Du, H., Fuh, R-C. A., Li, L., Corkan, A., Lindsey, J.S., 1998. PhotochemCAD: A computer-aided design and research tool in photochemistry. Photochem. Photobiol. 68, 141-142.

Elshorbagy, WE, Abu-Qadais, H., Elsheamy, M. K., 2000. Simulation of THM species in water distribution systems. Water Research, 2000, 34:3431–9. DOI: https://doi.org/10.1016/S0043-1354(00)00231-1

England A. J. Jr., Krenkel, P. A., 2003. Suspended Solids Removal. Book chapter. Encyclopaedia of Physical Science and Technology, 3rd Edition, 2003.

España. Real Decreto 140/2003, de 7 de febrero, por el que se establecen los criterios sanitarios de calidad del agua de consumo humano. Boletín Oficial del Estado, 7 de febrero del 2003, núm. 45.

España. Real Decreto 1620/2007, de 7 de diciembre, por el que se establece el régimen jurídico de la reutilización de aguas depuradas. Boletin Oficial del Estado, 7 de diciembre del 2007, núm. 294, pp. 50639 a 50661.

European Commission. Directive (EU) 2020/2184 of the European Parliament and of the Council of 16 December 2020 on the quality of water intended for human consumption. Official Journal of the European Union, L435/1 of December 2020.

European Commission. Directive (EU) 98/83/EC of the European Parliament and of the Council of 3 November 1998 on the quality of water intended for human consumption. Official journal of the European Communities, L330/32, 5.12.98., 1998.

European Commission. Directive (EU) 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy.

European Commission. Commission decision of 12 August 2002 implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results (2002/657/EC). Off. J. Eur. Commun. 2002; 221: 8–36

European Commission. Regulation (EU) 2020/741 of the European Parliament and of the Council of 25 May 2020 on minimum requirements for water reuse. L 177/32.

Fabbricino, M., and Korshin, G. V., 2009. Modelling disinfection by-products formation in bromide-containing waters. Journal of Hazardous Materials, 168(2–3), 782–786. DOI: https://doi.org/10.1016/j.jhazmat.2009.02.078

Fitzpatrick, C.S.B., Fradin, E., Gregory, J., 2004. Temperature effects on flocculation, using different coagulants. Water, Science and Technology, Vol. 50, No 12. pp 171 – 175. DOI: https://doi.org/10.2166/wst.2004.0710

France. LEMA, 2006-1772. Loi sur l'eau et les milieux aquatiques du 30 décembre 2006, LEMA.

Franceschi, M., Girou A., Carro-Diaz A.M., et al., 2002. Optimisation of the coagulation-flocculation process of raw water by optimal design method. Water Research 36 (2002) 3561–3572. DOI: http://dx.doi.org/10.1016/S0043-1354(02)00066-0

Garrido Baserba, M., Corominas, L., Cortes, U., Rosso, D., Poch, M., 2020. The fourth-revolution in the water sector encounters the digital revolution. Environ. Sci. Technol., 50 (2020), pp. 4698-4705. DOI: http://doi.org/10.1021/acs.est.9b04251

Gauchi, J.P., Chagnon, P., 2001. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. Chem Intell Lab Sys 2001, 58:171-93. DOI: https://doi.org/10.1016/s0169-7439(01)00158-7

Geladi, P., 2003. Chemometrics in spectroscopy. Part 1. Classical chemometrics. Spectrochimica Acta – Part B Atomic Spectroscopy, 58(5), 767-782. DOI: https://doi.org/10.1016/S0584-8547(03)00037-5

Gernaey, K.V., van Loosdrecht, M. C. M., Henze, M., Lind, M., Jorgensen, S. B., 2004. Activated sludge wastewater treatment plant modelling and simulation: state of the art. Environmental Modelling & Software. Vol 19(19), 2004, pp.763-783. DOI: https://doi.org/10.1016/j.envsoft.2003.03.005

Giraud, C., 2021. Introduction to High-Dimensional Statistics, 2nd Edition. CRC Press.

Godó-Pla, L., Emiliano, P. Poch, M., Valero, F., Monclús, H., 2021. Benchmarking empirical models for THMs formation in drinking water systems: An application for decision support in Barcelona, Spain. Science of Total Environment 763 (2021) 144196. DOI: https://doi.org/10.1016/j.scitotenv.2020.144197

Golfinopoulos, S., Arhonditsis, G., 2002. Multiple regression models: a methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics. Chemosphere 2002, 47:1007–18. DOI: https://doi.org/10.1016/s0045-6535(02)00058-9

Gregory, J., 2004. Monitoring floc formation and breakage. Water Science and Technology, Vol. 50, No 12. pp 163–170. DOI: https://doi.org/10.2166/wst.2004.0709

Gruber G., Bertrand-Krajewski, J. L., De Benedits, J., Hochedlinger, M., 2005. Practical aspects, experiences and strategies by using UV/Vis sensors for long-term sewer monitoring. Water Practice and Technology, 2005. 10th ICUD – International Conference on Urban Drainage, Copenhagen, Denmark.

Gurney K., 2004. An introduction to Neural Networks. Taylor & Francis e-Library.

Hanrahan, G., Zhu, J, Gibani, S., Patil, D. G., 2005. Chemometrics and Statistics: Experimental Design, Editor(s): P. Worsfold, A. Townshend, C. Poole, Encyclopedia of Analytical Science (2nd Edition), Elsevier, 2005, pp. 8–13. ISBN: 9780123693976.

Harrell, F. E., 2015. Regression modelling strategies: With applications to linear models, logistic regression, and survival analysis. Springer-Verlag, New York.

Hastie, T. J., Pregibon, D., 1992. Generalised linear models. Chapter 6 of Statistical Models in S. Eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Heinze, G., Wallisch, C., Dunkler, D., 2020. Variable selection – A review and recommendations for the practising statistician. Biometrical Journal, 2017; 1 – 19. DOI: 10.1002/bimj.201700067

Hernandez-Ramirez, A. G., Matinez-Tavera, E., Rodriguez-Espinosa, P.F., Mendoza-Pérez, J. A., Tabla-Hernandez, J., Escobedo-Urías, D.C., Jonathan, M. P., Sujitha, S. B., 2019. Detection, provenance and associated environmental risks of water quality pollutants during anomaly events in River Atoyac, Central Mexico: A real-time monitoring approach. Science of The Total Environment, vol. 669(15), 2019, pp. 1018-1032. DOI: https://doi.org/10.1016/j.scitotenv.2019.03.138

Hsu, C.W., Chang, C.C., and Lin, C.J., 2016. A practical guide to support vector classification.

Huang, Z., Wang, Y., Jiang, L., et al., 2018. Mechanism and performance of a self-flocculating marine bacterium in saline wastewater treatment. Chem. Eng. J. 334, 732–740.

Hue, J., Dupoy, M., Bordy, T., Rousier, R., Vignoud, S., Schaerer, B., Tran-Thi, T. H., Rivron, C., Mugherli, L., Karpe, P., 2013. Benzene and xylene detection by absorbance in the range of 10-100 ppb application: Quality of indoor air. Sensors and Actuators, B: Chemical, 189, 194–198. DOI: https://doi.org/10.1016/j.snb.2013.03.047

IUPAC, 2006. The Gold Book. Compendium of Chemical Terminology, 2nd Edition. Online corrected version, 2006. "Beer–Lambert law". DOI: 10.1351/goldbook.B00626

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. An introduction to Statistical Learning: With Applications in R. 2nd Ed. Springer

Jarvis, P., Jefferson, B., S.A. Parsons, S. A., 2006. Floc structural characteristics using conventional coagulation for a high doc, low alkalinity surface water source. Water Res. 40 (2006) 2727-2737. DOI: http://dx.doi.org/10.1016/j.watres.2006.04.024

Jiang, Q. J. 2015. The role of coagulation in water treatment. Curr. Opin. Chem. Eng. 8:36–44. DOI: http://dx.doi.org/10.1016/j.coche.2015.01.008

Jones, R. N., 1941. The ultraviolet absorption spectra of Aromatic Hydrocarbons. Chemical Reviews, 32 (1), 1-46. DOI: https://doi.org/10.1021./cr60101a001

Karlowatz, M., Kraft, M., Mizaikoff, B., 2004. Simultaneous Quantitative Determination of Benzene, Toluene, and Xylenes in Water Using Mid-Infrared Evanescent Field Spectroscopy. Analytical Chemistry, 76(9), 2643–2648. DOI: https://doi.org/10.1021/ac0347009

Kaur, G., Singh, H., Singh, J., 2021. UV-Vis spectrophotometry for environmental and industrial analysis. Chapter. Green Sustainable Process for Chemical and Environmental Engineering and Science. Analytical Techniques for Environmental and Industrial Analysis, 2021, pp. 49-68. DOI: https://doi.org/10.1016/B978-0-12-821883-9.00004-7

Ke, J., Liu, X., 2008. Empirical analysis of optimal hidden neurons in Neural Network modeling for stock prediction. Pacific – Asia Workshop on Computational Intelligence and Industrial Application (PACIIA). DOI: https://10.1109/PACIIA.2008.363

Keith, T., 2015. Multiple Regression and Beyond. Book. 2nd Edition. Routledge.

Khan S., Newport, D., Le Calvé, S., 2021. A Sensitive and portable Deep-UV Absorbance Detector with a Microliter Gas Cell Compatible with Micro GC.

Korshin, G. V., Benjamin, M. M., Chang, H. S., Gallard, H., 2007. Examination of NOM chlorination reactions by conventional and stop–flow differential absorbance spectroscopy. Environmental Science and Technology, 41(8), 2776–2781. DOI: 10.1021/es062268h

Korshin, G.V., Sgroi, M., Ratnaweera, H., 2017. Spectroscopic surrogates for real-time water quality monitoring in wastewater treatment and water reuse. Current opinion in Environmental Science & Health, (2) 12-18. DOI: https://doi.org/10.1016/j.coesh.2017.11.003

Kröse, B., Van der Smagt, Patrick., 1996. An introduction to neural networks. Journal of computer science, 48. University of Amsterdam: Amsterdam, The Netherlands, 1996.

Kuhn, M., 2008. Building predictive models in R using the CARET package. Journal of Statistic Software. 28, 5. DOI: https://doi.org/10.18637/jss.v028.0i5

Kuhn, M., Johnson K., 2013. Applied Predictive Modelling. 1st Edition. Springer. DOI: https://doi.org/10.1007/978-1-4614-6849-3.

Kumar, N., Bansal, A., Sarma, G. S., Rawal, R. K., 2014. Chemometrics tools used in analytical chemistry: An overview. Talanta, 123, 186–199. DOI: https://doi.org/10.1016/j.talanta.2014.02.003

Kutner, M. H.; Nachtsheim, C. J.; Neter, J., 2004. Applied Linear Regression Models. 4th Edition. McGraw-Hill Irwin.

Lameski, P., 2015. SVM parameter tuning with grid search and its impact on reduction of model overfitting. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC 2015), LNAI 9437, pp. 464–474, 2015. DOI: https://doi.org/10.1007/978-3-319-25783-9_41

Langergraber, G., Fleischmann, N., Hofstaedter, F., Weingartner, A., Lettl, W., 2003. Detection of (unusual) changes in wastewater composition using UV/Vis spectroscopy. 9th IWA Specialized Conference Design, Operation and Economics of Large Wastewater Treatment Plants, IWA Prag 2003, September 1 - 4, 2003, Prague, Czech Republic

Li, X., 2012. A simulation evaluation of Backward Elimination and Stepwise variable selection in regression analysis. Shandong Polytechnic University.

Lima, Kássio M.G., Raimundo, I. M., Pimentel, M. F., 2011. Simultaneous determination of BTX and total hydrocarbons in water employing near infrared spectroscopy and multivariate calibration. Sensors and Actuators B: Chemical, 160(1), 691–697. DOI: https://doi.org/10.1016/j.snb.2011.08.050

Lin, J., Chen, X., Ansheng, Z., Hong, H., Liang, Y., Sun, H., Lin, H., Chen, J., 2018. Regression models evaluating THMs, HAAs and HANs formation upon chloramination of source water collected from Yangtze River Delta region, China. Ecotoxicol. Environ. Saf. 160, 249–256. DOI: https://doi.org/10.1016/j.ecoenv.2018.05.038.

Maier, H. R., Dandy, G. C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental

Modelling and Software. 15(1), 101–124. DOI: https://doi.org/10.1016/S1364-8152(99)00007-9

Malekian, A., Chitsaz, N., 2021. Concepts, procedures, and applications of Artificial Neural Networks models in streamflow forecasting. Advances in Streamflow Forecasting, from traditional to modern approaches, 2021, pp 115-147. 1st Edition, Elsevier.

MASE, 2015. Aspectos hidrológicos, ambientales, económicos, sociales y éticos del consumo de reservas de agua subterránea en España. Preparado por E. Custodio para UPC and AQUALOGY–CETAQUA. Universidad Politécnica de Cataluña, Barcelona. e-books: 1–487. http://hdl.handle.net/2117/111272.

Matilainen, A., Sillanpää M., 2010a. Removal of natural organic matter from drinking water by Advanced oxidation processes. Chemosphere 80. 351–365. DOI: https://doi.org/10.1016/j.chemosphere.2010.04.067

Matilainen, A., Vepsäläinen, M., Sillanpaä M., 2010b. Natural organic matter removal by coagulation during drinking water treatment - A review. Advances in Colloid and Interface Science 159. 189–197. DOI: https://doi.org/10.1016/j.cis.2010.06.007

Matsché, N., Stumwöhrer, K., 1996. UV absorption as control-parameter for biological treatment plants. Water Science and Technology, Vol. 33(12), pp. 211-218. DOI: https://doi.org/10.1016/0273-1223(96)00471-1

Mayer, B. K., Ryan, D. R., 2017. Impact on Disinfection By-products Using Advanced Oxidation Processes for Drinking Water Treatment. The handbook of Environmental Chemistry, (chapter 82). DOI: https://doi.org/10.1007/698_2017_82

Mayerhöfer, T. G., Pahlow, S., Popp, J., 2020. The Bouguer-Beer-Lambert law: Shining light on the obscure. ChemPhysChem 21 (18). DOI: https://doi.org/10.1002/cphc.202000464

Mehmood, T., Liland-Hovde, K., Snipen, L., Sabo, S., 2012. A review of variable selection methods in Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems 118 (2012) 62-69. DOI: https://doi.org/10.1016/j.chemolab.2012.07.010

Menze, B.H., Kelm, M.B., Masuch, R., Himmelreich, U., Bachert , P., Petrich W., Hamprecht, F. A., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. 10(1), 213–0. DOI: https://doi.org/10.1186/1471-2105-10-213

Milot, J., Rodríguez, M. J., Sérodes, J. B., 2002. Contribution of neural networks for modelling trihalomethanes occurrence in drinking water. J Water Resour. Plann Manag. ASCE 2002, 128:370–6.

Montero-Montoya, R.D., Lopez, R., Arellano-Aguilar, O., 2018. Volatile Organic Compounds in Air: Sources, Distribution, Exposure and Associated Illness in Children. Annals of Global Health 84 (2): 225-238. DOI: https://doi.org/10.29024/aogh.910

Mudge, S.M., Ball, A.S., 2006. Sewage. In: Morrison, R., Murphy, B. (Eds.), Environmental Forensics: A Contaminant Specific Approach. Elsevier Academic Press, New York, pp. 35–54.DOI: https://doi.org/10.1016/B978-012507751-4/50025-2

Neter, J., Wasserman, W., Kutner, M., 2004. Applied Linear Statistical Models. 5th Edition. McGraw-Hill Irvin.

Nickson, R.T., et al., 2005. Arsenic and other drinking water quality issues, Muzaffargarh District, Pakistan. Applied geochemistry, 20 (1), 55–68.

Nieuwenhuijsen, M. J., Toledano, M. B., Eaton, N. E., Fawell, J., Elliott, P., 2000. Chlorination disinfection by-products in water and their association with adverse reproductive outcomes: A review. Occup. Environ. Med. 2000, 57, 73–85. DOI: https://doi.org/10.1136/oem.57.2.73

Olanrewaju, R., Salawundeen, T. O., 2012. An intelligent modeling of coagulant dosing system for water treatment plants based on artificial neural networks. Journal of Basic and Applied Sciences, 6 (1); 93–99, 2012.

Patel, H, Vash, R. T., 2013. Comparison of naturally prepared coagulants for removal of COD and colour from textile wastewater. Glob NEST J 15(4):522–528.

Platikanov, S., Martí J., Tauler, R., 2012. Linear and non-linear chemometric modelling of THM formation in Barcelona's water treatment plant. Science of the Total

Environment 432 (2012) 365–374. DOI: https://doi.org/10.1016/j.scitotenv.2012.05.097

Platikanov, S., Puig, X., Martín, J., Tauler, R., 2007. Chemometric modelling and prediction of trihalomethane formation in Barcelona's waterworks plant. Water Research 41, 3394–3406. DOI: https://doi.org/10.1016/j.watres.2007.04.015.

Poleneni, S., 2020. Disinfection By-Products in Drinking Water. Global disinfection by-products regulatory compliance framework overview. Disinfection By-Products in Drinkinw Water: detection and treatment. Pp. 305-335. DOI: 10.1016/B978-0-08-102977-0.00014-7

Poleneni, S., 2017. Management of DBP Formation Using Enhanced Treatment Technologies and an Array of Prediction Tools. University of Missouri-Columbia, Columbia, Missouri. DOI: https://doi.org/10.32469/10355/62334

Poleneni, S., Inniss, E., 2015. Small water distribution system disinfection by-product control: water quality management using storage systems. Int. J. Geotechn. Construct. Mater. Environ. (GEOMATE). 1-17 (9), 13651369. From: https://geomatejournal.com/geomate/article/view/1954

Poleneni, S., Inniss, E., 2013. Small water distribution system operations and disinfection by product fate. J. Water Res. Protect. 5 (8A), 3541. DOI: https://doi.org/10.4236/jwarp.2013.58A005.

Putri, M.S.A., Lou, C-H., Syai'in, M., Ou, S-H., Wang, Y.C., 2018. Long-Term River Water Quality Trends and Pollution Source Apportionment in Taiwan. Water, 2018, 10, 1394. DOI: https://doi.org/10.3390/w10101394

Qi, W., Zhang, H., Hu, C., Liu, H., Qu, J., 2018. Effect of ozonation on the characteristics of effluent organic matter fractions and subsequent associations with disinfection by-products formation. Science of The Total Environment 2018, 610: 1057–1064. DOI: https://doi.org/10.1016/j.scitotenv.2017.08.194

Quina, F. H., Carroll, F. A., 1976. Radiative and nonradiative transitions in solution. First excited singlet state of benzene and its methyl derivatives. J. Am. Chem. Soc. 98, 6-9.

Raich-Montiu, J., Barios, J., Garcia, V., Medina, M. E., 2014. Integrating membrane technologies and blending options in water production and distribution systems to improve organoleptic properties. The case of Barcelona Metropolitan Area. Journal of Cleaner Production, 69, pp 250-259. DOI: https://doi.org/10.1016/j.jclepo.2014.01.034

Raich-Montiu, J., 2014. Review of sensors to monitor water quality. European Reference Network for Critical Infrastructure Protection (ERNCIP Project). Chemical & Biological Risks in the Water Sector.

Redondo-Hasselerharm, P. E., Cserbik, D., Flores, C., Farré, M. J., Sanchís, J., Alcolea, J. A., Planas, C., Caixach, J., Villanueva, C.M., 2022. Insights to estimate exposure to regulated and non-regulated disinfection by-products in drinking water. Journal of Exposure Science & Evironmental Epidemiology. DOI: https://doi.org/10.1038/s41370-022-00453-6

Rencher, A. C.; Christensen, W. F., 2012. Methods of Multivariate Analysis. 3rd Edition. Wiley Series in Probability and Statistics, 709, John Wiley & Sons, p. 19, ISBN 9781118391679.

Richardson, D., Postigo, C., 2015. Formation of DBPs: State of the Science. American chemical society, Washington DC, 2015.

Richardson, S. D., Plewa, M. J., Wagner, E. D., Schoeny, R., DeMarini, D. M., 2007. Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: A review and roadmap for research. Mutat. Res. 2007, 636, 178–242. DOI: https://doi.org/10.1016/j.mrrev.2007.09.001

Riedmiller, M., Braun, H., 1992. RPROP – A fast adaptative learning algorithm. Proceedings of ISCIS VII.

Rivadeneyra, A., García–Ruiz, M.J., Delgado–Ramos, F., González–Martínez, A., Osorio, F., Rabaza, O., 2014. Feasibility study of a simple and low-cost device for monitoring trihalomethanes presence in water supply systems based on statistical models. Water (Switzerland) 6, 3590–3602. DOI: https://doi.org/10.3390/w6123590.

Roda, A., Mirasoli, M., Michelini, E., Magliulo, M., Simoni, P., Guardigli, M., Curini, R., Sergi, M., Marino, A., 2006. Analytical approach for monitoring endocrine-disrupting compounds in urban wastewater treatment plants. Analytical and Bioanalytical Chemistry 385, 742-752 (2006). DOI: https://doi.org/10.1007/s00216-006-0473-7

Rodríguez, M. J., Sérodes, J. B., 2001. Spatial and temporal evolution of trihalomethanes in three water distribution systems. Water Research, 2001, 35:1572–86. DOI: https://doi.org/10.1016/S0043-1354(00)00403-6

Rodríguez, M. J., Sérodes, J. B., 2004. Application of back-propagation neural network modelling for free residual chlorine, total trihalomethanes and trihalomethanes speciation. Journal of Environmental Engineering Science, 2004, 3: S25–34. DOI: https://doi.org/10.1139/s03-069

Sadiq, R., Rodriguez, M. J., Mian, H. R., 2019. Empirical models to predict disinfection by-products (DBPs) in drinking water: an updated review. Encyclopaedia of Environmental Health, 2nd ed Elsevier Inc. DOI: https://doi.org/10.1016/B978-0-12-409548-9.11193-5.

Sadiq, R., Rodriguez, M.J., 2004. Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review. Sci. Total Environ. 321, 21–46. DOI: https://doi.org/10.1016/j.scitotenv.2003.05.001.

Sanz, J., Suescun, J., Molist, J., Rubio, F., Mujeriego, R., Salgado, B., 2015. Reclaimed water for the Tarragona petrochemical park. Water Science & Technology: Water Supply. 15.2. 2015. DOI: https://doi.org/10.2166/ws.2014.114

Saritha, V., Srinivas, N., Srikanth-Vuppala, N. V., 2017. Analysis and optimization of coagulation and flocculation process. Appl. Water. Sci (2017). 7:451-460. DOI: https://doi.org/10.1007/s13201-014-0262-y

Savitz, D. A., Singer, P. C., Hartmann, K. E., Herring, A. H., Weinberg, H. S., Makarushka, C., Hoffman, C., Chan, R., Maclehose, R., 2005. Drinking-Water Disinfection By-Products and Pregnancy Outcome. AWWA Research Foundation: Denver, CO, 2005, p-212. DOI: DOI: https://doi.org/10.1093/aje/kwj300

Serrano, A., Gallego, M., 2007. Rapid determination of total trihalomethanes index in drinking water. Journal of Chromatography, 2007, 1154:26–33. DOI: https://doi.org/10.1016/j.chroma.2007.03.101

Sharp, E. L., Parson, S. A., Jefferson, B., 2006. Coagulation of NOM: linking character to treatment. Water Sci Technol. 53 (7), 67–76. DOI: https://doi.org/10.2166/wst.2006.209

Sillanpää, M., Ncibi, M. C., Matilainen, A., et al., 2018. Removal of natural organic matter in drinking water treatment by coagulation: a comprehensive review. Chemosphere 190, 54–71. DOI: http://dx.doi.org/10.1016/j.chemosphere.2017.09.113

Singer, P. C., 1994. Control of disinfection by-products in drinking water. Journal of Environmental Engineering–ASCE, 120(4), 727–744. DOI: https://doi.org/10.1061/(ASCE)0733-9372(1994)120:4(727)

Sun, Y., Zhou, S., Chiang, P. C., Shah, K. J., 2020. Evaluation and optimization of enhanced coagulation process: Water and energy nexus. Water-Energy Nexus 2. 25–36. DOI: https://doi.org/10.1016/j.wen.2020.01.001

Swietlik, J., Dabrowska, A., Raczyk-Stanislawiak, U., Nawrocki, J., 2004. Reactivity of natural organic matter fractions with chlorine dioxide and ozone. Water Res. 38, 547–558. DOI: https://doi.org/10.1016/j.watres.2003.10.034

Tchobanoglous, G., Stensel, H. D., Tsuchihashi, R., Burton, F. L., 2013. Wastewater Engineering: Treatment and Resource Recovery. McGraw Hill. 5th Edition.

Thayer, J. D., 1990. Implementing variable selection techniques in regression. Paper presented at the annual meetings of the American Educational Research Association, Boston, MA.

Thayer, J. D., 2002. Stepwise regression as an exploratory data analysis procedure. Paper presented at the annual meetings of the American Educational Research Association, New Orleans, LA.

Thomas, D. R., Hughes, E., Zumbo, B. D., 1998. On variable importance in linear regression. Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, 45, 253-275.

Thomas, O., Causse, J., 2017. From spectra water quality monitoring. UV-Visible Spectrophotometry of Water and Wastewater. 2nd Edition. Elsevier. ISBN: 9780444639004

Toroz, I., Uyak, V., 2005. Seasonal variations of trihalomethanes (THMs) in water Distribution networks of Istanbul City. Desalination 2005, 176:127–41. DOI: https://doi.org/10.1016/j.desal.2004.11.008

Torres, A., Bertrand-Krajewski, J. l., 2008. Partial Least Squares local calibration of a UV-Visible spectrometer used for in situ measurements of COD and TSS concentrations in urban drainage systems. Water Sci. Technol. 2008, 57 (4):581-8. DOI: https://doi.org/10.2166/wst.2008.131

UNESCO, 2006. Water, a shared responsibility. The United Nations World Water Development Report 2. ISBN UNESCO: 92-3-104006-5. UN-WATER/WWAP/2007/02

United Nations. Goal 6th: Ensure access to water and sanitation for all. Water Action Decade, 2018-2028. www.un.org/sustainabledevelopment/water-and-sanitation/

US EPA, 2010. National Primary Drinking Water Standards. US Environmental Protection Agency, Office of Water, USA

US EPA. EPA Method 501.3, (EPA 500-Series, November 1979). Measurement of Trihalomethanes in Drinking Water with Gas Chromatography/Mass Spectrometry and Selected Ion Monitoring. Cincinnati, Ohio. U.S. Environmental Protection Agency (US EPA).

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY, 1995.

Venables, W. N., Ripley, B. D., 2002. Modern Applied Statistics with S. 4th Edition. New York: Springer

Vesilind, P., 2003. Wastewater Treatment Plant design. 1st Edition. Water Environment Federation, US. ISBN: 1572781777

Vigil, K. M., 2003. Clean Water: An introduction to Water Quality and Water Pollution Control. 2nd Edition. Oregon ST. Univ Pr. ISBN: 0870714988

Villanueva, C. M., Castaño-Vinyals, G., Moreno, V., Carrasco-Turigas, G., Aragonés, N., Boldo, E., 2012. Concentrations and correlations of disinfection by-products in municipal drinking water from an exposure assessment perspective. Environ Res. 2012; 114:1–11. DOI: https://doi.org/10.1016/j.envres.2012.02.002. 4.

Villanueva, C. M., Cordier, S., Font-Ribera, L., Salas, L. A., Levallois, P., 2015. Overview of disinfection by-products and associated health effects. Curr. Environ Health Rep. 2015; 2:107–15. DOI: https://doi.org/10.1007/s40572-014-0032-x

Villanueva, C. M., Cantor, K. P., Cordier, S., Jaakkola, J. J. K., King, W. D., Lynch, C. F., Porru, S., Kogevinas, M., 2004. Disinfection by-products and bladder cancer: A pooled analysis. Epidemiology 2004, 15, 357–367. Source: https://www.jstor.org/stable/20485906

Vörösmarty, C. J., Green, P., Salisbury, J. y Lammers, R. 2000. Global water resources: Vulnerability from climate change and population growth. Science. Vol. 289, pp. 284-88. DOI: https://doi.org/10.1126/science.289.5477.284

Vujicic, T., Matijevic, T., Ljucovic J., Balota, A., 2016. Comparative analysis of methods for determining the number of hidden neurons in Artificial Neural Network. Central European Conference on Information and Intelligent Systems (CECIIS), 219-250.

Walczak, S., Cerpa, N., 2003. Artificial Neural Networks. Encyclopaedia of Physical Science and Technology, 2003, pp. 631-645. 3rd Edition, Elsevier.

Waller, K., Swan, S. H., DeLorenzo, G., Hopkins, B., 1998. Trihalomethanes in drinking water and spontaneous abortion. Epidemiology 1998, 9, 134–140. DOI: https://doi.org/10.2307/3702950

WHO (World Health Organization), 1996. Guidelines for drinking water quality. 2nd ed. Geneva, Switzerland: World Health Organization.

WHO (World Health Organization), 1997. Guideline for drinking water quality, health criteria. 2nd Edition. Vol. 2. Geneva, Switzerland: World Health Organization.

WHO (World Health Organization), 2011. Guidelines for drinking-water quality. 4th ed. Geneva, Switzerland: WHO. Yagoub, S.O. and Ahmed, R.Y., 2009. Microbiological evaluation of the quality of tap water distributed at Khartoum State. Research journal of microbiology, 4 (10), 355–360.

WHO (World Health Organization), 2018. Strengthening operations and maintenance through water safety planning: A collection of case studies. Geneva: World Health Organization; 2018. Licence: CC BY-NC-SA 3.0 IGO.

Wold, S., Ruhe, A., Wold, H., Dunn III, W. J., 1984. The collinearity problem in Linear Regression. The Partial Least Squares (PLS) approach to generalized inverses. Siam Journal of Scientific Statistic Computing, 5(3), 735–743. DOI: https://doi.org/10.1137/0905052

Xu, Z., Dong, Q., Otieno, B., Liu, Y., Williams, I., Cai, D., Li, Y., Lei, Y., Li, B., 2016. Real-time in situ sensing of multiple water quality related parameters using micro-electrode array (MEA) fabricated by inkjet-printing technology (IPT). Sensors and Actuators B: Chemical. Vol. 237, 2016, pp. 1108-1119. DOI: https://doi.org/10.1016/j.snb.2016.09.040

Yamashita, T., Yamashita, K., Kamimura R., 2007. A Stepwise AIC Method for Variable Selection in Linear Regression. Communications in Statistics – Theory and Methods, 36:13, 2395-2403, DOI: https://doi.org/10.1080/03610920701215639

Yan, M., Wang, D., Qu, J., et al., 2008. Enhanced coagulation for high alkalinity and micropolluted water: the third way through coagulant optimization, Water Res. 42. 2278–2286. DOI: https://doi.org/10.1016/j.watres.2007.12.006

Shi, Z., Chow, C. W. K., Fabris, R., Liu, J., Jin, B., 2022. Applications of Online UV-Vis Spectrophotometer for Drinking Water Quality Monitoring and Process Control: A Review. DOI: https//doi.org/10.3390/s2282987. ISBN: 1424-8220

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., Ye, L., 2022. A review of the application of Machine learning in water quality evaluation. Eco-Environment & Health. Vol 1(2), 2022, pp. 107-116. DOI: https://doi.org/10.1016/j.eehl.2022.06.001

Zoccolillo, L., Alessandrelli, M., Felli, M., 2001. Simultaneous determination of benzene and total aromatic fraction of gasoline by HPLC-DAD. Chromatographia, 54(9–10), 659–663. DOI: https://doi.org/10.1007/BF02492195